UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

Aportación a la predicción de palabras en castellano y su integración en sistemas de ayuda a personas con discapacidad física

TESIS DOCTORAL

Sira Elena Palazuelos Cagigas

Ingeniero de Telecomunicación

2001

DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

Aportación a la predicción de palabras en castellano y su integración en sistemas de ayuda a personas con discapacidad física

Sira Elena Palazuelos Cagigas

Ingeniero de Telecomunicación

Director: Santiago Aguilera Navarro
Doctor Ingeniero de Telecomunicación

2001

TESIS DOCTORAL: Aportación a la predicción de palabras en castellano y su integración en sistemas de ayuda a personas con discapacidad física.

AUTORA: Sira Elena Palazuelos Cagigas

DIRECTOR: Santiago Aguilera Navarro

El tribunal nombrado para juzgar la Tesis arriba indicada, compuesto por los siguientes doctores:

PRESIDENTE: D.

VOCALES: D.

D.

D.

SECRETARIO: D.

acuerdan otorgarle la calificación de:

Madrid

El Secretario del Tribunal

A mis padres,
mi hermana y Javi

Agradecimientos

Muchas son las personas a las que tengo que agradecer haber llegado hasta aquí. En primer lugar, a Santiago Aguilera, director de este trabajo, por haber apostado por mí, dándome la oportunidad de trabajar en este campo, que ha resultado ser apasionante, y por la enorme paciencia y el apoyo que me ha dado a lo largo de estos años.

Esta tesis me ha permitido conocer a gran cantidad de gente, que ha tenido mucha influencia en mi vida, tanto profesional como personalmente. El Laboratorio de Tecnología de Rehabilitación me ha proporcionado todo su apoyo profesional, además de conseguir convertir en agradables los largos días de trabajo. Gracias a todos los proyectandos Alberto, Luis, los Franciscos, Juan, Antonio, Tobar, Juanjo, Elena y David etc., por aportar su granito de arena.

Gracias a todos mis amigos y compañeros del Departamento de Ingeniería Electrónica de la ETSI de Telecomunicación de la UPM, donde se ha elaborado esta tesis, que han compartido conmigo estos años apoyándome.

Especialmente enriquecedoras han resultado mis estancias en el extranjero, en la Universidad de Sheffield; en la Universidad de Dundee, con Alan Newell, Ian Ricketts, Peter Gregor y todos los demás; en el KTH, en Estocolmo, donde disfruté de la calidad científica y humana de Sheri Hunnicutt, Hakan Melin y Alice Carlberger, entre otros; y en el GRIL de la Universidad de Clermont Ferrand, donde tuve tantas fructíferas discusiones con Gabriel Bes y José L. Rodrigo, que posteriormente fueron continuadas en España con Francisco Aliaga.

A lo largo de los años, he disfrutado de la comprensión del Departamento ICS de la EUIT de Telecomunicación de la UPM, y por supuesto, del Departamento de Electrónica de la Universidad de Alcalá, donde no ha pasado un sólo día sin que alguno de sus miembros tuviera para mí palabras (y muchas obras) de ánimo. Gracias a Manuel Mazo, José Luis Lázaro y todos los demás y, sobre todo, a los habitantes de los pasillos 21 y 22.

Esta tesis no habría sido la misma sin el CEAPAT, sus proyectos, sus profesionales (Cristina Rodríguez, Cristina Larraz, Lola Abril, Margarita Sebastián y tantos otros) y la valiosa información que ha puesto a mi disposición, que espero haya proporcionado los frutos deseados, aportando un granito de arena a la mejora de la calidad de vida de las personas con discapacidad. Muchas gracias también a Amparo Candelas y Manuel Lobato. También me gustaría agradecer todas las críticas, sugerencias, comentarios e informes recibidos, especialmente a David Carreres y Sergio, del Centro Maset de Frater, y a Montse, que me han ayudado a mejorar mi trabajo, a la vez que me animaban a seguir adelante.

En el plano personal, he tenido mucha gente cerca. Quiero dar las gracias a las familias Baena-Moreno y Romero-Marín, por haberme permitido compartir un trocito de sus vidas. A Javier Romañach, por sus estimulantes charlas sobre lo divino y lo humano. A Fernando González, por estar ahí cuando hace falta. A Pepe, porque su entusiasmo tiene parte de culpa de que yo esté aquí. A Javier Gamo, que aparece oportunamente en los momentos más insospechados de mi vida. A Jero, por los ratos de trabajo codo con codo. A Javi Rubio, Jorge, Cris y Olga, por las estupendas veladas que hemos disfrutado juntos. A Juan Ignacio (¡ánimo, ahora te toca a ti!), compañero de fatigas con quien he compartido tantos buenos y malos momentos.

A Javi y Mónica, por recordarme lo que significa la palabra amistad, y a Jose, porque siempre Está dispuesto a ofrecer una mano llena de pAz.

Por último, muchas gracias a toda mi familia y a Javi, por su cariño y apoyo incondicional y por soportar todas mis ausencias, que espero no se vuelvan a repetir.

Por supuesto, son todos los que están, pero no están todos los que son.

Muchísimas gracias a todos...

Índice General

Índice de figuras **vii**

Resumen **ix**

Abstract **xi**

1 Introducción **1**

2 Encuadre científico-tecnológico **5**

2.1 Introducción 5

2.2 Modelos de lenguaje 6

2.2.1 Modelos de lenguaje estadísticos 6

2.2.2 Modelos categoriales 7

2.2.3 Modelos formales 8

2.2.3.1 Gramáticas formales 8

2.2.3.2 Analizadores 12

2.2.3.3 Tipología de errores en el análisis del lenguaje natural . . . 15

2.2.4 Gramáticas de unificación 16

2.2.5 Gramáticas basadas en marcos de subcategorización 17

2.2.6 Gramáticas asociativas por la izquierda 17

2.3 Interfaces adaptativas orientadas a usuarios con discapacidad 18

2.3.1 Criterios generales de diseño de interfaces de usuario 18

2.3.2 Diseño de interfaces de usuario orientadas a personas con discapacidad física 19

2.3.3 Interfaces adaptativas 20

2.4 Evaluación 21

2.4.1 Evaluación cuantitativa 21
2.4.2 Evaluación cualitativa 21
2.5 Sistemas que utilizan predicción de palabras 22
2.5.1 Aurora 3.0 22
2.5.2 Co:Writer 23
2.5.3 EZ Keys 23
2.5.4 PAL 24
2.5.5 PredictAbility 24
2.5.6 Predice 25
2.5.7 Prophet 26
2.5.8 SAW (Switch Access to Windows) 27
2.5.9 SoothSayer Word Prediction 27
2.5.10 Telepatic 28
2.5.11 TPV 28
2.5.12 Write Away 2000 28

3 Predicción de palabras 31
3.1 Arquitectura general del sistema 31
3.2 Diccionarios 33
3.2.1 Diccionario general 34
3.2.1.1 Estructura del diccionario general 34
3.2.1.2 Textos de entrenamiento del diccionario general 36
3.2.1.3 Generación del diccionario general 36
3.2.1.4 Elección del conjunto de categorías y rasgos 39
3.2.1.5 Mecanismos de gestión del diccionario general 48
3.2.1.6 Limitaciones del diccionario general 50
3.2.2 Diccionario personal y diccionarios temáticos 51
3.2.2.1 Textos de entrenamiento del diccionario personal y de los temáticos 52
3.2.2.2 Contenido del diccionario personal y de los temáticos . . . 52
3.2.2.3 Generación del diccionario personal y de los temáticos . . . 52
3.2.2.4 Gestión del diccionario personal y de los temáticos 53
3.2.2.5 Problemática del diccionario personal y de los temáticos . . 54

3.3 Métodos de predicción 55
3.3.1 Métodos estadísticos básicos 56
3.3.1.1 Limitaciones de los métodos estadísticos básicos 57
3.3.2 Modelos estadísticos categoriales 58
3.3.2.1 Mejoras introducidas en los modelos categoriales básicos . 61
3.3.2.2 Limitaciones de los métodos categoriales 69
3.3.3 Gramáticas formales 69
3.3.3.1 Selección del analizador 70
3.3.3.2 Funcionamiento del analizador modificado 71
3.3.3.3 Aumento de la potencia del analizador 74
3.3.3.4 Reglas de reescritura 93
3.3.3.5 Gestión de errores 93
3.3.4 Métodos adicionales a la predicción de palabras 95
3.3.4.1 Filtro de las palabras rechazadas 95
3.3.4.2 Predicción de terminaciones 95
3.3.4.3 Escritura automática de caracteres 96
3.4 Módulo de coordinación 96
3.5 Interfaz de usuario 99
3.5.1 Adaptaciones de la interfaz de *Predice* para personas con distintas discapacidades 100
3.5.2 Adaptación para personas con discapacidades físicas graves 101
3.5.3 Inclusión de mecanismos adaptativos en una interfaz de barrido . . . 104
3.6 Arquitectura detallada del sistema 106

4 Evaluación de la predicción de palabras 109
4.1 Evaluación automática 109
4.1.1 Consideraciones acerca de la evaluación automática 109
4.1.1.1 Textos de entrenamiento y de prueba 109
4.1.1.2 Configuración de la predicción 111
4.1.1.3 Métricas 111
4.1.2 Descripción del sistema de evaluación automática 116
4.1.3 Resultados 117
4.1.3.1 Límite máximo teórico de la predicción 119

4.1.3.2 Evaluación de los métodos estadísticos básicos 120
4.1.3.3 Evaluación de los modelos estadísticos categoriales 124
4.1.3.4 Evaluación del modelo formal 140
4.1.3.5 Evaluación del diccionario personal y de los temáticos . . . 148
4.1.3.6 Efecto de los mecanismos auxiliares de aceleración 158
4.2 Evaluación subjetiva 160
4.2.1 Evaluación a corto plazo 160
4.2.2 Evaluación a largo plazo 162
4.2.2.1 Informe 1 164
4.2.2.2 Informe 2 166
4.3 Comparación con otros sistemas 168

