FastType

Dear collegues of Intersteno,

I have the pleasure to send you herewith documentation about an Italian project carried out by Synthema s.r.l. a company based in Pisa who exhibited their software for speech recognition at the Intersteno Congress in Rome.

Dott Carlo Aliprandi, who is the responsible of special projects of this Company, has kindly informed me about this final release of this project which was announced for the first time on a booklet (Tèlema) specifically devoted to reporting. This booklet was issued shortly after the Italian Congress and a copy of it was released to the members of the Board.

Dott. Carlo Aliprandi, at the same Congress, spoke about the possible use of speech-technologies for preparing transcripts (see the Final Report of Intersteno Congress 2003 pag. 82). Carlo Aliprandi is not related to Giuseppe Aliprandi family, who was the first President of Accademia Aliprandi, I say this because many persons are asking about this possible, but not existing, relationship.

FastType project is specifically intended for disabled people, who have motor problems and get frustrated in writing texts on traditional keyboard. The project was developed by Synthema, in cooperation with the Informatics Department of the University of Pisa, and with the financial support of Fondazione Cassa di Risparmio di Pisa.

Due to this specific goal, the software has an approach completely different from the ones experienced in the past by inventors of steno-keyboarding systems as well as different from the to-day experiences which aim to use short-forms of word processors for increasing keyboarding speed.

It is useless that I give you full details of this software, since the documentation here below gives a lot of information.

I would only underline the following points:

1 - this software is specifically foreseen for the Italian language
2 - the research showed that the number of strokes to write a text is about 46 % less then traditionally needed
3 -the saving in time for preparing a document can be considered arount 25 %
4 - the software uses of a word prediction tools, similar but much more accurate to those used in handy phones
5 - the software includes linguistic controls and rules so that for example the final vowels of a word are automatically inserted in a correct way
6 - text to speech technology helps user to understand better what he has to do.

Synthema is willing to exhibit and present this project at the next Intersteno Congress in Prague.

I think this is a very interesting project who can stimulate additional thoughts and is a sound result of connections created by Intersteno Congress with research and industries.

Gian Paolo Trivulzio

7 December 2006

I vantaggi della scrittura assistita con FastType

- **Risparmio del 48% delle battute di tastiera**
- **Risparmio del 25% nel tempo di composizione di un documento**

La finestra principale di Donley

- **Integrazione automatica ed istantanea con qualsiasi applicazione di videoscrittura sotto Windows.**
- **Interfacce semplificate per garantire la massima facilità d'uso e di configurazione**

Una collaborazione:

Synthema s.r.l. (Pisa)

Dipartimento di Informatica dell'Università di Pisa

Università di Pisa

Fondazione Cassa di Risparmio di Pisa

Synthema: via Malasoma 24, Ospedaletto (Pisa)
Dipartimento di Informatica: via F. Buonarroti 4, Pisa
Fondazione Cassa di Risparmio di Pisa: Lungarno Sonnino 20, Pisa

E-mail: rubino@di.unipi.it
aliprandi@synthema.it
p.mancarella@di.unipi.it

Scrittura Assistita

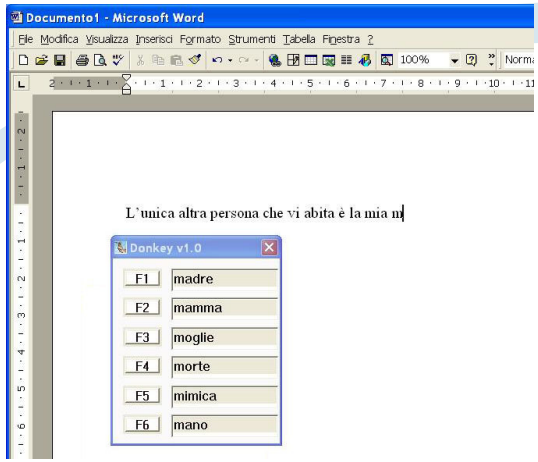
FastType

Il programma "intelligente" di predizione delle parole. Semplice e versatile

Completamento Automatico

La scrittura assistita con FastType

FastType è un sistema di scrittura assistita basato sulla sintassi. Il suo principale scopo è quello di rendere più semplice, rapida e corretta la scrittura per utenti con disabilità motorie o cognitive. FastType suggerisce all'utente una serie di completamenti per la parola che sta digitando e basta un clic o la



Durante la scrittura con Word, FastType completa la parola

semplice pressione di un tasto funzione per completare automaticamente una parola anche di 10 caratteri già dopo aver scritto i primi 2! FastType si compone di due parti: una libreria dinamica di predizione e un'interfaccia chiamata Donkey, entrambe progettate per funzionare sotto Windows.

