International Journal of Advanced Robotic Systems

OPEN ACCESS
ARTICLE

On Combining Language Models to Improve a Text-based Human-machine Interface

Regular Paper

Daniel Cruz Cavalieri[1]*, Teodiano Bastos-Filho[1], Sira Elena Palazuelos-Cagigas[2] and Mario Sarcinelli-Filho[1]

1 Department of Electrical Engineering, Federal University of Espirito Santo, Vitoria, Brazil
2 Department of Eletronics, University of Alcalá, Madrid, Spain
*Corresponding author(s) E-mail: dcruzcavalieri@gmail.com

Received 05 July 2014; Accepted 14 October 2015

DOI: 10.5772/61753

Abstract

This paper concentrates on improving a text-based human-machine interface integrated into a robotic wheelchair. Since word prediction is one of the most common methods used in such systems, the goal of this work is to improve the results using this specific module. For this, an exponential interpolation language model (LM) is considered. First, a model based on partial differential equations is proposed; with the appropriate initial conditions, we are able to design a interpolation language model that merges a word-based *n*-gram language model and a part-of-speech-based language model. Improvements in keystroke saving (KSS) and perplexity (PP) over the word-based *n*-gram language model and two other traditional interpolation models are obtained, considering two different task domains and three different languages. The proposed interpolation model also provides additional improvements over the hit rate (HR) parameter.

Keywords Human-machine Interfaces, Word Prediction Systems, Language Modelling, Communication Aid

1. Introduction

One of the major faculties that most people with severe disabilities lose is the ability to speak. The ability to participate in complex communication enables the exchange of ideas and concepts, as well as helping in social integration. Once this connection between the mind and the outside world is broken, frustration, loneliness and a lack of confidence will inevitably be felt.

Besides speech problems, these people may have other disabilities that affect their motor skills. This can make the communication process slow and challenging, often requiring the use of specialized keyboards or other input devices.

One alternative for people with disabilities is the use of augmentative and alternative communication (AAC) devices within the human-machine interface (HMI) field. According to [1], ACC refers to the use of methods or devices to supplement the communication skills of a person with disabilities, and can be a dedicated device and/or a computing solution with simple output sound messages,

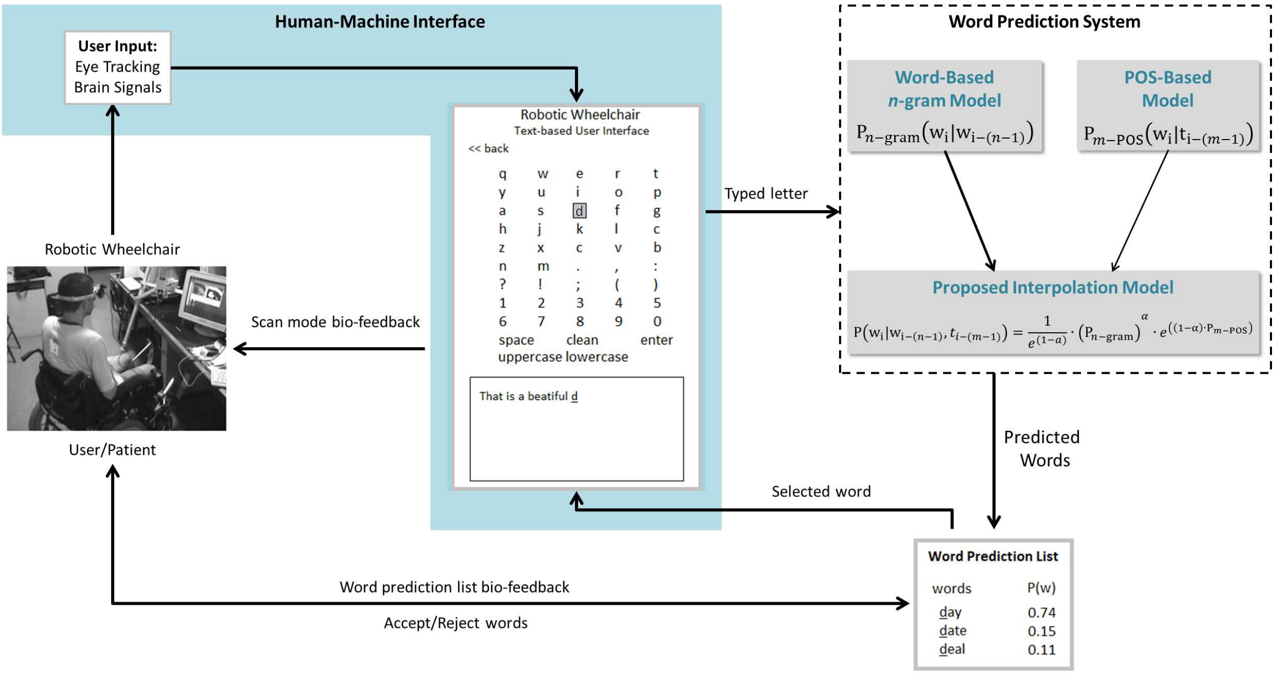

Figure 1. General architecture of the HMI with a text-based user interface within a robotic wheelchair

or more complex systems, using techniques such as eye tracking and brain signals [2, 3]. In this context, the use of word prediction systems (WPS) in the communication process, such as automatic speech recognition (ASR) [4, 5], becomes essential.

Formally, WPSs were developed as a communication aid method, in order to increase message composition rate for people with severe motor and speech disabilities [6, 7]. Nowadays, text prediction methods – if adequately integrated within user interfaces – can benefit anyone trying to produce text messages or commands [8]. Generally, "word prediction" refers to those systems that guess which letters, words, or phrases are likely to follow a given segment of a text [9]. In [10] and [11], the WPS is considered an important element within the context of natural language processing (NLP), whereby the correct word is predicted given a particular context. In all these cases, the main goal of these systems is to improve the keystrokes saved (KSS) value, which is the percentage of keystrokes that the user saves by using WPS, while ensuring a good quality of produced text. Thus, this work focuses on the improvement of this aspect in order to achieve a better performance in a text-based user interface located within a robotic wheelchair. To do this, a WPS is considered. The WPS, thoroughly described in [12], aims to enable people to interact with other people, using a written language in a natural way. Figure 1 shows the general architecture of the HMI with a text-based model incorporated.

The proposed system shown in Figure 1, as with the work developed by [2] and [13] to control a robotic wheelchair, uses myoelectrical eye blinks, iris-tracking and a brain-computer interface to choose letters in a personal digital assistant (PDA). The PDA provides a graphic interface containing possible letters and actions. Once a specific letter has been selected using a scan mode, the WPS displays a number of possible words. Thus the user only has to accept – or reject – the suggested words. This kind of interface decreases the effort of the user when writing the text, delegating the writing effort to the WPS.

Since the WPS plays a fundamental role in the proposed system, improvements should be made in this direction. In this regard, we are proposing an interpolation language model (LM), based on the exponential combination of a word-based n-gram language model and a part-of-speech-based (POS) language model. To corroborate the methodology, results obtained by the proposed interpolation model were compared with the linear and geometric interpolations tested in three different languages: English, Portuguese and Spanish.