5 Conclusiones 171

6 Líneas de trabajo futuras 175
6.1 Diccionarios 175
6.2 Métodos de predicción 176
6.3 Métodos de evaluación 177
6.4 Aplicación de la arquitectura a otros idiomas y nuevos entornos de aplicación 178

A Descripción de los textos utilizados para entrenamiento y evaluación 179
A.1 Criterios de selección de textos 179
A.2 Características generales 180
A.3 Características cuantitativas de los textos de entrenamiento 181
A.4 Características cuantitativas de los textos de evaluación 182
A.5 Distribución frecuencial de las pulsaciones 183
A.6 Distribución estadística categorial de los textos de PAIS 184

B Comparación entre los textos TESIS y TESIS2 187

C Conjunto de reglas de reescritura del sistema 195

D Cartas escritas por usuarios de *Predice* 215

Bibliografía 221

Índice de Figuras

1.1 Pulsador adaptado a la barbilla 2

1.2 Pulsador neumático 2

2.1 Ejemplo de máquina de estados finitos. 14

2.2 Ejemplo de red de transiciones recursivas. 14

2.3 Ejemplo de red de transiciones aumentadas. 15

2.4 Pantalla principal de Aurora. 23

2.5 Pantalla principal de PredictAbility. 24

2.6 Pantalla principal de *Predice*. 25

2.7 Pantalla principal de Prophet para sueco. 26

2.8 Pantalla principal de Profeta para castellano, con emulador de teclado. . . . 26

2.9 Pantalla principal de SAW. 27

2.10 Pantalla principal de Telepatic 2000. 28

2.11 Pantalla principal de Write Away 2000. 29

3.1 Arquitectura general de la predicción de palabras. 32

3.2 Diagrama de bloques del diccionario general. 35

3.3 Diagrama de bloques del diccionario personal y los temáticos. 51

3.4 Pantalla principal de *Predice*. 100

3.5 Matriz de letras de *Predice*. 103

3.6 Diagrama de bloques detallado del sistema de predicción de palabras. . . . 107

A.1 Distribución de palabras en función de pulsaciones necesarias para el texto TESIS 184

A.2 Distribución de palabras en función de pulsaciones necesarias para el texto QUIJOTE15-13K 184

A.3 Distribución de palabras en función de pulsaciones necesarias para el texto FEB100K 185

Resumen

La presente tesis doctoral aborda el estudio de la inclusión de información lingüística en la predicción de palabras en castellano, con el objetivo de mejorar los sistemas de ayuda a la escritura de personas que pueden tener distintos tipos de discapacidad.

Para la inclusión de la información lingüística proponemos una arquitectura novedosa que permite desarrollar una metodología original de combinación de las diversas fuentes de información exploradas (fundamentalmente en los niveles léxico, morfológico y sintáctico), gracias a la inclusión de un módulo de gestión y combinación de los flujos de información utilizados y a la separación estricta entre los distintos diccionarios (general, personal y temático) y los métodos de predicción en sí.

Los métodos de predicción que se han incluido utilizan dos estrategias fundamentales de modelado de la información lingüística: el modelado estocástico (unigramas, bigramas, bipos y tripos) y el formal (que hace uso de una gramática probabilística independiente del contexto a la que hemos dotado de potencia adicional).

En cada uno de los módulos que realizan la inclusión del conocimiento lingüístico se han realizado también aportaciones específicas, tanto en el diseño y organización de la información (fundamentalmente orientada a su uso en la gramática formal), como en la metodología particular de uso de la misma de cara a la predicción de palabras y a la mejor colaboración con otros módulos.

Consideramos también una contribución destacable de esta tesis el criterio de diseño y la definición de las categorías, para conectar mejor con los comportamientos sintácticos reales observados, junto con el diseño de un conjunto de rasgos hacia los que se desplaza parte de la carga expresiva y, por supuesto, los originales mecanismos de manejo de los mismos que hemos incluido en el diseño de la gramática formal.

Con respecto al modelado formal, el estudio detallado de los fenómenos lingüístico tanto teórica como empíricamente, nos ha llevado a diseñar una gramática probabilística independiente de contexto que contempla una original imbricación de mecanismos (concordancia, imposición y prohibición de rasgos en símbolos terminales, potente gestión de rasgos también en símbolos no terminales, imposición y prohibición tanto de lemas como de significantes y la presencia de elementos opcionales) que la dotan de una importante potencia descriptiva del lenguaje, al tiempo que el número de reglas se mantiene controlado dentro de unos márgenes razonables para su tratamiento computacional.

En este trabajo no nos hemos limitado a realizar un planteamiento teórico, sino que hemos implementado y evaluado un sistema construido según la arquitectura propuesta,

en la que, además, se han tenido en cuenta consideraciones específicas para el diseño de su interfaz de usuario.

Se aporta igualmente un estudio detallado de los diversos factores que influyen en la evaluación cuantitativa (sobre los cuales sería necesario realizar un esfuerzo de normalización, dada la ausencia de estándares definidos al respecto), proponiendo métricas que analizan la potencia de las fuentes de información y que permiten elegir la mejor estrategia de combinación de métodos que represente de hecho una ayuda para los usuarios de esta tecnología. En dicha combinación se da prioridad a las palabras que aporta el modelo basado en bigramas de los diccionarios temático y personal. A continuación se utilizan los modelos estocásticos categoriales, aplicados primero al diccionario temático correspondiente y, posteriormente, con una adecuada ponderación, a los diccionarios personal y general.

En cuanto al método de predicción basado en la gramática formal, el conjunto de aportaciones realizadas ha permitido obtener resultados cercanos a los de los métodos basados en modelos estocásticos categoriales, quedando para las líneas de trabajo futuras el completar su capacidad descriptiva. La modularidad y flexibilidad con que se ha dotado a la arquitectura permitirán realizar esta investigación aprovechándose del gran esfuerzo ya invertido aquí.

Abstract

This Ph.D. thesis is aimed at the study of including linguistic information in word prediction for Spanish, with the main objective of improving the writing aids available for people with different kind of disabilities.

In order to include linguistic information, we propose a novel architecture that allows the development of an original methodology in order to combine the different sources of information we have explored (mainly in the lexical, morphological and syntactic levels), thanks to the inclusion of a management module, able to deal with and combine the different information flows used, and to the strict separation between the lexicons (main, custom and subject) and the prediction methods themselves.

The prediction methods included use two main modeling strategies for the linguistic information: stochastic modeling (unigrams, bigrams, bipos and tripos) and formal modeling (using a probabilistic context free grammar strengthened with additional characteristics).

In every module including linguistic knowledge, we have made specific contributions, both in the design and organization of the information (mainly oriented to be used in the formal grammar) and in the particular methodology of using this information when facing word prediction and the adequate cooperation with other modules.

We also consider a significant contribution of this thesis the design criterion and the definition of the grammatical parts-of-speech (pos) used, in order to better connect with the observed syntactic behavior, along with the design of a feature set towards which we have shifted part of the expressive content. In order to deal with both pos and features, we also propose some original mechanisms included in the design of the formal grammar.

With respect to the formal model, the detailed study of linguistic phenomena (both theoretically and empirically) has led us to design a probabilistic context free grammar that uses an original interweaving of different mechanisms (terminal symbol feature concordance, imposition and prohibition; powerful feature management also in non terminal symbols; lemma and word imposition and prohibition; and the possibility of dealing with optional symbols) that endow it with a significant descriptive power of the language, while keeping the number of rules and the search process computationally tractable.

Our work is not only limited to a theoretical study. We have also implemented and evaluated a working system, built following the proposed architecture in which, additionally, we have taken into account specific considerations on the user interface design.

We also contribute a detailed study on the different factors that affect the quantitative evaluation (where a normalization effort should be done, given the lack of defined standards

on this topic), proposing metrics able to analyze the power of the information sources that allows us to select the best combination strategy leading to actual improvements for the users of this technology. In this combination, we prioritize the words coming from the subject and custom lexicons using a bigram model. After this, we use the stochastic pos models, applied first to the subject lexicon and afterwards, with an adequate weighting, to the custom and main lexicons.

With respect to the word prediction method based in the formal grammar, the overall set of contributions allowed us to get results close to those obtained with the stochastic pos models, leaving for future research the completion of its descriptive capabilities. The modularity and flexibility of the architecture will allow us to carry out this research work taking great advantage of the effort already invested here.

Capítulo 1

Introducción

Los ordenadores están cobrando cada vez mayor importancia en nuestra sociedad. Está aumentando vertiginosamente el número de ordenadores en los hogares y su inserción en el mundo laboral es indudable. La integración de Internet en la vida cotidiana está cambiando paulatinamente la forma de acceso a la información, comunicación, compras, reserva de billetes, etc. Lo que en principio era solamente una herramienta, está modificando sustancialmente la forma de vida de muchas personas.

Hay un sector de la sociedad para el que los ordenadores son especialmente importantes, por la trascendencia de las tareas que les permite realizar: las personas con discapacidad [Vila98]. Dependiendo del tipo de discapacidad y de su grado, estas personas pueden tener dificultades e incluso estar totalmente imposibilitadas para moverse, comunicarse o controlar su entorno por sí mismas. Los problemas en actividades tan básicas como las mencionadas dificultan en mayor o menor medida su vida diaria. Se puede pensar, por ejemplo, en las dificultades con las que se encuentran las personas con parálisis cerebral, esclerosis lateral amiotrófica, las víctimas de accidentes, etc. [Agui99].

