UNIVERSIDADE DO ALGARVE
Faculdade de Ciências Humanas e Sociais

Universitat Autònoma de Barcelona
Facultat de Filosofía i Lletres

An Evaluation of Automatic Speech Recognition in the Spanish Version of Windows 7: Effects of Language Variety, Speaking Style and Gender.

María Soledad López Gambino

Mestrado Internacional em Processamento de Linguagem
Natural e Indústrias da Língua

International Masters in Natural Language Processing and
Human Language Technology

Faro, 2012

NLP-HLT
NATURAL LANGUAGE PROCESSING
HUMAN LANGUAGE TECHNOLOGY

UNIVERSIDADE DO ALGARVE
Faculdade de Ciências Humanas e Sociais

UNIVERSITAT AUTÒNOMA DE BARCELONA
Facultat de Filosofía i Lletres

An Evaluation of Automatic Speech Recognition in the Spanish Version of Windows 7: Effects of Language Variety, Speaking Style and Gender.

María Soledad López Gambino

Orientador/Supervisor: **Dr. Joaquim Llisterri**
(Universitat Autònoma de Barcelona)

Orientador/Supervisor: **Dr. Thomas Pellegrini**
(Universidade do Algarve) (L²F – INESC ID Lisboa)

Dissertation submitted as part of the study program for the award of the Master degree in Natural Language Processing and Human Language Technology, and supported by a grant from the European Commission, Education & Culture, under the Erasmus Mundus Master Courses Program (ref. EMMC 2008-0083).

Faro, 2012

ABSTRACT

This study consists in an evaluation of the Spanish version of the automatic speech recognizer embedded in what is currently one of the most widespread operating systems: Microsoft's Windows 7. Emphasis is placed upon the effects of gender, language variety and speaking style on system performance. Two groups of subjects were included in the tests: one of them was composed of 20 speakers of a Peninsular variety (Spanish as spoken in Catalonia) and the second one, of 20 speakers of a Latin American variety (Spanish as spoken in Buenos Aires), 10 female and 10 male speakers within each group. The test set consisted of three different tasks aimed at evaluating command recognition as well as automatic dictation. These tasks were carried out in one-to-one meetings with each of the selected subjects.

Results revealed higher error rates for the group of Latin American speakers in comparison to Peninsular speakers. Word error rate (WER) in the dictation tasks was 28.2% for the former group and 23.1% for the latter. Regarding the task on commands, 88% of these were correctly recognized for the Peninsular group, whereas the group from Buenos Aires obtained a recognition percentage of 82.5%. With respect to speaking style, the system performed worse for speech exhibiting a higher degree of spontaneity and informality (WER = 30.7%) than for semi-scripted speech on relatively formal topics (WER = 22.8%). In contrast, results corresponding to the speech of men and women only showed slight differences which in general did not prove significant. For male speakers, 86.5% of the commands were correctly recognized, compared to 84% for female speakers, and WER for the automatic dictation tasks was 24.9% for the former group and 26.6% for the latter.

Keywords: *speech recognition, speech variability, gender, language variety, speaking style.*

RESUMO

Este estudo consiste em uma avaliação da versão espanhola do reconhecedor automático de fala, incluído no que é atualmente um dos sistemas operativos mais comuns: o Windows 7, da empresa Microsoft. A ênfase é colocada sobre os efeitos do gênero, a variedade da língua, o estilo de fala e o tipo de tarefa de reconhecimento no desempenho do sistema.

O impacto da variedade da língua na precisão do reconhecimento é uma questão particularmente interessante, dado que o sistema de reconhecimento da fala de Windows 7 (Windows 7 ASR) é apresentado como um sistema desenvolvido para "espanhol da Espanha", um termo que é geralmente usado para se referir às variedades do espanhol falado na Península Ibérica e nas ilhas Baleares e Canárias. Isto tem implicações importantes. Por um lado, é provável que os modelos acústicos tenham sido treinados com corpora dessas variedades. Analogamente, o conjunto de fones para os modelos foi, provavelmente, selecionado de acordo com os sons utilizados por essas comunidades linguísticas. Além disso, o léxico do sistema também deveria refletir o vocabulário utilizado nas variedades peninsulares de espanhol.

A enorme popularidade do sistema operativo Windows tem causado, por todo o mundo, o aparecimento de um imenso mercado de usuários potenciais do Windows ASR, entre os quais uma grande parte são falantes de variedades de espanhol não peninsular. Dado que existe apenas uma versão disponível para o espanhol, é muito provável que essa seja a versão empregue por estes falantes. Este facto leva-nos à interessante questão do impacto que as características lexicais e fonológicas distintivas das variedades do espanhol não peninsular terão no desempenho do sistema. Por outras palavras, se essas características exercem, ou não, um efeito prejudicial sobre o reconhecimento.

Um dos principais objetivos desta avaliação é procurar uma resposta para esta pergunta. Duas variedades linguísticas foram selecionadas: o espanhol falado na

Catalunha, Espanha, como exemplo de uma variedade peninsular, e o espanhol falado em Buenos Aires, Argentina, como exemplo de uma variedade latino-americana. Esta última tem sido objeto de muita investigação linguística, devido a certas peculiaridades fonológicas e morfossintáticas que a distinguem de todas as outras variedades, tanto ibéricas como latino-americanas. Visto que tais características são susceptíveis de apresentar dificuldades no reconhecimento, uma análise comparativa do desempenho do reconhecedor para os falantes das duas variedades surge como um desafio altamente interessante.

Para além destas considerações, outras duas fontes da variabilidade na fala são tidas em conta, juntamente com os seus efeitos sobre o desempenho do sistema. Trata-se do sexo do falante e do estilo de fala. O objetivo com relação ao primeiro é observar se as taxas de reconhecimento para mulheres e para homens apresentam disparidades, dadas as diferenças entre a fala de ambos os sexos que têm sido descritas na literatura. Isto não se refere apenas a diferenças de origem biológica, como a frequência fundamental (Martínez Celdran & Fernández Planas, 2007, p. 148) (Gil, 1988, p. 39) e o ponto de articulação (Simpson, 2009, p. 625), mas também a diferenças que resultam de padrões de comportamento aprendidos através da imersão num ambiente sociocultural (Foulkes, Scobbie & Watt, 2010, pp 711-712;. Simpson, 2009, p 621). A outra fonte de variabilidade da fala que é considerada está ligada à distinção entre o estilo formal e o informal, como também à noção de graus de espontaneidade na fala. Neste caso, o interesse reside nas defluências e processos co articulatórios que caracterizam a fala espontânea informal e nas suas consequências em matéria de reconhecimento.

Dois grupos de falantes foram selecionados: o primeiro inclui falantes da Catalunha, e o segundo, falantes de Buenos Aires. Por razões de homogeneidade, tendo em conta a natureza bilíngue da comunidade catalã, um perfil linguístico específico foi definido para a pesquisa. Somente falantes bilíngues espanhol-catalão dominantes em catalão, nascidos e criados na Catalunha, foram selecionados. Para o segundo grupo, o perfil linguístico foi o de falantes nativos de espanhol que nasceram e foram criados em Buenos Aires. Outros aspetos foram também considerados na seleção. A faixa etária foi definida entre 18 e 35 anos. Na tentativa de reduzir a heterogeneidade sociocultural dentro das possibilidades, bem como por razões de disponibilidade dos sujeitos, apenas estudantes ou graduados universitários foram selecionados. A familiaridade com o uso de computadores e sistemas de reconhecimento da fala também foi tida em conta. Os falantes selecionados usam computadores diariamente para diferentes tarefas, mas têm pouca ou nenhuma experiência com sistemas de reconhecimento da fala.

A fim de determinar a elegibilidade em relação a esses critérios, os potenciais candidatos preencheram um questionário. Dois questionários diferentes foram utilizados: um para cada grupo.

O teste consistiu de três tarefas. A primeira teve como objetivo testar o reconhecimento de comandos. A segunda e a terceira foram tarefas de ditado automático. Na tarefa 2, os falantes receberam uma série de palavras e frases que deviam ditar ao reconhecedor em forma de parágrafos inteiros. O objetivo desta atividade foi a obtenção de amostras de fala semi-espontânea baseada em notas, assim como de fala semi-formal, já que as notas tratavam de temas acadêmicos. Na tarefa 3, por outro lado, buscou-se obter exemplos de fala informal quase espontânea, já que os sujeitos deviam ditar um correio eletrónico para um amigo a partir de um conjunto de orientações gerais.

Durante o desenvolvimento das tarefas incluíram-se deliberadamente fenômenos que poderiam apresentar dificuldades especiais ao reconhecimento. Alguns deles eram de natureza lexical, como a inclusão de números, datas, palavras estrangeiras, nomes de lugares e outros nomes próprios. Também se consideraram aspetos fonéticos através da inclusão de certas formas ortográficas que são pronunciadas de forma diferente na variedade de espanhol falada na Catalunha e na variedade falada em Buenos Aires.

O corpus de fala foi coletado em quarenta reuniões, uma com cada um dos falantes selecionados. A duração média aproximada das reuniões foi de quarenta e cinco minutos. O equipamento utilizado consistiu de um computador portátil Compaq Presario CQ40-705LA com um processador Intel Pentium T4300, 2 GB de RAM e 320 GB HDD, e fones de ouvido H110 Logitech com supressão de ruído. A versão do sistema operativo utilizada foi Windows 7 Home Basic.

As reuniões consistiram de cinco etapas. A primeira foi a configuração do microfone. Em seguida teve lugar a familiarização do utilizador com o sistema, que consistiu na leitura de frases que apareceram sequencialmente no ecrã. Após estas duas fases, as três tarefas de teste foram realizadas. Antes de cada tarefa, os participantes receberam breves instruções orais, e posteriormente tiveram um minuto para ler as instruções e os conteúdos da tarefa antes de realizá-las. As gravações e o reconhecimento foram realizados simultaneamente. Para a tarefa 1, os percentuais de reconhecimento foram calculados utilizando informações registradas numa grade durante as reuniões. Para as tarefas 2 e 3, as transcrições de referência das gravações foram elaboradas manualmente e, subsequentemente, alinhadas com as hipóteses produzidas pelo reconhecedor, a fim de calcular a taxa de erro de palavra (Word Error Rate - WER) usando um programa chamado SCLITE.

Os resultados revelaram taxas de erro significativamente maiores para o grupo de falantes de Buenos Aires em comparação com o grupo de falantes da Catalunha. A taxa de erro de palavra nas tarefas de ditado foi de 28,2% para o primeiro grupo e 23,1% para o último. Em relação à tarefa de comandos, 88% deles foram reconhecidos corretamente para o grupo da variedade Peninsular, enquanto o grupo da variedade latino-americana obteve uma percentagem de reconhecimento de 82,5%. Com relação ao estilo de fala, o sistema mostrou mais dificuldade para fala com um maior grau de espontaneidade e informalidade (WER = 30,7%) do que para fala semi-planeada sobre temas relativamente formais (WER = 22,8%). Em contraste, os resultados correspondentes ao discurso de homens e mulheres mostraram diferenças que em geral não foram significativas. Para os falantes do sexo masculino, 86,5% dos comandos

foram correctamente reconhecidos, em comparação com 84% para as falantes do sexo feminino, e as taxas de erro para as tarefas de ditado automático foram de 24,9% para o primeiro grupo e 26,6% para o segundo.

Keywords: reconhecimento da fala, variabilidade da fala, género, variedade linguística, estilo de fala.

CONTENTS

1. INTRODUCTION

The existence of machines with communicative capabilities which resemble the human ones is an idea which has captivated mankind for ages. One of the reasons of this appeal might be the human being’s curiosity towards his own nature: therefore, perhaps, our interest in building ‘artificial replicas’ of ourselves and interacting with them. Regardless of its philosophical grounds, this attitude seems to have been present since times in which the idea of such creations was considered as fascinating as impossible. Nonetheless, the progress achieved in the areas of artificial intelligence and speech technologies in the last decades has not only challenged this assumption, but it has in fact produced language-understanding and language-generating creations which can now be regarded as an intrinsic part of our everyday life. Although it must be acknowledged that human-like performance is still a distant goal, such developments could easily be viewed as the first important steps in this quest.

The research conducted within the field of speech technologies has given birth to a number of applications which can be grouped into different categories. **Speech synthesizers** convert text into artificially-generated speech, whereas **speech recognizers** perform the inverse operation, that is, they turn sound signals corresponding to human speech into text. **Spoken dialogue systems**, on the other hand, execute a somewhat more complex task, since they take a human utterance as input and provide a suitable verbal response as output. In order to achieve this, they employ both a speech synthesizer and a speech recognizer, alongside with other components, such as a language-understanding module, a language-generating module and a dialogue manager.

As stated above, the aim of automatic speech recognition, often abbreviated ASR, is to faithfully and efficiently transform sound input corresponding to human speech into a sequence of words. The output of this operation may be subsequently shown to the user in the form of text, or other actions might be executed as a result of the process, based on the type of ASR application and its intended use. ASR is already widespread in a number of fields. One of the most active of these fields is telephony, where speech recognition applications are employed for automatic dialing, customer service, information enquiry, phone sales, ticket reservations, etc. It is also employed in augmentative communication, that is, to fulfill the needs of people with certain pathologies resulting in partial or total inability to move and/or type (Jurafsky & Martin, 2009). A somewhat less extended but equally promising use is related to the realm of education, in particular of foreign language learning, for training and automatic correction of pronunciation (Neri, Cucchiarini & Strik, 2003). ASR also constitutes a central contribution to the development of disciplines such as building automation and domotics, as it enables voice-control of building and home appliances. Additionally, the use of automatic dictation is becoming more popular in professional environments. Some examples are the elaboration of medical reports and legal documentation, as well as certain types of office work such as letter-writing (Llisterri, 2011b). Anusuya and Katti (2009, p. 183) provide a detailed account of the uses of ASR, in which they also include computer and video games, as well as applications related to the military sector.

Such a technology clearly offers numerous advantages. One of its most evident assets is that it requires neither tactile nor visual interaction, which enables the user to simultaneously operate machinery which must be controlled manually and/or entails continuous use of a visual interface (what Jurafsky and Martin (2009) call "hands-busy" or "eyes-busy" applications). This is also the feature which makes ASR suitable for the needs of people with physical disabilities. On the other hand, automatic dictation can prove an efficient time-saving tool in professional environments, reducing the duration of the task with respect to the time that the process of typing or handwriting documents usually requires. Apart from these aspects, which are basically related to individual users, corporations also benefit enormously from ASR. Rabiner and Juang (2006) explain that this happens in three ways: a) It is a technology which lowers costs, since it allows the replacement of humans for machines and a consequent reduction in staff expenses, b) it allows enterprises to offer twenty-four seven automatic customer service, creating new revenue opportunities, and c) it enables customization of services and goods, which in turn increases user satisfaction. For this last point, the authors provide the example of a voice controlled automobile which recognizes the driver's voice and adjusts the car features automatically according to his preferences.

This paper will present an evaluation of the speech recognizer embedded in what is currently one of the most widespread operating systems, *Windows 7*. Chapter two will provide an account of the state of the art for this technology, as well as the available procedures and metrics for its evaluation. The notion of speech variability will also be discussed. Chapter three will describe the methodology employed in the selection of subjects for the experiment, as well as the criteria contemplated during task design. The results obtained will be presented and discussed in chapter 4. Finally, chapter 5 will be devoted to conclusions.

2. THEORETICAL CONSIDERATIONS

2. 1) Types of speech recognizers

Different criteria can be employed to classify speech recognizers. In the first place, these applications differ in the type of speech that they process: **isolated word recognizers** on the one hand, are frequently used for recognition of digits or commands, whereas speech structured in the form of connected sentences requires the use of a **continuous speech recognizer.** It is also possible to classify them according to the size of the vocabulary they are prepared to handle. In this respect, different authors propose different criteria for division, but in general terms, a distinction between **small vocabulary recognizers** and **large vocabulary recognizers** can be made, the latter comprising a lexicon of 20,000 to over 60,000 words (Jurafsky & Martin, 2009). Additionally, recognizers can be **speaker dependent** or **speaker independent**; in other words, they might be trained to recognize the speech of only one particular user or a previously defined group of users, or they can be conceived to be used by potentially any speaker. Certain authors also propose other classifications, based on criteria such as transmission channel (telephone network or microphone) and response time (real-time or delayed performance) (Tapias, 2002). The degree of difficulty of a recognition task will be highly dependent on the characteristics of the recognizer in connection to the aspects mentioned above, as will be discussed in section 2.5.

2.2) The recognition process

As stated previously, the process of recognition begins when a human speaker utters a word or a series of connected words. The recognizer takes the corresponding signal and digitalizes it. The resulting signal is divided into a series of frames or spectral feature vectors whose duration usually ranges from 10 to 20 ms (Jurafsky & Martin, 2009) and, for each of these segments, information about the patterns of energy distribution at the different frequency levels is extracted. This stage receives the name of **feature extraction**.

Once this information has been obtained, the **decoding** process begins. This operation consists in searching among a set of possible sentences in the target language, in order to find the best match for the input sentence, that is, the sequence of words with the highest probability of representing the speaker’s utterance. This is done by calculating the product of what authors often call the **prior probability** or ***a priori* probability** and the **observation likelihood** or **acoustic likelihood** of the word string (Jurafsky & Martin, 2009; Rabiner & Juang, 2006). The prior probability, symbolized as *P(W)*, represents the probability of a certain sentence *W* being a possible sentence in the target language, whereas the observation likelihood, *P(O/W)*, is the likelihood of *W* producing the observed series of acoustic segments *O*. The sentence which maximizes this product is selected as the best match and constitutes the output of the recognition process.

The linguistic and acoustic information required for this selection is contained in three blocks or modules which aid the system during the decoding process: the acoustic model, the language model and the lexicon (see Figure 1). The **acoustic model** is an inventory of linguistic units together with information about their respective acoustic features. The nature of these units varies depending on the type of system. In small-vocabulary applications, such as those which recognize digits, “yes-no” answers or a reduced number of isolated command words, the acoustic model may consist of whole words. In LVCSR (large vocabulary continuous speech recognizers), however, subword units (usually phones or diphones) are employed, due to the impracticality inherent in having to train one model for each of the thousands of lexical items required. This module contributes the information needed to compute P(O/W), namely, the observation likelihood mentioned above. The other probability which must be calculated, P(W), is determined using the information in the **language model.** This module is usually an N-gram grammar (Jurafsky & Martin, 2009) which contains rules underlying possible combinations of words in the target language, together with their associated probability of occurrence. Finally, the words which are relevant to the recognition task are comprised in the **lexicon**, a set of vocabulary items in connection with their respective pronunciations, which are represented using the basic acoustic units of the recognizer (Lamel & Gauvain, 2003).

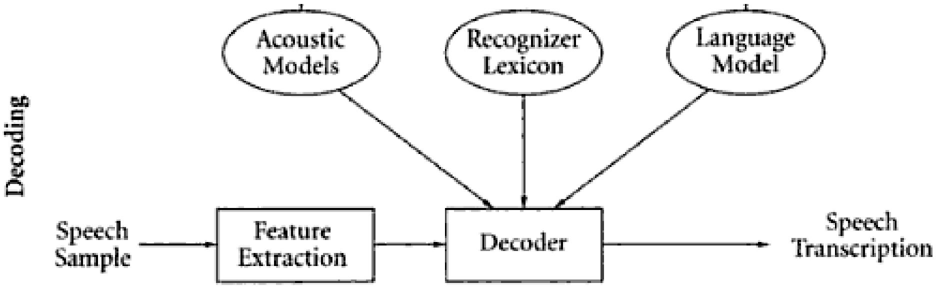


Figure 1: System diagram of the decoding process of an ASR system (Lamel & Gauvain, 2003, p. 306)

2.3) Speech variability

In a hypothetical universe in which all individuals spoke the same language, and they did so in an identical way, speech recognition would be a fairly easy task. The main difficulty which has until now prevented researchers from developing unlimited vocabulary speaker independent applications which perform flawlessly for any user is, precisely, speech variability. This is a concept which describes two types of phenomena. On the one hand, if the same sentence is uttered by two speakers, the outcomes might nevertheless differ dramatically: this phenomenon is known as **interspeaker variation**. Furthermore, even when the same person pronounces an utterance twice, the realizations are likely to exhibit significant differences; this is commonly referred to as **intraspeaker variation**.

Let us now analyze the causes of intraspeaker variability. Strik and Cucchiarini (1999) postulate the nature of connected speech as one of the main factors responsible for the changes in the speech of the same individual. The fact that words are, as Kaisse (1985) puts it, “strung together” (as cited in Strik & Cucchiarini, 1999, p. 226) “[results] in the application of various phonological processes such as assimilation, co-articulation, reduction, deletion and insertion” (Strik & Cucchiarini, 1999, p. 226). According to these authors, this is closely linked to stylistic considerations: when analyzing formal speech, we are less likely to encounter a vast amount of these occurrences than in casual conversation. This issue is discussed in more detail in section 2.3.2. On the other hand, Strik and Cucchiarini (1999, p. 226) also mention **free variation**, a term which describes those instances in which a speaker is able to choose between two equally valid realizations of a word, syllable or phoneme. Finally, the authors also consider the interlocutor to be a source of variation, since human beings show a tendency to adapt their speech according to the characteristics of the listener and/or the environment (Strik & Cucchiarini, 1999, p. 226). Regarding the last point, it is worth mentioning Lindblom’s (1990) H&H theory, which explains intraspeaker variation by focusing on the relation between the content of the speech signal itself and the information provided by the context. Lindblom claims that speakers “tune their performance according to communicative and situational demands, controlling the interplay between production-oriented factors [...] and output-oriented constrains [...] (Lindblom, 1990, p. 403). He further states that an individual’s speech oscillates between “hyper- and hypo-articulation” (hence the name of the theory) depending on the speaker’s perception of the degree of sufficiency of complementary information within the communicative setting (Lindblom, 1990, p. 404).