The rest of the paper is organized as follows: Section 2 presents a brief overview of the HMI within the robotic wheelchair; Section 3 gives a general overview of the word-based model and the POS-based model, as well as introducing our proposed interpolation method for combining these language models; Section 4 reports the outcomes of the experimental evaluations conducted using a WPS in English, Portuguese and Spanish; and finally, Section 5 gives conclusions and outlines for future research.

2. Overview of the HMI System

The main objective of the HMI is for the user/patient to be capable of writing a text by means of his/her biological signals. In this work, the robotic wheelchair – developed at the Federal University of Espírito Santo (Brazil) – is able to use different HMIs, such as eye tracking, brain signals and sip-and-puff. All these HMIs involve acquisition systems, which include the amplification, filtering, digitization, recording and processing of the different kinds of signals provided by the wheelchair user [3]. The signals are recorded and classified, sending the identified command to a PDA. This PDA can be used to control the robotic wheelchair or, as in our case, write a text. Once a valid command is identified, a voice player confirms the option chosen, providing feedback to the user, as well as allowing communication with the people around.

3. Language Models to Word Prediction

There are several WPSs that have been and are being developed using distinct methods for different languages [9, 14]. Traditionally, these systems have been based on statistical n-gram language modelling. Recently, more sophisticated language models have been developed in order to improve the performance of these traditional language models [15]. In many cases, these language models explore and capture separately specific phenomena of natural language. Here, a question naturally arises regarding how to build more powerful and complex language models, capable of integrating all language components (such as syntactic, semantic and morphological structures).

To answer this question, the most efficient method would be to combine them in some optimal sense [16]. A simple method that can combine a broad range of models is that of linear interpolation (Equation 1), which takes into account a weighted sum of the probabilities given by the component language models. Normally, it is used to add a part-of-speech (POS) cache-component to a word-based n-gram model [17], taking into account the semantic structure of the language [18], or both POS and semantic structures [19]. Nonetheless, according to [20], even if the perplexity of the linear combined model is minimized, this type of methodology does not guarantee optimal use of the different information sources. This way, [21] proposes a method based on the latent maximum entropy principle, which extends the basic principle of maximum entropy proposed in [22], incorporating a hidden dependence structure. Distinct from linear interpolation, this approach generates probabilistic models capable of capturing all the information from the different sources, but with computational limitations in the estimation of the model parameters. In order to improve this, [21] has also proposed a methodology based on directed Markov random fields, first developed in [23]. With this model, the authors were able to combine a word trigram model, a probabilistic model based on context-free grammar, and a probabilistic model based on latent semantic analysis.

$$
\begin{aligned}
P_{interpolation}^{(linear)} & \left(w_i \mid w_{i-(n-1)}, t_{i-(m-1)}\right) = \\
& \alpha \cdot \left[P_{n-gram}\left(w_i \mid w_{i-(n-1)}\right)\right] + \\
& +(1-\alpha) \cdot \left[P_{m-POS}\left(w_i \mid t_{i-(m-1)}\right)\right]
\end{aligned}
\tag{1}
$$

As can be seen, the above-mentioned language models reach a high level of mathematical (and computational) complexity, despite the efforts exerted to minimize it. Such models are widely used in applications such as automatic speech recognition (ASR) and machine translations (MT). However, when dealing with word prediction, it is recognized that the syntactic structure of the language plays a key role – in many cases, a primary one – in the composition of the language model. According to [24], the problem of word prediction in English and similar languages can be seen as a combination of two problems: the prediction of *function words* on the one hand, and *content words* on the other. Function words are the words used to make sentences grammatically correct. Pronouns, determiners, prepositions, and auxiliary verbs are examples of function words. Content words are words such as nouns, most verbs, adjectives, and adverbs, which refer to some object, action, or other non-linguistic meaning. In [24], it has also been shown that the performance of word prediction systems is mainly correlated with the ability to predict very common words, i.e., function words. In this context, word-based n-gram models are very effective. To predict content words, topic-based or POS-based language models can be used.

Thus, this work is motivated by the assumption that n-gram models could be more effective in WPSs, performing in combination with POS-based language models, and therefore proposes a novel exponential approach, based theoretically on partial differential equations as a way of combining them. As in many natural processes, once the differential equations that characterize a particular system have been determined, it is possible to extract relevant information about them. Figure 1 gives a general overview of the proposed methodology.

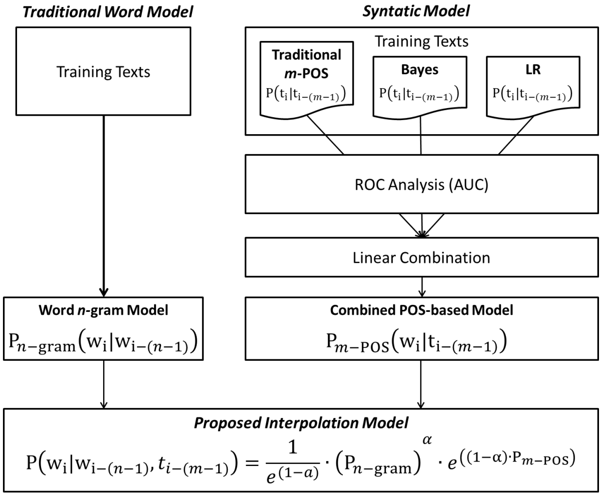

Figure 2. General overview of the proposed methodology

3.1 Word-Based Language Model

In word prediction, a statistical language model tries to predict the next word based on the history of previously used words. This idea of word prediction is formalized by probabilistic models called *n*-gram models, which in turn predict the next word from the *n*-1 previous words. In its simplest version, the unigram model only considers the absolute frequency of the word. When using this model, at each moment the most frequent words that begin with the written letters of the word in progress are predicted. By considering the sequence of words, or the probability that each word will follow the previous words, there arises bigram, trigram,..., *n*-gram models. In other words, suppose a sentence in which the sequence of words given so far is

$$w_{i-(n-1)} \cdots \; w_{i-2} \;\; w_{i-1} \;\; cw_i,$$

where w_{i-n-1} are the $n-1$ previous words and cw_i is the current word prefix typed by the user. In this case, the bigram model ($n=2$) is given by

$$P_{bigram}\left(w_i \mid w_{i-1}\right) = \frac{F\left(w_{i-1}w_i\right)}{\sum_{cw_i} F\left(w_{i-1}w_i\right)}. \tag{2}$$

This equation can be simplified, since the sum of all the bigrams beginning with the w_{n-1} must be equal to the count of the word unigram. Therefore,

$$P_{bigram}(w_i \mid w_{i-1}) = \frac{F(w_{i-1}w_i)}{F(w_{i-1})}, \tag{3}$$

which can be easily extended to the *n*-gram model, or

$$P_{n-gram}(w_i \mid w_{i-(n-1)}) = \frac{F(w_{i-(n-1)} \ldots w_i)}{F(w_{i-(n-1)} \ldots w_{i-1})}, \tag{4}$$

where $F(w_{i-(n-1)} \ldots w_i)$ and $F(w_{i-(n-1)} \cdots w_{i-1})$ are the frequencies of the *n*th and $(n-1)$ th previous word sequences, respectively.