Es ahora, con la flexibilidad de los ordenadores, cuando se pueden construir sistemas capaces de ayudar a estas personas de forma eficiente, con soluciones a la vez sencillas de utilizar y suficientemente potentes. La versatilidad de los ordenadores permite adaptarlos para que personas con muy diversas necesidades los puedan utilizar, aumentando su calidad de vida y su grado de autonomía personal. Piénsese, por ejemplo, en las escasas posibilidades que hubiera tenido hace unos años Stephen Hawking, aquejado de esclerosis lateral amiotrófica, que le provoca graves problemas físicos y de comunicación. En la actualidad, utilizando su ordenador personal con un solo dedo, es capaz de comunicarse, impartir conferencias e, incluso, escribir libros por sí mismo.

La característica que hace al ordenador especialmente adecuado para servir de base a las ayudas técnicas, es su posibilidad para adaptarse a las necesidades de los usuarios. Habitualmente el ordenador se utiliza manejando el teclado y el ratón, visualizando datos en la pantalla y escuchando mensajes sonoros. Sin embargo, para las personas con discapacidad que tienen problemas con algunos o todos estos dispositivos, es posible modificar dicha interacción, para evitar las partes del sistema que el usuario no pueda manejar.

Las adaptaciones que se deben realizar dependen de la discapacidad del usuario: a las

personas sordas se les muestran por escrito todos los mensajes audibles; para las personas ciegas se transforma la información visual en auditiva (mediante síntesis de voz) o táctil (utilizando líneas braille), y para las personas con discapacidades físicas (con dificultades para utilizar el ratón o el teclado), se buscan formas de entrada alternativas.

Las dificultades de las personas con discapacidad física a la hora de utilizar un ordenador son muy variadas, tanto en su tipología como en su grado. Dependiendo de las particularidades de su discapacidad, la persona puede tener problemas en el uso del teclado, del ratón o en ambos, y es necesario buscar una solución alternativa para sustituirlos, manteniendo su funcionalidad. En [Cand97] se presentan las ayudas técnicas más comunes que pueden utilizar las personas con diversas discapacidades para el acceso al ordenador y en [Basi98] se ofrece una perspectiva más amplia, tratando desde las ayudas en sí (especialmente en su capítulo 5 [Suar98]) hasta las estrategias de intervención necesarias para su uso en un contexto social. En los casos extremos, el usuario puede que sólo sea capaz de controlar con precisión una parte de su cuerpo, como puede ser un brazo, un dedo, una pierna, la lengua, un párpado o, incluso, su respiración. El problema a resolver es la transformación de ese movimiento controlado en una orden comprensible por el ordenador.

La solución habitual para estos casos extremos es instalar en el ordenador un pulsador (por ejemplo, conectándolo a alguno de sus puertos) que el usuario pueda controlar con su movimiento útil (cabeza, brazo, pierna, soplidos, etc.), y que envía al ordenador información cuando es pulsado. Para transformar esa información en una orden, se instalan programas específicos: por ejemplo, un editor de texto utilizable con pulsador tiene, además de los menús habituales de tratamiento de texto, ficheros, etc., una matriz con las letras que se barre automáticamente: un cursor se desplaza, resaltando secuencialmente las letras, de forma que el usuario solamente necesita presionar el pulsador cuando la letra que desea escribir es resaltada. Así, el usuario puede escribir sin necesidad de utilizar el teclado convencional. De la misma forma (barrido automático más pulsador) se accede a todos los demás menús del programa y de cualquier otra aplicación de este tipo.

Figura 1.1: Pulsador adaptado a la barbilla

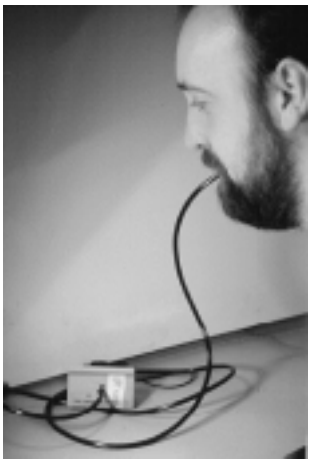


Figura 1.2: Pulsador neumático

El problema obvio que tiene este modo de acceso es la velocidad: el barrido automático tiene que ser lo suficientemente lento como para poder ser controlado por la persona, que incluso puede tener temblores. El tiempo que debe permanecer resaltada cada opción para

asegurar que el usuario sea capaz de pulsar y elegirla, suele ser superior al medio segundo, es decir, para escribir cada letra, es necesario esperar al menos medio segundo por cada una de las opciones que la precedan en el menú. Aunque hay diversos sistemas para intentar acelerar este acceso (por ejemplo, barrido por filas y columnas), una velocidad de escritura de cuatro o cinco palabras por minuto suele ser habitual en estos casos, según hemos encontrado en nuestros estudios [Pala96].

Por supuesto, no se puede comparar esta velocidad con las 200-400 pulsaciones por minuto (40-80 palabras por minuto) que puede alcanzar una persona experimentada que trabaja con ordenador, por lo cual, es necesario invertir todo el esfuerzo posible en intentar acelerarlo, especialmente porque hay un colectivo de personas para quienes el pulsador es su único modo de acceso al ordenador.

El objetivo de esta tesis es estudiar y desarrollar mecanismos para aumentar la velocidad de escritura con el ordenador, y reducir el esfuerzo necesario en dicho proceso para aquellas personas que sólo pueden utilizar un pulsador como medio de acceso.

Para conseguir este objetivo se hace uso de la predicción de palabras: el propio programa usado para escribir procesa el texto e intenta predecir cuál es la palabra que el usuario está intentando introducir en ese momento. Las candidatas predichas son mostradas y barridas junto con las letras. Si alguna de ellas es la adecuada, puede ser seleccionada e insertada con el mismo método usado para escribir una sola letra, ahorrando tiempo y esfuerzo. El ahorro depende de la calidad de los algoritmos de predicción y de la disposición de la interfaz de usuario.

Una predicción perfecta es aquella que siempre muestra al usuario la palabra que desea escribir en la lista de candidatas, en cuyo caso sólo es necesario seleccionar las palabras a medida que van apareciendo en el menú. Este caso ideal está muy lejos de la situación actual, aunque ahora pueden conseguirse velocidades de 8-10 palabras por minuto, lo que supone aproximadamente duplicar la velocidad con respecto a cuando no se usa predicción de palabras.

Además de lo visto, la predicción de palabras no es útil solamente para las personas con discapacidad física. Puede ser utilizada también como ayuda para personas que tengan problemas para escribir correctamente: como apoyo para el aprendizaje de la escritura o para reducir el número de faltas de ortografía, por algún tipo de discapacidad lingüística, dislexia, etc. También puede ser utilizada por estudiantes de una segunda lengua, aumentando la calidad del texto generado, la riqueza de vocabulario e, incluso, la cantidad, por la confianza que produce en el usuario el apoyo gramatical y ortográfico, como se explica en [Carl97a] y [Magn98]. Si se añade conocimiento lingüístico a la predicción de palabras, puede ayudar a personas con problemas para formar frases gramaticalmente correctas.

La predicción de palabras está enmarcada en los estudios sobre *Comunicación Alternativa y Aumentativa (Alternative and Augmentative Communication, AAC)*[1], encargada de todas las disciplinas destinadas a permitir la comunicación de personas con problemas: desde los sistemas más sencillos, utilizando gestos o lápiz y papel para comunicarse más fácilmente, hasta adaptaciones en la interfaz de los programas o el uso de sofistica-

[1] *Comunicación alternativa* es el conjunto de modos de comunicación (tales como lenguaje de signos, gestos y ayudas a la comunicación) utilizados para *reemplazar* el lenguaje oral. *Comunicación aumentativa* es el conjunto de modos de comunicación utilizados para *complementar* el lenguaje oral.

das técnicas de procesamiento de lenguaje natural para intentar predecir lo que se está intentando escribir. En [Baum90], [Chur92], [Dema89], [Dema92], [Dema94], [Lang99b] y [McCo95] se muestran investigaciones sobre la aplicación de diversas disciplinas a la comunicación alternativa y aumentativa.

En la actualidad, como se verá posteriormente, están disponibles varios sistemas de ayuda a personas con discapacidad que usan predicción de palabras. La mayoría están diseñados para el idioma inglés, aunque algunos utilizan técnicas automáticas de entrenamiento para predecir en cualquier idioma. La mayor parte de los sistemas de predicción disponibles en nuestro idioma han sido desarrollados en el Laboratorio de Tecnologías de Rehabilitación (LTR), en el cual se ha realizado esta tesis doctoral.

Capítulo 2

Encuadre científico-tecnológico

2.1 Introducción

La predicción de palabras es uno de los métodos más utilizados para apoyar la escritura de personas con diversos problemas. Sirve como ayuda para acelerar la escritura de las personas con discapacidades físicas y aquellas con problemas lingüísticos la utilizan para aumentar la calidad de sus textos, disminuyendo las faltas de ortografía y los errores gramaticales como se indica en [PAL00]. Los beneficios que se produzcan en ambos casos dependen de la calidad de la predicción de palabras, por lo cual es necesario utilizar los métodos más sofisticados a nuestro alcance, considerando la mayor cantidad posible de información a la hora de hacer las propuestas.

Además del ámbito de la discapacidad, la predicción de palabras puede utilizarse en otras áreas, como, por ejemplo, el reconocimiento de voz (para restringir la lista de palabras candidatas en un punto determinado del proceso de búsqueda), con lo que estas disciplinas afines pueden también beneficiarse de los avances realizados.

En este apartado se presenta el estado de la cuestión de diversos temas relacionados con la predicción de palabras, como base para abordar el trabajo de esta tesis. La bibliografía específica sobre la predicción de palabras no es tan extensa como la de otras áreas de investigación en procesamiento de lenguaje natural, con lo que nos apoyaremos en los estudios sobre disciplinas y materias relacionadas, cuyo impacto en la predicción es notable.