Variation between speakers, on the other hand, is a highly complex phenomenon, related to a wide range of factors (Stevens, 1972). The ones most frequently mentioned in the ASR literature are:

- **Age**: the speech of each age group exhibits distinctive features, attributable to generational causes as well as anatomic and developmental differences (Benzeghiba et al., 2007).

- **Sex and gender**: several differences exist between the speech of men and women, originating in biological as well as environmental factors. The effects of sex and gender in ASR are discussed in more detail in section 2.3.1.

- **Physical complexion**: even between individuals of the same sex, anatomical differences result in distinct speech characteristics (Hadman et al., 2011).

- **Socio-cultural background and level of formal education**: the type of environment in which an individual is raised and educated exerts an influence upon speech patterns.

- **Speaking style**: this aspect, which has already been mentioned in connection to intraspeaker variation, also plays a role between speakers. Although style is intrinsically related to situational factors (Labov, 1991), some individuals exhibit a general tendency towards more formal or more informal speech (see 2.3.2 for a discussion of speaking styles and their effects on ASR).

- **Geographic factors**: the same language may display significant phonetic variability when spoken by communities inhabiting different geographical areas (see 2.3.3 below).

In addition to these considerations, some authors mention the **environmental context,** which may lead a speaker to alter the volume or the quality of his voice. This is also connected to Lindblom’s H&H theory, mentioned above. **Rate of speech**, that is, the speed at which utterances are delivered, also bears upon phonetic realization: generally, more instances of reduction, assimilation, deletion, etc. are observed at fast rates than in slow speech, in which speakers tend to articulate more carefully (Martínez, Tapias, Álvarez & León, 1997). **Emotional state** also plays a crucial role, since feelings such as excitement, stress, anger, etc., are often reflected in how utterances are pronounced (El Ayadi, Kamel & Karray, 2011). In a similar way, aspects related to **health** must be taken into consideration: these might be temporary alterations or permanent conditions, and some examples are congestion, hoarseness, stuttering, among others. Rate of speech, environmental context and emotional and health state are aspects which may cause both variation between different speakers or within the speech of the same individual. Finally, when dealing with non-native speech, two additional factors may cause variability: **mother tongue** and **level of proficiency** in the target language[1].

Three of the aspects mentioned above in connection to speech variability will be central to the present evaluation. These are sex and gender, language variety and speaking style, and they are discussed in more detail below.

2.3.1) Sex, gender and ASR

Simpson claims that "gender is one of the most important factors that must be considered when trying to account for phonetic variation found within a speech community" (2009, p. 633). Indeed, the distinction between what is commonly referred to as "male speech" and "female speech" has attracted a great deal of attention throughout history, and writings on the subject go as far back as ancient Greek and Roman times (Rissel, 1981, p. 305). Furthermore, a number of beliefs exist which associate men and women with particular speech characteristics (García Mouton, 2000). One of the most widespread is probably the idea that women tend to exhibit a clearer speaking style than men, a concept which Labov (2000) attributes to their role as main providers of linguistic input during child-raising (as cited in Simpson, 2009, p. 636). These differences between male and female speech are particularly relevant in the area of speech recognition, since an understanding of their nature could enable development of more accurate models, which may in turn lead to improvements in performance. Analogously, for purposes of the present analysis, consideration of these differences may allow us to both predict and account for potential disparities in recognition rates.

[1] For more information on speech variability, see Stevens 1972.

Researchers in the area frequently draw a distinction between two main factors responsible for this kind of variation (Foulkes, Scobbie & Watt, 2010, pp. 711-712; Simpson, 2009, p. 621):

- Biologically-determined differences which stem from physical characteristics. These can be categorized according to the binary opposition "male-female".
- Socially-determined differences which result from behavioral patterns learned through immersion in a particular socio-cultural environment. These entail a psychological dimension which is linked to the individual's feeling of identification with a certain gender group, as well as to social conceptions and expectations[2.]

Although the terms "sex" and "gender" are sometimes used indistinctly by certain authors, the former is generally applied when describing the first type of phenomenon, whereas the latter is identified with the second type (Foulkes et al, 2010, pp. 711-712).

At first sight, this distinction may seem quite clear. It is well-known that men and women differ physically in aspects such as larynx, pharynx and vocal tract length, as well as vocal folds dimension. Nevertheless, when it comes to deciding whether the differences observed in speech respond to biological causes or to socially-determined considerations, controversy arises, as it is often difficult to establish a clear division. As an example, differences in pitch level between men and women appear to be easy to explain, given the fact that men's vocal folds are thicker and longer, which causes them to vibrate more slowly, thus producing lower-pitched sounds (Gil, 1988, p. 39;

[2] It should be noted that gender considerations alone may be insufficient to understand certain phenomena, since research has shown that this variable often interacts with other dimensions of psychological identification, such as membership within a socio-cultural group (Nichols, in Rissel, 1981, p. 4; Foulkes, Scobbie & Watt, 2010, p. 712). For purposes of the present study, however, efforts have been made to control the latter variable within the realm of possibility, by only selecting speakers with university studies.

Martínez Celdrán & Fernández Planas, 2007, p. 148). However, it is a well-documented fact that differences in average pitch level exist among different languages even between individuals of the same sex, which seems to indicate that even this feature may be partly learned (Simpson, 2009, p. 625). This difficulty holds for a number of phonetic features, over which contradictory research results exist. In view of this complexity, the present description will focus on the concrete phenomena which distinguish the speech of men from that of women and which might affect automatic recognition, without delving deeply into the nature of their origin.

Rissel (1981, p. 305) states that, in modern Western cultures, contrast between male and female speech may be found at the phonetic, lexical and discourse levels. Within the **phonetic** dimension, the distinction concerning pitch level mentioned above is probably the most noticeable difference. In technical terms, the superior length and thickness of male vocal folds results in a **lower fundamental frequency** (F0) than that of women, and the male voice is therefore perceived by the human ear as having a lower pitch. The average F0 values for male and female speakers usually found in the literature are 125 and 200 Hz respectively, with variations in the range of 80 – 300 Hz for men and 130 – 525 for women (Lieberman & Blumstein, 1988; Orlikoff & Kahane, 1996). Not only is phonation affected by the anatomical differences between men and women, but these also have articulatory implications. The female vocal tract is shorter than the male one (their average measures being 14-14,5 cm and 17-18 cm respectively) and this causes differences in the frequency configuration of vowel formants, which determines **vowel quality** (Simpson, 2009, p. 625).

Another phonetic aspect which has attracted the attention of researchers is related to word pronunciation variants. It has been noted in previous sections that, in actual speech production, it is possible to find word realizations which differ from the standardized norm. Adda-Decker and Lamel (2005) analyzed this phenomenon using French and American English corpora of broadcast news speech and spontaneous telephone speech, in order to detect potential correlations of **standard and non-standard forms** with the gender variable. Their findings revealed that, in the corpora analyzed, women produced standard pronunciations twice as often as men, which might be interpreted as a sign of female speech being “more conservative”. This is also confirmed by the results of a study conducted by Byrd (1992; as cited in Simpson, 2009, p. 631) which analyzed reduction of English vowels to the central vowel [ə] in the TIMIT database. Evidence to the contrary, however, was found in a study on the Spoken Dutch Corpus conducted by Binnenpoorte, Van Bael, Den Os and Boves (2005), who did not detect significant differences in the amount of phone substitutions, deletions and insertions between both gender groups.

On the other hand, agreement is found between Adda-Decker and Lamel (2005) and Binnenpoorte, Van Bael, Den Os and Boves (2005) with respect to **filled pauses and repetitions** since, in both studies, male speech was found to contain a larger number of occurrences of these phenomena. **Vowel duration** also appears to exhibit differences. Studies conducted on German, Quebecois French, American English and Swedish speech corpora have proved vowels uttered by women to be longer than those pronounced by men[3]. Additionally, interesting differences have been detected with respect to **speech rate**. Several studies seem to prove that, on average, men speak faster than women (Adda-Decker & Lamel, 2005; Byrd, in Simpson, 2009, p. 635). On the other hand, Binnenpoorte, Van Bael, Den Os and Boves (2005) did not find significant differences in speech rate, but observed that men exhibited a higher **articulation rate**, i.e. the percentage of words uttered per second without considering silent pauses.

Let us now analyze the implications of the phenomena described above for automatic speech recognition. Differences in F0 and vowel quality might, in principle, have a bearing upon automatic recognition rates. Therefore, numerous mechanisms have already been developed to address these specificities. These measures include the elaboration of gender-dependent acoustic models, as well as the implementation of techniques such as speaker adaptive training, unsupervised adaptation and vocal tract

[3] This tendency was found to apply to certain front vowel categories in the study performed on French, and to all vowel categories in the other studies.

length normalization[4] (Adda-Decker & Lamel, 2005, p. 2205). State-of-the-art recognizers are thus expected to be equipped with the tools required to provide high-level performance regardless of F0 and vowel formant differences.

The rest of the phenomena described deserve closer attention. The characteristics described in connection to female speech, i.e. a tendency towards standard pronunciations, fewer instances of repetitions and fillers, longer vowels and a lower speech rate, do not only appear to be in keeping with the popular belief that "women speak more clearly than men" (Simpson, 2009, p. 632), but could also facilitate automatic recognition. This would justify potential higher word error rates for male speakers. This is the case in the results obtained by Adda-Decker and Lamel (2005), where error rates for English- and French-speaking women were found to be 0,7 to 7% lower than those for men. Despite the existence of modeling techniques aimed at addressing this type of variation, these phenomena still appear to remain problematic for ASR. All these considerations will be taken into account for purposes of the present research, with the hope that the analysis might throw some light upon the issue.

2.3.2) Speaking style and ASR

Another factor on which this research will focus is speaking style and its effects on ASR performance with respect to the system under evaluation. There seems to be disagreement in the literature regarding speaking style categories. While Eskénazi (1993, p. 507) postulates that "the concept of speaking styles has to present been loosely defined with little theoretical basis", Aguilar and Machuca (1994, p.7) focus on the widely accepted dichotomy "laboratory speech" vs. "spontaneous speech", which is based on whether the samples are obtained under controlled conditions or not, as well as on whether they consist of texts which have been read aloud or, on the

[4] For a general discussion on techniques for modeling speech variation, see 2.3.2.

contrary, feature freely-occurring speech. They characterize this distinction as ambiguous, highlighting the need for a classification system which better reflects the complex interplay between the linguistic and extralinguistic elements of the communicative event. Finally, Llisterri (1992, p. 21) emphasizes the lack of clear correspondence between the multiple labels used by phoneticians dealing with stylistic variation on the one hand and the criteria employed in sociolinguistic research on the other, especially with reference to the continuum of speaking styles outlined by Labov (1972, pp. 79-85).

Despite the heterogeneity in the terminology employed, the literature written on the subject of speaking styles appears to present certain recurrent concerns. One of them is the interest which has been placed on the characteristics of the kind of speech which is made up while uttered (as it occurs in most real-world interactions) as opposed to speech which has been previously planned (as it happens when reading aloud). The former has been termed by different authors “spontaneous” or “unscripted”, whereas the latter has been categorized as “scripted”, “connected” or “read speech” (Llisterri, 1992, pp. 18-19). The strategy used in task 2 in our experiment might give rise to an intermediate style between purely spontaneous speech and read speech. Another distinction, closely related to the previous one, is based on level of formality. Such distinction is represented in the literature through the use of terms such as “casual” or “informal” speech as opposed to “formal” or “careful” speech (Eskénazi, 1993, p. 503).

Different acoustic and articulatory correlates have been assigned to these categories. At the segmental level, spontaneous speech has been found to exhibit a higher degree of hipo-articulation when compared to read speech (Aguilar, Blecua, Machuca, & Marín, 1993; Eskénazi, 1993, p. 504). Eskénazi (1993, p. 504) notes that “articulatory targets are reached much more often in clear, or read speech than in casual speech”, particularly in the case of consonants. Additionally, Llisterri (1992) reports a higher frequency of elisions and vowel reductions, as well as of coarticulatory processes, in spontaneous speech (p.13, p.17), . At the suprasegmental level, the author reports higher fundamental frequency averages and longer tone units in reading as compared to unscripted speech. Furthermore, Eskénazi (2006, p. 506) postulates the occurrence of more ungrammatical pauses as a feature of spontaneous speech. Another relevant aspect is the amount of disfluency features (Benzeghiba et al., 2007): as mentioned in 2.3.2, the need to decide in real time what is going to be said contributes an extra cognitive load, which is manifested in the appearance of false starts, repetitions, hesitations, filled pauses, etc.

Specifically in the case of Spanish, Aguilar et al. (1993) conducted a study aimed at analyzing reduction processes affecting consonants in spontaneous speech. They observed a higher degree of weakening and deletion in this speaking style than in reading style. They also focused on variation resulting from differences in the "degree of casualness" (p. 436) through the comparison of samples of conversational speech with samples of monologues. As a result, they claimed that the former style, which is usually associated with a more casual style than the latter, exhibits a larger number of processes affecting voiceless stops, which are either given a voiced realization, produced as unreleased or replaced by approximants. In contrast, in the case of monologues, voiceless stops tend to retain their canonical features.

All these issues have important implications for automatic speech recognition. As it has been suggested in section 2.3, one of the main hindrances to satisfactory performance are the challenges posed by speech variability. The above considerations illustrate how speaking style plays a key role in this respect. The degree of difficulty that this may offer a particular system is highly dependent on the characteristics of the training data. In this respect, Colley (2009) makes the following assertion:

> "Since no corpus can represent the totality of human language, speech recognition systems are always biased toward a particular style of language (usually written, since most large-scale corpora, such as the 100-million-word British National Corpus, consist predominantly of written texts)". (p. 3)

The core of Colley's observation consists in the idea that the corpora normally used for training the language models are not appropriate to prepare the system for the kinds of variation encountered in spontaneous speech. Analogously, the degree of similarity between the speech data used for training the acoustic models and the actual speech to be recognized is crucial in determining performance quality. Rodríguez and Torres (2006) compared two speech databases of human-human and human-computer dialogs, detecting a higher rate of disfluencies in the latter than in the former (p. 345), a disparity which may also result in difficulties for recognition if data such as the former is used for training and, subsequently, the actual speech uttered by the user resembles the latter.

Additionally, (Colley, 2009, pp. 10-12) emphasizes the crucial role that context plays on recognition, since phonetically-reduced words, as well as words which significantly deviate from their standard form, are often impossible even for humans to recognize when dislocated from their context of occurrence and presented in isolation. This indicates that, in continuous speech, words do not always contain the phonetic information needed for their identification, which in terms of automatic recognition would imply that well-trained acoustic models do not necessarily guarantee success in performance. Consequently, an ASR system intended for spontaneous speech needs to count on mechanisms to compensate for this lack.

2.3.3) Language varieties and ASR

As mentioned in 2.3, diatopic differences within a language, i.e. phenomena resulting from geographically-based linguistic heterogeneity, may also pose difficulties for automatic speech recognition. Consequently, state-of-the-art ASR systems include modeling methods which increase robustness to this kind of variation (see 2.3.4). The following sections will focus specifically on those language varieties which are relevant to the present evaluation: Spanish as spoken in Catalonia, Spain, and Spanish as spoken in Buenos Aires, Argentina (see 3.1).

2.3.3.1) Spanish in Spain or "Spanish from Spain"?

Despite the frequent use in non-specialized contexts of terms such as "Spanish from Spain", the idea that in this territory such language constitutes a homogeneous system presenting only minor differences among geographical areas is a misconstruction. Historical events and processes have conditioned and determined the evolution of this language (Zamora Vicente, 1967) and, as a result, the phonetic map of Spanish as spoken along Spain currently exhibits great diversity and richness of phenomena. Furthermore, the specificities which characterize the speech of the different linguistic communities do not only correspond to the phonetic-phonological level, but also concern morphosyntactic and lexical aspects[5].

The persistent use of the expression "Spanish from Spain" in spite of this heterogeneity seems to evoke the concept of *standard variety*, which refers to a linguistic variety accepted socially as a model of prestige (Carbó et al., 2003). In the case of Spanish, such standard has historically been represented by the Castilian norm (Navarro Tomás, 1999, p. 8) (Ávila, 2009, p. 1). In the last decades, however, alternative trends have emerged which view the standardization of Spanish in a different light. Carbó et al. (2003) state that, given the large number of Spanish speakers and the vast geographical dimensions of the Spanish speaking world, it is not possible to establish a unique standard variety, hence a standard for each linguistic area should be acknowledged. Fontanella (1983, p. 45, as cited in Rigatuso, 2004, p. 14) describes this situation through the use of the term *polycentric standardization*, formerly proposed by Stewart (1970, p. 534) to designate the simultaneous existence of different standard forms in a language. The *Real Academia Española* (RAE) also uses the term *policéntrico* to describe the situation of the Spanish language, claiming that all regional linguistic uses are fully legitimate, insofar as they are generalized among educated speakers from the area and they do not threaten the unity of the system as a whole (RAE, n.d.)

[5] For detailed information on phonetic/phonological, morphological, syntactic and lexical differences among Spanish varieties in Spain, see (Alvar, 1999)

These considerations bear a direct relation to the present research. The Spanish version of Windows 7 ASR, the system which will be evaluated, is presented as a recognizer for “Spanish from Spain”. In this context, the use of the term has probably been adopted for practical reasons, as a succinct way of expressing that the system can recognize speech from all the Spanish varieties spoken in the Iberian Peninsula and the Balearic and Canary Islands. This in turn should imply that such varieties were represented in the training corpus through the inclusion of a sufficiently large number of samples of each of them.

This has further implications. Although the cradle of Spanish can be found within the territory of what is now politically considered Spain, it is currently the official language in more than twenty countries and spoken by over 425 million people (Centro de noticias ONU, 2006). This has earned Spanish the name of “extended language”, in Guitarte’s words (1991), or as Hock (1986) puts it, “transplanted language”. Moreover, the significant distance which separates Spain from the other Spanish-speaking countries (all of them situated in the American continent except for Equatorial Guinea in Africa) has inevitably resulted in the emergence of substantial linguistic differences.

It is this reality that leads to the question of how Windows 7 ASR will perform when faced with non-Iberian users. To explore this issue, the experiment will include data from two native varieties of the language: Spanish as spoken in Catalonia, Spain, as an example of an Iberian variety, and Spanish as spoken in Buenos Aires, Argentina (sometimes referred to as “*Bonaerense*” or “*Porteño*” Spanish), as an example of a Latin American variety. The next sections will thus be devoted to a description of certain phonetic/phonological, morphological and lexical features of *Bonaerense* Spanish which distinguish it from the Iberian variety selected. The fact that the focus is on Spanish from Buenos Aires is based on the premise that, if the data used during system training covers a wide range of Iberian varieties, those characteristics which are specific of Spanish from Catalonia should not represent a hindrance to recognition, unlike those specific of *Porteño* Spanish, which could bear negatively upon system accuracy[6].

2.3.3.2) Phonetic differences

a) Palatal and pre-palatal consonants

Undoubtedly, the most salient characteristic of *Porteño* Spanish pronunciation is a phenomenon which has traditionally been referred to as "rehilamiento" or "yeísmo rehilado". In order to understand this process, it is first necessary to introduce some general concepts associated to other Spanish varieties, both peninsular and non-peninsular.

Most Spanish varieties include the phoneme /ʝ/ in their consonantal inventory[7]. It is normally realized as voiced, fricative and palatal, except for some cases in which it may be pronounced with affrication (D'Introno, Del Teso, & Weston, 1995, p. 305). This sound occurs as the phonetic manifestation of grapheme <y> in syllable-initial position: yo, [ˈʝo], ayuda, [aˈʝuð̞a].

At this point it is necessary to draw a distinction with another consonant, /ʎ/. This sound coincides with the one described above in terms of place and voicing, the difference lying in its lateral mode of articulation. /ʎ/ coincides with grapheme <ll> and also occurs in syllable-initial position: *calle [ˈkaʎe]*; *lluvia [ˈʎuβ̞ja]*.

Despite this distinction, it is nevertheless essential to notice that /ʎ/ has disappeared from a great number of peninsular varieties as well as from most Latin

[6] Details of the experiment variables can be found in chapter 3. The present section is limited to theoretical considerations which might shed light on posterior analysis.

[7] Although consideration of this segment as a phoneme has been the subject of much controversy, in this work it will be treated as such for practical purposes, following Hualde (2005, p. 172)

American ones as a result of a widespread phenomenon called *yeísmo*. This phenomenon consists in the replacement of /ʎ/ with /ʝ/ in all the contexts in which the former would occur. Consequently, in these areas, the word *calle* is pronounced *[ˈkaʝe], lluvia* is *[ˈʝuβja],* etc. These Spanish varieties are described as *yeístas,* in contrast to those referred to as *lleístas*, in which the opposition /ʎ/ - /ʝ/ is still retained (Quillis, 1993, pp. 315-324). Catalonia has been described as mainly *lleísta*, with the exception of certain areas, see (Calero Fernández, 2006, p. 208)[8].