Tecnologia: L'interfaccia

L'interfaccia utilizza le funzioni a basso livello di Windows per catturare le digitazioni dell'utente (e comunicarle alla libreria) e per completare automaticamente le parole. Questo gli permette di **funzionare con qualsiasi programma di videoscrittura per Windows**. L'interfaccia è in grado anche di assistere gli scrittori ipovedenti grazie ad effetti sonori e sintesi vocale compatibile con Microsoft SAPI, le più diffuse API di sintesi vocale.

Tecnologia: La libreria di predizione

La libreria di predizione è il cuore di FastType. Ogni volta che l'utente preme un tasto della tastiera, Donkey notifica la digitazione alla libreria, mettendo in moto i meccanismi di predizione della parola:

- Un Dizionario Generale per classificare le parole che l'utente va man mano digitando (o completando automaticamente grazie a FastType) ed assistere la predizione di parola basata sulla sintassi oltre a fornire all'utente i suggerimenti per il completamento automatico.
- Una Grammatica Generale per la lingua italiana, che **sceglie le parole da proporre rispettando le regole grammaticali e le concordanze morfosintattiche** rifacendosi agli schemi linguistici visti in fase di addestramento. Se ad esempio l'utente scrive "il mio g", verrà proposto "gatto" o "giovane", non "gatta" che è un femminile e nemmeno "gradisci" che è un verbo e non può mai comparire dopo un articolo e un aggettivo possessivo.
- Un efficace algoritmo di ranking che ordina le parole in base alla loro frequenza d'uso, alla correttezza grammaticale e alla concordanza morfosintattica, così da mettere **in cima alla lista di suggerimenti le parole effettivamente più probabili.**
- Un **Dizionario Personale con modulo di autoapprendimento**, che impara in poco tempo il lessico dell'utente per rendere la predizione ancora più sensibile alle necessità di chi scrive.

La libreria è completamente indipendente dall'interfaccia Donkey e perciò può essere incorporata in interfacce personalizzate (ad esempio tastiere virtuali) o in altri programmi di scrittura assistita per dare all'utente suggerimenti validi ed accurati che sfruttino le risorse linguistiche di FastType.

Synthema: via Malasoma 24, Ospedaletto (Pisa)
Dipartimento di Informatica: via F. Buonarroti 4, Pisa
Fondazione Cassa di Risparmio di Pisa: Lungarno Sonnino 20, Pisa

E-mail: rubino@di.unipi.it
aliprandi@synthema.it
p.mancarella@di.unipi.it

An Inflected-Sensitive Letter and Word Prediction System

Carlo Aliprandi[1], *Nicola Carmignani*[2] *and Paolo Mancarella*[2]

[1] Synthema Srl - Pisa, [2] Department of Computer Science - University of Pisa

Key words: *Natural Language Processing (NLP), Word Prediction, Assistive Technology, Alternative and Augmentative Communication, Computer Aided Language Learning.*

Abstract:

We present FastType, an innovative system for word and letter prediction for an inflected language, namely the Italian language. The system is based on combined statistical and lexical methods and it uses robust language resources. Word prediction is particularly useful to minimise keystrokes for users with special needs, and to reduce misspellings for users with linguistic difficulties. Word prediction can be effectively used in language learning, by suggesting correct and well-formed words to non-native users. This is significant, and particularly difficult to cope with, for inflected languages such as Italian, where the correct word form depends on the context. After describing the system, we evaluate its performances and, besides the high Keystrokes Saving, we show that FastType outclasses typical word prediction limitations getting outstanding results even over a very large dictionary of words.

1 Introduction

This paper introduces FastType, a word prediction system for the Italian language. At each keystroke, FastType suggests a list of meaningful predictions, among which hopefully the word the user is willing to type is included. Since the prediction is realized for incomplete sentences – unlike common text analysis which processes complete sentences – the amount of ambiguities (lexical, structural and semantic) is due to raise. FastType is then facing a very challenging Natural Language Processing (NLP) task.

FastType is particularly useful for users with motor impairments. The prediction of the most probable words helps to minimize the number of keystrokes needed to type a sentence and often becomes indispensable for users with speech and language disabilities or dyslexia: it has been proven that such systems assist users representing an Alternative and Augmentative Communication (AAC) technique, [6].