Así, en la primera parte de este capítulo se describen una serie de modelos de lenguaje que pueden ser empleados en la predicción de palabras, desde los métodos más sencillos, basados en estadísticas sobre secuencias de palabras, hasta los más potentes que van introduciendo información gramatical con diversos grados de complejidad.

Posteriormente se muestran las consideraciones a tener en cuenta en la etapa de evaluación, uno de los aspectos fundamentales para validar las propuestas realizadas, describiendo por último, un conjunto de sistemas disponibles en el mercado, de ayuda a personas con discapacidad lingüística o física, que integran la predicción de palabras como apoyo a la escritura.

El análisis de todo ello (fundamentalmente en lo que se refiere a las carencias observa-

das) y de la problemática concreta de la predicción de palabras en castellano (en lo que se refiere a exigencias de modelado para reflejar correctamente los comportamientos observados), es lo que fue conformando finalmente la evolución que aplicaremos en el estudio objeto de esta tesis y que describiremos detalladamente en el capítulo 3.

2.2 Modelos de lenguaje

La predicción de palabras se basa en un adecuado modelado del idioma, para lo que se pueden utilizar diversas técnicas a las que se dedican los próximos apartados.

Los modelos de lenguaje son formalismos que permiten expresar las relaciones entre las palabras. A continuación, se presentan varios modelos que se pueden utilizar para predecir, basados en diferentes fuentes de información, desde información personal del estilo del usuario, hasta probabilística y gramatical (con distintos grados de complejidad). Se puede encontrar más información en [Alle94], [Bunt96] y [Krul91], entre otros.

En [Bert95], [Boek96], [Gara94a], [Gara94b], [Hunn85], [Hunn86], [Hunn87] y [Vand91] se explican otros modelos utilizados para predecir. En algunos casos, como en [Hunn89], se ha intentado incluir también información semántica, pero para un vocabulario muy reducido, ya que es un método de predicción que precisa gran cantidad de información.

2.2.1 Modelos de lenguaje estadísticos

Son los más sencillos conceptualmente y consideran como única información la frecuencia absoluta y/o relativa de las palabras y las secuencias de palabras.

En su versión más simple, *modelo unigram*[1], sólo se considera la frecuencia absoluta de las palabras. Cuando se utiliza este modelo, en cada momento se predicen las palabras más frecuentes que comiencen por las letras escritas de la palabra en curso, sin aplicar ningún filtrado adicional. Esta información es claramente insuficiente, ya que ignora las restricciones impuestas por las palabras previamente insertadas en el texto.

Hay varias maneras de incluir la información aportada por las últimas palabras en el proceso de predicción. La más sencilla, como continuación natural del sistema probabilístico anterior, es realizar estadísticas sobre secuencias de palabras: en lugar de considerar la frecuencia absoluta de las palabras, se utiliza para predecir la probabilidad de que cada palabra siga a la anterior o las anteriores: *modelo bigram, trigram, ..., n-gram.* En estos casos, a la hora de predecir, se muestran las palabras que más frecuentemente han seguido a la o las últimas palabras escritas.

La información se introduce en el modelo mediante un proceso de entrenamiento previo: dado un corpus de texto, se contabiliza el número de veces que aparece cada secuencia de palabras. Para que el entrenamiento sea de calidad, el corpus utilizado debe ser lo suficientemente grande como para permitir obtener un número adecuado de ocurrencias

[1] En este documento se usará terminología inglesa en aquellos casos comúnmente aceptados en el entorno científico de habla castellana y en los que no se encuentre una mejor alternativa en nuestro idioma. Igualmente se ofrecerá la versión inglesa de palabras en castellano cuando se considere necesario.

de cada secuencia válida de palabras, de modo que las estimaciones estadísticas sean fiables. Debido a la cantidad de palabras diferentes que pueden aparecer en el texto de entrenamiento y las combinaciones que pueden presentar, es necesario utilizar gran cantidad de texto, especialmente si se desea modelar lenguaje natural escrito de la forma más completa posible. Esto, cuando el número de palabras de la secuencia considerada sea grande, exige gran cantidad de texto, que puede no estar disponible, en cuyo caso puede que no se obtengan buenos resultados.

Para solucionar este problema de escasez de datos de entrenamiento hay dos posibles soluciones (que, además, pueden combinarse):

- Uso de técnicas de *suavizado*, capaces de mejorar los resultados de matrices de probabilidades entrenadas con datos insuficientes, como se puede ver en [Ney91] y [Mart99].
- Agrupación de las palabras en *categorías gramaticales*, como se explica en la siguiente sección.

2.2.2 Modelos categoriales

En lugar de considerar estadísticas de palabras y de sus secuencias, como en el modelo anteriormente explicado, en este caso las palabras se agrupan en *categorías* (conjuntos de palabras con un comportamiento lingüístico homogéneo), y se realizan estadísticas sobre las secuencias de categorías. Se contabiliza el número de veces que aparece cada par y trío de categorías: *bipos* y *tripos* (del ingles, POS = *Part Of Speech*).

En este caso, es necesario resolver un nuevo problema, la clasificación de las palabras en categorías. El conjunto de categorías y el léxico incluido en cada una de ellas puede no coincidir necesariamente con las categorías tradicionales que se encuentran en los manuales de lingüística o en otros trabajos.

La selección de las categorías y la clasificación de palabras según dichas categorías puede realizarse:

1. *Por procedimientos automáticos* (técnicas de agrupamiento *(clustering)* automático), que son fiables a grandes rasgos, aunque no proporcionan información detallada. En [Jeli91] puede encontrarse un método de clasificación basado en la información mutua, por ejemplo.
2. *De forma manual:* se clasifica, manualmente, cada palabra en una o varias categorías, lo cual supone un gran esfuerzo que debe ser realizado por un experto.

Una vez decidida la clasificación, se incluye esa información en el sistema, creando un *diccionario*[2]. El diccionario es una de las fuentes de información más importantes del sistema, puesto que es el nexo de unión entre el texto del usuario y cualquier modelo de

[2]En el ámbito del tratamiento del lenguaje natural, se llama *diccionario* a la estructura (en formato electrónico) que contiene las palabras y toda la información asociada a cada una de ellas, como por ejemplo, su frecuencia o las categorías a las que pertenece.

lenguaje (en formato electrónico). Debe contener el mayor número posible de entradas, ya que las palabras que no aparecen en el diccionario son consideradas desconocidas, e interrumpen el correcto funcionamiento del modelo de lenguaje si el usuario las incluye en el texto.

Sin embargo, por muy completo que sea el diccionario, nunca contendrá todas las palabras que el usuario pueda utilizar, y será necesario diseñar mecanismos de recuperación frente a este problema.

Además, es necesario *etiquetar/categorizar* el texto de entrenamiento (asignar a cada una de sus palabras la categoría correspondiente) y *entrenar* el sistema, contabilizando la probabilidad de que cada categoría siga/preceda a las demás.

El entrenamiento se realiza de forma similar al de los modelos estadísticos de palabras, pero contabilizando las secuencias de categorías gramaticales en lugar de las de palabras. Como el número de categorías es muy inferior al de palabras, se necesita mucho menos texto para conseguir un número de suficiente ocurrencias de cada secuencia de categorías. Esto reduce los problemas de escasez de texto de entrenamiento, aunque este texto es más difícil de conseguir porque debe estar categorizado, y este proceso requiere cierta intervención manual, al no existir métodos automáticos que garanticen la perfección en el etiquetado.

Una vez entrenado el sistema, se predice utilizando las estadísticas sobre las secuencias de las categorías. Se categorizan las *n-1* últimas palabras escritas y se predicen palabras pertenecientes a las categorías que más probabilidad tengan de seguir a las categorías de las palabras anteriores.

Las técnicas vistas hasta ahora tienen el inconveniente de basarse solamente en información de la palabra anterior o, como mucho, de las dos o tres palabras anteriores, ya que para considerar una historia mayor se necesita una enorme cantidad de datos de entrenamiento y de recursos informáticos para tratarla. Esta información es insuficiente, ya que muchas veces la distancia entre palabras relacionadas es mayor.

Por esta razón se buscan métodos más complejos, que predigan considerando información de toda la frase, desde su comienzo, teniendo en cuenta relaciones de larga distancia entre las palabras: por ejemplo, permitiendo que un nombre prediga un verbo seis palabras después, independientemente de las palabras que aparezcan en medio. En las siguientes secciones se describen diversos formalismos que permiten expresar estas relaciones de dependencia.

2.2.3 Modelos formales

2.2.3.1 Gramáticas formales

A diferencia de los modelos estadísticos vistos hasta ahora, las gramáticas formales son mecanismos para describir los lenguajes mediante conjuntos de reglas. Una gramática formal está compuesta por:

1. *Símbolos terminales:* son las unidades básicas de la gramática. Dependiendo del ca-

so, pueden ser las palabras que componen el léxico (p. ej.: *azul*, *mesa*) o las categorías gramaticales (p. ej.: *artículo*, *nombre*).

2. *Símbolos no terminales:* son unidades superiores (constructores) resultantes de aplicar una regla determinada. P. ej.: SN (sintagma nominal) o SPrep (sintagma preposicional).

3. *Símbolo de comienzo:* es el símbolo no terminal a partir del cual se empiezan a generar frases (el axioma de partida). P. ej.: O (oración) o S (*sentence*).

4. *Conjunto de reglas de producción o reescritura:* permiten expresar las secuencias de símbolos (terminales y no terminales) válidas según nuestra gramática. P. ej.: SN → art n. Esta regla significa que un sintagma nominal (SN) está formado por un artículo (art) seguido por un nombre (n).

   Las cadenas válidas del lenguaje se generan aplicando las reglas de producción a partir del símbolo inicial. El *lenguaje* generado por la gramática es el conjunto de todas las secuencias generadas aplicando las reglas (todas las producciones válidas).