Although the concepts of *lleísmo* and *yeísmo* account for the behavior of these sounds in almost the whole Spanish-speaking world, the variety spoken in Buenos Aires exhibits a completely different panorama, since none of the sounds discussed above, i.e. neither /ʝ/ nor /ʎ/, is present in the consonantal inventory. Instead, a postalveolar[9] fricative consonant /ʒ/ replaces them in their contexts of occurrence. This sound exhibits a variety of allophonic realizations: It may be produced as a voiced fricative [ʒ], a voiceless fricative [ʃ] or a voiced postalveolar affricate [ʤ][10]. The fricative sounds occur in free variation, whereas affrication seems to be favoured by certain contexts, such as initial position after pause or emphatic realizations (Fernández Trinidad, 2010, pp. 267-268). Currently, the voiceless fricative realization is becoming increasingly widespread, particularly in the speech of the younger generations (Hualde, 2005, p. 56) This phenomenon is traditionally referred to as “yeísmo rehilado” or “rehilamiento” and, according to Fontanella (2004, p. 45) it is one of the two linguistic features whose combination results in the very peculiar

[8] This preservation of /ʎ/ in some areas seems to be attributable to the influence of the Catalan phonemic system (García Mouton, 1994, p. 45).

[9] Although this sound is described as postalveolar in the International Phonetic Alphabet (IPA), it is sometimes also referred to as prepalatal, particularly in the literature on Spanish phonetics (Hualde, 2005, p. 48). In this dissertation, the designation proposed by the IPA will be used.

[10] Several sociolinguistic studies have attempted to identify the factors which determine preference for each of these realizations, and the variables *speaking style*, *gender* and *socio-cultural group* have been signalled as relevant (Fernández Trinidad, 2010, p. 279, p. 289) (Guitarte, 1955, p. 270, as cited in Fontanella, 2004, p. 15).

character of *Porteño* Spanish, setting it apart from the rest of the Spanish varieties in the world. The second feature, called *voseo,* will be described in the section entitled “morphological differences”.

b) “Seseo”

In Northern-Central Peninsular Spanish varieties, an opposition exists between the alveolar voiceless fricative /s/ and the interdental voiceless fricative /θ/. This opposition is absent from the vast majority of non-Iberian varieties, including *bonaerense* Spanish, in which /s/ replaces /θ/ in all its contexts of occurrence. The replacement of the interdental fricative /θ/ with the alveolar fricative /s/ and the consequent loss of opposition receive the name *seseo*. This phenomenon is not exclusive of Latin American Spanish: it is also generalized in the autonomous communities of Andalusia and the Canaries (Navarro Tomás, 1999, pp. 93-94)[11]. Consequently, words such as *ceniza* or *difícil*, which receive the pronunciation *[θeˈniθa]* and *[diˈfiθil]* in Northern peninsular varieties, are realized as *[seˈnisa]* and *[diˈfisil]* in *seseante* varieties.

Although the amount of training data containing pronunciations which maintain the opposition /s/-/θ/ is likely to outnumber that of *seseante* varieties, this phenomenon should not, in principle, constitute an obstacle for the system under evaluation: given that the phenomenon is present in certain peninsular varieties, it would be natural to expect it to be represented in the training corpus.

c) Realization of /s/ in syllable-final position

D’Introno et al. (1995, p. 289-290) enumerate the different allophonic realizations of /s/ in syllable-final position in the Castilian area and in the North of

[11] Navarro also mentions the existence of instances of *seseo* among the working classes of Valencia, Mallorca, Catalonia and the Basque Country (1999, p. 94).

Spain. They postulate the following system, in which allophones occur in complementary distribution depending on the sound they precede:

/s/	[s]	voiceless fricative (before voiceless plosive or silence)
	[^{s}]	lax voiceless consonant (before voiceless fricative)
	[z] [^{z}]	voiced fricative, sometimes lax (before voiced consonant)
	[ɹ]	when followed by [r], /s/ is assimilated to an alveolar approximant rhotic sound

.*Figure 2.* Allophonic realizations of /s/ in syllable-final position in the Castilian area and in the North of Spain (D'Introno et al., 1995, p. 289-290).

The authors state that the realizations [s] and [^{s}] can also be found in the Spanish of Andalusia and Latin America, and they add two more possibilities for these varieties: an aspirated realization, [h], and the elision of the sound. They claim that, in contrast to the system in figure 2, the mentioned allophones occur in free variation (p. 291)[12].

The particular case of Spanish from Buenos Aires exhibits some specific characteristics. Among middle-class speakers, elision of syllable-final /s/ is considered a stigmatized feature (Lipski, 1994, p. 169); hence, it is usually avoided by speakers above a certain level of formal education. In prevocalic position the predominant realization is sibilant [s], whereas in preconsonantal contexts, aspiration, i.e. [h] (a

[12] The authors also include a fifth variant, a sound which is assimilated to the following consonant and which, for this reason, does not occur in free variation with the others.

voiceless glottal fricative), is observed in most cases (Fontanella, 2004, p. 47). Thus, whereas in Northern Peninsular Spanish varieties, words such as *espera* and *plasma* are pronounced *[esˈpeɾa]* and *[ˈplazma]* respectively, in Porteño Spanish the most frequent realization would be *[ehˈpeɾa]* and *[ˈplahma].* As in the case of *seseo* discussed above, although the number of instances of aspirated preconsonantal /s/ in the corpus might be more reduced than that of alveolar voiceless fricatives [s] or their voiced counterpart [z], aspiration of /s/ in preconsonantal contexts should not pose difficulties for recognition when using a system trained for Iberian Spanish. This is due to the fact that the phenomenon is represented in certain peninsular varieties, such as the ones spoken in Andalusia.

2.3.3.3) Morphosyntactic differences

The most striking morphosyntactic feature of the Spanish variety spoken in Buenos Aires is the complete replacement of the informal second person singular pronoun *tú* with the pronoun *vos.* This phenomenon, called ***voseo***[13], is generalized in oral as well as written discourse (Donni de Mirande, 1996, p. 215) (Fontanella, 2004, p. 50-52), and there is an ever increasing tendency towards the use of *vos* in contexts in which *usted,* its formal counterpart, was formerly used. This includes contexts such as broadcasting and advertising, as well as everyday language (Carricaburo, 1997, p. 24) (Fontanella, 2004, p. 52-54).

The relevance of these observations in terms of the evaluation described in this work lies in the fact that the phenomenon of *voseo* is absent from current European Spanish (Fontanella, 2004, p. 50); therefore, it is likely not to be represented in the training corpus. In this respect it should be noted that, although the verbal inflection paradigms corresponding to pronouns *tú* (*tuteo*) and *vos* (*voseo*) coincide for some

[13] The phenomenon of *voseo* is generalized over a number of Latin American countries, nevertheless presenting distinct characteristics in each of them (RAE, 2005)

tenses, they differ for others, as can be seen in table 1. Thus it follows that higher error rates might be observed when speakers employ these forms during automatic dictation.

TENSE	"TUTEO"	"VOSEO"
PRESENT - INDICATIVE	cantas comes vives	cantás comés vivís
SIMPLE PAST - INDICATIVE	cantaste comiste viviste	cantaste / cantastes[14] comiste / comistes viviste / vivistes
PRESENT - SUBJUNCTIVE	cantes comas vivas	cantes / cantés[15] comas / comás vivas / vivás
IMPERATIVE	canta come vive	cantá comé viví

Table 1– comparison of verbal forms corresponding to the pronoun tú (tuteo) and to the pronoun *vos* (voseo), the latter as employed in Buenos Aires (RAE, 2005)

Other morphosyntactic characteristics which differentiate *Porteño* Spanish from Iberian Spanish are the predominance of the periphrastic future over the synthetic one, and of the simple past forms over compound ones (Donni de Mirande, 1996, p. 217).

2.3.3.4) Lexical differences

The contrast between the Spanish varieties spoken in the Iberian Peninsula and the one spoken in Buenos Aires can also be appreciated at the lexical level. Some of these differences relate to words which are shared by several Latin American varieties,

[14] Only the first form of each pair, namely the one which coincides with the form of *tuteo*, is accepted within the standard variety. The second form is generally viewed as substandard.

[15] See note 13.

whereas others are specific of *Rioplatense* Spanish, i.e. the language variety spoken in the region surrounding *Río de la Plata* River, which includes the areas of Buenos Aires and Montevideo.

The existence of these differences may, at first, seem contradictory to the assumption that Peninsular Spanish is, in effect, the main source of *Bonaerense* lexicon. This can be explained through the fact that many of these terms are no longer used in everyday European Spanish. Some examples are the word pairs *lindo – hermoso*, *pollera – falda, vidriera – escaparate*, in which the first element is normally used in Buenos Aires, whereas the second one is the most habitual in Iberian Spanish (Fontanella, 2004, p. 61).

Other sources of *Porteño* lexicon are also mentioned in the literature. They include indigenous and African languages, which have contributed with words such as *choclo* (as opposed to *maíz,* "corn") or *banana* (*plátano*); Italian, from which high-frequency words such as *chau* (vs. *adiós*, "goodbye") and *pibe* (informal word for "boy") stem and, more recently, English (Fontanella, 2004, p. 62-66). The influence of the latter may be encountered in borrowings (*mouse, e-mail, CEO*) as well as adaptations (*tipear*, "to type").

2.3.4) Modeling speech variation

As it has been shown above, speech variability is a highly complex phenomenon, determined by a variety of factors, and each of them may affect oral performance in several ways. It follows that, in order to achieve satisfactory recognition rates, ASR systems must include models which take all of these factors into account. Firstly, it must be noted that the linguistic information required to build these models can be obtained in two different ways (Benzeghiba et al., 2007; Strik & Cucchiarini, 1999). **Knowledge-based methods** rely on information which is available *a priori*, usually derived from linguistic studies or pronunciation dictionaries. **Data-driven methods**, on the other hand, consist in the *ad hoc* extraction of information from speech corpora through an analysis of the speech signal and the subsequent elaboration of transcriptions of the variants encountered. The transcriptions can be produced manually or automatically, in the latter case using a phone recognizer or through forced alignment. Details about knowledge-based and data-driven methods can be found in (Strik and Cucchiarini, 1999) and in (Colley, 2009, pp. 2-6).

One of the objections to the knowledge-based approach is that, while pronunciation dictionaries and linguistic studies may provide valuable information regarding the phonetic form of word variants, frequency of occurrence (which constitutes crucial information for best-match selection during the decoding stage) is usually not contemplated in these works. Furthermore, the knowledge contained in many of these sources is not oriented towards spontaneous conversation, but it concerns other speaking styles. Therefore, the representations obtained might be inaccurate or insufficient for certain systems or tasks (Strik & Cucchiarini, 1999, p. 231).

Data-driven methods, on the other hand, may also pose problems. Firstly, the information obtained from the corpus is normally applied to one particular recognizer and cannot be used for other situations, which entails that every time a new ASR system is developed, a new corpus must be chosen and the operations of signal analysis and transcription must be performed again. Moreover, when using this type of method, special attention ought to be paid to the representativity of the corpus in order to avoid undercoverage, namely, insufficient coverage of the vocabulary required for that particular task (Strik & Cucchiarini, 1999, pp. 231 - 232).

Techniques to enhance ASR robustness to speech variability may be applied at different levels. Based on this distinction, Benzeghiba et al. (2007) classify them into three groups: *front-end techniques, acoustic modeling techniques and pronunciation modeling techniques*.

Front-end techniques focus on the feature extraction process. The authors cited above highlight the potential of these methods in addressing the obstacles imposed by the non-stationary nature of the speech signal. They also postulate their value for compensation of differences originated in speaker physiology, among other sources.

Acoustic modeling techniques rely on the assumption that "good performance is generally achieved when the model is matched to the task, which can be obtained through adequate training data" (Benzeghiba et al., 2007, p. 774). One possibility within this framework is acoustic model **adaptation**. When a set of conditions, such as a particular configuration of environment or speaker characteristics, can be considered relatively permanent for a certain application and/or task, adaptation of the acoustic models to these conditions may prove beneficial. Another alternative is **multiple modeling**, which, as the name indicates, consists in the elaboration of several separate models trained with subsets of data, instead of a general one trained with all the data present in the corpus. The data in each subset ought to be as homogeneous as possible, so that specialized models can be obtained, each of them suited for a specific set of conditions. Sex-specific models are frequently obtained in this way. Benzeghiba et al. (2007, p. 775) also mention two studies in which this technique was used to deal with regional variation.

Additionally, Strik and Cucchiarini (1999) propose two acoustic modeling strategies for the enhancement of previously existing acoustic models and their ability to deal with speech variability. The first one is **model optimization**. This is achieved by obtaining improved transcriptions of the speech signal through forced alignment and using them to re-train the models. Furthermore, the process can be iterated, resulting in an optimization procedure which these authors call **iterative transcribing**. The second strategy is **modification of** the **basic units** used by the acoustic models and selection of more appropriate ones. Although phones are chosen as basic units in most ASR systems, sub-phonemic units have been used in some studies in order to model pronunciation variation (Strik & Cucchiarini, 1999, p. 235). Furthermore, units larger than the phone (such as demi-syllables or syllables) may also be employed, and even whole words could be included, provided that their frequency of occurrence is high enough to justify this choice.

Pronunciation modeling techniques are applied at the level of the lexicon, and they are sometimes used to address variations caused by different types of speech (such as read vs. spontaneous speech), regional variants and foreign accent. The most common approach consists in the addition of alternative phonetic forms to the words in the lexicon, in order to reflect the possibility of some orthographic forms being realized in different ways in speech. Another strategy is the inclusion of multi-words in the dictionary (Strik & Cucchiarini, 1999, p. 229). Multi-words are strings of words which are added to the lexicon as whole, single units. This technique is employed in order to model cross-word processes which often take place in spontaneous speech (particularly at fast rates), such as assimilation and deletion, among others. The assumption underlying the use of pronunciation modeling techniques is that a lexicon which better describes the variability that characterizes actual speech will result in higher recognition rates. This concept, nevertheless, ought to be taken cautiously, as an increase in the volume of the dictionary may lead to more acoustic confusability and, consequently, have a negative impact on recognition accuracy (Strik & Cucchiarini, 1999, p. 233). A possible solution lies in careful selection of the units, so that only those phonetic forms which are frequent enough to improve the general performance of the recognizer are included; or as Strik and Cucchiarini express it, “adding only the set of variants for which the balance between solving old errors and introducing new ones is positive" (1999, p. 233).

2.4) Evaluation of ASR systems

The effectiveness of the modeling techniques presented above can be assessed through an appropriate evaluation process. Pallett and Fourcin (1996, p. 1) affirm that:

> *Assessment and evaluation are concerned with the global quantification and detailed measurement of system performance. Disciplined procedures of this type are at the heart of progress in any field of engineering. They not only make it possible to monitor change over time in a given system and meaningfully compare one approach with another; they also usefully extend basic knowledge.* (Pallett and Fourcin, 1996, p. 1)

Therefore, evaluation should be regarded as an inherent part of the process of creation of any software system, as it can provide the developer with valuable feedback on the strong and weak areas of the product, enabling him to introduce the necessary improvements. Apart from measuring performance at one particular moment, developers might also be interested in monitoring progress of the software during an extended time period, and the impact of all the modifications and adjustments introduced within that span, which can also be achieved through a suitable type of evaluation. Additionally, evaluation is sometimes used for comparison between the performances of two or more systems. Finally -and perhaps more importantly- the results obtained from evaluation processes extend the existing knowledge, contributing to scientific advancement.

2.4.1) Types of evaluation

Pallett (1985) establishes a distinction between **benchmark tests** and **application tests**. The former are considered to be more adequate for comparative evaluation, since they are characterized by previous definition of certain "benchmark conditions" which enable posterior comparison between the results corresponding to different systems or applications. As an example of these conditions, Pallett (1985, p. 374) mentions "use of a standard speech vocabulary and data base, and no use of syntax to actively control the recognition vocabulary". Nonetheless, if the objective is not to compare different recognizers, but rather to assess performance of one specific real-world system, an application test may be employed. In this type of test, the conditions ought to reflect the ones which characterize real use of the application (or as Pallett puts it, "simulate" the application (1985, p. 375)).

Jekat and Schultz (2004) provide further criteria for classification of tests. Table 2 lists and explains the categories proposed by these authors, together with Pallet's distinction discussed above.

Application tests vs. Benchmark tests	Application tests simulate the specific conditions of a real-world application. Benchmark tests are performed under pre-defined standard conditions.
User-oriented vs. Developer-oriented evaluation	User-oriented (also known as "adequacy oriented") evaluation takes into account the needs of the potential user of a specific software. Developer-oriented (also called "progress-oriented") evaluation addresses the needs of the system developer. Both types may be regarded as complementary and be used in combination.

Black-box vs. Glass-box tests	Black-box tests focus only on those aspects of the system which are accessible to the user. In glass-box tests, internal characteristics of the system, to which the user does not normally have access, are examined.
Static vs. Dynamic analysis	In static analysis, the program is analyzed manually or automatically, without being executed. Dynamic analysis is performed on the basis of system execution.
Field tests vs. Laboratory tests	In field tests, the program is handled by a member of the target user group, in the same environment in which it is normally used. Besides overall system performance, the influence of the environment on user behavior is taken into account in the analysis. This type of evaluation is normally applied to systems which are already fully-developed. Laboratory tests allow isolation and analysis of particular aspects of system performance. Applications which are still in development can nevertheless be tested through this method.

Table 2: Types of tests for speech processing systems (Pallett, 1985, pp. 374-375; Jekat & Schultz, 2004, pp. 574-575)

2.4.2) Assessment metrics

The results of the evaluation should ultimately be associated with a numeric value, in order to provide a more objective characterization of performance levels on the one hand, and to facilitate comparison on the other hand. This value is usually expressed in terms of a conventional assessment metric. The most frequently used metric is **word error rate** (WER). WER is calculated by comparing the **reference** text, which is a transcription of the actual words uttered by the speaker during the test tasks, and the **hypothesis**, namely the text produced by the recognizer as output, and subsequently calculating the **minimum edit distance** between them. This value represents the minimum number of modifications required to transform one of the texts into the other one. The procedure to obtain it is the following. First, the total number of errors in the hypothesis is calculated. Three error classes can be distinguished: **insertions**, **deletions** and **substitutions**. The sum of all the errors is then divided by the total amount of words in the reference text. The result obtained (usually expressed as a percentage) is the word error rate (Jurafsky & Martin, 2009, p. 362; Jekat & Schultz, 2004, p. 581):

$$\text{WER} = \frac{\text{deletions + substitutions + insertions}}{\text{words in reference text}} \;.\; 100$$

Another frequently used metric is **word accuracy** (WA), the difference between WER and 100 (Jekat & Schultz, 2004, p. 581):

$$\text{WA} = 100 - \text{WER}$$

An additional alternative is **sentence error rate**, which represents the percentage of sentences in the text with at least one error (Jurafsky & Martin, 2009, p. 362).

Some authors have pointed out certain drawbacks in connection to these metrics. Pallett (1985, p. 375) suggests that such measures in isolation are insufficient to provide a clear account of system performance, and he emphasizes the need for complementary mechanisms, such as confusion matrices, which offer information on the most frequently occurring confusion pairs. Jurafsky and Martin (2009, p. 364), discuss improving WER by associating a certain weight to words depending on their grammatical category, so that content words (such as nouns, adjectives, verbs and adverbs) are assigned a higher value than function words (prepositions, articles, conjunctions, etc).

2.5) Performance of ASR systems

A comparison in terms of WER between the first ASR systems and the ones available at present would clearly illustrate the dramatic progress made in the field during the last decades. Figure 3 shows the results obtained in the well-known DARPA-NIST benchmark tests between 1988 and 2009. It is noticeable how, as time passes, word error rates for each test type generally tend to decrease, particularly in the earliest evaluation programs.

As mentioned in section 2.1, the degree of difficulty of a speech recognition task is closely related to the type of speech which is to be processed and to the size of the vocabulary required for the task, as well as other factors such as speaker-dependence/independence and processing time. Recognition of isolated words tends to be more simple than continuous speech recognition, especially when the task vocabulary is reduced, as in the case of *yes-no* or *digit recognition*. The same can generally be claimed concerning recognition of one-word commands, in which the lexicon comprises a limited number of words. Besides vocabulary size, another element which makes isolated-word recognition easier is the absence of coarticulation effects. In the context of ASR, coarticulation is defined as a phenomenon which stems from the very essence of continuous speech, in which sequences of words are uttered as connected strings. This causes certain sounds, particularly those in word-initial or word-final position, to modify the phonetic realization of adjacent words, posing difficulties for recognition. Word error rates reported in 2003 for digit recognition were already lower than 1% (Lamel & Gauvain, 2003), and by 2006 they had decreased up to 0,5% (Jurafsky & Martin, 2009).