According to [25], some of the disadvantages of the word-based *n*-gram language model include its large number of parameters and its high dependence on the discourse domain, since it measures perplexity on a set of different texts belonging to (or outside of) the linguistic domain of the training corpus. As described by [25], a solution for overcoming the data sparseness problem and reducing the dependence on the discourse domain might consist of grouping words together into equivalence word classes (or POS in our case), instead of those of individual words. In this context, before detailing the POS-based language model itself, it is necessary to (briefly) describe the technique used to determine the POS tagset employed in this paper.

3.2 POS-Based Language Model

As shown in [26], the use of certain major POSs (noun, verb, adjective, etc) – along with inflections like gender (masculine, feminine, neuter), number (singular, plural, neuter) and person (1st, 2nd, 3rd, 1st/3rd) – can generate accurate POS-based word predictors with a relatively low-speed list of predicted words. Thus, an initial POS tagset was first derived by selecting the most functional POS tags corresponding to English, Spanish and Portuguese. Even though relatively low numbers of POS tags were chosen, this work also used the methodology developed in [27] to further reduce the number of POS tags in Spanish and Portuguese. Table 1 shows the POS tagset and the morphological analyser used for each language.

	Portuguese	**Spanish**	**English**
Morphological analyser	PALAVRAS[28]	HISPAL[29]	ENGCG[30]
Initial number of POS tags	259	282	129
Set of functional POS tags	71	82	54
Reduced POS tagset using [27]	66	66	54

Table 1. POS tagset for Portuguese, Spanish and English

As in [27], a syntactic predictor has access to the following sequence of words and POS tags to predict the current word:

$$\cdots \; w_{i-2} \,/\, t_{i-2} \;\; w_{i-1} \,/\, t_{i-1} \;\; cw_i,$$

where t_{i-2} and t_{i-1} are the POS tags of the previous words w_{i-2} and w_{i-1}, respectively, and cw_i is the current word prefix typed by the user. The algorithm predicts words starting with cw_i.

According to [1], there are different methods for incorporating the statistical POS tag information into the word predictor. As in [31], the *m*-POS predictor (with $m=2$) was here estimated by

$$\begin{aligned} P_{2-POS}\left(w_i \mid t_{i-1}\right) = & \\ & \sum_{\forall t_{i-1}^r \in T(w_{i-1})} \sum_{\forall t_i^s \in T(w_i)} P\left(t_i^s \mid t_{i-1}^r\right) \cdot \\ & \cdot P\left(w_i \mid t_i^s\right) \cdot P\left(t_{i-1}^r \mid w_{i-1}\right), \end{aligned} \tag{5}$$

where t_i^s is the sth tag for w_i, which varies from 1 to $\mid T(w_i) \mid$; $T(w_i)$ is the set of all possible POS tags that may be assigned to the word w_i ; $P(t_i^s \mid t_{i-1}^r)$ is the bigram POS tag probability (the probability of t_i^s being t_{i-1}^r the tag of the previous word); $P(w_i \mid t_i^s)$ is the conditional probability of the word w_i given t_i^s as its POS tag; and $P(t_{i-1}^r \mid w_{i-1})$ is the conditional probability of the previous word w_{i-1} to be tagged with its rth tag t_{i-1}^r.

This method can be extended to include in the prediction as many previous words as desired. The current system considers a maximum of two previous words in the prediction ($3-POS$ model), or

$$
\begin{aligned}
&P_{3-POS}\left(t_i \mid w_{i-1} t_{i-2}\right) = \\
&\sum_{t_{i-2}^p \in T(w_{i-2})} \sum_{t_{i-1}^r \in T(w_{i-1})} \sum_{t_i^s \in T(w_i)} P\left(t_i^s \mid t_{i-1}^r t_{i-2}^p\right) \cdot \\
&\cdot P\left(w_i \mid t_i^s\right) \cdot P\left(t_{i-1}^r \mid w_{i-1}\right) \cdot P\left(t_{i-2}^p \mid w_{i-2}\right).
\end{aligned} \tag{6}
$$

There is an important difference concerning the POS-based language model used within this paper and the ones used in works like [1]. The difference lies in the calculation of the POS tag probabilities $P(t_i^s \mid t_{i-1}^r)$ and $P(t_i^s \mid t_{i-1}^r t_{i-2}^p)$. Traditionally, these probabilities have been calculated using the frequencies of the previous word classes, or

$$
P_{m-POS}\left(t_i \mid t_{i-(m-1)}\right) = \frac{F\left(t_{i-(m-1)} \cdots t_i\right)}{F\left(t_{i-(m-1)} \cdots t_{i-1}\right)}. \tag{7}
$$

In this work, a traditional statistical POS-based language model (Equation 7), a logistic regression (LR) POS-based language model and a naive Bayes (NB) POS-based language model [32] using the area under the ROC curve (AUC) are combined.

The ROC curve[1] is a technique for visualizing, organizing and selecting classifiers based on their performance. According to [34], in order to compare classifiers, it is possible to reduce the ROC performance to a single scalar value representing the expected performance. A common method is to calculate the AUC, which has an important statistical property: the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [35].

The AUC is usually estimated in the same way as the error rate. To classify the accuracy of a classifier using this measure, the following equivalence is adopted:

- 90% to 100%: excellent;
- 80% to 90%: very good;
- 70% to 80%: good;
- 60% to 70%: fair;
- 50% to 60%: poor;
- <50%: fail.

In this way, a classifier that obtains an AUC of 86%, for example, will be considered a very good classifier, with a score of $\beta=0.86$. In the same way, if a classifier obtains an AUC of less than 50%, it will fail and the score will be approximated to $\beta=0$.

ROC analysis and the AUC are commonly employed in two-class problems; with more than two classes, as in our case, the situation becomes much more complex [33]. For handling this problem, different AUCs were calculated, one for each POS tag, using the *one-against-all* method. Thus, for a set of $\mid T(w_i) \mid$ POS tags, one can have $\beta_{(.)}=[\beta_1,\beta_2,\ldots\beta_{\mid T(w_i) \mid}]$ and the combined m-POS language model given by

$$
\begin{aligned}
&P_{m-POS}^{(combined)} = \\
&\beta_{(1)} \cdot P_{m-POS}^{(1)} + \beta_{(2)} \cdot P_{m-POS}^{(2)} + \beta_{(3)} \cdot P_{m-POS}^{(3)},
\end{aligned} \tag{8}
$$

where (1) represents the traditional statistical POS-based language model, (2) represents the LR POS-based language model, and (3) represents the NB POS-based language model.