Existen varios tipos de lenguajes basados en reglas de reescritura. La diferencia entre ellos es su capacidad generativa o potencia expresiva. En las siguientes secciones se describen las distintas gramáticas que componen la *jerarquía de Chomsky*, que es una clasificación de gramáticas en la que cada una es capaz de modelar los mismos casos que la anterior y alguno más (es decir, el lenguaje generado por cada una está incluido en el generado por la siguiente).

2.2.3.1.1 Gramáticas regulares

Son aquellas gramáticas cuyas reglas de reescritura son de la forma:

Símbolo no terminal → Símb. terminal [Símb. no terminal]

Es decir, cada símbolo no terminal se expande como un símbolo terminal seguido, opcionalmente, por un símbolo no terminal.

P. ej.:

SN1 → art SN2
SN2 → adj SN2
SN2 → n

que genera todas las cadenas de la forma: *art adj adj ... adj n*.

2.2.3.1.2 Gramáticas independientes del contexto

Las gramáticas independientes del contexto *(Context Free Grammars, CFG)* son aquellas gramáticas cuyas reglas de reescritura son de la forma:

Símbolo no terminal → secuencia de símbolos terminales y no terminales

Es decir, los símbolos no terminales se descomponen de la misma forma siempre, independientemente del contexto, y están compuestos por cualquier secuencia de símbolos terminales y no terminales. Son más flexibles que las anteriores y, por tanto, el conjunto de casos que son capaces de modelar es más amplio.

P. ej.:

O → SN SPred
SN → art n
SPred → verbo SN

Con la gramática del ejemplo se generan frases básicas de tipo: "art n verbo art n".

Estas gramáticas son muy importantes, porque, como se indica en [Alle94], son lo suficientemente:

- *Potentes* para describir parte de las estructuras del lenguaje natural (restringido).
- *Simples* como para hacer analizadores eficientes.

2.2.3.1.3 Gramáticas dependientes del contexto

Como contraposición a las anteriores, son aquellas en las que un símbolo no terminal se puede reescribir de forma distinta dependiendo del contexto.

P. ej.:

a X b → a Y b
c X d → c Z d

En este caso, el símbolo no terminal X se reescribe como secuencia diferente de categorías cuando está en el contexto "a..b" que cuando está en el contexto "c..d".

2.2.3.1.4 Gramáticas de tipo 0

Son aquellas que permiten reglas de reescritura arbitrarias.

Aunque todos los tipos de gramáticas formales vistos son capaces de generar *secuencias de categorías* válidas según su conjunto de reglas, es necesario ampliar su potencia para que consideren fenómenos adicionales, muy comunes en el lenguaje natural, que pueden suponer un gran beneficio a la hora de predecir:

1. *Gestión de rasgos:* las palabras no sólo pueden influir en la categoría de la palabra que se escribe a continuación, sino también en otras características, como el género y/o el número. Por ejemplo, tras la palabra *nosotros* es muy frecuente que se escriba un verbo, en cuyo caso es muy probable que sea en primera persona del plural, exigiendo concordancia de número y persona.

2. *Ambigüedad:* es común que una frase pueda ser analizada de varias maneras diferentes, bien porque las palabras que la componen puedan tener diferentes categorías gramaticales, o bien porque puedan agruparse de distintas formas. La calidad de la predicción de palabras es mejor si hay información sobre cuál es la interpretación correcta de la frase.

A continuación, se muestran algunas posibles modificaciones que se podrían realizar en las gramáticas para soportar estas características.

2.2.3.1.5 Gramáticas aumentadas

Las gramáticas mostradas en los apartados anteriores expresan el lenguaje como secuencia de categorías (de forma más o menos compleja), pero hay una parte muy importante que no consideran: la *gestión de los rasgos*, p. ej., *concordancia de género*, *número*, *persona*, *qué tipo de sintagma preposicional debe seguir a un verbo determinado*, etc.

Al incluir la gestión de rasgos, se restringen las cadenas consideradas gramaticalmente correctas: deben cumplir nuevas condiciones, como, por ejemplo, varias palabras deben concordar entre sí. Esto facilita el análisis del lenguaje natural, ya que reduce el grado de *sobregeneración*, es decir, el conjunto de cadenas consideradas gramaticalmente correctas por un analizador, pero que realmente no lo son.

Para poder considerar estas restricciones es necesario ampliar la cantidad de información del sistema: a cada palabra no solamente se le asignan la o las categorías gramaticales correspondientes, sino que se le debe asociar su *conjunto o estructura de rasgos*. Así, la palabra *casa* tiene 3 entradas en el diccionario, una como nombre femenino singular, otra como verbo en presente de indicativo, tercera persona del singular y otra como verbo en imperativo, segunda persona del singular.

Las reglas también se deben modificar para incluir las restricciones, por ejemplo:

$$\text{SN (gen g num n)} \rightarrow \text{art (gen g num n) nomb (gen g num n)}$$

Mientras que anteriormente, sin las restricciones de rasgos, se aceptaban como secuencias válidas aquellas formadas por cualquier artículo seguido por cualquier nombre (por ejemplo, *las niño*), ahora se exige que concuerden en género y número, por lo cual, a la palabra *las* sólo le pueden seguir nombres femeninos plurales.

La gestión de rasgos puede ser de gran ayuda en la predicción de palabras, ya que restringe mucho más las palabras válidas, aunque exige una mayor cantidad de información, tanto en las reglas como en el diccionario, y una mayor complejidad en el procesamiento. La gestión de rasgos puede añadirse a cualquiera de las gramáticas vistas anteriormente.

2.2.3.1.6 Gramáticas probabilísticas

Es común que un analizador encuentre varias posibles maneras de derivar un determinado fragmento de texto dado un conjunto de reglas. Aunque la interpretación válida esté

perfectamente clara para quien escribe, el analizador no dispone de información suficiente para deshacer la *ambigüedad*.

P. ej.: ser {nombre, verbo} humano {adjetivo, nombre}

Para elegir una de las interpretaciones, es necesaria información adicional que permita seleccionar cuál de ellas es la correcta. Una forma de hacerlo es añadir *probabilidades* a las reglas, indicando la prioridad que se debe dar a cada una ante un conflicto: de esta forma es posible tener una idea de cuál es la interpretación más probable (esto no garantiza la corrección de la solución, sino que asegura el acierto en un porcentaje mayor de ocasiones). En [Pala99a] y [Pala98a] mostramos la formulación básica para el tratamiento probabilístico a aplicar. Conviene mantener todas las interpretaciones durante todo el análisis, hasta que, en algún punto de la frase, la categoría de la palabra escrita invalide alguna de ellas.

Se han mostrado diversos tipos de gramáticas formales y las modificaciones que se pueden realizar para que soporten fenómenos adicionales. Este tipo de gramáticas se puede utilizar en la predicción de palabras si está disponible un modelo adecuado (suficientemente completo) del castellano. Para ello se averigua qué reglas está utilizando el usuario en cada momento y se predicen palabras que cumplan dichas reglas.

Los analizadores son los sistemas que averiguan las reglas que cumple una frase determinada, por lo cual, para predecir se necesitará un analizador que sea capaz de procesar la parte escrita de la frase en curso y predecir palabras correctas según la gramática disponible. A continuación se explican diversos analizadores y su posible aplicación para la predicción de palabras.

2.2.3.2 Analizadores

Los analizadores (*parsers*) son sistemas encargados de comprobar si una oración es correcta con respecto a una determinada gramática y de entregar información sobre su estructura. Los analizadores correspondientes a cada tipo de gramática (tipo 0, independiente del contexto, etc.) son diferentes, tanto más complejos cuanta mayor sea la flexibilidad de la gramática. Además, es posible construir diferentes analizadores para cada tipo de gramática, cada uno utilizando una estrategia de análisis distinta:

1. *Analizador de estrategia descendente (top-down parser)*: parte del símbolo inicial de la gramática *(S: sentence)* e intenta reescribirlo aplicando las reglas gramaticales hasta encontrar una combinación de reglas que se corresponda con la estructura de la entrada. Dependiendo del orden en el que se realiza la reescritura de cada símbolo, es posible seguir dos estrategias:

   (a) *búsqueda en profundidad (depth first strategy)*: es aquella que intenta reescribir cada símbolo no terminal hasta que llega a los símbolos terminales antes de seguir con otro símbolo,

   (b) *búsqueda en anchura (breadth first strategy)*: es aquella en la que se van reescribiendo todos simultáneamente.

   Ambas aproximaciones encuentran la solución, pero varía el orden de búsqueda.

2. *Analizador de estrategia ascendente (bottom-up parser)*: parte de la secuencia de símbolos que forma la frase y asocia sus valores a la parte derecha de las reglas, para obtener la parte izquierda de aquellas reglas que se cumplan. Este proceso se aplica recursivamente hasta llegar al símbolo inicial (S u O).

Otra posible clasificación, dependiendo del orden en el que se analizan los datos de entrada, sería:

1. *De izquierda a derecha:* se analiza la frase comenzando por los símbolos de la parte izquierda, avanzando hacia la derecha.
2. *De derecha a izquierda:* se analiza en orden inverso a la anterior.

En las siguientes secciones se explican diferentes ejemplos de analizadores para algunas de las gramáticas vistas previamente.

2.2.3.2.1 Autómatas

Con respecto a los autómatas o máquinas de estados, hay discrepancia en la bibliografía sobre si considerarlos gramáticas por sí mismos (por ejemplo, en [Alle94]) o analizadores de las gramáticas formales (como se indica en [Alfo90]). Existen diferentes tipos de autómatas, con potencias expresivas diferentes, por lo cual, si se decide considerarlos analizadores, basta elegir para cada gramática la máquina de estados con potencia expresiva equivalente: por ejemplo, para las gramáticas regulares se utiliza un autómata finito y para gramáticas de tipo 0 se puede utilizar una Máquina de Turing. A continuación, se muestran varios tipos de autómatas y, además, cómo se pueden ampliar para aumentar su potencia y soportar los fenómenos vistos anteriormente, como la concordancia.