However, FastType has an additional didactic use: by suggesting correct and well formed words, it may simplify language learning, especially for non-native users. This is significant, and particularly difficult to cope with, for inflected languages such as Italian, where the correct form of a word depends on the context in which the word is included.

Usually, systems use word frequencies to suggest words that the user has started typing. Suggestions may often appear inaccurate since word frequencies do not take the sentence

structure into account. More sophisticated methods based on previous words (sentence context) or syntactic rules, permit to enhance the predictive accuracy. FastType, allows further Keystroke Saving providing orthographic and grammatical cues that respect the syntactic structure of the language. Details not presented here can be found in [5].

FastType is currently under development and receives support and funding from *Fondazione Cassa di Risparmio di Pisa*. FastType is the result of a research collaboration between Synthema and USID, the University of Pisa Disability Office.

The remainder of this paper is organized as follows. Section 2 outlines the description of some state of the art in research and commercial applications. In Section 3 we describe some background for word prediction. Section 4 introduces FastType and its architecture. In Sections 5 and 6 we describe in detail FastType components. In Section 7 we evaluate the performance of our system and present results. Finally, Section 8 presents our conclusions and ideas for future work.

2 Word Prediction – State of the Art

Word prediction is a research area where a very challenging and ambitious task is faced, basically with methods coming from Artificial Intelligence, Natural Language Processing and Machine Learning. The main goal of word prediction is guessing and completing what word a user is willing to type in order to facilitate and speed-up the text production process. Word predictors are intended to support writing and are commonly used in combination with assistive devices such as keyboards, virtual keyboards, touchpads and pointing devices.

Prediction methods have become quite known as largely adopted in mobile phones and PDAs, where multitap is the input method. Commercial systems as Tegic Communications T9[1], Zi Corporation eZiText[2], and Motorola Lexicus iTAP[3] are all successful systems that adopt a very simple method of prediction based on dictionary disambiguation. At each user keystroke the system selects the letter between the ones associated with the key guessing it from a dictionary of words: hence they are commonly referred to as letter predictors. Letter predictors bring a Keystroke Saving (KS) but it has been proven to be not completely free from ambiguities that are more frequent for inflected languages. So it is not surprising that these methods had a great success for non inflected languages such as English: the limited number of inflectional forms lead to very high KS that, at the moment, are above 40%.

Word prediction is a more sophisticated technique within recent flourishing research; once again most of word prediction methods apply to the English language. Differently from letter predictors, word predictors typically make use of language modelling techniques, namely stochastic models that are able to give context information in order to improve the prediction quality. Profet [4] is a statistically based adaptive word prediction program incorporating a lexicon of 10.000 words unigrams and a set of *n*-gram language models: a word is predicted and ranked depending on the last *n* typed words. Prophet II introduced Markov models also for word classes (word class unigrams, bigrams and trigrams) and, as shown in [10], KS improves on average by at least 22,7% due to grammatically ranked predictions.

This paper presents FastType, a system based on a novel prediction method for handling inflected languages, showing preliminary results of its application to the Italian language. Italian is a very morphologically rich language with a high rate of inflected forms that make it hard for any predictor to give accurate predictions. In literature several reviews have been presented and most of them assert that word prediction is not so effective for inflected languages as it is for non-inflected languages as English. In [8] several methods are

[1] http://www.t9.com.
[2] http://www.zicorp.com.
[3] http://www.motorola.com.

investigated, concluding that the high number of inflected forms make standard methods hardly applicable, especially when no syntactic information is available for the prediction engine. FastType deals with this issue by enriching the language model with deep morphological information and on-the-fly syntactic analysis to generate proper predictions. Another interesting research describes the application of word prediction to Russian, a language very rich in inflectional word forms: in [9] a morphological component in a two step procedure is used to compose inflections for root forms of verbs. FastType is able to predict correct inflected words for a broaden set of word classes.

A research approach close to the one presented in this paper is perhaps shown in [11], which employed syntactic information in the form of Part-of-Speech (POS) *n*-grams promising more effective predictions mostly with compound words for German language. Our approach introduces even more rich information to POS *n*-grams in the form of deep morpho-syntactic tags that we claim to be useful in order to cope with the very wide lexicon underlying FastType.

Dictionary coverage is another dominant factor affecting KS especially for inflected languages. Typically, for inflected languages reasonable KS is obtained limiting the number of words. In this work we outclass this limitation showing that a significant KS above 30% can be offered and that, surprisingly, slightly better results have been proven for a dictionary of about 1.200.000 words than for a limited dictionary of 250.000 words. Unlike existing state of the art related work, we employed a large morpho-syntactic tagged corpus to train the language model and a Part-of-Speech tagger that annotates on-the-fly words with their POS and related morpho-syntactic information.