3.3 Proposed Method

3.3.1 Mathematical Formulation

Before presenting the interpolation model proposed in this paper, the work presented in [36], which uses a model based on geometric interpolation (Equation 9), is discussed in detail, to better explain the relationship between the independent models and the mathematical model merging them.

$$
\begin{aligned}
&P_{interpolation}^{(geometric)}\left(w_i \mid w_{i-(n-1)}, t_{i-(m-1)}\right) = \\
&\left[P_{n-gram}\left(w_i \mid w_{i-(n-1)}\right)\right]^{\alpha} \cdot \\
&\cdot \left[P_{m-POS}\left(w_i \mid t_{i-(m-1)}\right)\right]^{(1-\alpha)}
\end{aligned} \tag{9}
$$

Mathematical models seek to quantitatively and qualitatively explain natural phenomena, and usually employ differential equations to describe the dynamic evolution of systems [37]. By solving such equations, it is possible to extract relevant information about such systems and possibly to predict their behaviour.

The main challenge involved in modelling using differential equations is that of formulating the equations describ-

[1] For a more detailed explanation, see [33].

ing the problem from a set of limited information about the general behaviour of the system. However, since a possible solution to the proposed modelling problem (Equation (9)) is available, it is possible to analyse, based on a number of assumptions, the opposite way, and thus determine the differential equations that represent our language model problem.

Considering that the overall goal of the interpolation model is to integrate the benefits of each language model. and assuming language modelling to be a natural system problem, it is possible to treat this problem as the solution of a partial differential equation, i.e.,

$$\frac{\partial u(x,y)}{\partial x}+\frac{\partial u(x,y)}{\partial y}=u(x,y), \tag{10}$$

where $u(x,y)$ is the interpolation model, and x and y are the independent variables, representing the word-based n-gram and POS-based language models, respectively.

By analysing (9), it is possible to see that this equation is not a particular solution of (10), but a particular solution of another partial differential equation, namely

$$\begin{cases} x\dfrac{\partial u(x,y)}{\partial x}+y\dfrac{\partial u(x,y)}{\partial y}=u(x,y), \\ u(1,1)=1. \end{cases} \tag{11}$$

To solve (3.3.1), it is possible to use the technique of separation of variables, which reduces the partial differential equation to several ordinary differential equations [37]. In this case, it is assumed that a solution can be expressed as the product of two unknown functions, where each one is only a function of the respective independent variable. This assumption seems very reasonable, after a study considering each language model as an independent problem. Thus, it is possible to have

$$u(x,y)=X(x)\mathrm{Y}(y)\Rightarrow\begin{cases} \dfrac{\partial u(x,y)}{\partial x}=X'Y \\ \dfrac{\partial u(x,y)}{\partial y}=XY' \end{cases} \tag{12}$$

$$x\frac{\partial u(x,y)}{\partial x}+y\frac{\partial u(x,y)}{\partial y}=xX'Y+yXY'=XY. \tag{13}$$

Dividing (11) by XY, it follows that

$$x\frac{X'}{X}+y\frac{Y'}{Y}=1. \tag{14}$$

Since X is only a function of the variable x and Y of y, each term in (12) should be constant and equal to α (known as the *separation constant*[37]), or

$$x\frac{X'}{X}=1-y\frac{Y'}{Y}=\alpha, \tag{15}$$

which is implied by two ordinary equations, namely

$$x\frac{X'}{X}=\alpha\Rightarrow X(x)=\gamma_1\cdot x^{\alpha} \tag{16}$$

and

$$1-y\frac{Y'}{Y}=\alpha\Rightarrow Y(y)=\gamma_2\cdot y^{(1-\alpha)}, \tag{17}$$

where γ_1 and γ_2 are two constants.

Finally, substituting (14) and (15) into (3.3.1), one obtains

$$u(x,y)=\gamma_1\cdot x^{\alpha}\cdot\gamma_2\cdot y^{(1-\alpha)}=\gamma\cdot x^{\alpha}y^{(1-\alpha)}, \tag{18}$$

where $\gamma=\gamma_1\cdot\gamma_2$ is also a constant.

Applying the condition $u(1{,}1)=1$ to (16), it follows that

$$u(1,1)=\gamma=1, \tag{19}$$

and the solution will be

$$u(x,y)=x^{\alpha}\cdot y^{(1-\alpha)}, \tag{20}$$

which is the same as the geometric interpolation presented in 9.

Even when presenting satisfactory results to word prediction, as can be seen in Section 4, the interpolation model based on geometric interpolation has some negative characteristics that occur when, for example, any independent language model has a zero (or very low) probability, causing the interpolation model to present a zero probability output.

In such a context, considering that word-based n-gram models play a fundamental role in predictive modelling systems and can be improved by their combination with a POS-based language model, we here propose a modified partial differential equation, given by

$$\begin{cases} x\dfrac{\partial u(x,y)}{\partial x}+\dfrac{\partial u(x,y)}{\partial y}=u(x,y), \\ u(1,1)=1 \end{cases} \tag{21}$$

The modelling performed in (3.3.1) seeks to create a natural exponential function to interpolate the word-based n-gram model and the POS-based language model. This methodology, common in the modelling of natural processes (radioactive decay and population decay, among others), reduces the negative characteristics presented by the interpolation model based on geometric interpolation.

In a similar way to the geometric interpolation modelling, the new interpolation model can be found by solving (3.3.1), the solution of which is

$$u\left(x,y\right)=\gamma\cdot x^{\alpha}\cdot e^{(1-\alpha)y}. \tag{22}$$

Again, applying the condition $u(1{,}1)=1$ to (19), one has

$$u\left(1{,}1\right)=1=\gamma\cdot e^{(1-\alpha)}\Rightarrow\gamma=\frac{1}{e^{(1-\alpha)}}, \tag{23}$$

and, finally,

$$u\left(x,y\right)=\frac{1}{e^{(1-\alpha)}}\cdot x^{\alpha}\cdot e^{(1-\alpha)y}. \tag{24}$$

Rewriting (21) in terms of the language models, it follows that

$$\begin{aligned}&P_{interpolation}^{(proposed)}\left(w_i\mid w_{i-(n-1)},t_{i-(m-1)}\right)=\\&\quad\frac{1}{e^{(1-\alpha)}}\cdot\left[P_{n-gram}\left(w_i\mid w_{i-(n-1)}\right)\right]^{\alpha}\cdot\\&\quad\cdot e^{(1-\alpha)\cdot[P_{m-POS}(w_i\mid t_{i-(m-1)})]},\end{aligned} \tag{25}$$

where α can be empirically obtained.

It is worth noting that (22) has a form that is similar to the conventional maximum entropy model first developed in [22]. In the latter, the authors confront two of the essential tasks of statistical modelling: to determine a set of statistics that captures the behaviour of a random process, and to combine these facts into an accurate model of the process – a model capable of predicting the future process output. In attempting to solve this problem, the proposed methodology, based on differential equations, addresses the problem of building interpolation models and thus opens the way for the use of different mathematical tools to construct and analyse natural language comprehension.