Las *máquinas de estados finitos* consisten en un conjunto de *nodos o estados* unidos por *arcos* etiquetados. Se comienzan a recorrer por el *estado inicial* y, a partir de él, se avanza por el camino indicado por los arcos: si en un estado determinado la entrada coincide con la etiqueta de alguno de los arcos de salida del nodo, se avanza hacia el nodo apuntado por dicho arco y así hasta el último. Si en un estado determinado la categoría no coincide con la etiqueta de ninguno de sus arcos de salida se interrumpe el recorrido.

Se pueden aplicar a la predicción de palabras etiquetando los arcos con categorías gramaticales. Las entradas al sistema son las categorías de las palabras ya escritas y las categorías a predecir son las etiquetas de los arcos de salida del nodo en el que se encuentre la máquina, ya que son las transiciones válidas. En este caso (etiquetando los arcos con categorías), las máquinas de estados finitos tienen la misma potencia expresiva que las gramáticas regulares.

Con la red de la figura 2.1 se modelan todos los sintagmas nominales que consisten en un artículo seguido por un número indeterminado de adjetivos (o ninguno), más un nombre, más uno o más adjetivos.

Para conseguir la potencia expresiva de las gramáticas independientes del contexto, es necesario incluir la noción de *recursividad*. Las *redes de transiciones recursivas (recursive*

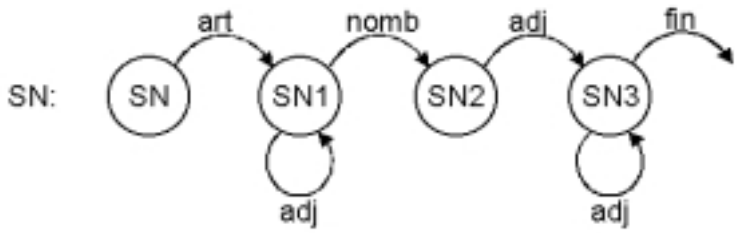


Figura 2.1: Ejemplo de máquina de estados finitos.

transition networks, RTN) son máquinas de estados en las cuales los arcos de la red pueden estar etiquetados, no sólo con categorías, sino con saltos a otras redes o, incluso, a ella misma.

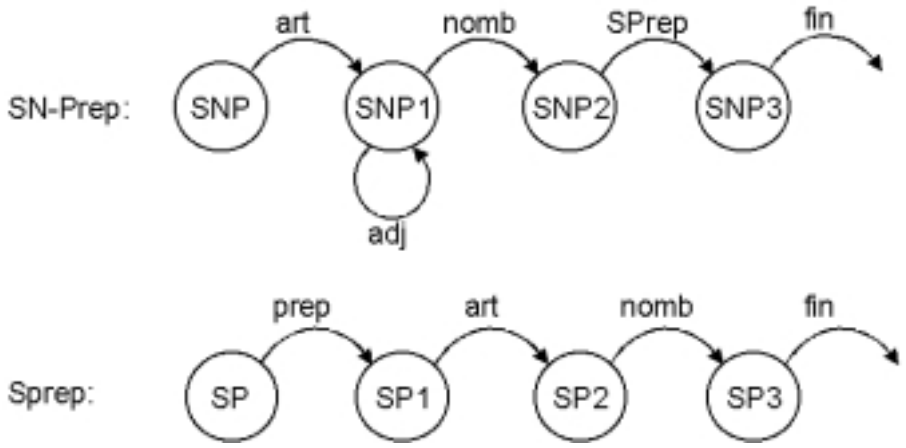


Figura 2.2: Ejemplo de red de transiciones recursivas.

Con la red de la figura 2.2 se modelan todos los sintagmas nominales que consisten en un artículo seguido por un número indeterminado de adjetivos (o ninguno), más un nombre, más un sintagma preposicional, Sprep, que es el símbolo no terminal modelado por la red de la parte inferior de la figura.

Las *redes de transiciones aumentadas (augmented transition networks, ATN)*, son redes de transiciones recursivas, en las cuales, además, se tienen en cuenta las restricciones de *concordancia*, por lo cual tienen la misma potencia expresiva que las gramáticas independientes del contexto aumentadas. En ese caso, para cambiar de estado no sólo se considera la categoría de cada palabra, sino las restricciones de concordancia impuestas a las estructuras de rasgos que acompañan a cada categoría.

El sintagma nominal del ejemplo de la figura 2.3 es muy sencillo, un artículo seguido por un nombre, aunque se han añadido restricciones de concordancia, de forma que, para avanzar por el arco 2, es necesario que los rasgos del nombre concuerden con los del artículo del arco 1. Esto asegura, por ejemplo, que a un artículo masculino singular (*el*) le sigue un nombre masculino singular. Para obtener más información véanse [Bate78] y [Wood69]. En [Vand91] se puede encontrar una aproximación al uso de ATN para predicción de palabras.

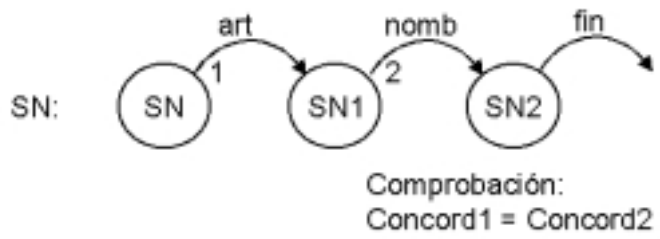


Figura 2.3: Ejemplo de red de transiciones aumentadas.

2.2.3.2.2 Analizador de Earley

Un analizador de Earley es un analizador de tipo chart (*chart parser*) para gramáticas independientes del contexto que realiza un análisis de arriba a abajo. Una descripción detallada del algoritmo se puede encontrar en [Earl70].

Por el procedimiento que utiliza, recorre la entrada una sola vez, optimizando el análisis y, además, es capaz de mantener en paralelo todas las posibles interpretaciones del fragmento de texto analizado, eliminando caminos solamente cuando aparece una categoría gramatical no esperada por la regla.

En [Gara94a] se explica la aplicación de un analizador de Earley a la predicción de palabras, utilizando una gramática independiente del contexto con información probabilística.

2.2.3.3 Tipología de errores en el análisis del lenguaje natural

Uno de los problemas a resolver en un analizador es el procedimiento a seguir cuando falla el análisis, es decir, en las situaciones en las cuales la frase escrita no sigue la gramática.

Estas situaciones anulan al analizador y, si no se implementan mecanismos alternativos, interrumpen el análisis hasta que comience la siguiente oración (en el siguiente punto), degradando en gran medida el rendimiento del sistema total. Por esto es muy importante incluir mecanismos de recuperación de errores en el algoritmo de análisis. La tipología de errores a tratar es la siguiente:

1. Inserción de *palabras desconocidas* en el texto: cuando en el texto a analizar hay una palabra que no está incluida en el diccionario (por ser muy poco frecuente, muy técnica, un neologismo, una falta de ortografía etc.). El analizador no puede averiguar a qué categoría pertenece y el análisis se detiene.

2. La palabra es conocida, pero ninguna de sus categorías gramaticales es esperada por ninguna regla, bien por falta de *cobertura* de la gramática (por muy completa que sea la gramática es muy difícil cubrir el 100% de los casos) o porque la frase *no* sea *gramaticalmente correcta*. En [Kong95] se puede encontrar la siguiente clasificación de errores:

   (a) *Errores de sustitución*: una de las palabras de la frase tiene una categoría diferente a la esperada, pero todas las demás coinciden con la secuencia de la

regla.

(b) *Errores de inserción*: en un punto de la frase se intercala una palabra, pero el resto de la secuencia de categorías sigue la regla.

(c) *Errores de borrado*: se elimina una palabra de la secuencia válida.

2.2.4 Gramáticas de unificación

Es una generalización del empleo de los rasgos, como se explica en [Cald88] y [Cald87]. En las gramáticas explicadas en las secciones anteriores, los rasgos sólo se emplean en las redes de transición aumentadas y en las gramáticas formales aumentadas y restringen los conjuntos o estructuras de rasgos que acompañan a la categoría gramatical. Sin embargo, el uso de los rasgos puede ser generalizado hasta eliminar el resto de la estructura formal y considerar la gramática entera como un conjunto de restricciones entre estructuras de rasgos.

Como se explica en [Alle94], las gramáticas de unificación se basan en las siguientes relaciones entre estructuras de rasgos:

1. *Extensión:* se dice que la estructura de rasgos F1 extiende o es más restrictiva que otra estructura de rasgos F2 si todos los valores de rasgos de F1 están especificados (y tienen el mismo valor) en F2. P. ej.: F1 = (*Categoría: verbo*, *Lema: llorar*) extiende la estructura de rasgos F2 = (*Categoría: verbo*) ya que F1 incluye el rasgo *Categoría: verbo* y añade *Lema: llorar*, siendo más restrictiva que F2.

2. *Unificación:* dos estructuras de rasgos se pueden unificar, si hay otra estructura que sea extensión de ambas (es decir, si no tienen rasgos contradictorios). P. ej. no se pueden unificar F1 = (*Persona: 3*, *Número: singular*) F2 = (*Persona: 3*, *Número: plural*) ya que contienen los rasgos *Número: singular* y *Número: plural*, que no son unificables.

   El *unificador más general* es la estructura de rasgos mínima que sea extensión de 2 estructuras de rasgos.

Si las operaciones utilizadas para gestionar la concordancia se amplían, extendiéndose para controlar todas las relaciones de la gramática, se obtiene una gramática de unificación.