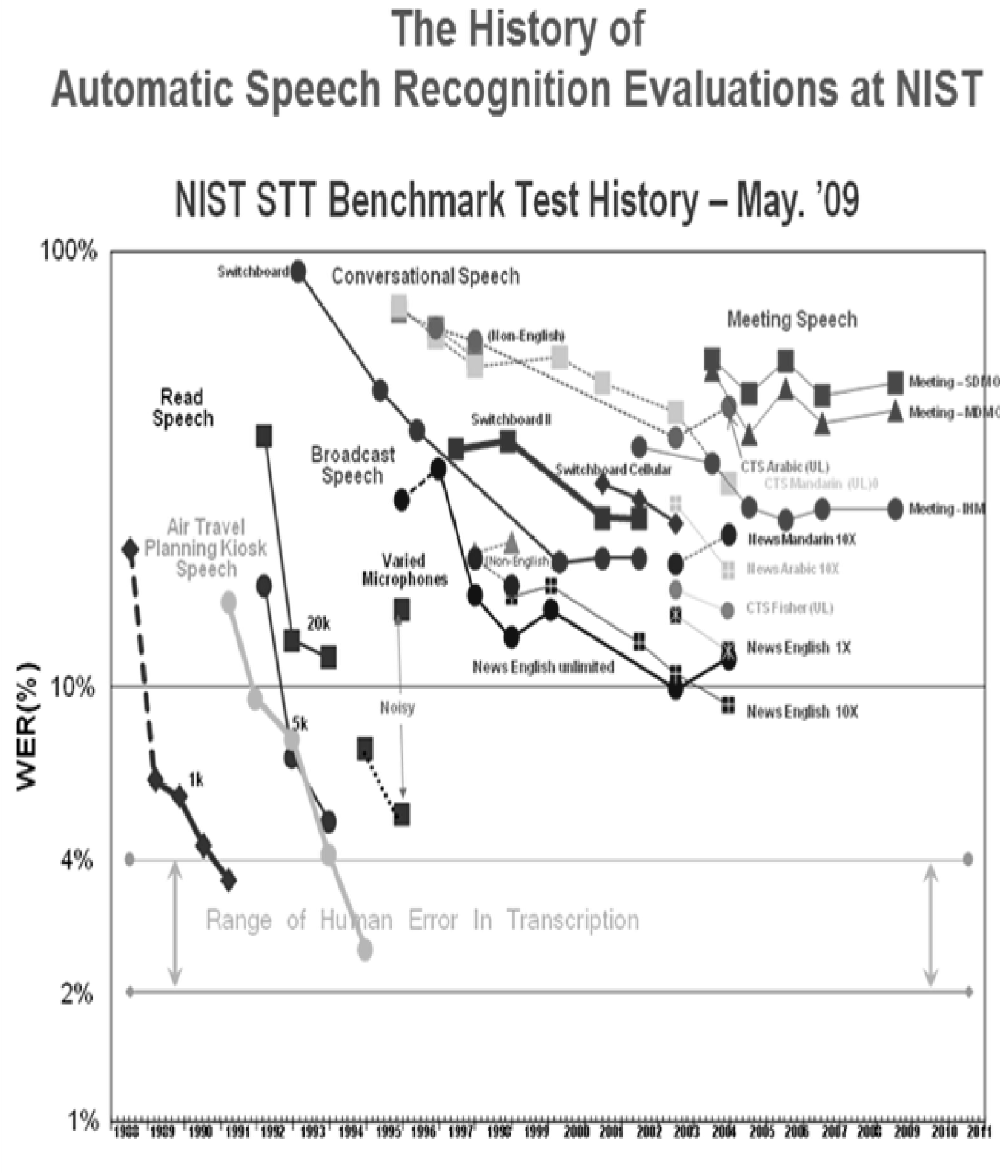


Figure 3. History of NIST ASR evaluations (NIST IAD, 2009)

In the realm of continuous speech recognition it is possible to encounter different levels of difficulty. This variation results from a number of factors (some of which have been briefly discussed in section 2.3). A central distinction is the one between planned and unplanned speech. In the latter, the need to decide in real time what is going to be said contributes an extra cognitive load, which is manifested in the appearance of disfluency features: false starts, repetitions, hesitations, filled pauses (Benzeghiba et al., 2007), etc. These occurrences can be a significant hindrance to recognition, and ASR systems intended for spontaneous speech ought to be equipped with the necessary mechanisms to handle such phenomena. On the contrary, in speech which is not spontaneous, such as that resulting from reading aloud tasks, disfluencies tend to be much less frequent. Additionally, certain speech types could be placed at a half-distance between these two extremes: for instance, the outcome obtained when delivering a lecture or elaborating a message in real time while reading from reminder notes. We shall return to this issue later, since one of the tests in the evaluation which will be presented in this paper includes such a task.

Another aspect that might bear upon recognition success (briefly mentioned in section 2.3) is style. Informal conversational speech is without doubt one of the biggest challenges for ASR for several reasons. In general this style is associated with faster speech rates, which in turn gives rise to a greater number of disfluencies and coarticulation effects (Benzeghiba et al., 2007, p. 766). On the other hand, the vocabulary, expressions, and even the grammatical patterns employed in casual conversation may sometimes differ radically from what is considered canonical in the language, hence the need for appropriate training corpora which resemble the type of speech for which the recognizer is intended as faithfully as possible.

All the above considerations can be clearly appreciated in the WER data available for different types of systems and tasks. Results published around 2006 for recognition of read articles from the *Wall Street Journal* reflected an error rate of 3%. For transcription of broadcast news, a genre which could be placed between planned and unplanned speech and whose level of formality may vary, word error rates around 10% were registered. Finally, for conversational telephone speech, WER rose up to 20% (Jurafsky & Martin, 2009, p. 320)[16]. In the same year, Burger, Sloane and Yang (2006) conducted an evaluation of commercially available speech recognizers in multiple languages, using samples of both read and spontaneous speech. Overall error percentages ranged between 10% and 46%, the best performing system being a recognizer for Japanese. A system for Spanish obtained a word error rate slightly above 20% (Burger et al., 2006, p.812). In 2009, Serrahima (2009) used read speech to evaluate two automatic dictation systems; one of them was the in-built ASR system in Windows Vista, predecessor to Windows 7. After user enrollment had taken place, both systems exhibited WERs around 5% or below for different dictation tasks (Serrahima, 2009, pp. 78-79).

2.6) Windows 7 ASR

Windows 7 was released by *Microsoft* in October 2009. This operating system includes a speech recognizer which enables both desktop and application management using voice commands, as well as creation of text documents through automatic dictation. Unfortunately, it has not been possible to find academic articles reviewing the system and the technical information made available by the company is mainly addressed to programmers. For this reason, the only description that can be provided is one which focuses on those features which are observable from the user's point of view. Regarding internal system features, only information corresponding an older version of the OS, Windows Vista, has been encountered. This information will be nevertheless presented, on the premise that it might allow us to make better informed inferences about how the system under evaluation functions, given that basic features, such as system architecture or the characteristics of the core speech recognizer, could

[16] It should be noted, however, that differences in vocabulary size for these tasks make direct comparisons misleading: For the broadcast news and the conversational speech tests, a 64,000-word vocabulary was employed, whereas for read speech the set was more reduced (both a 5,000-word and a 20,000 word test took place). Furthermore, transmission channel in the conversational telephone speech tests is a crucial, potentially detrimental factor which should not be overlooked.

present similarities from one version to the following. Section 2.6.1 below summarizes Odell & Mukerjee's description of Windows Vista ASR in (Odell & Mukerjee, 2007).

2.6.1) Windows Vista ASR internal features

2.6.1.1) Basic system architecture – Windows Vista (Odell & Mukerjee, 2007)

The main components of the ASR system embedded in Windows Vista are:[17]

The Speech UX (user experience), responsible for disambiguating recognition results, generating recognition grammars and providing Graphical User Interface (GUI) elements to assist users.

The Speech Application Programmer Interface (SAPI), which routes events based on which grammar was involved in the recognition. It also manages the audio channel as well as the user lexicon.

The Speech Recognition Engine (SRE), which processes the input detecting the sounds which constitute speech, subsequently matches them against the active grammars and returns recognition results to SAPI (which redirects them to UX or to the corresponding application).

The audio subsystem.

2.6.1.2) Characteristics of the core speech recognizer

- It is based on continuous density Hidden Markov Models using cross-word context-dependent tied state triphones.

[17] The authors also mention a speech synthesis engine, for those applications which need to convert text to speech (Odell & Mukerjee, 2007, p. 1160)

- The decoder supports trigram-word and bigram-class-based language models, as well as context-free grammars. It is dynamic, since it performs lexicon and language model updates based on error correction by the user.
- The conversion of sound into written forms is carried out through a finite-state transducer.

2.6.1.3) Model adaptation mechanisms

- **Language model**

A subsystem called LMA (Language Model Adaptation) enables the model to learn new words and language patterns, learn from mistakes and corrections and reinforce correctly recognized language patterns. This is achieved through the following process: after receiving the speech input, the data is retained in the history cache for 90 seconds; this period is called *rollback window*, since it allows the system to “roll back” the phrase if the user signals an error; if no error is signalled, the data is sent to LMA, which increments the probabilities for the corresponding trigrams.

- **Acoustic model**

The AMA (Acoustic Model Adaptation) subsystem performs both supervised and unsupervised adaptation. Supervised adaptation can be conducted at start-up time through the interactive speech tutorial and the training wizard (see 2.6.2.1). Unsupervised acoustic model adaptation is performed during normal use, through a mechanism analogous to the rollback window procedure for LMA described above. Maximum A Posteriori (MAP) and Maximum Likelihood Linear Regression (MLLR) are used for this purpose.

2.6.2) Windows 7 ASR features

The description below will focus on those characteristics of the system under evaluation which are accessible to the user. Firstly, those features aimed at optimizing performance through personalization and model adaptation will be presented. Secondly, aspects pertaining to actual system use will be described.

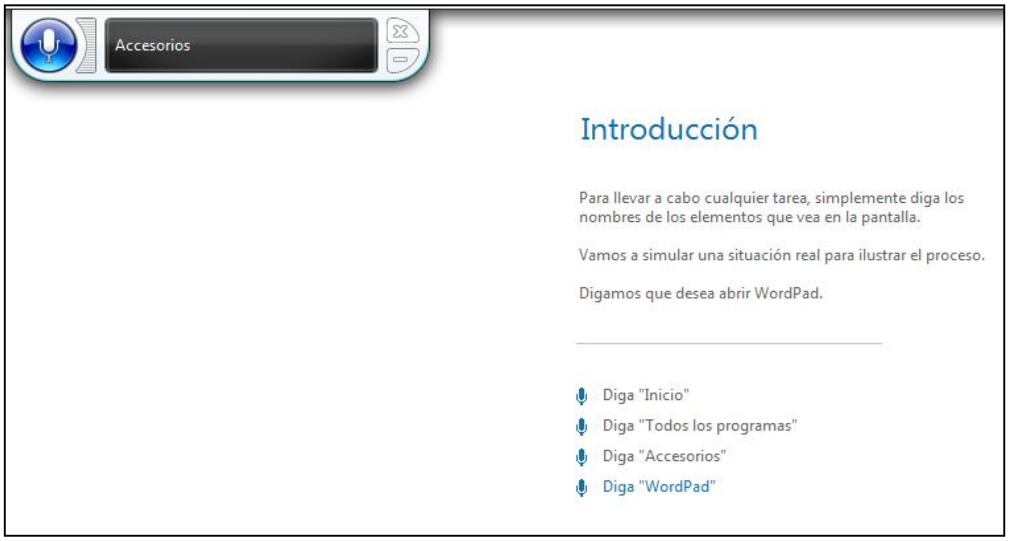
2.6.2.1) Performance optimization

- Tutorial and training

The interactive tutorial provides the user with basic knowledge required to work with the system, as well as with practical exercises. It consists of seven stages, entitled *introduction, basic concepts, dictation, commands, working with Windows* and *conclusion.* Although skipping the tutorial does not impede system use, its completion is recommended by the developers.

Another available resource, in this case labelled “optional”, is the training wizard. This consists of a sequence of sentences for the user to read aloud. After each sentence is read, the next one appears automatically on the screen. The sentences provide general information regarding automatic speech recognition, as well as hints for the use of this particular system.

The purpose of the interactive tutorial and the training wizard is twofold. These resources function as what Pallett (1985, p. 377) calls “user training”, namely a mechanism for the user to familiarize with the system through information and exercises. Additionally, they constitute an instance of speaker enrollment (Pallett, 1985, p. 377), since they are used to obtain speech data in order to further train the acoustic models and tune them for a particular user. Evidence of this is a pop-up window which appears when attempting to quit the tutorial before its completion, which encourages the user to continue, on the grounds that this may improve system precision. Furthermore, the description of the training wizard on the control panel explicitly states: “Train the computer for better understanding: Read texts to the computer in order to improve its ability to understand your voice. This is not necessary, but it can improve precision during dictation”.



Introducción

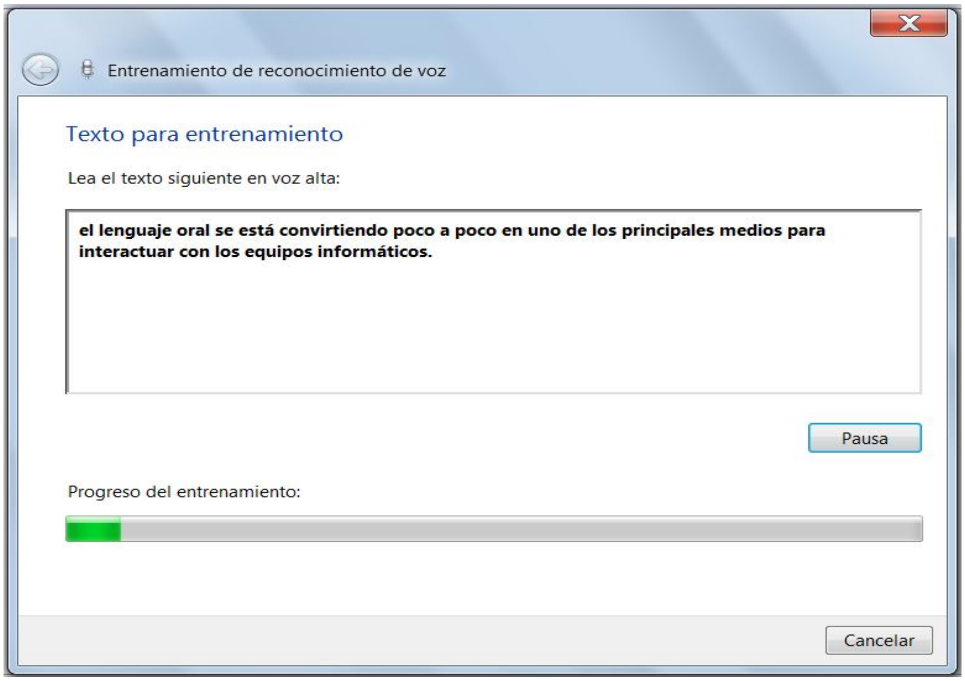
Para llevar a cabo cualquier tarea, simplemente diga los nombres de los elementos que vea en la pantalla.

Vamos a simular una situación real para ilustrar el proceso.

Digamos que desea abrir WordPad.

- Diga "Inicio"
- Diga "Todos los programas"
- Diga "Accesorios"
- Diga "WordPad"

Figure 4. Tutorial. The right column shows instructions for the user.


Entrenamiento de reconocimiento de voz

Texto para entrenamiento

Lea el texto siguiente en voz alta:

el lenguaje oral se está convirtiendo poco a poco en uno de los principales medios para interactuar con los equipos informáticos.

Progreso del entrenamiento:

Figure 5. Training box showing the sentence to be read in the center and the degree of completion of the training process at the bottom.

- Document harvesting

This is the name used by Odell & Mukerjee (2007, p. 1164) to refer to a feature which is present in both Windows Vista and Windows 7 ASR, by which the user may authorize the system to mine documents and e-mails stored in the hard drive searching for frequently occurring items and patterns. This may provide valuable information regarding the user's writing style (Odell & Mukerjee, 2007, p. 1164) and the lexicon corresponding to his areas of interest. High-frequency out-of-vocabulary words could also be detected and added to the lexicon, which might prove extremely useful for items such as proper nouns.

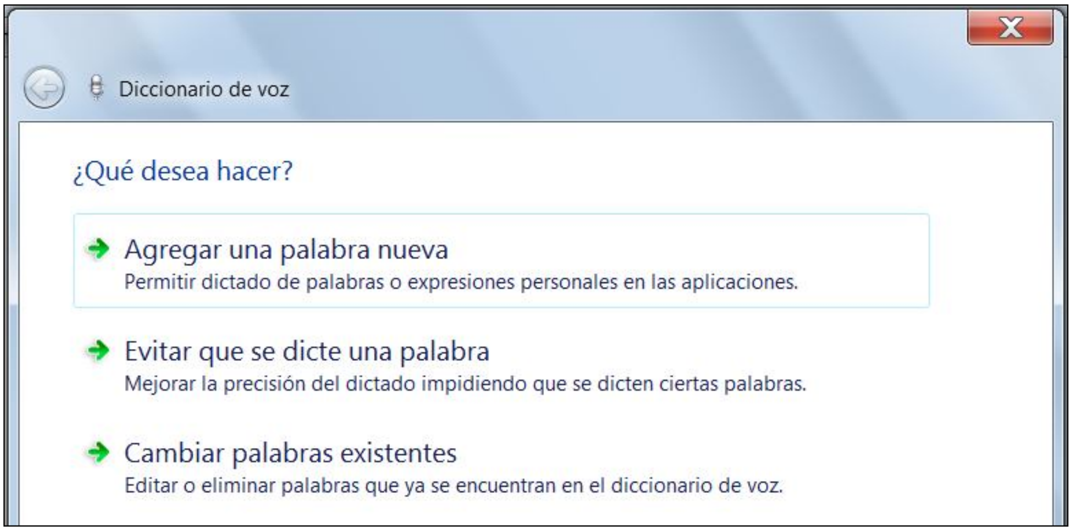
In Windows Vista, once such files have been read, the relevant text is passed on to the LMA subsystem of the SRE via a SAPI interface (Odell & Mukerjee, 2007, p. 1164).

- Multiple profile creation

The system supports the creation of multiple speech profiles for different users and/or different environmental conditions. Hence, it is possible to create a new profile and train the system, for instance, in an environment with a distinct background noise configuration, in order to obtain better results when the recognizer is used under those conditions.

- User lexicon

In addition to the in-built lexicon, a customized lexicon can be created during system use. This enables the speaker to add new words, as well as to block words so that they do not appear in the dictation output. Optionally, the speaker's pronunciation can be recorded and associated to the lexical forms added by the user.



Diccionario de voz

¿Qué desea hacer?

Agregar una palabra nueva
Permitir dictado de palabras o expresiones personales en las aplicaciones.

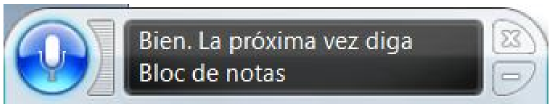
Evitar que se dicte una palabra
Mejorar la precisión del dictado impidiendo que se dicten ciertas palabras.

Cambiar palabras existentes
Editar o eliminar palabras que ya se encuentran en el diccionario de voz.

Figure 6 – User lexicon menu. Options: “add a new word”, “block a word from being dictated” and “change existing words”.

2.6.2.2) System use

- Speech control panel: It is a GUI device intended to facilitate user-system interaction. It signals when the recognizer is ready to receive input. It also warns the user when the system has been unable to recognize the uttered command, so that it can be attempted differently. Occasionally, it may provide feedback on an uttered command, such as shorter or more practical ways of performing the same operation. Finally, the command “¿Qué puedo decir?” (“What can I say?”) opens a reference card with a list of useful commands.



Bien. La próxima vez diga
Bloc de notas

Figure 7 – Speech control panel showing a more efficient way of expressing the command just uttered.

- Number grid: A useful tool for desktop management is a grid containing numbers 1 to 9 which covers the whole screen and can be called through the command “cuadrícula de mouse” (mouse grid). It can be employed in order to click on an element whose name the user does not know, as it enables him to direct the mouse pointer to the square where the element is located by uttering the corresponding number. This strategy simplifies desktop management significantly since, as explained in 2.5, digit recognition is nowadays one of the least challenging ASR tasks (Odell & Mukerjee, 2007, p. 1159).

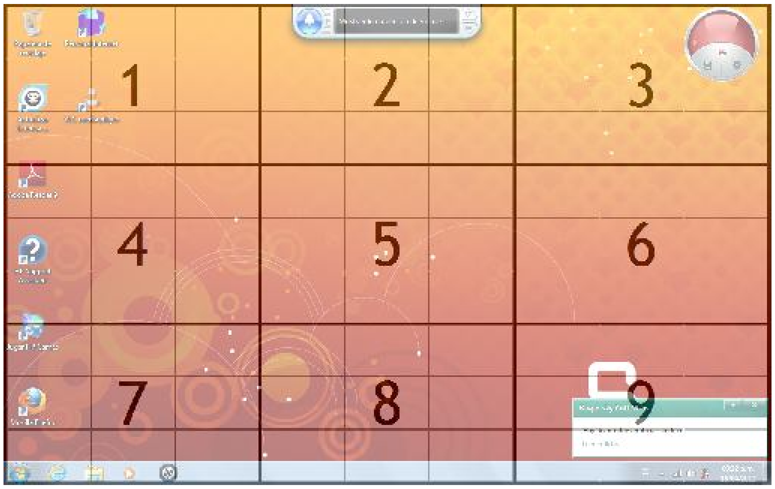


Figure 8 – Mouse grid.

2.6.3) Windows 7 ASR and Spanish varieties

Windows 7 ASR is available in English, French, Spanish, German, Japanese, Simplified Chinese, and Traditional Chinese. In the language menu, the option “Spanish” reads “Español (España)”, which indicates that the system is designed for speakers of those varieties of Spanish spoken in the Iberian Peninsula and the Balearic and Canary Islands. This has important implications for our evaluation. The acoustic models are likely to have been trained using speech corpora representative of such varieties. Therefore, the phone set for the models has probably been selected to reflect the sounds employed by these language communities. 2.3.3.2 above illustrated the phonetic particularities of Spanish as spoken in Buenos Aires when compared to Peninsular varieties: it could be hypothesized that those sounds may not be contemplated in the phone set. On the other hand, an analogous conjecture can be made with respect to the lexicon. Section 2.3.3.4 briefly described some specificities of the vocabulary used by speakers from Buenos Aires. A recognizer designed for Iberian Spanish would not necessarily include such lexical items in its dictionary. This might even have morphosyntactic repercussions, as section 2.3.3.3 showed how verbal paradigms may exhibit different forms across language varieties.