4. Experimental Framework

The tests and subsequent analyses needed to confirm the assumptions made are presented in this section. Firstly, to construct possible comparisons between the methodology adopted here and the existing state-of-the-art methodology, it was necessary to select appropriate training and test sets, as well as proper procedures for testing each methodology.

The text set used for training and testing is one of the key aspects of the evaluation step, since it may significantly influence the results. Thus, texts from newspapers and text transcripts from spoken language were chosen to compose the test set. Texts from newspapers were adopted because they employ a language directed at a great number of readers, providing a reasonable contextual diversity in terms of vocabulary and grammatical constructions. Text transcripts from spoken language were adopted, in turn, because they are more spontaneous, less rigid, and cover daily communication in general.

The evaluation procedures for the word prediction were conducted using automatic methods, whereby all language models were incorporated into a WPS for the experiments, a common approach in these types of systems.

It is also important to consider that any changes in the configuration parameters of the experiment (language, test and training texts, WP interface system, etc.) can lead to significant variations in the results. This variability makes it very difficult to compare the results presented here with others already established (in [31], the main factors that can affect the prediction results of a given system are exposed and discussed).

4.1 Word Prediction Engine

In order to evaluate our method, the PredWin software was used. This software was first developed for Spanish by [31] and was adapted here for Portuguese and English. This system has some important blocks, such as the following:

- **user model:** This is the automatic algorithm used by the word prediction system to emulate a real user. For each letter in the test text, the prediction system shows a list of predicted words. If the desired word is in this list, the user model selects the word. If not, it selects the next letter. This loop is repeated until the test text ends.
- **coordination module:** This controls the flow of information between the user interface/user model and the dictionaries and prediction methods.
- **dictionaries:** These contain the words, along with the information about each word required to support the various word prediction methods, such as POS tags and word frequencies.

As in [31], it is also important to highlight certain system parameters that will affect the word prediction process:

- **coefficient α :** This can be defined as a variable for combining language models. This variable can take values between 0 and 1, giving more weight to any one of the language models. As shown in [38] and [1], this parameter has been experimentally determined by varying its value from 0.1 to 0.9, with increments of 0.1.

- **number of previous POSs used:** The number of previous POSs used has a major influence on the system, since these are directly related to the structure of the language. The works presented in [38, 31] and [39], which used POS training sets with sizes similar to the set used in this paper, obtained the best results when using up to two previous POSs. Thus, the models in this work were evaluated using only one or two previous categories.
- **number of words in the prediction list:** According to [31], *seven* is the maximum number of words that the user can process, maintaining the balance between increased cognitive load and the increased processing time needed to select from the predicted words. However, other studies, such as [40] and [38], have shown that the optimal number of words presented to the user must be at most *five* words. In this work, the prediction systems were tested using *one* and *five* words in the prediction list. With just *one* word in the prediction list, it is possible to simulate an automatic system, of the type that is common in mobile devices, for example.
- **number of words in the dictionaries:** The general dictionaries play a fundamental role in the word prediction system. Once categorized texts are needed, the same texts used for training the POS models (presented in Table 3) are also considered. Table 2 shows, for each language, the number of words in the general dictionaries:

Language	#Words
Portuguese	118,878
Spanish	92,482
English	57,711

Table 2. Number of words in each general dictionary

- **repeated suggestions in consecutive predictions:** Since we are working with an ideal user, it is possible to drop the predicted words not previously selected by the user. According to [39], this methodology enables an increase in the probability of suggesting the most appropriate words.
- **automatic insertion of blank spaces:** Apart from word prediction, there are certain characters that can be inserted automatically to improve the KSS. An example of this technique is the insertion of blank spaces after punctuation symbols (commas, full stops and colons, for example).
- **back-off regression:** The language models also have certain limitations when the frequency of the words or POS sequences are null. To overcome such limitations, certain techniques can be used as back-off techniques. As initially proposed in [41], when the language model has zero frequency, it can be approximated by the immediately previous model, which would then continue until reaching the unigram model. Due to its simplicity of implementation and satisfactory results it achieved, as presented in [31] for the Spanish language, this methodology is also used in this work.
- **test of significance:** This work attempts to validate the methods using a statistical test based on the calculation of confidence intervals for proportions, given by

$$p = \sigma \cdot \sqrt{\frac{KSS_{LM} \cdot \left(1 - KSS_{LM}\right)}{N}}, \tag{26}$$

where p is the confidence interval, KSS_{LM} is the keystrokes saved by each language model, N is the total number of keystroke needed to write the text without word prediction, and σ is a constant parameter that depends on the confidence interval, usually set to 1.96 (or 95% of confidence). Thus, the KSS in the experiment is within the range

$$KSS_{LM} = KSS_{LM} \pm p, \tag{27}$$

and it will be considered significantly better (with respect to the baseline) if the results are better, and moreover, if the confidence intervals do not overlap.

4.2 Training Set

The training sets used to train the language models and to generate the dictionaries for each language here addressed (Portuguese, Spanish and English) are shown in Table 3.

Language	#Words		Corpus
	n-gram	*m*-POS	
Portuguese	17,599,914	505,412	CHAVE [42]
Spanish	17,601,472	502,800	Spanish Gigaword First Edition [43]
English	17,763,503	511,066	English Gigaword [44]

Table 3. Number of words used in the training set for each language model

4.3 Validation Set

The validation sets used to find the coefficients α needed to combine the language models (*n*-gram and *m*-POS) are shown in Table 4, with Table 5 showing the values of α used in the WPS, after considering the best results achieved when analysing the KSS parameter.

Language	#Words	Corpus
Portuguese	80,259	Rhetalho [45]
Spanish	67,722	Corpus92 [46]
English	81,631	MASC [47]

Table 4. Number of words in the validation sets used to find the best α values

Language	#Words in Prediction List	Interpolation Model		
		Linear	Geometric	Proposed
Portuguese	1	0.9	0.9	0.6
	5	0.9	0.9	0.9
Spanish	1	0.9	0.9	0.7
	5	0.9	0.9	0.8
English	1	0.9	0.9	0.8
	5	0.9	0.9	0.9

Table 5. Optimized values of α for each interpolation model considering 1 and 5 words in the prediction list.

4.4 Test Set

For the test sets, it is important to use texts that were not used in the training or validation sets. Therefore, for Portuguese, texts were chosen from the journalistic corpus TeMário [48], as well as the "Português Falado" corpus [49], which contains transcript texts from audio recordings of the language spoken in Brazil. To compose the test set for Spanish, the Europarl corpus [50] was used, which contains transcript texts from speeches in the European Parliament, along with texts from the HC corpus [51], consisting of newspaper articles from different sources. Finally, for English, the test set was extracted from the Brown corpus [52] and the Uppsala Student English corpus [53], consisting of newspaper articles and transcript texts, respectively. Table 6 shows the domain, number of words and keystrokes needed (without word prediction) to write the texts in each test set.