P. ej.: según una gramática formal, una posible regla que modela la exigencia de un sintagma nominal por parte de los verbos transitivos es: SPred $\rightarrow$ verbo nombre. Una gramática de unificación expresa esta relación con rasgos, para lo cual es necesario asignar al verbo un rasgo más, que es "necesito un objeto directo" y a los nombres el rasgo "puedo ser objeto directo", que concuerda con el anterior. A partir de este momento todo se controla por concordancias, haciendo innecesaria la regla.

El problema fundamental para utilizar esta gramática para predecir palabras es gestionar el orden, ya que sería necesario codificar esa información mediante rasgos, porque a priori no está controlada por la gramática.

2.2.5 Gramáticas basadas en marcos de subcategorización

Los marcos de subcategorización son las distintas estructuras sintácticas que un verbo admite: qué complementos acepta, formas impersonales, formas reflexivas, combinaciones de sintagmas nominales y/o preposicionales, etc.

Los complementos asociados al verbo son *argumentos*. Existen, además, *adjuntos*, que son los complementos adverbiales, que se pueden aplicar a todos los verbos, porque sólo modifican las circunstancias. Conociendo exactamente los argumentos que puede tomar cada verbo, se puede predecir con mucha mayor precisión, aunque puede no resultar muy restrictivo si no proporciona información sobre orden o si algún complemento aparece antes que el propio verbo, por ejemplo.

2.2.6 Gramáticas asociativas por la izquierda

Analizan la frase de izquierda a derecha, como se explica en [Haus87]. La gramática está compuesta por un conjunto finito de reglas de la forma:

$r_i : [C_1\ C_2] \rightarrow [rp_i\ C_3]$

donde:

$\mathbf{r}_i$ = regla número i.

$\mathbf{C}_1$ = principio de cadena (la parte analizada de la cadena en curso).

$\mathbf{C}_2$ = nuevo elemento (el siguiente elemento que se debe analizar).

$\mathbf{rp}_i$ = paquete de reglas que hay que aplicar en el siguiente paso.

$\mathbf{C}_3$ = nuevo principio de cadena (se obtiene de forma recursiva a partir de C_1 y C_2).

Es decir, una gramática asociativa por la izquierda utiliza un conjunto finito de reglas que se aplican a un principio de cadena y a un nuevo elemento, para obtener el nuevo principio de cadena y el nuevo conjunto de reglas que se deben aplicar. El flujo de control/orden lo forman las reglas. El proceso sigue hasta que se encuentre una regla que no pueda aplicarse porque no se cumplen los requisitos (falla el análisis: la frase no es gramaticalmente correcta) o cuando se llega a la regla final sin más elementos para procesar.

En [Gome94] se muestra el uso de las gramáticas asociativas por la izquierda para describir lenguajes naturales. Además, se indica que se ha escrito una gramática del castellano con la cual se han formalizado ciertos fenómenos lingüísticos, como son la valencia, la concordancia y el orden de las palabras.

Su posible uso en algoritmos de predicción de palabras se basa en el análisis de la parte escrita de la frase en curso, tomándolo como principio de la cadena. Los nuevos elementos se añaden al análisis a medida que se escriben y se predicen las palabras cuya categoría pertenece a las esperadas por cualquiera de las reglas que se pueden aplicar en ese punto.

2.3 Interfaces adaptativas orientadas a usuarios con discapacidad

A continuación, se muestran los criterios generales de diseño de interfaces de usuario, para pasar posteriormente a considerar las restricciones impuestas por las personas con discapacidades físicas y, por último, a la inclusión de mecanismos adaptativos en la interfaz.

2.3.1 Criterios generales de diseño de interfaces de usuario

En esta sección se muestran los criterios generales de diseño de las aplicaciones que son utilizadas con los dispositivos de entrada y salida convencionales: pantalla, teclado, impresora, etc. Aunque estos criterios son aplicados también en el diseño de las interfaces orientadas a personas con alguna discapacidad, éstas últimas requieren un diseño específico, debido a las restricciones adicionales impuestas por las necesidades del usuario.

Los criterios que se deben utilizar para mejorar la interfaz de entrada se basan en un buen diseño de los menús, ya que la práctica totalidad de los programas los emplean en algún momento. Un sistema de menús bien construido permite al usuario crear los modelos mentales adecuados sobre manejo del programa que le permitan el acceso fácil a las opciones de los menús, reduciendo la carga cognitiva necesaria para utilizar la aplicación: la memoria, el esfuerzo, etc.

Los tres aspectos fundamentales que se deben considerar en el diseño de los menús son la organización, el ahorro de recursos y la comunicación:

1. La correcta *organización* de los menús sirve para dar al usuario una estructura conceptual del programa consistente y clara. La *consistencia* debe darse tanto a nivel *interno* del programa (por ejemplo, las opciones que realicen acciones similares deben tener el mismo aspecto y estar situadas cerca), como a nivel *externo* (utilizando en los iconos imágenes fácilmente reconocibles según las convenciones culturales o los mismos que se utilizan en otros programas o incluso en el mundo real, como, por ejemplo, utilizar el icono de una impresora para indicar *imprimir*).

   La *disposición* de las opciones en los menús también debe hacerse de forma cuidadosa, en matrices horizontales, verticales o rectangulares, sin presentar al usuario demasiadas opciones en cada momento, puesto que necesitaría emplear demasiado tiempo en leer y analizar todas las alternativas hasta encontrar la adecuada.

   También es importante el criterio según el cual se decide la *agrupación* de opciones en los menús: se deben colocar en un mismo menú los elementos relacionados y separar los no relacionados. Por ejemplo, en un editor de texto, las opciones *cortar*, *copiar* o *pegar* deben estar en el mismo menú. Del mismo modo, al ser acciones de edición, pertenecen al menú *edición*.

2. Además de la organización de los menús, es importante dar toda la funcionalidad del programa con los *mínimos recursos posibles*. Son importantes conceptos como la *simplicidad* (incluir sólo los elementos importantes para la comunicación), la *claridad* (utilizar elementos con significado no ambiguo), la *distinción* entre las diferentes

propiedades de los elementos, el *énfasis* que se debe dar a los elementos más importantes o la *eliminación de redundancias*. La *jerarquía de los menús* cobra aquí una gran importancia, dividiendo la funcionalidad del programa en grupos y subgrupos, proporcionando una visión simple y clara al usuario final.

3. Otro factor importante es la facilidad con la que el usuario obtiene la información sobre la funcionalidad de un menú. Algunas de las estrategias utilizadas para la *comunicación* óptima de información son: uso de *formatos* fáciles de identificar y capaces de captar la atención del usuario y el uso moderado de diferentes *tipos y tamaños de letra* en las opciones (3 como máximo) y de combinaciones adecuadas de *colores*.

Estos criterios, que se aplican en las interfaces de usuario convencionales, deben tenerse especialmente en cuenta en las interfaces de los programas para personas con discapacidad física, ya que la interacción usuario-programa en este caso puede ser particularmente difícil, agravando las consecuencias de un mal diseño.

2.3.2 Diseño de interfaces de usuario orientadas a personas con discapacidad física

El diseño de estas interfaces está condicionado fundamentalmente por las necesidades de los usuarios, que tienen problemas, más o menos graves, con los *dispositivos físicos* convencionales (ratón, teclado, discos, etc.). Dependiendo del tipo y grado de la discapacidad física, puede que para la persona sea muy difícil o incluso imposible utilizar algunos de estos dispositivos o, incluso, todos. En [Cand97] se muestra un completo conjunto de ayudas para el acceso al ordenador de personas con diferentes discapacidades. En la norma [AENO98a] se pueden encontrar los requisitos que debe cumplir el soporte físico del sistema (botones, reguladores de sonido o luminosidad, periféricos, etc.) para que una persona con problemas físicos pueda utilizarlo, bien directamente o con la asistencia de alguna ayuda técnica.

Una vez seleccionado el dispositivo físico alternativo que resulte más adecuado para la persona, es necesario modificar la interfaz de la aplicación para permitir el acceso a toda su funcionalidad con los recursos que el usuario pueda manejar. Por ejemplo, si sólo es posible utilizar teclado, hay que asegurar que ninguna opción exige el uso de ratón. Al contrario, si no se puede utilizar el teclado, todo debe ser accesible con el ratón, lo cual implica, entre otras cosas, añadir al sistema un simulador de teclado en pantalla.

Los casos anteriormente mencionados son especialmente favorables (considerando la existencia de una discapacidad física), ya que permiten *acceso directo* a las opciones, con teclado o con ratón y con ellos pueden conseguirse velocidades medianamente rápidas.

Si la discapacidad es más severa, es posible que el usuario no pueda manejar ni el teclado ni el ratón, porque su movimiento controlado sea mínimo, como, por ejemplo, el movimiento de un dedo, la rodilla, la respiración o el cuello. En estos casos, el usuario solamente puede manejar un pulsador y necesita una interfaz que incluya barrido de las opciones para su acceso. La interacción con estas interfaces es especialmente lenta, por la propia naturaleza del barrido, y hace necesario un diseño de la interfaz muy cuidadoso

y el empleo de las técnicas más sofisticadas a nuestro alcance para intentar acelerarla: situación de las opciones de forma óptima, caracteres añadidos automáticamente en la escritura, diferentes modalidades de barrido, etc.

Además, una persona con discapacidad física puede tener otro tipo de problemas, auditivos o visuales, que hagan necesaria la modificación de la interfaz para adaptarse también a esta nueva condición. Por ejemplo, las personas con problemas graves de visión necesita realimentación auditiva de todo lo que se muestre en pantalla y las personas con problemas auditivos necesitan que se les muestren en pantalla todos los mensajes sonoros que emita el sistema. En la norma [AENO98b] se pueden encontrar los requisitos que deben cumplir las aplicaciones para poder ser utilizadas por personas con problemática muy diversa.