In view of these considerations, a stimulating question arises. The enormous popularity of the operating system Windows around the globe indicates the existence of an immense Spanish-speaking market of potential Windows ASR users, among whom a large proportion are speakers of non-Iberian Spanish varieties. Given that only one speech recognizer is available for Spanish, it follows that it will be the one employed by all Windows ASR users, including this group. This leads to the interesting issue of how the distinctive lexical and phonological characteristics of non-Iberian Spanish varieties will impact on system performance: in other words, whether these characteristics will exert a detrimental effect on recognition. As stated previously, one of the objectives of this research is to seek an answer to this question.

3. METHOD

3.1) Subject selection

The first group of subjects consisted of speakers of a Peninsular variety of Spanish, whereas the second was made up of speakers of a Latin American variety. In the former case, only Spanish-Catalan bilingual speakers presenting Catalan dominance, born and raised in Catalonia, were selected. This entails several considerations. Firstly, they all speak Catalan as mother tongue and have learnt it at home before the age of five. Secondly, they speak Catalan to their immediate circle of relations, or at least to the majority of its members. Finally, they all share a preference for expressing themselves in Catalan in their daily interactions. This linguistic profile was preferred over one with Spanish dominance for reasons of homogeneity, since many bilingual speakers who present this second profile come from migrant families from different parts of Spain or other countries. For the group of Latin American speakers, the variety selected was *Porteño* Spanish, i.e. the one spoken in Buenos Aires, Argentina. In this case, speakers were sought within a monolingual community, which simplified the problem of heterogeneity of linguistic profiles to a certain extent.

Forty speakers were selected, ten female speakers and ten male speakers for each of the two language variety groups. All of them were university students or graduates between eighteen and thirty-five years of age who used computers on a daily basis but had little or no experience with ASR. In order to determine eligibility with respect to these requirements, as well as to the linguistic criteria mentioned in the previous paragraph, a questionnaire was administered to potential candidates before selection. Two different questionnaires were designed: one for each language variety (appendixes 1 and 2).

3.2) Test design

3.2.1) Task overview

It has been stated before that two main tasks can be performed using Windows 7 ASR: a) desktop and application control using voice commands and b) automatic dictation. For the evaluation, an activity set was designed which aims at testing both uses (appendix 3). The set consists of three tasks.

Task 1 is a list of commands which speakers are expected to read aloud. These are instructions to perform basic operations, such as opening and closing programs or files, minimizing windows, scrolling down, selecting words in a text, etc. The aim of the task is to test system performance within the domain of isolated-word and isolated-phrase recognition.

Task 2 is a dictation activity in which speakers are given a series of notes made up of isolated words and phrases to read in silence, and they are afterwards expected to dictate whole paragraphs based on such notes. The aim of the task is to obtain samples of semi-spontaneous speech in which the subjects need to perform certain cognitive operations while dictating, in order to transform the isolated notes into cohesive paragraphs. The two main assumptions underlying this task are:

a) The effort caused by the simultaneity between the actions of *speaking* and *planning how the ideas will be expressed* will result in the appearance of some of the disfluencies which are characteristic of spontaneous speech: repetitions, pauses, hesitations, false starts, “fillers” (*eh... hm...*), etc.

b) In contrast to the isolated commands of task 1, this exercise elicits connected speech. Thus, the outcome is expected to include a greater number of coarticulation effects, such as assimilation (Gil, 1988, p. 127; Farnetani & Recasens, 2010, p. 320-321), weakening, substitution and deletion of certain sounds (see 2.3.2).

Abbreviated forms and other time-saving devices which are often used by some individuals in note-taking were deliberately included, on the one hand for the sake of authenticity and, on the other hand, to increase the cognitive load required by the task. These forms include ***q'*** for the word *qué*, ***hab*** for *habitantes,* the sign "+" instead of the word *más* and the abbreviation ***km***^{2} for *kilómetros cuadrados*.

The procedure employed is the following. To create a scenario, the subjects are told that they have attended three lectures; a history class, a literature class and a geography class, where they have made short notes by hand, and they subsequently wish to store the information in their computers in the form of three well-organized paragraphs. They are given a minute to read the notes and, afterwards, they use automatic dictation to create their texts.

In ***task 3***, the subjects use automatic dictation to create an e-mail message to a friend. The aim of the task is to elicit continuous, quasi-spontaneous informal speech. The cumbersome term "quasi-spontaneous" has been chosen because, although these are not samples of real-world, naturally-occurring speech, the subjects have considerable freedom to develop their own ideas and decide how to express them within the limits imposed by the instructions. This task differs from the previous one in two respects. First of all, the outcome is less constrained in terms of language, since no keywords or phrases are provided and, secondly, the communicative situation thereby created calls for a more informal speaking style.

For this activity, the subjects are told to imagine that they have forgotten a friend's birthday, and they must create an e-mail:

- apologizing and justifying themselves
- asking what kind of birthday present their friend would like, and providing two or three ideas for him/her to choose
- suggesting a day, time and place to meet

- reminding their friend to return them a CD which they have lent him/her, and saying why they need it back

3.2.2) Task design

3.2.2.1) Problematic expressions

Several factors were taken into consideration during the process of task design. In choosing the types of information, vocabulary and structures which would be elicited from the speakers through the activities, deliberate decisions were made in order to include instances of phenomena which might offer special difficulties during recognition. This was done with a view to imposing a greater challenge to the tested system, based on Llisterri's (2007) discussion of features which generally pose difficulties for speech processing systems. Although the article centers mainly on speech synthesis, the ideas presented were here taken as a basis and adapted to the specificities of the speech recognition domain, and a set of items were consequently selected and included in the tests:

- **Numbers**: *200.000 personas, 7.793.000 hab.* (task 2).
- **Measures**: *112.492 km²* (task 2).
- **Dates**: *15 enero 1929, 4 abril 1968* (task 2).
- **Foreign words**: *Windows, Paint* (task 1), *high school* (task 2).
- **Names of places**: *Atlanta, Memphis, Washington, Birmingham, Honduras, Tegucigalpa, Comayagüela* (task 2).
- **Other proper names**: *Martin Luther King, Michael King Junior, Booker T. (Washington High School), Yolanda Bauzá, Roald Dahl* (task 2).

The inclusion of both Spanish and foreign words in the categories “names of places” and “other proper names” was a deliberate choice to increase the difficulty of the recognition task. Furthermore, two different types of foreign words were included: those with rather “transparent” spelling, that is, those which would be pronounced in a similar way if they were read using Spanish pronunciation rules (*Atlanta, Martin, King*) and those whose pronunciation cannot be predicted from spelling using such rules (*high school, Michael, Booker*).

3.2.2.2) Phonetic considerations

In order to further increase the degree of difficulty for the recognizer, the test tasks contain certain words whose phonetic realization differs considerably between the language varieties selected. These differences are related to the phenomena discussed in 2.3.3. Some examples of words containing such forms are *hacer, C, cerrar, inicio, accesorios, minimizar, aceptar, lloviendo, licencia, seleccionar* (task 1), *civiles, paz, ayuda, pacíficas, Yolanda, Bauzá, policía, llama, halla, Centroamérica, superficie, Tegucigalpa, Comayagüela, población* (task 2).

3.2.2.3) Sentence length

Variety in sentence length was another criterion considered during task design, in order to make provisions for the possibility that this factor might have a bearing upon recognition results. The sections “historia” and “geografía” in task 2 were specifically designed to elicit long and short sentences respectively.

3.3) Data collection

Approximately five hours of speech data were collected and stored in 200 *.wma* files (see CD ROM attached). The speech data required for the experiment was collected in forty one-to-one meetings: one with each of the selected subjects. Although it was not possible to conduct all meetings in the same room (given that half of the data was collected in Barcelona and the other half, in Buenos Aires) all the rooms used presented similar characteristics in terms of size. The average duration of the meetings was approximately forty-five minutes. The equipment used consisted of a Compaq Presario CQ40-705LA portable computer with an Intel Pentium T4300 processor, 2 GB RAM and 320 GB HDD, and a Logitech H110 headset with noise suppression. The version of the operating system used was Windows 7 Home Basic. A user account for each speaker was created before the interviews[18].

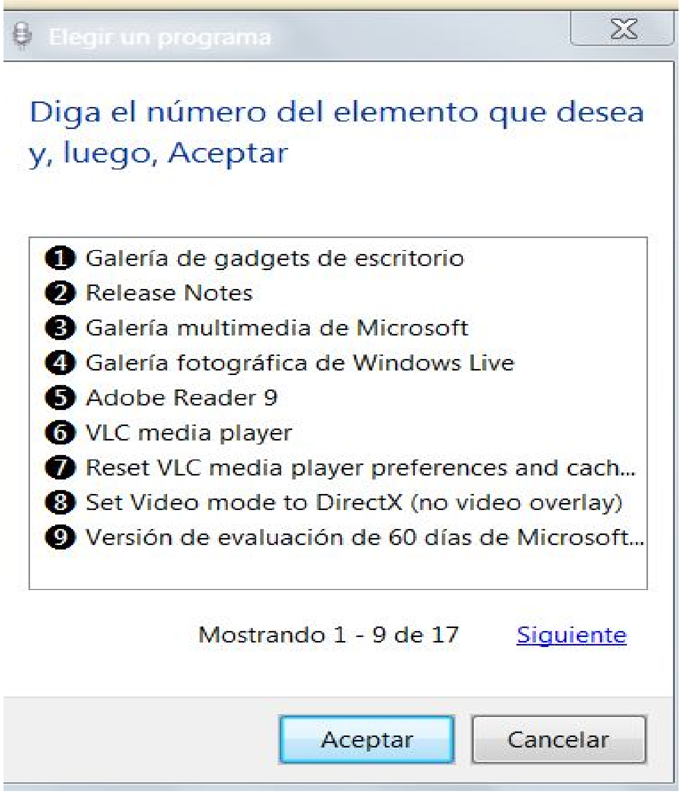
The meetings consisted of five stages. The first one was microphone configuration. Secondly, minimal user enrollment through use of the training wizard (see 2.6.2.1) took place. After these two stages, the three test tasks were carried out. All the tests were recorded in separate *.wma* audio files, using the sound recorder included with the operating system. As the task on commands did not involve a written output, the response of the recognizer to each uttered command was registered by the interviewer in a chart (appendix 4) which included the options “recognized”, “not recognized” and “confirmation requested”. For task 2, dictation of the three texts was carried out using Microsoft Word 2007 and the outcome was stored in a *.docx* file. Finally, dictation of the e-mail in task 3 was done in Windows Live Mail and stored both as a draft within that program and in *.docx* format.

Before each task, the subjects received brief oral instructions, and they were subsequently given a minute to read the directions and contents of the task (appendix 3) before starting to record. The recordings and the recognition/dictation were performed simultaneously.

[18] This was done for purposes of practicality regarding organization and storage of the data, although it was not necessary from the point of view of speech recognition, since Windows 7 allows the creation of multiple speech profiles within the same user account.

3.4) Treatment of punctuation and correction

Since the speakers had little or no experience with automatic speech recognition and automatic dictation, a number of decisions had to be made. The system includes certain recovery features for cases in which recognition breaks down. One of them is the *correction dialogue box* (figure 9), which can be used when a command is not properly recognized. This chart displays a list of alternatives among which the user can choose. Additionally, during automatic dictation, another dialogue box makes it possible to select written words and erase or modify them (figure 10). The value of these tools is self-evident, especially during unplanned or "semi-planned" dictation (as in tasks two and three), since they could be employed to eliminate some of the errors generated by the disfluencies and connected speech processes mentioned in previous sections (see 3.3.2). Nonetheless, the decision not to use them during our tests was made, since the process of training speakers for their use would have been excesively time-consuming, and the constraints in the subjects' time availability would have rendered it impossible. Moreover, as the use of these correction tools without appropriate training is bound to produce more confusion than solutions, it would have proved detrimental to recognition.

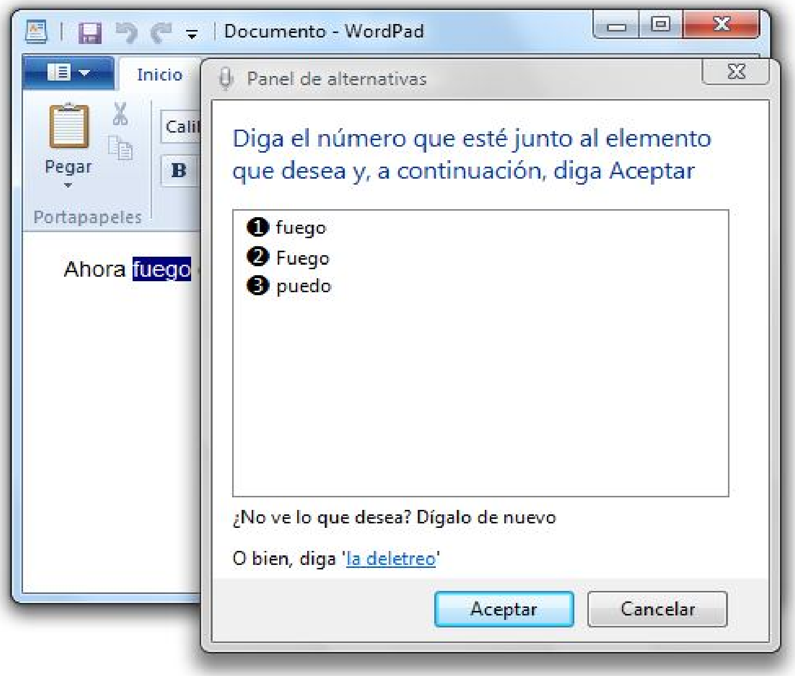


Diga el número del elemento que desea y, luego, Aceptar

1. Galería de gadgets de escritorio
2. Release Notes
3. Galería multimedia de Microsoft
4. Galería fotográfica de Windows Live
5. Adobe Reader 9
6. VLC media player
7. Reset VLC media player preferences and cach...
8. Set Video mode to DirectX (no video overlay)
9. Versión de evaluación de 60 días de Microsoft...

Mostrando 1 - 9 de 17 Siguiente

Figure 9: Correction dialogue box for unrecognized commands.



Ahora fuego

Diga el número que esté junto al elemento que desea y, a continuación, diga Aceptar

1. fuego
2. Fuego
3. puedo

¿No ve lo que desea? Dígalo de nuevo

O bien, diga 'la deletreo'

Figure 10: Correction dialogue box for automatic dictation.

Another area in which decisions had to be made was the treatment of punctuation during dictation. As the system does not automatically recognize punctuation marks, these have to be verbalized, which might feel unnatural for new users. Furthermore, the commands to insert them are not always intuitive or easily predictable[19]. Hence, in the second task, speakers were only instructed to use commas and periods, and in the third one, exclamation and question marks were added as well, together with the command to start a new paragraph. Some speakers also spontaneously attempted to use other marks intuitively, such as the colon. Furthermore, when speakers forgot to verbalize a punctuation mark, the recording was not stopped, since it was assumed that this would also happen naturally to first-time users in a real context.

[19] For instance, in order to introduce simple quotation marks (‘’), “comillas” will not be recognized, since “abrir comillas simples” (open simple quotation marks) or “cerrar comillas simples” (close simple quotation marks) is required.

4. RESULTS AND DISCUSSION

This chapter presents the recognition results obtained in the tests described in chapter 3. In the first section, overall results for each of the tasks are presented. In the second section, results for male and for female speakers are shown. Finally, accuracy and error percentages are organized by language variety (speakers from Catalonia vs. speakers from Buenos Aires).

Results corresponding to commands for different gender and language variety groups are compared in terms of percentages of successful recognition, and significance of differences is tested using the chi-squared test. In the case of automatic dictation, performance is compared in terms of WER, and significance is determined based on the relative difference between the percentages: differences are considered significant when the relative difference between percentages is higher than 10% (T. Pellegrini, personal communication, May 1, 2012).

4.1) Overall results

4.1.1) Commands

As mentioned in chapter 3, results corresponding to recognition of isolated commands (task 1) were registered during the tests by means of a chart (appendix 4) which included the options “recognized”, “not recognized” and “confirmation requested”. The category “recognized” represents the instances in which the system performed the uttered command. “Not recognized” contemplates those cases in which the command was not properly executed: this includes lack of system response as well as replacement with a different action. The third category, “confirmation requested”, refers to the appearance of a dialog box providing a list of options for the user to choose the intended action. After the tests, the percentages corresponding to each of these outcomes were calculated. These are illustrated in figure 11 below. 85.25% of the commands were correctly recognized, 14.13% were not recognized and, for 0.63% of them, confirmation was requested.

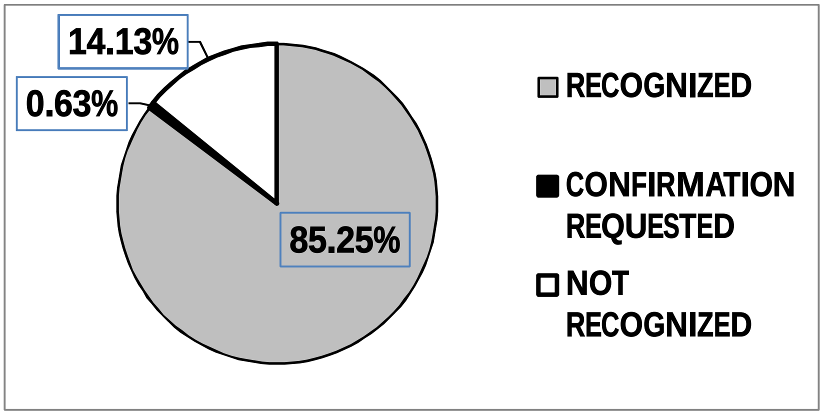


Figure 11. Recognition percentages for commands.

An analysis of recognition results for individual commands (see figures 12 and 13 below) shows that every command was recognized correctly for at least 50% of the speakers, with the exception of command number 15, "Lloviendo"[20], which was recognized for 47.5% (19 speakers out of a total of 40). Another command which stands out as having a lower recognition rate than the rest is number 8, "Paint", successfully recognized for half of the speakers. Among the other 18 commands on the list, 15 were recognized for more than 75% of the speakers.

[20] "Lloviendo" was the name of a text file which speakers were expected to open. In Windows 7 ASR, users can click on an element by saying its name, therefore speakers were required to say the word "lloviendo" in order to select the file.

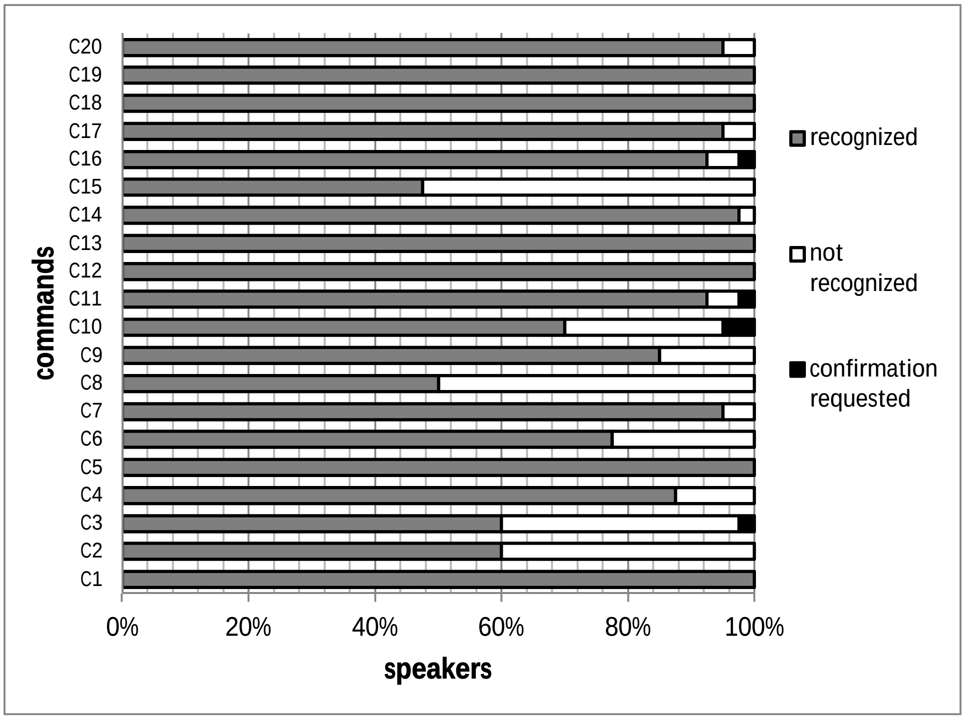


Figure 12– Recognition results for each command – all speakers

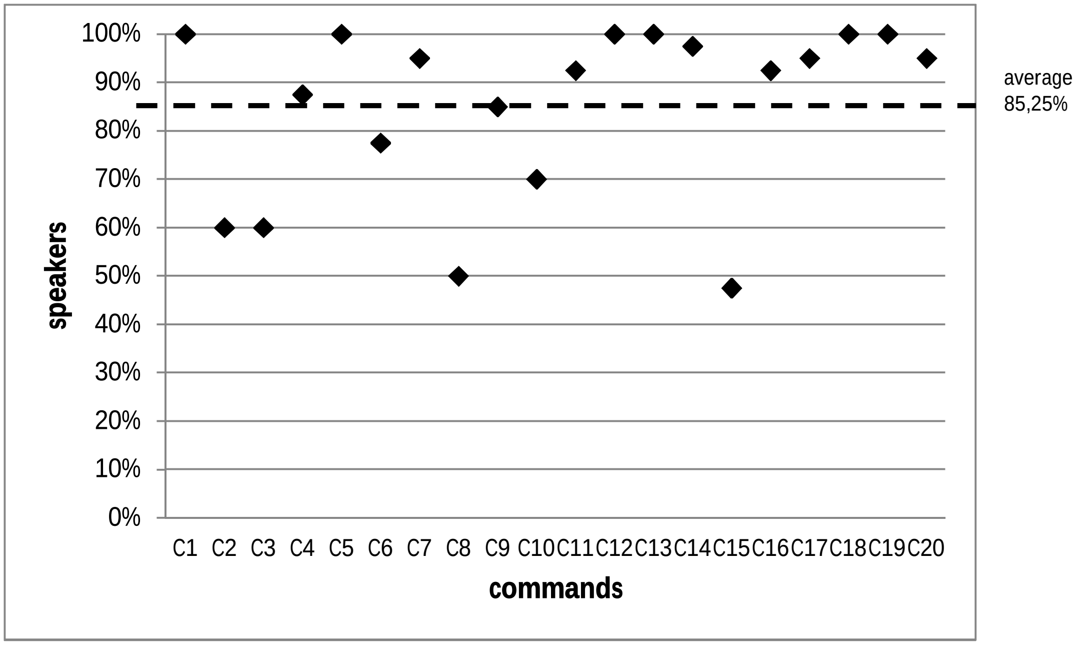


Figure 13 – Percentage of successful recognition for each command – all speakers

4.1.2) Dictation

As it would be expected, the type of treatment required for the analysis of results corresponding to automatic dictation differs from the one employed in command recognition. One of the main differences lies in the need to elaborate reference transcriptions of the speech data. Section 2.4.2 presented the usual procedure for calculating word error rate (WER) in continuous speech recognition: a **reference** text with the actual words uttered by the speaker is compared with a **hypothesis** text, namely the output of the system. For the effects of the present research, WERs for automatic dictation were obtained using SCLITE, a scoring tool used to evaluate the output of ASR systems[21]. SCLITE operates by aligning the hypothesized text with a manually elaborated reference transcription, and it subsequently calculates WER, as well as a variety of other reports (NIST, n.d.).