Language	Topic	#Words	#Keystrokes Needed
*Portuguese	Newspaper	51,724	323,857
	Spoken	20,498	110,724
Total		**72,222**	**434,581**
*Spanish	Newspaper	49,354	300,029
	Spoken	53,627	337,866
Total		**102,981**	**637,895**
*English	Newspaper	63,158	369,621
	Spoken	27,928	151,302
Total		**90,931**	**520,923**

Table 6. Domain and number of words used in the test set

It is important to mention that about 3% of the words in each test set were not categorized. For the moment, particular attention has been paid to the known words, at the expense of the unknown words. This was treated as a separate problem and smoothing techniques were used to avoid null probabilities for any unseen events in the test set.

4.5 Performance Measures

The WPS was evaluated according to four different criteria: keystrokes saved (KSS), hit rate (HR), words predicted (WP) and perplexity (PP).

The KSS refers to the percentage of keystrokes that the user saves when using the word prediction system. It is calculated by comparing two measures: the total number of keystrokes needed to type the text (K_T) without the help of word prediction, and the effective number of keystrokes needed when using word prediction (K_E). Hence,

$$KSS = \frac{K_T - K_E}{K_T} \times 100. \tag{28}$$

The higher the KSS value, the better the system performance.

The HR is defined as the percentage of instances in which the suggestion list contains the correct word before any letters of the following word have been entered. In other words, it is the relation between the number of times that a word is guessed without any letters being known and the total number of words in the test text. Again, a higher HR means a better performance.

The PP can be defined as the average number of potential choices/words after a given string of words [38]. Therefore, the lower the PP value, the better the language model.

4.6 Results

In order to evaluate the WPS with different interpolation models, the experiments have been conducted with the texts shown in Table 3. The results are presented in Table 7, considering the word-based *n*-gram model as a baseline. The relative improvements were also evaluated and the results are presented in Table 8, along with the test of significance for each model.

4.7 Discussion

From Table 7, it can be noted that all the interpolation models show improvements with respect to the number of KSS, compared to the word-based *n*-gram model. These results are in agreement with [54] and support one of the assumptions made in this work: that the problem of word prediction can be solved by finding linguistically relevant factors, and that one efficient method is the combination of a POS-based and word-based language models. In some cases, especially when the results obtained from language models with *fine* suggested words are considered, the impact of the POS model on the improvement in KSS (i.e., even with the decrease in the number of words predicted) can be clearly seen.

It is also important to note, from Table 7, that the proposed interpolation model shows the best results in all parameters

Language	Model	#Words in Prediction List							
		One				Five			
		HR(%)	WP(%)	KSS(%)	PP	HR(%)	WP(%)	KSS(%)	PP
*Spanish	*n*-gram	21.30	86.93	41.32	95.6	38.00	**95.37**	**54.09**	**240.2**
	Linear	21.20	87.43	42.30	90.1	38.10	95.22	54.64	218.4
	Geometric	21.20	87.30	42.13	111.1	38.00	95.21	54.58	253.0
	Proposed	**21.40**	**87.66**	**42.44**	**88.7**	**38.30**	**95.28**	**54.76**	**190.5**
*Portuguese	*n*-gram	23.30	86.16	38.02	188.5	**35.61**	**95.17**	**50.35**	**416.3**
	Linear	23.30	86.43	39.43	159.4	35.50	94.90	51.14	331.6
	Geometric	23.30	86.41	39.27	186.1	35.50	94.89	51.08	373.3
	Proposed	**23.31**	**86.83**	**39.53**	**156.2**	**35.60**	**94.93**	**51.19**	**308.4**
*English	*n*-gram	21.10	88.17	38.88	121.2	35.10	**97.24**	**52.01**	**283.7**
	Linear	21.10	88.50	39.46	115.8	35.10	97.17	52.28	263.6
	Geometric	21.00	88.06	39.18	141.0	35.10	97.22	52.21	304.7
	Proposed	**21.11**	**88.63**	**39.52**	**111.8**	**35.20**	**97.20**	**52.31**	**244.4**

Table 7. Word prediction results for the different interpolation models, with *one* and *five* words in the prediction lists

Language	Model	#Words in Prediction List					
		One			Five		
		KSS(%)	Relative Improvement(%)	Significant	KSS(%)	Relative Improvement(%)	Significant
*Spanish	*n*-gram	41.32(±0.12)	-	-	54.09(±0.12)	-	-
	Linear	42.30(±0.12)	2.39	Yes	54.64(±0.12)	1.02	Yes
	Geometric	42.13(±0.12)	1.98	Yes	54.58(±0.12)	0.91	Yes
	Proposed	**42.44(±0.12)**	**2.72**	**Yes**	**54.76(±0.12)**	**1.24**	**Yes**
*Portuguese	*n*-gram	38.02(±0.14)	-	-	50.35(±0.15)	-	-
	Linear	39.43(±0.14)	3.69	Yes	51.14(±0.15)	1.57	Yes
	Geometric	39.27(±0.14)	3.28	Yes	51.08(±0.15)	1.45	Yes
	Proposed	**39.53(±0.14)**	**3.96**	**Yes**	**51.19(±0.15)**	**1.67**	**Yes**
*English	*n*-gram	38.88(±0.13)	-	-	52.01(±0.14)	-	-
	Linear	39.46(±0.13)	1.49	Yes	52.28(±0.14)	0.52	No
	Geometric	39.18(±0.13)	0.79	Yes	52.21(±0.14)	0.37	No
	Proposed	**39.52(±0.13)**	**1.66**	**Yes**	**52.31(±0.14)**	**0.57**	**Yes**

Table 8. Test of significance and relative improvement for the results shown in Table 7

related to word prediction (WP, HR and KSS) when only *one* word is considered in the prediction list. However, when considering *fine* words in the prediction list, the word-based *n*-gram models present, in all languages, the best results for the WP parameter. These results clearly show the importance of word-based *n*-gram models in the prediction of *function words* – consisting mainly of pronouns, determiners, preposition and auxiliary verbs – and words with a lower number of letters, as opposed to *content words*: normally nouns, verbs, adjectives and adverbs.

When the PP values in Table 7 are analysed, the proposed interpolation model also shows the best results for both *one* and *five* words in the prediction list. When *five* suggested words were considered, the model achieved a 20.69%, 25.92% and 13.85% relative reduction for Spanish, Portuguese and English, respectively. With *one* word in the prediction list, the proposed model reached 7.21%, 17.14% and 7.76% relative reduction in the PP. This measure was consistent for almost all models, i.e., it can also be noted from Table 6 that the PP obtained by the geometric model for the English and Spanish languages is higher than the value obtained by the word-based *n*-gram model, even with higher values of KSS. Such inconsistencies have been already presented and discussed in other works, such as [38].

As Table 8 shows, the relative improvements in English using the proposed interpolation model (1.66% and 0.57% for *one* and *five* words in the prediction list) are lower than those obtained for the other languages. These results illustrate the statement that English is *grammatically poor* when compared to Portuguese and Spanish, and are consistent with the work in [39]. Using a linear combination (with $\alpha = 0.8$) to combine a word bigram model with a POS-based model considering two previous POSs – trained using 81 million words and 5.8 million categorized words, respectively – [39] has presented a total of 53.14% KSS against the 52.90% KSS obtained by the word bigram model itself, when considering *five* words in the prediction list, a test set of 951, 932 words and a POS tagset with 79 classes.