In this paper we present a new approach to prediction, which constitutes a novel enhancement both in letter prediction and in word prediction: FastType performs combined letter and word prediction: the initial results are, in our opinion, interesting and quite promising, especially in reference to the dictionary size which is no more limited to few thousands words.

3 Background

In the literature, stochastic modelling strategies have been widely exploited in NLP tasks, for syntactic, morphological and even semantic analysis.

A basic assumption in word prediction is that contextual information affects the environment where the word has to be entered. In order to predict the most likely word it is necessary to have a high-order representation of the context. This assumption has been consolidated into stochastic methods that are based on Markov models and *n*-gram word models, [3] and [7].

The task of word prediction can be modelled as the estimation of the probability to guess the n^{th} word (w_n) given the current sequence of $n-1$ previous words ($w_1, w_2, ..., w_{n-1}$), denoted by

$$\Pr(w_n \mid w_1, w_2, \ldots, w_{n-1})$$

So, the probability of a sentence $W = w_1, \ldots, w_n$ is estimated using the Bayes rule as the product of conditional probabilities

$$\Pr(W) = \prod_{i=1}^{n} \Pr(w_i \mid w_1, w_2, \ldots, w_{i-1})$$

In practice, it is possible to incorporate some syntactic and semantic information due to the dependencies between words that are captured and estimated by *n*-grams words models; usually, as the vocabulary size is very large, it is necessary to approximate *n*-gram models to *unigrams* ($n = 1$), *bigrams* ($n = 2$) and *trigrams* ($n = 3$):

$$\Pr(w_i \mid w_1, \ldots, w_{i-1}) \approx \Pr(w_i \mid w_{i-n+1}, \ldots, w_{i-1})$$

The *n*-gram models of words (particularly, for $n = 3$) have been used successfully for many NLP tasks but they suffer the well-known drawback of being inadequate for inflected languages [8], since the parameters space becomes too wide, both for the vocabulary size and the training corpus. Thus different new models where introduced with the purpose of reducing the parameters space. These models generalize the *n*-gram model making use of *n*-grams of Part-of-Speech: the context is forced into an equivalence class determined by a function φ

$$\Pr(w_i \mid \varphi[w_{i-n+1}, \ldots, w_{i-1}])$$

Part-of-Speech tags are considered as function φ to restrict exponential increase of the context. Such tags capture many different word forms, so contextual dependencies are represented in smaller set of *n*-grams. Using POS tags, a larger surrounding information may be taken into consideration but there is a loss in semantics since different words may be captured in one word class and tags only inform about sequences of words classes and not which particular word is typically connected with previous words or words classes.

4 Architecture of the Word Predictor

A word prediction environment, Assistive Writing Environment (AWE), was set up in [2] and a prototype to asses its potential for Italian was developed in [1]. FastType, the new version of the word predictor, introduces improvements, namely to the revisited modular architecture, to the linguistic resources and to the prediction algorithm. FastType prediction is now more accurate and its underlying components are more flexible. We committed some efforts in separating the core system in different adaptable modules, in order to make the whole system more usable both as a standalone system and as a component to be integrated into an external system, e.g. a virtual keyboard or a word processor. Consequently some parts of the system, as the kernel Prediction Engine, have been re-implemented; some others, as the User Interface, have been separated and, for the time being, kept apart for future improvements.

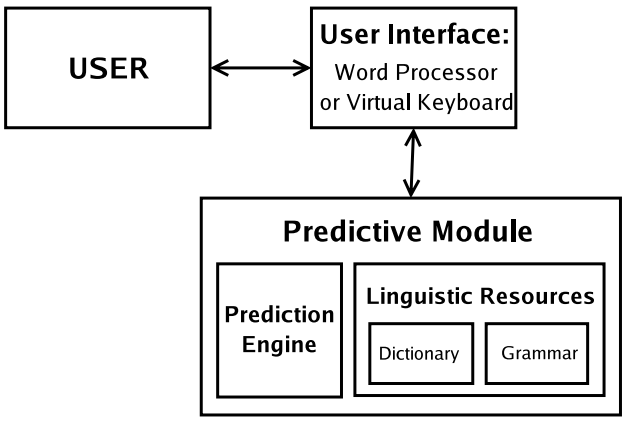


Figure 1: FastType Architecture

As shown in Figure 1, FastType has three main components: the *User Interface*, the *Predictive Module* and the *Linguistic Resources*. As the User Interface is not object of research at this stage of the project, we opted to improve the simple custom word processor developed in [1], that was also used for testing and assessing the results here presented.