Overall WER for automatic dictation tasks (tasks 2 and 3) was 25.7%. For dictation from class notes (task 2) the system obtained a WER of 22.8% whereas, for quasi-spontaneous dictation of an e-mail to a friend (task 3), the error rate was significantly higher: 30.7% (relative difference = 34.65%).

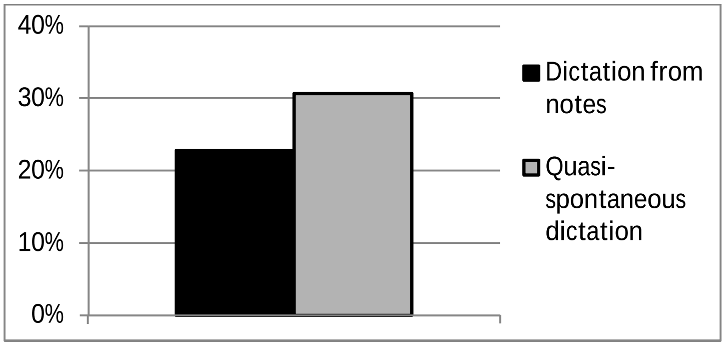


Figure 14. WERs for dictation from notes (task 2) and quasi-spontaneous dictation (task 3)[22]

[21] SCLITE is included in the NIST SCTK scoring toolkit http://www.icsi.berkeley.edu/Speech/docs/sctk-1.2/sctk.htm

[22] In all graphs representing automatic dictation results, the *y*-axis will only show the range between 0% and 40%, in order to allow clearer visualization of the differences between series.

4.2) Gender

4.2.1) Commands

Accuracy in command recognition was higher for male than for female speakers (86.5% and 84% respectively), a difference which is not statistically significant (χ2= 0.9941, *df =1, p =.32)*. These results are illustrated in figure 14. Percentages of confirmation requests were low: 1% for female speakers and 0.25% for male speakers. Figure 15 shows the number of instances of correct recognition corresponding to each command for male and female speakers.

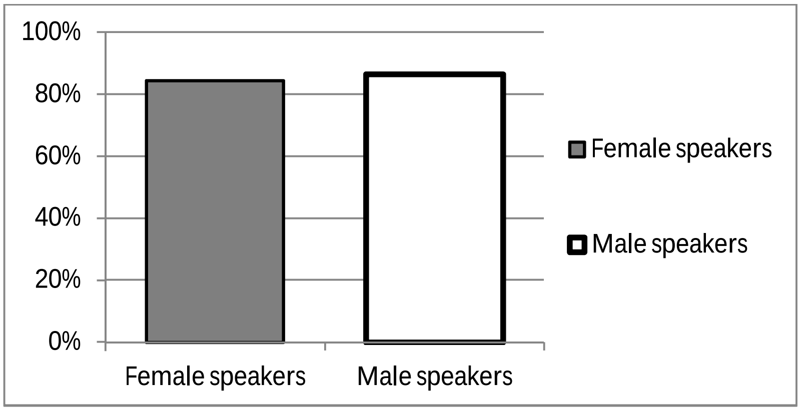


Figure 14. Commands: percentages of successful recognition for male and female speakers

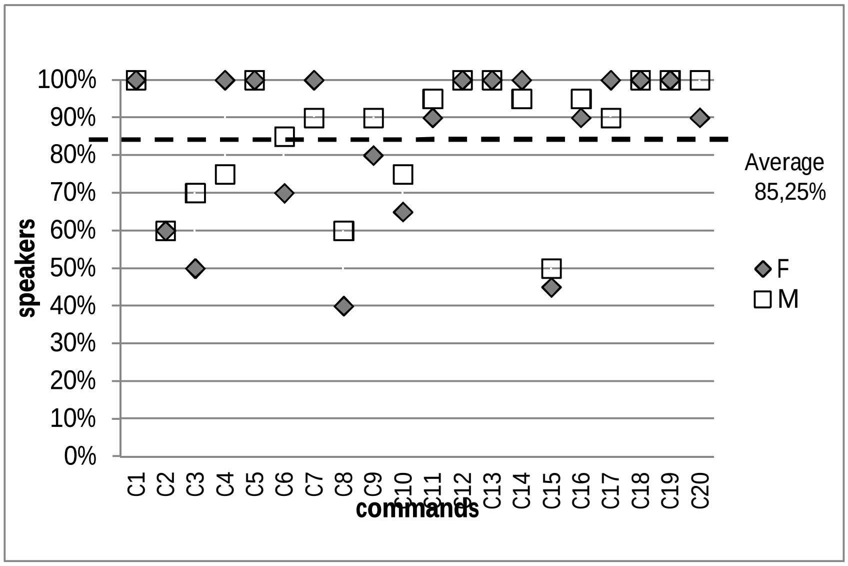


Figure 15 – percentage of successful recognition for each command: male and female speakers.

4.2.2) Dictation

No significant difference was found in overall results between male and female speakers. WER was slightly lower for the former (24.9%) than for the latter (26.6%) (relative difference = 6.83%)

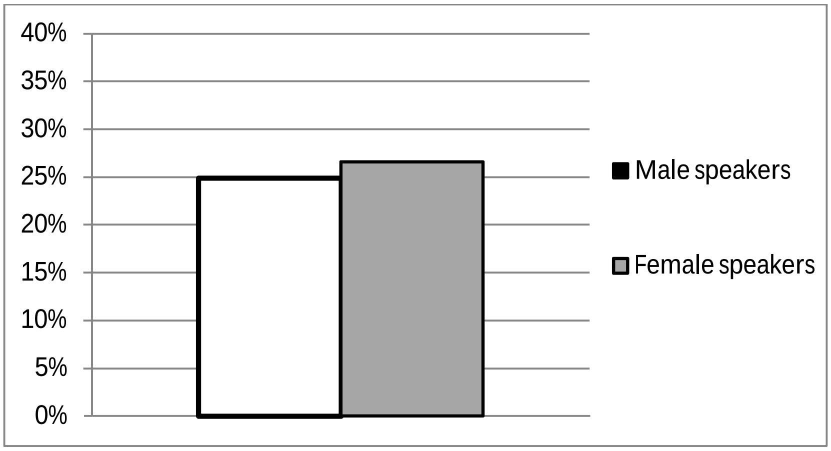


Figure 16. Dictation: WER for male and female speakers

Analogously, no significant differences were found with respect to results for task 2 on dictation from notes. In this case, performance proved slightly better for female than for male speakers (WER for male speakers = 23.4%, WER for female speakers = 22.2%, relative difference = 5.41%). In contrast, significantly better results were obtained for male speakers in task 3 on quasi-spontaneous dictation (WER for male speakers = 27.5%, WER for female speakers = 33.9%, relative difference = 23.27%)

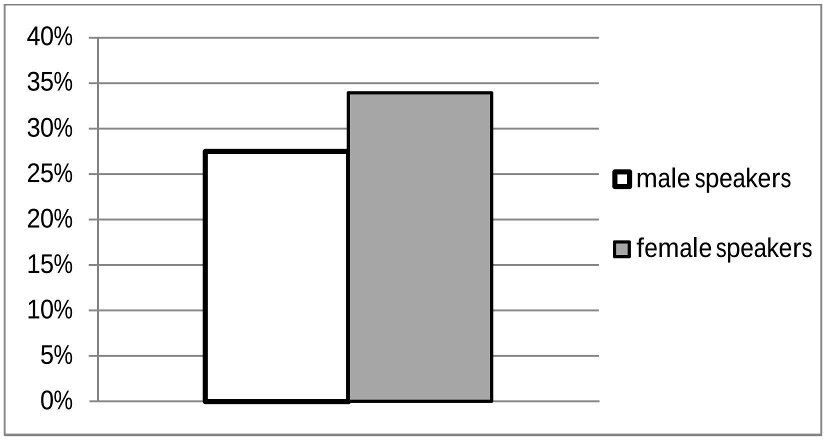


Figure 17. Quasi-spontaneous dictation: WERs for male and female speakers.

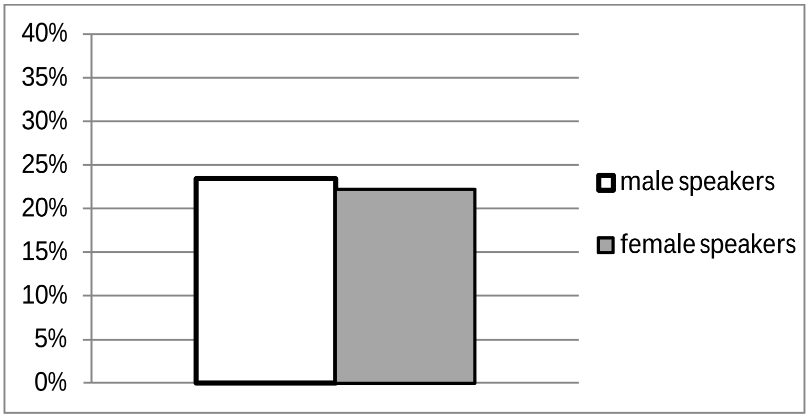


Figure 18. Dictation from notes: WERs for male and female speakers.

4.3) Language variety

4.3.1) Commands

As expected, command recognition exhibited higher accuracy for speakers from Catalonia than for speakers from Buenos Aires. For the former group, the percentage of successfully recognized commands was 88%, whereas for the latter it was 82.5% (See figure 19). This difference proved statistically significant (χ2= 4.8114, *df =1, p = .028)*. Confirmation requests were 0.75% and 0.50% respectively.

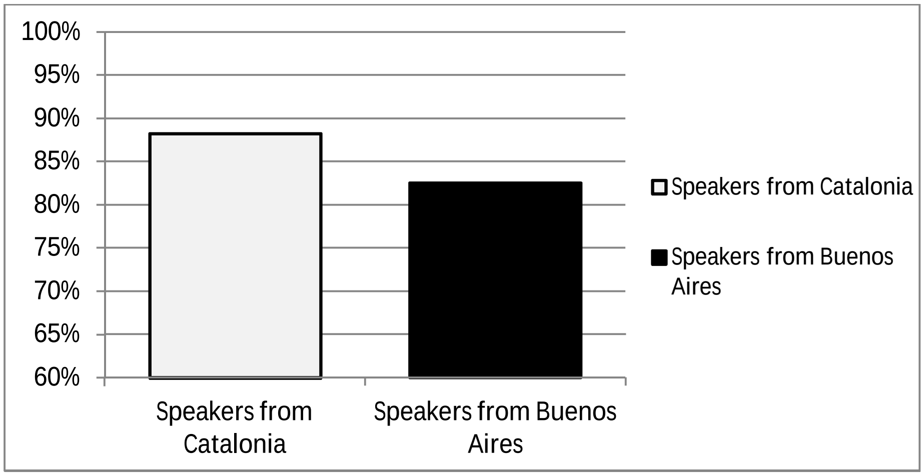


Figure 19. Commands: percentage of successful recognition for each language variety group

A closer look at individual commands reveals a very noticeable difference between groups regarding recognition of command number 15, “Lloviendo”, since it was correctly recognized for 90% of Iberian speakers (18 out of 20 speakers) but only for 5% of non-Iberian speakers (1 out of 20 speakers). This issue will be taken up again in section 5.2.

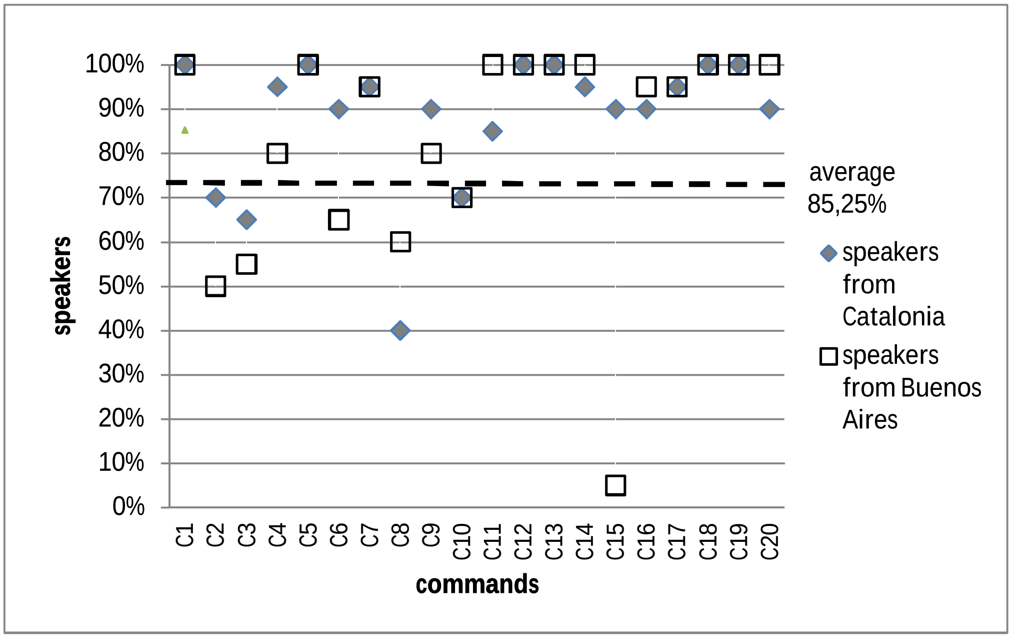


Figure 20 - Percentage of successful recognition for each command: speakers from Catalonia and from Buenos Aires.

4.3.2) Dictation

The results obtained for recognition of continuous speech considering both dictation tasks together are comparable to those observed in connection to command recognition in that the system achieved significantly better performance for the Peninsular variety than for the Latin American variety: WER was 23.1% for speakers from Catalonia and 28.2% for speakers from Buenos Aires (relative difference = 22.08%). The same observation holds for task 2 on dictation from notes, in which the former group obtained 19.4% and the latter, 25.9% (relative difference = 33.51%). Finally, results for the quasi-spontaneous speech task were 29.3% and 32.1% respectively. In this case, the relative difference is 9.56%, situated slightly below the significance threshold. This issue is revisited in section 5.2.

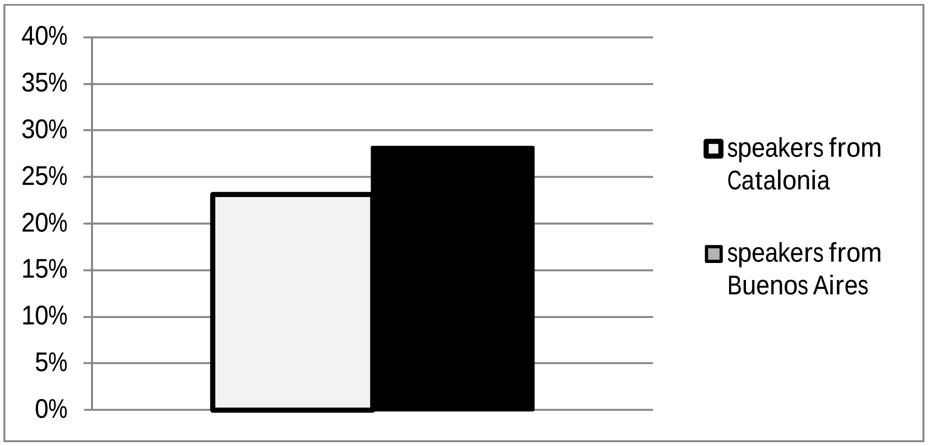


Figure 21. Dictation: WERs for both language variety groups

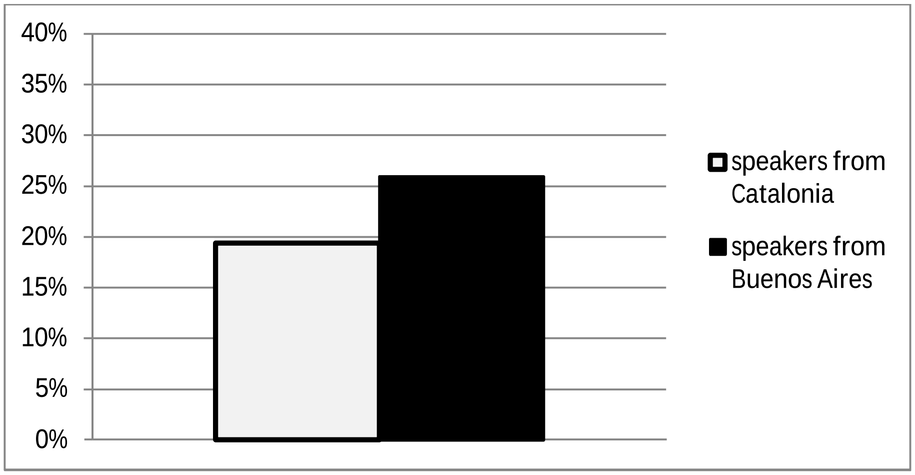


Figure 22. Dictation from notes: WERs for both language variety groups

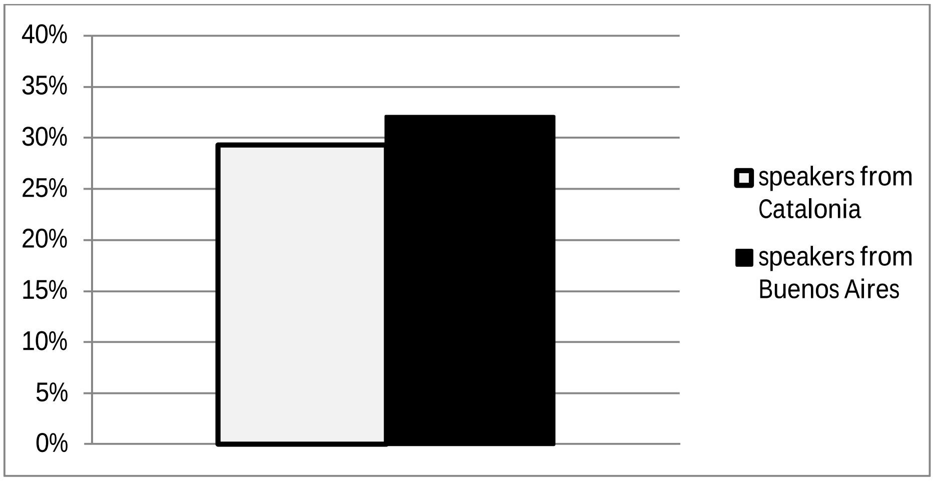


Figure 23. Quasi-spontaneous dictation - WERs for both language variety groups

4.4) Factors influencing performance

The results presented in the previous chapter raise some interesting issues regarding the robustness of the system under evaluation to certain types of interspeaker variability. An in-depth analysis of all the intervening phenomena would exceed the scope of this dissertation; therefore, only certain aspects will be selected for further discussion. Given that noticeable differences have been obtained between results for different language variety groups, the following analysis will focus on this variable.