It is also worth analysing, using Table 8, the relative improvements obtained by each interpolation model, considering the number of words in the prediction list. In this case, it can be seen that increasing the number of suggested words reduces the relative improvement of each model, yet it also increases the total number of KSS. In addition, it should be noted that some of the results for English, excluding the results obtained by the proposed model, were not significant. A possible solution to this could be to increase the number of words in the test set.

The empirically optimized values for the coefficient α in Table 5 also deserve discussion. It can be observed that the values found for the linear and geometric interpolation model were almost equal to 0.9 (the linear model showed a value of 0.8 in Table 5 for Spanish with *five* suggested words). These results are consistent with [38] and [36]. However, when analysing the values of α obtained by the proposed model, this parameter showed different values for each language and domain analysed, which means that the proposed model could be more *susceptible* to the relationship between the word-based *n*-gram model and the POS-based model. This characteristic is quite interesting and points the way forward in the search for the optimal method to determine the separation coefficient α.

Finally, even though real users were not used in the test set, the results show good prospects for the application of the proposed system within the text-based user interface in the robotic wheelchair. Besides presenting better results in the KSS parameters, the proposed method was able to improve the percentage of instances in which correct words appear in the suggestion list before any letter has been entered (HR), a parameter that is directly related to avoiding extraneous effort on the part of the user when generating the HMI signals. Moreover, the PP results show that the proposed interpolation LM is a better predictor than the state-of-the-art interpolation models.

5. Conclusion

The goal of this work was to explore text-based human-machine interactions by considering a word prediction system installed in a robotic wheelchair. Since the word prediction system plays a fundamental role in improving the writing of text, this work proposed an exponential interpolation model, which combines a traditional word-based *n*-gram language model with a POS-based language model. We addressed this problem by first finding a differential partial equation to represent the modelling of the language, which would then be used to derive the interpolation model.

The proposed methodology was evaluated for two different domains (journalistic and spoken texts) and three different languages (Portuguese, Spanish and English). The results reported in this paper show improvements in the KSS, HR and PP parameters, with both *one* and *five* words in the prediction lists.

Future efforts could concentrate on testing the proposed model with real users in the robotic wheelchair. Moreover, we could improve the proposed differential partial equation by adding more information through a semantics-based model, as part of the search for a more sophisticated language model. Finally, we also plan to test our current interpolation language model on another task: that of automatic language recognition.

6. References

[1] Afsaneh Fazly and Graeme Hirst. Testing the efficacy of part-of-speech information in word completion. In *TextEntry '03: Proceedings of the 2003 EACL Workshop on Language Modeling for Text Entry Methods*, pages 9–16, Budapest, Hungary, 2003. Association for Computational Linguistics.

[2] Alexandre Santos Brandao, Leonardo Bonato Felix, Daniel Cruz Cavalieri, Antonio Mauricio Ferreira Leite Miranda de Sa, Teodiano Freire Bastos-Filho, and Mario Sarcinelli-Filho. Controlling devices using biological signals. *International Journal of Advanced Robotic Systems*, 8(3):22–33, March 2011.

[3] Teodiano Bastos, Fernando Cheein, Sandra Muller, Wanderley Celeste, Celso Casano, Daniel Cavalieri, Mario Sarcinelli-Filho, Paulo Amaral, Elisa Perez, Carlos Soria, and Ricardo Carelli. Towards a new modality-independent interface for a robotic wheelchair. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2013.

[4] Santiago Omar Caballero Morales, Gladys Bonilla Enríquez, and Felipe Trujillo Romero. Speech-based human and service robot interaction: An application for mexican dysarthric people. *International Journal of Advanced Robotic Systems*, 10(11), October 2013. doi: 10.5772/54001.

[5] Raquel Justo, Oscar Saz, Antonio Miguel, M. Inés Torres, and Eduardo Lleida. Improving language models in speech-based human-machine interaction. *International Journal of Advanced Robotic Systems*, 10(87):1–11, December 2013. doi: 10.5772/55407.

[6] J. A. Amott J. L., Picketin. An adaptive and predictive communication aid for the disabled that exploits the redundancy in natural language. In *In RESNA 7th Annual Conference*, pages 349–350, Ottawa, Canada, 1984. RESNA.

[7] Lynda Booth, Corinne Morris, Ian Ricketts, and Alan Newell. Using a syntactic word predictor with language impaired young people. In H.J. Murphy, editor, *In Proceedings of the California State University, Northridge (CSUN), 7th Annual Conference on Technology and Persons with Disabilities*, pages 57–61, Los Angeles, California, USA, March 1992. California State University, Northridge.

[8] Nestor Garay-Vitoria and Julio Abascal. Text prediction systems: a survey. *Univers. Access Inf. Soc.*, 4(3):188–203, February 2006.

[9] M. Ghayoomi and S. Momtazi. An overview on the existing language models for prediction systems as writing assistant tools. In *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, pages 5083–5087, San Antonio, Texas, 11-14 October 2009. ISSN: 1062-922X.

[10] Pertti Väyrynen. *Perspectives on the Utility of Linguistic Knowledge in English Word Prediction*. PhD thesis, University of Oulu, Linnanmaa, November 19th 2005.

[11] Hisham Al-Mubaid. A learning-classification based approach for word prediction. *Int. Arab J. Inf. Technol.*, 4(3):264–271, 2007.

[12] S. E. Palazuelos-Cagigas, S. Aguilera, J. Rodrigo, and J. Godino. Grammatical and statistical word prediction system for spanish integrated in an aid for people with disabilities. In *V International Conference on Spoken Language Processsing*, pages 381-384, 1998.

[13] Fernando A. Auat Cheein, Fernando di Sciascio, Juan Marcos Toibero, and Ricardo Carelli. *Robot Manipulator Probabilistic Workspace Applied to Robotic Assistance, Robot Manipulators New Achievements*, volume 1 of 1. InTech, 1 edition, December 2011.

[14] Nestor Garay-Vitoria and Julio Abascal. Modelling text prediction systems in low- and high-inflected languages. *Comput. Speech Lang.*, 24(2):117–135, 2010.

[15] John L. Arnott and Norman Alm. Towards the improvement of augmentative and alternative communication through the modelling of conversation. *Computer Speech and Language*, 27(6):1194–1211, 2013. Special Issue on Speech and Language Processing for Assistive Technology.

[16] Kadri Hacioglu and Wayne Ward. On combining language models: oracle approach. In *Proceedings of the first international conference on Human language technology research*, HLT '01, pages 1–4, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

[17] Diego Linares, José-Miguel Bened, and Joan-Andreu Sánchez. A hybrid language model based on a combination of n-grams and stochastic context-free grammars. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(2):113–127, 2004.

[18] Tomás Brychcín and Miloslav Konopík. Semantic spaces for improving language modeling. *Computer Speech and Language*, (0), 2013. In Press. doi: 10.1016/j.csl.2013.05.001.