The *Prediction Engine* is the kernel of the *Predictive Module*; it manages the communication to and from the User Interface, keeping trace of the prediction status and of typed words. At each keystroke it predicts suggestions, in the form of list of word completions and of list of next-letters, by assuring accordance (gender, number, person, tense and mood) with the syntactic sentence context. The *Predictive Module* functionalities, such as the morpho-syntactic agreement and the lexicon coverage, are provided by the statistic language model based on POS tag *n*-grams and on morphological information provided by the *Linguistic Resources*. This module can also optionally learn the user writing style, mainly tuning a personal Dictionary and a personal Grammar. Thus prediction is adapted to the user producing more effective results.

5 Linguistic Resources

In order to make predictions, FastType has to realize the preceding context of the current word, so a language model is needed. We studied and elaborated a complex framework for language modelling in AWE [2], that was used to develop and tune the *Linguistic Resources* underlying FastType.

The FastType language model has been trained from a large tagged corpus to acquire context knowledge in sentence construction. As models relying on *n*-grams of word forms have been proven to be not sufficient for inflected languages, we concentrated the modelling process on POS *n*-grams, like (NOUN, VERB, VERB), that can capture a sentence as "*sport può aiutare*" ("*sport can help*").

The Italian POS *n*-grams, approximated to $n = 2$ (bigrams) and $n = 3$ (trigrams), have been trained from a large balanced corpus (approximately 2.000.000 word forms) created from newspapers, magazines, documents, commercial letters and emails. The corpus was cleaned, standardised (punctuations, capitalisations) and then parsed using the Italian POS tagger Synthema Lexical Parser (SLP), a rule-based parser (see Lexical Data Base Management System – LDBMS [12]). The result was a corpus tagged with syntactic and morphological knowledge, that was automatically extracted from Italian dictionaries available in LDBMS. An example of such a tagged corpus is shown in Figure 2. Each word in the corpus is tagged: as an example **`lo[RDMS(lo)]`** means that "*lo*" is a "definite article" (**RD**), "masculine" (**M**) and "singular" (**S**). A subset of the Tag Set is shown in Table 1.

POS	Tag	Morpho-Syntactic Features
Verb	V	Transitive (T), Intransitive (I), Modal (M), …
Noun	S	Gender (F or M), Number (S or P)
Def. Article	RD	Gender (F or M), Number (S or P)
Indef. Article	RI	Gender (F or M), Number (S or P)
Preposition	P	Simple (S), Compound (A), …
Adjective	G	Possessive (G), Qualifier (E), …

Table 1: A subset of the Tag Set

To grant a high lexical coverage, predicted words are extracted from tagged lexicon dictionaries. The Prediction Engine relies on a large Italian dictionary, comprising 43.000 lemmas, 876.000 forms and 1.165.000 tagged forms.

Since a word form may be ambiguous, i.e. it may adhere to more than one Part-of-Speech, and mistakes can be done by the SLP POS tagger, a partial hand-review was necessary. The

tagged corpus was then used to train bigrams and trigrams, using Perl scripts; *n*-grams falling below a threshold were discarded, thus reducing sparse data. The language model was eventually loaded into the LDBMS via a native implementation, the result being a very compact Lexical Data Base quickly accessible by the Prediction Engine via LDBM primitive APIs.

```
Lo[RDMS(lo)] sport[SCMI(sport)] può[VMN3IN(potere)]
aiutare[VTNIFN(aiutare)] in[PS(in)] tanti[GIMP(tanto)]
modi[SCMP(modo)] a[PS(a)] rendere[VBNIFN(rendere)]
la[RDFS(lo)] vita[SCFS(vita)] più[AC(molto)]
sopportabile[GENS(sopportabile)] e[CC(e)]
accettabile[GENS(accettabile)] "[@H(")] .[@E(.)]
```

Figure 2: Sample of the Tagged Corpus

We remark that, as the corpus has been tagged with deep morpho-syntactic knowledge, the language model is highly descriptive, much more than standard POS models (as for example the one employed in [10]).

The main benefit is that FastType can better cope with a very inflected language as Italian, producing similar and even better results than word predictors for non inflected languages.