The discussion of the characteristics of Peninsular and Porteño Spanish varieties in section 2.3.3 mentioned that a combination of two linguistic features distinguishes the latter from all other Spanish varieties. These features are: a) the replacement of palatal consonants /ʝ/ and /ʎ/ with a postalveolar fricative, a phenomenon sometimes referred to in the literature as *"yeísmo rehilado","rehilamiento porteño" or just "rehilamiento"* (see 2.3.3.2), and b) the complete replacement of the informal second person singular pronoun *tú* with the pronoun *vos,* known as *voseo* (see 2.3.3.3). Consequently, the impact of these two phenomena on the recognition percentages obtained is analyzed and discussed below.

4.4.1) Palatal vs. postalveolar consonants ("rehilamiento")

This phenomenon was taken into account during the test design stage, specifically for task 2 (see appendix 3). In this task, the subjects received a set of notes and they were required to produce and dictate sentences using the information provided in them. These notes deliberately included five occurrences of words whose realizations in *Porteño* Spanish contain a postalveolar consonant. The words were: *llama, halla, Yolanda* and *ayuda*, the last item appearing twice in the notes. Nearly all 40 speakers produced these five words; however, in 10 cases they were avoided, either through the use of a synonym or through paraphrasing. For the sake of spontaneity of the speech samples, a crucial issue in the study of the effects of style on recognition, speakers were not corrected in these cases. As a result, 190 occurrences of these words were obtained: 96 corresponding to Peninsular speakers, and 94, to *Porteño* speakers.

Table 3 shows the error percentages in recognition of these words for each language variety group. These percentages represent all cases in which the uttered word was either substituted by another word or deleted. Initially, these percentages were calculated for all occurrences of the words listed above, without distinction between those containing grapheme <ll> and those including grapheme <y>. This decision was made due to the fact that in *Porteño* Spanish there is no phonetic distinction between the realization of both orthographic forms (see 2.3.3.2). The resulting percentages are shown in row 1. Subsequently, error percentages were also calculated separately for words *halla* and *llama* on the one hand, and for words *ayuda* and *Yolanda* on the other hand, with a view to detecting any possible disparities caused by the different spelling forms. These results are shown in rows 2 and 3.

	Peninsular speakers	Porteño speakers
All words	28.1%	88.3%
Words with <ll>	60.53%	91.89%
Words with <y>	6.9%	85.96%

Table 3. Error percentages found in task 2 for words *halla, llama, ayuda* and *Yolanda*.

Several observations can be made. The percentages in the first row show that, for the whole set of selected words, the system performed considerably better for Peninsular than for *Porteño* speakers. This statement also holds with respect to specific error rates calculated separately for the words with <ll> and for the words with <y>, as can be appreciated in the second and third rows. It is worth noticing, however, that the difference in error percentages between language variety groups for words *halla* and *llama*, although very noticeable (31.36%) is remarkably inferior to that for words *ayuda* and *Yolanda* (79.06%). This issue will be taken up again later.

These interesting results are, however, not sufficiently informative in isolation, since they do not consider whether misrecognition should be attributed to the occurrence of the postalveolar fricatives or to other reasons. It is thus necessary to examine the output word for each error case. The examples of substitutions corresponding to speakers of *Porteño* Spanish found in table 4 illustrate this need.

	Input word	Output word(s)
1	llama (pronounced *[ʃ]ama*)	llamaba
2	ayuda (pronounced *a[ʃ]uda*)	A sudar

Table 4. Examples of substitution errors detected for words *ayuda* and *llamaba*

Although both cases in table 4 constitute recognition errors, it is clear that, in number two, the failure is connected to the segment [ʃ], which has been interpreted by the system as [s]. In contrast, in example number one, the problem does not involve the syllable where [ʃ] occurs: thus, misrecognition cannot be attributed to the presence of this segment in the uttered word. Table 5 shows the percentage of cases comparable to number two, i.e. cases in which the sound uttered as phonetic realization of graphemes <y> or <ll> is clearly involved in the error.

	Peninsular speakers	Porteño speakers
All words	18.75%	84.04%
Words *halla* and *llama*	42.11%	81.08%
Words *ayuda* and *Yolanda*	3.45%	85.96%

Table 5. Errors which involve the phonetic realization of <y> or <ll>

These results seem clearly revealing: Percentages for speakers from Buenos Aires are substantially higher than those corresponding to speakers from Catalonia. It should be noted, however, that the differences between the percentages obtained for both language variety groups are much less marked for the words containing grapheme <ll> (row 2) than for the words containing grapheme <y> (row 3). This is comparable to the observation made for table 3 above.

All these considerations raise a series of relevant issues. In the first place, the superiority of phone confusion percentages for *Porteño* Spanish over those for the Peninsular variety suggests that pre-palatalization of palatal fricatives exerts a highly detrimental effect upon recognition, which in turn hints at the idea that the lexicon does not include an alternative pronunciation with a pre-palatal (postalveolar) fricative for the words selected. On the other hand, the question arises of whether these postalveolar consonants are present among the acoustic models. It could be argued that a recognizer for Peninsular Spanish would not require them, given that these sounds are not included in the phonemic sets corresponding to Peninsular varieties (see 2.3.3.2). Nonetheless, models for these sounds could be included in order to aid recognition of proper nouns of foreign origin (e.g. Catalan names such as *Joan* - *[ʒ]oan*, *La Caixa – La Cai[ʃ]a*), borrowings (*show - [ʃ]ow)*, etc. In this respect, it is interesting to notice that, for one of the speakers of *Porteño* Spanish, the word *hallaron* (realized as ha[ʃ]aron) was transcribed by the system as “a Sharon”, a case in which [ʃ] has apparently been correctly recognized. Although this occurrence could perhaps suggest that at least the voiceless postalveolar fricative is contemplated among the system’s acoustic models, it is evident that more data would be required in order to elaborate any valid conclusions.

Another salient aspect, briefly mentioned above, is connected to recognition of the two words containing the grapheme <ll> when uttered by speakers from Catalonia. As table 3 shows, error percentages for this group are noticeably higher for the words *halla* and *llama* (60.53%) than for *ayuda* and *Yolanda* (6.9%). Analogously, the percentage of cases in which the phonetic realization of grapheme <ll> by these speakers is involved in misrecognition of the word (42.11%) is also appreciably higher than for grapheme <y> (3.45%), as table 5 illustrates. This phenomenon would be worthy of a more detailed analysis. A possible hypothesis is related to the possibility that Peninsular realizations of the words *halla* and *llama* exhibit instances of *yeísmo*, already defined in section 2.3.3.2 as the replacement of /ʎ/ with /ʝ/ in words containing the grapheme <ll>. The effects of *yeísmo* on system performance constitute a highly interesting issue, given the vast extension of this phenomenon in Peninsular as well as Latin American Spanish varieties. Nonetheless, as the focus of the present evaluation consists in the effects of an extra-Peninsular Spanish variety on a system trained for Peninsular Spanish, this type of *yeísmo*, which does not occur in *Porteño* Spanish (see 2.3.3.2)[23], will not be taken into account in the present analysis, remaining as a possible line for future research.

4.4.2) Voseo

For a discussion of the effects of *voseo* on system performance, attention will be focused on task 3 (see appendix 3). As described in section 3.2, this task consisted in spontaneous dictation of an e-mail to a friend. In order to encounter instances of *voseo*¸ the communicative scenario should involve the figure of a listener/reader who can be addressed directly using informal style: in this case, such a figure is represented by the receiver of the e-mail.

An analysis of this phenomenon needs to consider, on the one hand, the occurrences of the pronoun *vos* in the speech data and, on the other hand, the verb forms which correspond to it. Nevertheless, not all verb forms associated with the pronoun *vos* are likely to pose difficulties to the system. As already shown in table 1, in standard *Porteño* Spanish, only the present indicative and the imperative *voseo* forms differ from those corresponding to the pronoun "tú" in other varieties.

[23] The term *yeísmo* (/ʎ/ > /ʝ/) must not be confused with "*yeísmo rehilado*" *(/ʎ/ > [ʒ] or [ʃ])*, since the latter is an alternative name for "*rehilamiento porteño*".

Additionally, although the e-mail scenario creates the conditions for the use of such forms, these may not be as frequent in this kind of discourse as they would be in dialogue, where there is a more direct interaction between participants. These factors result in the available data being more scarce than desirable. Consequently, this section will be approached as a discussion based on the phenomena observed, rather than as a detailed analysis.

4.4.2.1) Pronoun *"vos"*

In the 20 e-mails dictated by speakers from Buenos Aires, only nine occurrences of the pronoun *vos* were found. Table 6 shows the input and output words for each of these instances.

UTTERED WORDS	SYSTEM OUTPUT
vos qué	vocablo
pensé en vos	pensemos
vos	boes
decime vos	decimeba
vos	voz
vos	sus
vos	voz
vos	vos
vos	vos

Table 6. Recognition results for the pronoun *vos* in task 3

As table 6 shows, the pronoun was correctly recognized in two cases, and substituted by a different word (or sometimes interpreted as a syllable within a longer word) in all other instances. In two of these instances, the word hypothesized by the system was *voz ("voice"),* which directs us back to our discussion of the phonetic phenomenon of *seseo* in 2.3.3.2. The two instances of successful recognition constitute sufficient proof that the word is included in the recognizer lexicon. This could be attributed to the fact that the pronoun exists in Peninsular Spanish, although it is practically not used, unless in extremely formal discourse. The last consideration could hence account for the high number of errors obtained, given the low frequency of the word in Peninsular Spanish.

4.4.2.2) Verb forms corresponding to *voseo*

Forty-one verbs were encountered which correspond to the paradigm of *voseo* and at the same time differ from the forms for the pronoun "tú". Only three of them (7.32%) were recognized correctly. For the remaining 38 (92.68%), the system selected an incorrect word.

It is worth noting a phenomenon which was observed in connection to the unrecognized verbs. In 15 cases (39.47% of the total number of errors), the correct verb in the right tense, but with the form corresponding to the pronoun *tú*, was selected. Some examples are listed in table 7.

UTTERED VERB	SYSTEM OUTPUT
perdoname	perdóname
sabés	sabes
traeme	tráeme
avisás	avisas
necesitás	necesitas

Table 7. Instances of correct choice of verb and tense with incorrect conjugation

The relevance of these cases lies on the fact that, although in a strict sense they constitute errors, the substitution involved can be considered harmless in communicative terms. Such cases might be attributable to a combination of two factors: a) the high degree of acoustic similarity between the realizations of both forms, only differing in the position of the lexically stressed syllable and b) the fact that they would be expected to occur in practically identical contexts: therefore, the language models which include *tú* forms may occasionally allow these very similar *vos* forms during actual decoding.

Notwithstanding these considerations, the error percentages obtained with the data available seem to suggest that verb forms corresponding to the *voseo* paradigm constitute an important source of difficulty for the recognizer. It is highly likely that these are not included in the system lexicon, given their absence from Peninsular Spanish.

5. CONCLUSIONS

The results obtained in the test tasks and presented in the previous chapter will be summarized below. They will be structured according to the three experiment variables: gender, language variety and task.

5.1) Gender

As stated in section 4.2.1, no statistically significant differences were found between results for female and male speakers regarding recognition of commands (task 1). Accuracy rates proved slightly better for male (86.5% correct) than for female speakers (84% correct). Neither were significant differences found between dictation results for task 2 (dictation from notes). Performance was slightly better for female (WER = 22.2%) than for male speakers (WER = 23.4%), absolute and relative differences being very reduced (1.2% and 5.41% respectively). In view of these considerations, it may be claimed that variability due to gender did not show an effect on system performance for command recognition and for dictation from notes, therefore suggesting that the system would include techniques which allow it to address this issue satisfactorily.

Only dictation task number 3, i.e. quasi-spontaneous dictation of an e-mail to a friend, showed a significant difference between the results obtained for female and for male speakers. Performance was better for men than for women (WER for men = 27.5%, WER for women = 33.9%, relative difference = 23.27%). It would be interesting to conduct other similar experiments focusing specifically on the effects of gender on this type of dictation, since these results could perhaps suggest that the kind of phenomena encountered in female quasi-spontaneous speech offer the system more difficulty than those produced by male speakers. If this were the case, the lack of significant differences in previous tasks could be explained by the characteristics of the speaking style in each of them. A more hyperarticulated style and a smaller number of disfluency phenomena might be expected in isolated commands in comparison to semi-spontaneous speech. Regarding task 2, speakers were given notes with key words and phrases, which facilitated the dictation task, probably reducing the appearance of disfluencies. The presence of other segmental and suprasegmental features usually found in spontaneous speech may also have been limited by the more formal scenario featured in this task, in which subjects spoke about semi-academic topics.

5.2) Language variety

As expected, the percentage of successfully recognized commands was higher for Peninsular (88%) than for *Porteño* speakers (82.5%), a difference which proved statistically significant (see 4.3.1). One particular command offered enormous difficulty to the system when pronounced by the latter group: “Lloviendo”, which was properly recognized for 90% of Peninsular speakers but only for 5% of *Porteño* speakers. This may be connected to the presence of the postalveolar fricative in the *Porteño* pronunciation of the word. It would be interesting, therefore, to conduct a similar experiment in which all items on the list of commands contain the graphemes <ll> and <y>, in order to see if performance deteriorates with respect to the results presented above.

Performance was also better for Peninsular (23.1%) than for *Porteño* speakers (28.2%) when both dictation tasks were considered together. Given that the relative difference between the results is higher than 10% (22.08%), the difference can be considered significant. It seems clear that the specific characteristics of *Porteño* Spanish constitute a considerable source of difficulty for the recognizer. Two of the phenomena present in Porteño Spanish which could be partly responsible for this have been described from a theoretical point of view (section 2.3.3) and subsequently revisited in the light of the results obtained (section 4.4): the morphosyntactic phenomenon called *voseo* on the one hand, and the ocurrence of postalveolar fricative consonants on the other. Our brief analysis of their effect on system performance seemed to indicate that these have a detrimental effect upon recognition. It would perhaps be possible to cast more light upon the issue by conducting studies specifically tailored for testing the effects of these two features on performance. Analogously, it could prove insightful to inquire into other specificities of *Porteño* Spanish which may be responsible for the increase in error rates, such as the ones described in section 2.3.3.

Finally it is worth noting that, whereas the difference in error rates for the task on semi-spontaneous dictation from notes (task 2) is clearly significant, as shown by the relative difference of 33.51% between results for both groups, in the case of quasi-spontaneous dictation (task 3), the corresponding value (9.56%) is situated slightly below 10%, which makes the significance issue questionable. This could perhaps be explained by the fact that recognition of this kind of speech always offers a high degree of difficulty, independently of which language variety is spoken, a fact which may cause the difference between both groups to be less pronounced. The effects of task and speech style on performance are discussed in the next section.

5.3) Task

When comparing results obtained for the task on dictation from notes (task 2) with those for quasi-spontaneous dictation (task 3), a significantly higher error rate can be observed for the latter (WERs = 22.8% and 30.7% respectively, relative difference = 34.65%). This may be attributed to a series of factors, already discussed in previous sections.

Task 3 creates a more informal, conversational scenario than task 2, a circumstance which facilitates the appearance of disfluences and other segmental and suprasegmental processes commonly associated with spontaneous speech (see 2.3.2). Furthermore, the speech in task 3 is unscripted, practically spontaneous, whereas in task 2 it could be viewed as “semi-scripted”, since the words and phrases in the notes provided constitute a kind of supporting structure for speech: the speaker only needs to “fill in the gaps” in order to connect the ideas contained in the notes. Therefore, the cognitive load involved in planning what is going to be said is greater in task 3 than in task 2, which results in a greater number of disfluencies in the former, increasing WER.

Regarding the task on isolated commands (task 1), the system showed a percentage of correct recognition of 85.25%, as stated in 4.1.1. It would not be possible to establish a direct comparison with the results obtained for automatic dictation, given that different kinds of measures are used for each task type (percentage of recognized commands is used for task 1, whereas WER is used for tasks 2 and 3).

5.4) Closing remarks

The evaluation presented in this work aimed at examining the impact of gender, language variety and speaking style on the performance of the speech recognizer embedded in the operating system Windows 7 with respect to two types of recognition tasks: command recognition and automatic dictation. Performance was measured and general conclusions were drawn. The need to focus on different variables and tasks resulted in the impossibility to conduct an in-depth analysis of each of them individually, which meant that a number of interesting aspects had to be left aside. Some of them are enumerated below:

- Two features of the Spanish variety spoken in Buenos Aires and their effects on recognition were discussed. One of them was the occurrence of postalveolar fricative consonants, which appeared to have a detrimental effect upon system performance. The analysis was carried out through selection of a list of words which contain postalveolar fricatives [ʒ] and [ʃ] in *Porteño* Spanish, and subsequent comparison of recognition success for those words between language variety groups. A deeper study of the phenomenon could benefit from phonetic transcriptions of the actual words uttered, with the aim of producing a confusion matrix in order to visualize which sounds were involved in errors more frequently, as well as with which sounds these were generally replaced. This might help us gain insight into the causes of the problem and enable us to suggest possible ways in which it could be solved.

The second feature was the phenomenon of *voseo*, analyzed through spontaneously occurring instances of the pronoun *vos,* as well as of relevant verb forms. Since the task from which the data was extracted was not particularly designed for this purpose, the number of occurrences encountered was relatively reduced. A study in which the design of the tasks elicited a larger number of *voseo* forms could make it possible to draw more solid conclusions about the effects of this phenomenon on system performance, as well as about its causes.

- Other features of *Porteño* Spanish were discussed in chapter 2. Among these were the phenomenon of *seseo* and the phonetic realizations of preconsonantal <s>. An analysis of the impact of these features on recognition is likely to contribute very valuable knowledge due to the widespread nature of this phenomenon, since detection of potential problems could improve the system for numerous Spanish varieties.

- As explained in section 3.2.2.1, certain choices were made during task design in order to include instances of phenomena which might offer special difficulties during recognition, such as numbers, measures, dates, foreign words, proper names, etc. The study of their effects on system performance remains a stimulating issue for future research.

- It has been shown that recognition was significantly weaker for the task in which the speech involved exhibited the highest degree of spontaneity (task 3). Therefore, the effects of spontaneous speech phenomena on system performance constitute a further aspect worthy of analysis.

Finally, it should be noted that the work performed can be described as a user-oriented black-box evaluation carried out by means of field tests (Jekat & Schultz, 2004) with specifically designed tasks. The fact that detailed technical documentation on the system under consideration is not publicly available, as is usally the case in commercial products, certainly limits the possibility of explaining our findings and suggesting improvements. Nevertheless, this kind of evaluation is still relevant, since it may help potential users to assess the suitability of the application guided by a better knowledge of the strengths and weaknesses of the recognizer as far as language variety and tasks are concerned. The work above presented intends to be a first step in this direction.

6. ACKNOWLEDGMENTS

Years ago, in my hometown Buenos Aires, I came across a few unfamiliar words that I had never heard before and which arose my interest. They were terms such as “acoustic model”, “automatic dictation” and “speech recognizer”, among others. Driven by curiosity, I started a search, discovering, as a result, the inspiring works which would later on lead me to embark on this project. Little did I know at that time that the author of those works would continue to inspire me throughout the process of writing this thesis, now not only through his writings but also –and especially– through his hard work and dedication, through his passion for science and his tenacious pursuit of knowledge. At the end of this journey, I only have words of gratefulness to Dr. Joaquim Llisterri for the confidence he placed on me right from the start, for his constant encouragement and his unflagging support throughout the process.

Halfway through my journey, I met two other professionals whose contributions were crucial as I walked the path towards my aim. Dr. Thomas Pellegrini, whose assistance and suggestions constituted an enormous source of help, and Dr. Jorge Baptista, who was always an outstanding example of hard work, responsibility, commitment and dedication to his students.

I am deeply grateful to Prof. Llisterri, Prof. Pellegrini and Prof. Baptista for willing to share their vast expert knowledge with me, each one in his own field, enabling me to broaden my perspectives and making the process of writing this dissertation an extraordinary learning experience.

This research would of course never have been possible without forty very special people who donated their time to my project. Some of them did so without even knowing me. Some during exam period, some after spending ten hours in an office, even early morning after working the night shift. No matter the circumstance, all of them received me with a smile. At times, interviews would take considerably longer, because they were eager to know more about what I was doing, and because they all had an interesting story to tell. And it was the same on both sides of the ocean. All of this was an unbelievable source of learning for me, just as enriching as all the articles and books. I will always have the best memories of those forty meetings and of the wonderful people who participated in them.

I would also like to thank two friends who assisted me enormously with subject recruitment: Regina Call Daví and Gaia Paixão. I feel indebted to them for the disinterested help that they provided me, as well as for their friendship and support. And I wish to thank another friend, Dr. Diane Uber, for her guidance and assistance with bibliography and with language doubts.

Finally, I thank the most important sources of inspiration in my life: Jorge López and Carolina Gambino, for three decades of unwavering support and unconditional trust.

7. REFERENCES

Adda-Decker, M., & Lamel, L. (2005) Do speech recognizers prefer female speakers? In *Proceedings of the 9th European Conference on Speech Communication and Technology - Interspeech 2005*, pp. 2205-2208. Lisbon, Portugal, September 4-8, 2005. Retrieved from http://www.iscaspeech.org/archive/interspeech_2005/i05_2205.html

Aguilar, L., Blecua, B., Machuca, M. J., & Marín, R. (1993). Phonetic reduction processes in spontaneous speech. In *Proceedings of the 3rd European Conference on Speech Communication and Technology - Eurospeech 1993,* pp. 433-436. Berlin, Germany, September 21-23, 1993. Retrieved from http://liceu.uab.cat/publicacions/Aguilar_et_al_93_Phonetic_Reduction.pdf

Alvar, M. (1999). *Manual de dialectología hispánica. El español de España*. Barcelona: Ariel.