[19] Manzoor Ahmad Chachoo and S. M. K. Quadri. Adaptive hybrid pos cache based semantic language model. *International Journal of Computer Applications*, 39(13):7–10, February 2012. Published by Foundation of Computer Science, New York, USA.

[20] Chuang-Hua Chueh, Jen-Tzung Chien, and Hsin-Min Wang. A maximum entropy approach for semantic language modeling. *Computational Linguistics and Chinese Language Processing*, 11(1):37–56, March 2006.

[21] Shaojun Wang, Dale Schuurmans, and Yunxin Zhao. The latent maximum entropy principle. *ACM Trans. Knowl. Discov. Data*, 6(2):8:1–8:42, July 2012.

[22] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71, March 1996.

[23] Shaojun Wang, Shaomin Wang, Russell Greiner, Dale Schuurmans, and Li Cheng. Exploiting syntactic, semantic and lexical regularities in language modeling via directed markov random fields. In *In Proceedings of ICML 2005*, pages 948–955, 2005.

[24] Van A. den Bosch. Scalable classification-based word prediction and confusible correction. *Traitement Automatique des Langues*, 46(2):39–63, 2006.

[25] Giulio Maltese, P. Bravetti, H. Crépy, B. J. Grainger, M. Herzog, and Francisco Palou. Combining word- and class-based language models: a comparative study in several languages using automatic and manual word-clustering techniques. In Paul Dalsgaard, Børge Lindberg, Henrik Benner, and Zheng-Hua Tan, editors, *INTERSPEECH*, pages 21–24. ISCA, 2001.

[26] Aliprandi, Carlo, Carmignani, Nicola, Mancarella, Paolo, and Rubino, Michele. A word predictor for inflected languages: system design and user-centric interface. In *Proceedings of the Second IASTED International Conference on Human Computer Interaction*, IASTED-HCI '07, pages 148–153, Anaheim, CA, USA, 2007. ACTA Press.

[27] Daniel Cruz Cavalieri, Teodiano Freire Bastos-Filho, Mário Sarcinelli-Filho, Sira E. Palazuelos-Cagigas, Javier Macas Guarasa, and José Luis Martn

Sánchez. A part-of-speech tag clustering for a word prediction system in portuguese language. *Procesamiento del Lenguaje Natural*, 47:197–205, 2011. ISSN: 1135-5948.

[28] Eckhard Bick. *The Parsing System " Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Aarhus University, Aarhus, Denmark, November 2000.

[29] Eckhard Bick. A constraint grammar parser for spanish. In *Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006 - 4th Workshop in Information and Human Language Technology (TIL 2006)*, Ribeirão Preto, Outubro 27-28 2006. ISBN: 85-87837-11-7.

[30] A. Voutilainen and J. Heikkila. An English constraint grammar (EngCG): a surface-syntactic parser of English. In U. Fries, G. Tottie, and P. Schneider, editors, *Creating and Using English Language Corpora*, volume 13, Amsterdam, 1994. Editions Rodopi.

[31] S. E. Palazuelos-Cagigas. *Contribution to word prediction in Spanish and its integration in technical aids for people with physical disabilities*. PhD thesis, Universidad de Alcalá de Henares, Alcalá de Henares, Madrid, Spain, 2001.

[32] Daniel C. Cavalieri, Sira E. Palazuelos-Cagigas, Teodiano F. Bastos-Filho, and Mário Sarcinelli-Filho. Evaluation of machine learning approaches to portuguese part-of-speech prediction. In *Computational Processing of the Portuguese Language, 9th International Conference, Proceedings (PROPOR 2010)*, Porto Alegre, Brasil, 27-30 April 2010.

[33] Tom Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874, June 2006.

[34] Tom Fawcett. Roc graphs: Notes and practical considerations for researchers. Technical report, HP Laboratories Palo Alto, 2004.

[35] Rodrigo C. Barros, Márcio P. Basgalupp, André C.P.L.F. de Carvalho, and Alex A. Freitas. Towards the automatic design of decision tree induction algorithms. In *Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation*, GECCO '11, pages 567–574, New York, NY, USA, 2011. ACM.

[36] Tonio Wandmacher and Jean-Yves Antoine. Methods to integrate a language model with semantic information for a word prediction component. *CoRR*, abs/0801.4716, 2008.

[37] W.E. Boyce and R.C. DiPrima. *Elementary Differential Equations*. Wiley, 2012.

[38] Keith Trnka. *Word prediction techniques for user adaptation and sparse data mitigation*. PhD thesis, University of Delaware, Newark, DE, USA, 2011. ISBN: 978-1-124-48009-1.

[39] Afsaneh Fazly. The use of syntax in word completion utilities. Master's thesis, University of Toronto, Department of Computer Science, 2002.

[40] D. K. Anson, P. Moist, M. Przywara, H. Wells, H. Saylor, and H. Maxime. The effects of word completion and word prediction on typing rates using on-screen keyboards. *Proceedings RESNA'05*, 2005.

[41] Slava M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, pages 400–401, 1987.

[42] Diana Santos and Paulo Rocha. The key to the first clef in portuguese: Topics, questions and answers in chave. In *5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004*, pages 821-832, Bath, UK, September 15-17 2004.

[43] David Graff. Spanish gigaword first edition. Linguistic Data Consortium, Philadelphia, 2006.

[44] David Graff and Christopher Cieri. English gigaword. Linguistic Data Consortium, Philadelphia, 2003.

[45] Nelson Neto, Carlos Patrick, Aldebaro Klautau, and Isabel Trancoso. Free tools and resources for brazilian portuguese speech recognition. *J. Braz. Comp. Soc.*, 17(1):53–68, 2011.

[46] Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada (IULA). Corpus92 corpus. http://hdl.handle.net/10230/20054, 2012. Accessed on 05 July 2014.

[47] Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. The manually annotated subcorpus: A community resource for and by the people. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 68–73, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[48] T. A. S. Pardo and L. H. M. Rino. TeMário: Um corpus para sumarização automática de textos. Série de Relatórios do NILC NILC-TR-03-09, Núcleo Interinstitucional de Lingüística Computacional (NILC), São Carlos-SP, October 2003. 11 p.

[49] Maria Fernanda Bacelar do Nascimento, Luísa Pereira, and João Saramago. Portuguese corpora at clul. In *Second International Conference on Language Resources and Evaluation*, pages 1603-1607, Atenas, Grécia, 2000.

[50] Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT.

[51] Hans Christensen. Hc corpora. http://corpora.heliohost.org/, 2012. Accessed on 05 July 2014.

[52] W. N. Francis and H. Kucera. Brown corpus manual. Technical report, Department of Linguis-

tics, Brown University, Providence, Rhode Island, US, 1979.

[53] Margareta Westergren Axelsson. Project use (uppsala student english). ASLA Information, 1999.

[54] Charlotte Wilson. *Combining Part of Speech Induction and Morphological Induction*. PhD thesis, University of Melbourne, Melbourne, Australia, November 2004.