6 The Prediction Algorithm

By modelling the language with Part-of-Speech categories, the system predicts the next POS tag to be produced in the current sentence and narrows the amount of words and single letters when each letter of the word is entered. FastType employs a second order Markov model to take into account the conditional probability of the next POS tag given the probability estimation of the two previous POS tags: the Prediction Algorithm produces inflected suggestions in accordance with the context.

Suppose the following sequence of words has been typed so far:

$$\ldots\ w_{n-2}\ \ w_{n-1}\ \ p_n$$

where w_{n-2} and w_{n-1} are completed words, and p_n is a partially typed word (*prefix*). If W is the set of all the words in the dictionary beginning with p_n, the FastType Prediction Algorithm selects W_{best}, a set of the most appropriate words from W that are in accordance with the context. Moreover, if L is the set of all the letters, the FastType Prediction Algorithm selects L_{best}, the subset of next-letters from L that fit with p_n.

As discussed in the previous sections, Italian is a very inflected language, with many variations and morphological forms that make it difficult to predict appropriate words. The main idea is to parse the sentence and, applying a context-aware sieve, to provide only inflected completions in accordance with the context for p_n. Specifically, as shown in Figure 3, given a candidate Part-of-Speech POS_n, p_n completions are extracted from the dictionary and filtered by the Morpho-syntactic Sieve. Words which do not agree with the candidate POS_n are discarded, the others are on-the-fly inflected and ranked into the list of Inflected Suggestions.

When a word is completed, either by selecting it from the suggestion list (W_{best}) or by typing it, the Prediction Algorithm classifies it (using the Synthema Lexical Classifier – SLC), producing a new POS unigram for the language model.

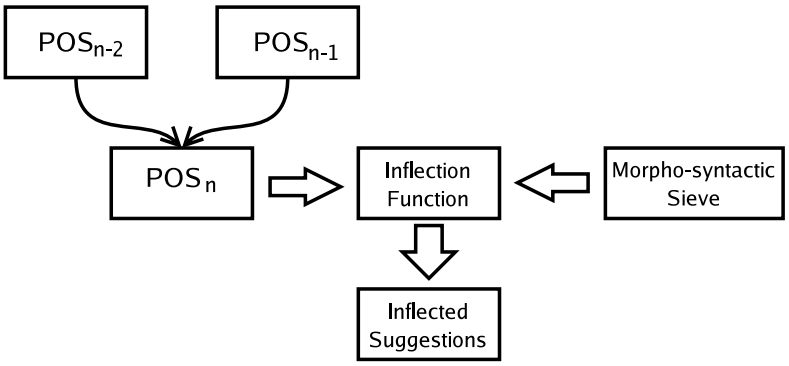


Figure 3: The Prediction Process of FastType

Suppose, for example, the user is willing to type "*la ragazza scrive*" ("*the girl writes*"). By typing "*l*", *la* is the first word in the suggestion list. Then, the Prediction Algorithm classifies it, producing a new unigram corresponding to the Italian definite feminine singular article (**RDFS**, according to the notation in Table 1). As the next predicted POS tag is NOUN, every suggestion will be only a feminine singular noun. So, typing the user "*rag*", "*ragazza*" will be one of the suggestions, since it is the singular form (while "*ragazze*" is the plural one): morpho-syntactic information of the context allows the Inflection Function to properly inflect the word baseform (or lemma) "*ragazzo*". Notice that "*ragazzo*" corresponds also to the masculine noun, but it is not suggested by the Prediction Module, since the expected POS is feminine singular. The next predicted POS tag is VERB, so the Inflection Function can properly apply the verb inflectional paradigm producing the 3rd singular person: typing "*sc*", "*scrive*" is a correct element in the suggestion list (while "*scriviamo*", the 1st plural person, is not).

7 Evaluation and Results

It is difficult to evaluate a word prediction system, in particular since varying user difficulties, a specific metric may be more appropriate than another one. Since subjective estimations cannot be effectuated, we had to perform a general system evaluation.

The system evaluation has been performed by using the well-known metric for word prediction, *Keystroke Saving* (KS), which estimates the saved effort percentage. KS is calculated by comparing two kinds of measures: the total number of keystrokes needed to type the text (K_T) and the effective number of keystrokes using word prediction (K_E). Hence,

$$KS = \frac{K_T - K_E}{K_T} \times 100$$

Keystroke Saving is affected by different parameters, as the training set and the test set features, the lexicon coverage, the prediction speed, the number of words in the suggestion list, the method for selecting prediction cues and user interface, and others as described in [4].