Anusuya, M. A., & Katti, S. K. (2009). Speech recognition by machine: A review. *International Journal of Computer Science and Information Society, 6*(3), 181-205. Retrieved from http://arxiv.org/pdf/1001.2267

Ávila, R. (2009). La pronunciación del español: Medios de difusión masiva y norma culta. *Nueva Revista de Filología Hispánica 1,* 57-90. Retrieved from http://www.colmex.mx/academicos/cell/ravila/docs/Pronunciacion.pdf

Benzeghiba, M., Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., . . . Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication, 49*(10-11), , 763-786. doi:10.1016/j.specom.2007.02.006

Binnenpoorte, D., Van Bael, C., Den Os, B., & Boves, L. (2005) Gender in everyday speech and language: a corpus-based study. In *Proceedings of the 9th European Conference on Speech Communication and Technology - Interspeech 2005,* pp. 2213-2216. Lisbon, Portugal, September 4-8, 2005. Retrieved from http://www.isca-speech.org/archive/interspeech_2005/i05_2213.html

Burger, S., Sloane, Z. A., & Yang, J. (2006). Competitive evaluation of commercially available speech recognizers in multiple languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation – LREC 2006.* Genoa, Italy, May 24-26, 2006. Retrieved from http://pages.cs.brandeis.edu/~marc/misc/proceedings/lrec-2006/pdf/802_pdf.pdf

Calero Fernández, M. A. (2006). El español hablando en Cataluña: una revisión sociolingüística. In A. M. Cestero Mancera, I. Molina Martos, & F. Paredes García (Eds.), *Estudios sociolingüísticos del español de España y América* (pp. 205-211). Madrid: Arco Libros.

Carbó, C., Llisterri, J., Machuca, M. J., de la Mota, C., Riera, M., & Ríos, A. (2003). Estándar oral y enseñanza de la pronunciación del español como primera lengua y como lengua extranjera. *ELUA, Estudios de Lingüística de la Universidad de Alicante, 17*, 161-180. Retrieved from http://liceu.uab.cat/~joaquim/publicacions/Carbo_et_al_ELUA03.pdf

Carricaburo, N. (1997). *Las fórmulas de tratamiento en el español actual*. Madrid: Arco Libros.

Centro de noticias ONU (2006). Países de habla hispana promueven uso del español en la ONU. Retrieved from http://www.un.org/spanish/News/fullstorynews.asp?newsID=6370

Colley, M. (2009). The role of speaking style and context in the accuracy of automated speech recognition. *American Speech* (submitted). Retrieved from http://michaelcolley.com/Speech_Recognition_paper_Michael_Colley.pdf

Donni de Mirande, N. (1996). Argentina-Uruguay. In M. Alvar (Ed.), *Manual de dialectología hispánica: El español de América* (pp. 209-219). Barcelona: Ariel.

D'Introno, F., Del Teso, E., & Weston, R. (1995). *Fonética y fonología actual del español*. Madrid: Cátedra.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572-587. doi:10.1016/j.patcog.2010.09.020

Farnetani, E., & Recasens, D. (2010). Coarticulation and connected speech processes. In W. Hardcastle, J. Laver, & F. Gibbon (Eds.), *The Handbook of Phonetic Sciences (*2nd ed.) (pp. 316-354). Oxford: Blackwell.

Fernández Trinidad, M. (2010). Variaciones fonéticas del yeísmo: un estudio acústico en mujeres rioplatenses. *Estudios de fonética experimental*, *XIX*, 263-292. Retrieved from http://www.raco.cat/index.php/EFE/article/view/218611/298349

Fontanella de Weinberg, M. B. (1992). *El español de América.* Madrid: MAPFRE.

Fontanella de Weinberg, M. B. (2004). El español bonaerense. In *El español de la Argentina y sus variedades regionales (*2nd ed.) (pp. 45-73). Bahía Blanca: Asociación Bernardino Rivadavia.

Foulkes, P., Scobbie, J., & Watt, D. (2010). Sociophonetics. In W. Hardcastle, J. Laver, & F. Gibbon, (Eds.), *The handbook of phonetic sciences (*2nd ed.) (pp.703-754). Oxford: Blackwell.

García Mouton, P. (1994). *Lenguas y dialectos de España*. Madrid: Arco Libros.

García Mouton, P. (2000). *Cómo hablan las mujeres (*2nd ed). Madrid: Arco/Libros.

García, C., & Tapias, D. (2000). La frecuencia fundamental de la voz y sus efectos en reconocimiento de habla continua. *Procesamiento del Lenguaje Natural*, *26*, 163-167. Retrieved from http://www.sepln.org/revistaSEPLN/revista/26/garcia.pdf

Gil, J. (1988). *Los sonidos del lenguaje*. Madrid: Síntesis.

Guitarte, G. (1955). El ensordecimiento del žeísmo porteño. *Revista de Filologia Española,39*, 261-283.

Guitarte, G. (1991). Del español de España al español de veinte naciones: La integración de América al concepto de lengua española. In *El español de América: Actas del III Congreso Internacional de el español en América. Vol. 1* (pp. 65-86). Valladolid: Junta de Castilla y León.

Hamdan, A. -L., Al-Barazi, R., Tabri, D., Saade, R., Kutkut, I., Sinno, S., & Nassar, J. (2011). Relationship between acoustic parameters and body mass analysis in young males. *Journal of Voice*, in press, corrected proof. doi:10.1016/j.jvoice.2011.01.011

Hock, H. (1986). *Principles of historical linguistics*. Berlin: Mouton de Gruyter.

Hualde, J. I. (2005). *The sounds of Spanish*. Cambridge: Cambridge University Press.

Hutchinson, B. (2001). A functional approach to speech recognition evaluation. In *Eurospeech 2001 Scandinavia. Proceedings of the 7th European Conference on Speech Communication and Technology, 2nd Interspeech Event*, pp. 1683-1686. Aalborg, Denmark, September 3-7, 2001. Retrieved from http://perso.telecomparistech.fr/~chollet/Biblio/Congres/Audio/Eurospeech01/CDROM/papers/page1683.pdf.

Jekat, S., & Schultz, T. (2004). Evaluation sprachverarbeitender Systeme. In K.U. Carstensen, C. Ebert, C. Endriss, S. Jekat, R. Klabunde. & H. Langer (Eds.), *Computerlinguistik und Sprachtechnologie: eine Einführung (*pp. 573-590). Munich: Elsevier-Spektrum Akademischer Verlag.

Jurafsky, D., & Martin, J.H. (2009). *Speech and language processing. An introduction to natural language processing, computational linguistics and speech recognition (*2nd ed.). New Jersey: Prentice Hall.

Kaisse, E., (1985). *Connected speech: The interaction of syntax and phonology*. Orlando: Academic Press.

Labov, W. (1972). The isolation of contextual styles. In *Sociolinguistic patterns (*pp. 70-109). Oxford: Basil Blackwell.

Lamel, L., & Gauvain, J.L. (2003). Speech recognition. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics (*pp. 305-322). Oxford: Oxford University Press.

Lieberman, P., & Blumstein, S. E. (1988). *Speech physiology, speech perception and acoustic phonetics.* Cambridge: Cambridge University Press.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H and H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling (*pp. 403-439). Dordrecht: Kluwer.

Lipski, J. M. (1994). *Latin American Spanish*. London & New York: Longman.

Llisterri, J. (2007). Màquines que parlen i que escolten: El paper de la fonètica en el desenvolupament de les tecnologies de la parla. In J. Carrera & C. Pons (Eds.), *Aplicacions de la fonètica. Catorzè Col·loqui Lingüístic de la Universitat de Barcelona* (pp. 139-169). Barcelona: Promociones y Publicaciones Universitarias. Retrieved from http://liceu.uab.cat/~joaquim/publicacions/Llisterri_07_Fonetica_Tecnologies_Parla.pdf

Llisterri, J. (2008). La representación ortográfica de corpus orales. Universitat Autònoma de Barcelona. Retrieved from http://liceu.uab.cat/~joaquim/language_resources/spoken_res/Repres_ortog_corp_oral.html

Llisterri, J. (2011a). La fonación. Universitat Autònoma de Barcelona. Retrieved from ttp://liceu.uab.cat/~joaquim/phonetics/fon_produccio/fonacion.html#fonacion_habla

Llisterri, J. (2011b) Las aplicaciones del reconocimiento automático del habla. Universitat Autònoma de Barcelona. Retrieved from http://liceu.uab.cat/~joaquim/speech_technology/tecnol_parla/recognition/applications_recognition/aplicaciones_reconocimiento.html.

Martínez Celdrán, E., & Fernández Planas, A. M. (2007). *Manual de fonética española: articulaciones y sonidos del español*. Barcelona: Ariel.

Martínez, F., Tapias, D., Álvarez, J., & León, P. (1997). Characteristics of slow, average and fast speech and their effects in large vocabulary continuous speech recognition. In *roceedings of the 5th European Conference on Speech Communication and Technology, Eurospeech 1997*, pp. 469-472. Rhodes, Greece. September 22-25, 1997. Retrieved from http://www.iscaspeech.org/archive/eurospeech_1997/e97_0469.html

Navarro Tomás, T. (1999). *Manual de pronunciación española (*27th ed). Madrid: Consejo Superior de Investigaciones Científicas.

Neri, A., Cucchiarini, C., & Strik, H. (2003). Automatic speech recognition for second language learning: How and why it actually works. In *Proceedings of the 15th International Congress of Phonetic Sciences, ICPhS 2003*, pp. 1157-1160. Barcelona, Spain, August 3-9, 2003. Retrieved from http://lands.let.kun.nl/literature/neri.2003.1.pdf

NIST. (n.d.). NIST Sclite Scoring Package. Retrieved from http://www.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm#sclite_name_0

NIST IAD (2009) The history of automatic speech recognition evaluations at NIST. Retrieved from http://itl.nist.gov/iad/mig/publications/ASRhistory/index.html

Orlikoff, R. F., & Kahane, J. C. (1996). Structure and function of the larynx. In N. J. Lass (Ed.), *Principles of experimental phonetics* (pp. 112-184). St. Louis: Mosby.

Pallett, D. S. (1985). Performance assessment of automatic speech recognizers. *Journal of Research of the National Bureau of Standards, 90*(5), 371-387. Retrieved from http://nvl.nist.gov/pub/nistpubs/jres/090/5/V90-5.pdf#page=41.

Pallett, D. S., & Fourcin, A. (1996). Speech input: Assessment and evaluation. In R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, & V. Zue (Eds.), *Survey of the state of the art in human language technology* (pp. 495-499). Cambridge: Cambridge University Press. Retrieved from http://cslu.cse.ogi.edu/HLTsurvey/ch13node8.html
.

Quillis, A. (1993). *Tratado de fonología y fonética españolas*. Madrid: Gredos.

Rabiner, L. R., & Juang, B. H.. (2006) Speech recognition: Statistical methods. In K. Brown (Ed.), *Encyclopedia of language & linguistics* (pp. 1-18). Amsterdam: Elsevier. doi:10.1016/B0-08-044854-2/00907-X.

RAE. (n.d.). La política lingüística panhispánica. Retrieved from http://www.rae.es/rae/Noticias.nsf/Portada4?ReadForm&menu=4

RAE (2005). Voseo. In *Diccionario panhispánico de dudas*. Madrid: Santillana. Retrieved from http://buscon.rae.es/dpdI/

Rissel, D. (1981) Diferencias entre el habla femenina y la masculina en español. *Thesaurus, XXXVI* (2), 305-322. Retrieved from http://cvc.cervantes.es/lengua/thesaurus/boletines/1981.htm

Rodríguez, L. J., & Torres, M. I. (2006). Spontaneous speech events in two speech databases of human-computer and human-human dialogs in Spanish. *Language and Speech*, *49*(3), 333-366. doi:10.1177/00238309060490030201

Serrahima, L. (2009). Reconocimiento de voz de Windows Vista: ¿Mejor, igual o peor que Dragon Naturally Speaking? *Panace@*, *10*(29), 76-79. Retrieved from http://medtrad.org/panacea/IndiceGeneral/n29_tribuna-Serrahima2.pdf

Simpson, A. P. (2009) Phonetic differences between male and female speech. *Language and linguistics compass,3*(2), 621–640. doi: 10.1111/j.1749-818x.2009.00125.x.

Stevens, K.N. (1972). Sources of inter- and intra- speaker variability in the acoustic properties of speech sounds. In *Proceedings of the 7th International Congress of Phonetic Sciences, ICPhS 1972 (*pp. 206-232). The Hague: Mouton.

Strik, H., & Cucchiarini, C. (1999) Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, *29*(2-4), 225-246. Retrieved from http://lands.let.kun.nl/TSpublic/strik/a64b.html.

Tapias , D. (2002) Interfaces de voz con lenguaje natural. In M.A. Martí & J. Llisterri (Eds.), *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita* (pp. 189-207). Barcelona: Edicions de la Universitat de Barcelona - Fundación Duques de Soria.

Zamora, A. (1967). *Dialectología española (*2nd ed). Madrid: Gredos.

8. APPENDIX

APPENDIX 1: SPEAKER ELIGIBILITY QUESTIONNAIRE – PENINSULAR SPEAKERS

1) Nombre y apellidos:

2) Edad:

3) Teléfono de contacto:

4) Correo electrónico:

5) Lugar de nacimiento (localidad y provincia):

6) Lugar de nacimiento de la madre (localidad y provincia):

7) Lugar de nacimiento del padre (localidad y provincia):

8) Tiempo de residencia en el lugar de nacimiento:

9) Lugar de residencia actual (localidad y provincia):

10) Tiempo de residencia en el lugar de residencia actual:

11) Lugares de residencia anteriores (localidad y provincia):

12) Tiempo de residencia en los lugares de residencia anteriores:

13) ¿Qué lengua aprendiste en casa?

□ Castellano □ Catalán □ Las dos

14) ¿A qué edad comenzaste a aprender el castellano?

□ 0-5 años □ 6-10 años □ 11-20 años □ 21-adelante

15) ¿A qué edad comenzaste a aprender el catalán?

□ 0-5 años □ 6-10 años □ 11-20 años □ 21-adelante

16) ¿En qué lengua te impartían las clases en la escuela?

□ Siempre en castellano □ Más en castellano que en catalán □ Más en catalán que en castellano □ Siempre en catalán □ Tanto en castellano como en catalán

17) ¿En qué lengua te diriges a tu madre?

□ Siempre en castellano □ Más en castellano que en catalán □ Más en catalán que en castellano □ Siempre en catalán □ Tanto en castellano como en catalán

18) ¿En qué lengua te diriges a tu padre?

□ Siempre en castellano □ Más en castellano que en catalán □ Más en catalán que en castellano □ Siempre en catalán □ Tanto en castellano como en catalán

19) ¿En qué lengua te diriges a tus hermanos/as?

□ Siempre en castellano □ Más en castellano que en catalán □ Más en catalán que en castellano □ Siempre en catalán □ Tanto en castellano como en catalán

20) ¿En qué lengua te diriges a tu pareja?

□ Siempre en castellano □ Más en castellano que en catalán □ Más en catalán que en castellano □ Siempre en catalán □ Tanto en castellano como en catalán

21) Si puedes elegir, ¿en qué lengua prefieres expresarte?

□ Castellano □ Catalán □ Tanto en castellano como en catalán

22) ¿Qué otras lenguas hablas? Indica tu nivel en cada una de ellas.

....

23) ¿Con qué frecuencia utilizas el ordenador?

□ Nunca □ Menos de una vez a la semana □ Una o dos veces a la semana

□ Tres o cuatro veces a la semana □ Cinco o más veces a la semana

24) ¿Qué tipo de tareas realizas con el ordenador?

....

APPENDIX 2: SPEAKER ELIGIBILITY QUESTIONNAIRE – *PORTEÑO* SPEAKERS

Tus datos personales se tratarán de forma confidencial. Se solicita tu número de teléfono y tu dirección electrónica únicamente para poder contactarte si, eventualmente, fuera necesario repetir alguna grabación.

- Nombre y apellidos:

- Correo electrónico:

- Teléfono:

- Edad:

- Lugar de nacimiento (localidad y provincia):
- Tiempo de residencia en tu lugar de nacimiento
- Lugar de residencia actual (localidad y provincia)
- Tiempo de residencia en tu lugar de residencia actual
- Lugares de residencia anteriores (localidad y provincia) y tiempo de residencia en cada uno de ellos)
- Lugar de nacimiento de tu madre (localidad y provincia)
- Lugar de nacimiento de tu padre (localidad y provincia)
- Lugar de nacimiento de tu pareja (localidad y provincia)
- Si hablás otras lenguas además de castellano, enumeralas e indicá tu nivel en cada una de ellas:

- Estudios cursados
    - completos:
    - en curso:
    - incompletos:

- ¿Con qué frecuencia utilizás la computadora?
- ¿Para qué tareas la utilizás?
- ¿Utilizás algún sistema de reconocimiento del habla? SI/NO

  - En caso afirmativo, indicar:

    a) nombre del sistema:

    b) ¿cuánto tiempo hace que lo utilizás?

    c) ¿con qué frecuencia lo hacés?

    d) ¿qué tipo de tareas realizás mediante el mismo? (manejo de comandos, dictado de e-mails, etc.)

- Por favor, indicá tu disponibilidad horaria para realizar una única reunión de 30-40 minutos (días y horarios):

¡Muchas gracias por tu tiempo!

Soledad ☺ - soledadlopezg@yahoo.com

APPENDIX 3: TEST TASKS

Tarea 1

Abrir explorador de Windows

Hacer clic en “equipo”

Hacer doble clic en “disco local C”

Cerrar

Menú inicio

Todos los programas

Accesorios

Paint

Minimizar

Inicio bloc de notas

Archivo

Mostrar números

2

Aceptar

Lloviendo

Abrir

Ir a “licencia”

Seleccionar las 10 palabras siguientes

Bajar 20

Cambiar a Paint

Tarea 2

Acabas de salir de la universidad y deseas pasar a tu ordenador la información que has apuntado a mano en tres clases, pero no tienes ganas de teclear. Lee las notas de cada clase y díctaselas al ordenador en forma de textos, pensando en que los utilizarás después para redactar varios trabajos que te han pedido.

Importante: Cuando necesites utilizar signos de puntuación, debes nombrarlos en voz alta.

Ejemplo:

Si la oración es: *Me gusta escuchar música, bailar y salir.*

Tú dices: *Me gusta escuchar música* **COMA** *bailar y salir* **PUNTO**.

HISTORIA

Martin Luther King (Atlanta, 15 enero 1929 – Memphis, 4 abril 1968)

Nombre verdadero: Michael King Junior - pastor iglesia Bautista – líder movimiento por los derechos civiles – Premio Nobel Paz.

Asistió a Booker T. Washington High School.

Acciones recordadas: - Ayuda en campaña de Birmingham: boicots, protestas en restaurantes, marchas pacíficas, famoso discurso frente a + de 200.000 personas.

LITERATURA:

Prof. Yolanda Bauzá presenta cuento de Roald Dahl “La pata de cordero”.

Marido dice a mujer q’ la dejará.

Ella: saca pata de cordero del congelador – golpe en la nuca – marido muerto.

Luego: - cocina el cordero - disimula: llama a policía y pide ayuda - les invita cordero –

Policía: come con gusto - investiga escena crimen - nunca halla arma asesina.

GEOGRAFÍA

Honduras - (Centroamérica)

Capitales: Tegucigalpa y Comayagüela

Superficie: 112.492 km²

Población: 7.793.000hab.

Forma de gobierno: república democrática representativa.

Tarea 3:

¡Uno de tus amigos cumplió años y no lo llamaste! 😲 Envíale un correo electrónico:

-Pidiéndole perdón por el olvido y justificándolo con un buen motivo.

- Propón un encuentro, sugiere un lugar, un día y una hora, y pregúntale si le iría bien.

- Dile que no sabes qué regalarle, dale 2 o 3 opciones de regalos que se te ocurre que le podrían gustar, y pregúntale qué prefiere.

- Recuérdale que te lleve el CD que le prestaste, y dile por qué lo necesitas.

Recuerda que para usar **signos de puntuación** debes usar sus nombres: "punto", "dos puntos", "coma", etc. Además, puedes agregar los siguientes.

- ***Abrir/cerrar signo de interrogación/exclamación***
- ***Nueva línea*** (para escribir en una nueva línea).

Estas notas pueden ayudarte mientras dictas:

✓ **PERDÓN – JUSTIFICACIÓN**

✓ **ENCUENTRO: DÍA, LUGAR, HORA. ¿BIEN?**

✓ **REGALO: OPCIONES – ¿QUÉ PREFIERE?**

✓ **CD – LO NECESITAS PORQUE...**

APPENDIX 4: RESULTS GRID FOR TASK 1

COMANDO	RECONOCIDO	NO RECONOCIDO	PIDE CONFIRMACIÓN
1) Abrir explorador de Windows			
2) Hacer clic en “equipo”			
3) Hacer doble clic en “disco local C”			
4) Cerrar			
5) Menú inicio			
6) Todos los programas			
7) Accesorios			
8) Paint			
9) Minimizar			
10) Inicio bloc de notas			
11) Archivo			
12) Mostrar números			
13) 2			
14) Aceptar			
15) Lloviendo			
16) Abrir			
17) Ir a “licencia”			
18) Seleccionar las 10 palabras siguientes (marca hasta “etc”.)			
19) Bajar 20			
20) Cambiar a Paint			