A cognitive load is added since users have to shift their gaze from the keyboard to the screen and back, then they have to choose the right suggestion from the list. This drawback is highly connected to the specific user interface and, for this reason, the corresponding load has not been taken into consideration at this stage of the project.

The selection of the training set and of the test set is a key factor in evaluating the system efficiency. To build the training set we used a balanced tagged corpus (approximately 2.000.000 word forms) of Italian texts from newspapers, magazines, documents, commercial letters and emails. We selected a disjoint subset of 40 balanced texts from the training set as the test set.

To measure FastType performances in relation to the lexicon coverage we run two trials on different dictionaries: an Italian General dictionary and a smaller Italian Most Frequent Words dictionary, consisting of 1.165.000 and 245.000 words. Our test bench produced an average KS of 30% using the General dictionary and an average KS of 20% using the Most Frequent Words dictionary. This result, that was at first glance surprising, is mostly due to the effect of unknown words on typing and, especially, on the Prediction Engine. It is of course clear that unknown words imply more typing, but also that they affect the Prediction Algorithm, since the corresponding POS tags cannot be produced. As a consequence, the language model looses its efficacy and the suggested words may be incorrect, thus limiting the KS. Nevertheless, for the Most Frequent Words dictionary, less keystrokes are needed for known words, bringing a higher prediction speed.

In our trials, each of the 40 texts was typed using the word processor developed in [1] and detailed logs were produced, recording among others:

- the number of keystrokes,
- the total number of characters,
- the number of words.

Figure 4 shows the results for the General dictionary; FastType reduces the average number of keystrokes per text from 315 to 222, resulting in an average 30% KS. Top results are achieved with KS peaks of 40% to 43%; the best outcome, reducing the number of keystrokes from 218 to 124, is presented in Figure 5 (where the underlined text represents saved keystrokes).

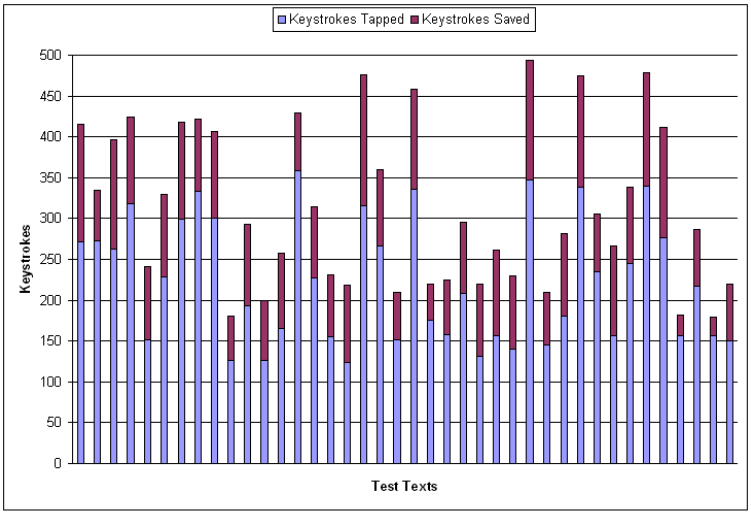


Figure 4: Results for the test set using the General dictionary

It has been proven in the literature that a dominant factor affecting KS is the dictionary coverage and that, typically, for inflected languages a reasonable KS is obtained limiting the number of words. We believe that our results outclass these limitations showing that the FastType approach can offer a significant KS for inflected languages and, even more significantly, better results can be provided with a very large dictionary of words than with a limited dictionary. Unlike existing state of the art related work, we employed a large morpho-syntactic tagged corpus to train the language model and a POS tagger that annotates words on-the-fly with morpho-syntactic information.

Credeva di compiere un gesto di solidarietà ed ha invitato a cena due giovani vagabondi. Ma il gesto si è trasformato in un incubo: i soggetti le hanno somministrato del sonnifero e poi hanno derubato l'appartamento.

Figure 5: Sample of a test text

8 Conclusions

We described a combined statistical and lexical word prediction system for handling inflected languages. By making use of POS tags we built a language model and an innovative prediction method that resulted in a Keystroke Saving comparable to the one achieved with standard methods for non inflected language, thus outclassing some typical word prediction limitations.

Our conclusions are consistent with state of the art literature, for example with [8], who claimed that a word prediction method without syntactic information are not applicable to inflected languages; we additionally enriched the language model with morpho-syntactic information and provided the prediction method with an on-the-fly Part-of-Speech word tagger and large lexicon dictionaries.

We showed that FastType reduces the number of keystrokes, achieving KS peaks of 40% to 43% and an average KS of 30% on the balanced test set. FastType performances are above all significant when measured in relation to the lexicon: better results are provided with a very large dictionary of words than with a limited dictionary.

For future work we have plans for building a new User Interface, more simple and friendly, which may be used by disabled people as a plug-in for any existing application. The prediction method is also currently under improvement, as new language resources are being explored. We are willing to improve language models integrating also *n*-grams of words that will help in ranking word completions. We also intend to investigate the system adaptability to the user style, enriching FastType with a personal language model and a personal dictionary. Combining existing methods with the user grammar and lexicon, we expect to obtain, after a starting training period, a more effective prediction user by user.

In conclusion, FastType has peculiarities and potential advantages since, using very large lexical resources and statistically based techniques, an effective word prediction can be performed in real domains. We claim that, apart from the Assistive domain we have focussed on, the application of this technology are wide, bringing the benefits of fast text typing to particular devices as virtual keyboards, smart phones and PDAs.

Acknowledgements:

We acknowledge the financial support of the *Fondazione Cassa di Risparmio di Pisa*, partially funding the FastType Project.

References:

[1] Aliprandi C.; Barsocchi D.; Fanciulli F.; Mancarella P.; Pupillo D.; Raffaelli R.; Scudellari C.: *AWE, an Innovative Writing Prediction Environment*, HCI International 2003 – 10th International Conference on Human-Computer Interaction, (Adjunct Proceedings), pp. 237–238, Greece, 2003.

[2] Barsocchi D.: Disabilità, Informatica, Linguistica: un'istanza del trinomio, Master's Thesis in Computer Science, Department of Computer Science, University of Pisa, 2002 (In Italian).

[3] Brown P.; Della Pietra P.; DeSouza P.; Lai J.; Mercer R.: *Class-based n-gram Models of Natural Language*, Computational Linguistics, 18, pp. 467–479, 1990.

[4] Carlberger A.; Carlberger J.; Magnuson T.; Hunnicutt M. S.; Palazuelos-Cagigas S.; Navarro S. A.: *Profet, A New Generation of Word Prediction: An Evaluation Study*, Proceedings of the ACL Workshop on Natural Language Processing for Communication Aids, Madrid, Spain, pp. 23–28, 1997.

[5] Carmignani N.: *A Word Prediction System for People with Disabilities based on Part-of-Speech Tagging*, Master's Thesis in Computer Science, Department of Computer Science, University of Pisa, 2005.

[6] Copestake A.: *Augmented and Alternative NLP Techniques for Augmentative and Alternative Communication*, Proceedings of the ACL Workshop on Natural Language Processing for Communication Aids, Madrid, Spain, pp. 37–42, 1997.

[7] Even-Zohar Y.; Roth D.: *A Classification Approach to Word Prediction*, NAACL-2000, The 1st North American Conference on Computational Linguistics, pp. 124–131, 2000.

[8] Garay-Vitoria N.; Abascal J.: *Word Prediction for Inflected Languages. Application to Basque Language*, Proceedings of the ACL Workshop on Natural Language Processing for Communication Aids, Madrid, Spain, pp. 29–36, 1997.

[9] Hunnicutt S.; Nozadze L.; Chikoidze G.: *Russian Word Prediction with Morphological Support*, 5th International Symposium on Language, Logic and Computation, Tbilisi, Georgia, 2003.

[10] Magnuson T.; Hunnicutt S.: *Measuring the Effectiveness of Word Prediction: The Advantage of Long-Term Use*, Technical Report TMH-QPSR Volume 43, Speech, Music and Hearing, KTH, Stockholm, Sweden, 2002.

[11] Matiasek J.; Baroni M.; H. Trost: *FASTY: A Multi-Lingual Approach to Text Prediction*, In Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler (eds.), Proceedings of the 8th International Conference on Computers Helping People with Special Needs (ICCHP), Dordrecht: Springer, pp. 243–250, 2002.

[12] Raffaelli R.: *Lexical Data Base Management System – LDBMS*, Synthema Internal Report, Pisa, 2000.

Authors:

Carlo Aliprandi, Dr.
Synthema Srl
Via Malasoma 24, Pisa
carlo.aliprandi@synthema.it

Nicola Carmignani, Dr. PhD student
Department of Computer Science,
University of Pisa,
Largo B. Pontecorvo 3, Pisa
carmigna@di.unipi.it

Paolo Mancarella, Prof.
Department of Computer Science,
University of Pisa,
Largo B. Pontecorvo 3, Pisa
paolo@di.unipi.it