En la actualidad, la accesibilidad de las aplicaciones informáticas es un tema que está siendo ampliamente debatido, como lo demuestra la diversidad de recomendaciones que están surgiendo en las principales empresas de programación, véanse, por ejemplo, las recomendaciones de diseño accesible de Microsoft [Micr99], IBM [IBM00] o SUN [Berg00]. Las normas [AENO98a] y [AENO98b] son las primeras normas mundiales de accesibilidad a las plataformas informáticas, y han sido desarrolladas en el AEN/CTN139/SC8/GT1 de AENOR, del que forma parte la autora de esta tesis.

2.3.3 Interfaces adaptativas

Además de las soluciones planteadas para cada discapacidad, se pueden considerar en el diseño de la interfaz una serie de estrategias que también tienen como efecto acelerar y facilitar el acceso a la funcionalidad del programa, dando lugar a una interfaz adaptativa.

La *adaptabilidad* consiste en introducir un cierto grado de inteligencia en la interfaz que analiza las interacciones del usuario. Tras este análisis, se proponen modificaciones de la interfaz de forma que los accesos sean óptimos: que las opciones más utilizadas sean más cómoda y rápidamente accesibles y que el usuario necesite realizar un esfuerzo menor para llevar a cabo las acciones habituales. En [Lang97], [Lang99a] y [Suka92] se pueden encontrar las estrategias utilizadas para construir una interfaz adaptativa y los tipos de adaptaciones que pueden realizarse, como, por ejemplo:

- Modificación de la posición de las opciones en menús y cuadros de diálogo.
- Elaboración automática de macros que ejecutan las acciones más comunes del usuario.
- Agentes de ayuda que observan las actuaciones del usuario y le sugieren la forma óptima de realizar las tareas.

Los cambios en la interfaz modifican la interacción con la aplicación, lo que puede producir rechazo por parte del usuario. Es necesario introducir métodos que permitan al usuario controlar dichos cambios, pudiendo aceptar o rechazar las adaptaciones. Además, es previsible un tiempo de adaptación a la nueva interfaz, durante el cual su utilización puede no ser óptima y pueden cometerse un número de errores superior al normal. Superada esta fase, los estudios teóricos sobre el uso de la interfaz (si este se mantiene homogéneo)

muestran que mejora el rendimiento, tanto en velocidad como en número de interacciones necesarias. En [Lang98] se presentan los factores a considerar en la evaluación de las interfaces adaptativas y se plantea teóricamente cómo llevar a cabo dicha evaluación.

El concepto de interfaz adaptativa no está ligado a programas para personas con discapacidad, aunque puede resultar especialmente importante en estos casos, debido a la trascendencia de un mal diseño de la interfaz, que puede reducir significativamente la velocidad de escritura, comunicación, etc., del usuario.

2.4 Evaluación

La evaluación de los efectos de la predicción de palabras puede hacerse desde dos puntos de vista: cualitativa y cuantitativamente.

2.4.1 Evaluación cuantitativa

La *evaluación cuantitativa* consiste en un estudio de los resultados de la predicción, evaluados fundamentalmente respecto a dos criterios: esfuerzo y tiempo teóricamente ahorrados al usuario. Por supuesto, son estudios del máximo beneficio que la predicción de palabras puede producir, ya que suponen un aprovechamiento máximo de sus posibilidades (usuario perfecto). Sin embargo son muy útiles ya que una mejor calidad de la predicción produce mejores resultados para todos los usuarios.

Otras posibles medidas a realizar son, por ejemplo, la cantidad de palabras correctamente predichas y el número de letras antes de que fueran predichas o la influencia de métodos auxiliares de predicción, como puede ser el empleo de filtros para eliminar las palabras que ya se hayan mostrado al usuario y se hayan rechazado o el ahorro producido por añadir automáticamente un espacio y la primera letra mayúscula tras un punto.

La evaluación cuantitativa puede ser realizada automáticamente, contabilizando los parámetros que reflejan el esfuerzo (número de pulsaciones) y el tiempo (considerando barrido automático y una determinada interfaz). Estos parámetros se miden para distintas configuraciones: con y sin predicción de palabras, variando la combinación de métodos de predicción que se consideran, etc., hasta conseguir averiguar la combinación óptima. Por supuesto, ésta es la combinación óptima teórica, ya que al configurar un sistema para un usuario, se deben tener en cuenta otras consideraciones: por ejemplo, un método de predicción que almacene y prediga las palabras nuevas es totalmente contraproducente para un niño aprendiendo a escribir y que comete muchas faltas de ortografía. En [Pala99b] describimos los factores a tener en cuenta a la hora de evaluar de forma automática la predicción de palabras.

2.4.2 Evaluación cualitativa

La *evaluación cualitativa* consiste en un estudio exhaustivo de la calidad de los textos generados con y sin predicción de palabras. Para ello es necesario establecer unos criterios que permitan juzgar la calidad del texto generado (número de faltas de ortografía, número

de palabras nexo utilizadas, número de frases gramaticalmente correctas, etc.). Estos criterios son diferentes para cada usuario, dependiendo de su discapacidad particular: si tiene un problema físico, se valora más la aceleración, mientras que si el problema es lingüístico, es más importante que la palabra predicha sea correcta, procedente de un diccionario seguro, como se explica en [Morr98].

Entre los esfuerzos realizados a nivel internacional para evaluar de forma cualitativa la influencia de la predicción en la escritura, en [Magn97a] y [Magn97b] se muestran los resultados de un estudio sobre su influencia, tanto en el plano lingüístico, como de ahorro de tiempo y de pulsaciones necesarias para escribir un texto. En él se puede apreciar que el número de errores producidos al escribir un texto con predicción es mayor que sin ella, cuando el usuario está escribiendo texto que le es dictado. Sin embargo es menor cuando está escribiendo libremente.

Por otra parte, se muestran los resultados de una prueba de lectura, evaluando la *impresión general* producida por el texto a los lectores. Dicha impresión es ligeramente peor en los textos escritos utilizando la predicción de palabras, posiblemente porque los usuarios no corrigen las palabras incorrectamente escritas. Sin embargo, los usuarios prefieren escribir asistidos por la predicción, porque les resulta más fácil.

Además de los efectos de la predicción sobre la calidad del texto escrito, también es importante el punto de vista del usuario final, es decir, la valoración subjetiva de la persona que utiliza el sistema: si le facilita la tarea de escribir, si aumenta su confianza en la calidad de su texto, o si prefiere no utilizarlo, por el mayor esfuerzo requerido para escribir (aumento de carga cognitiva).

2.5 Sistemas que utilizan predicción de palabras

A continuación, se hace una presentación de diversos sistemas actuales que utilizan predicción de palabras, sus métodos de predicción y sus características. La mayor parte de ellos son programas que muestran ventanas que se superponen al procesador de texto habitual y que contienen solamente listados con las palabras predichas. Otras herramientas son más completas (editores o procesadores de texto) e incluyen emuladores de teclado para utilizar con el ratón (acceso directo) o que implementan barrido automático, para poder ser utilizadas con pulsadores.

Estos sistemas son comerciales y el acceso a la información específica de cada uno está, lógicamente, restringido a la que sus desarrolladores y distribuidores publican. La escasez de información en algunos casos puede limitar el rigor deseado en la presentación.

2.5.1 Aurora 3.0

Aurora 3.0 es un programa para Windows© comercializado por Aurora Systems, que incluye predicción de palabras. La aplicación muestra una ventana que aparece en la parte superior derecha de la pantalla, con el listado de palabras predichas, como se puede ver en la figura 2.4. Se puede utilizar con la mayoría de los programas de procesamiento de texto.

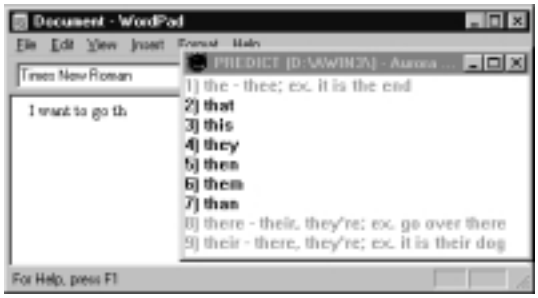


Figura 2.4: Pantalla principal de Aurora.

Para predecir utiliza métodos estadísticos. Su diccionario consta de 100.000 entradas y gestiona sufijos para modificar la terminación de las palabras. Además, está incluida la posibilidad de predecir macros. Puede encontrarse más información en [Auro99].

2.5.2 Co:Writer

Según [CoWr97] y [CoWr00], Co:Writer (Don Johnston, Inc.) es un programa tanto para Windows© como para Apple Macintosh©, que puede trabajar con cualquiera de sus procesadores de texto. Consiste en una aplicación que muestra un listado de palabras predichas, diseñada para ayudar a personas con diferentes grados de discapacidad, física o lingüística (dislexia u otras dificultades del lenguaje).

Con respecto a la predicción de palabras, incluye 3 diccionarios de inglés británico, de 2000, 10000 y 40000 palabras que utiliza para predecir. La elección entre uno u otro depende de la edad y capacidad del usuario. Gestiona un diccionario personal donde cada usuario puede añadir sus propias palabras: nombres propios de amigos, términos específicos, etc. Incluye barrido automático y puede manejar dispositivos de entrada tales como pulsadores y teclados alternativos. La salida de voz permite trabajar y comunicarse desde el ordenador a las personas con problemas de fonación.

2.5.3 EZ Keys

EZ Keys es un programa de Words+, Inc. para Windows©, que apoya en la escritura y comunicación a usuarios con capacidad para leer y escribir con las siguientes técnicas:

- *Predicción de palabras,* situada en una lista especialmente diseñada para minimizar la búsqueda visual y el tiempo de decisión que el usuario necesita para seleccionar una palabra. Como se explica en [Lawr98], para ello se limita la lista a un máximo de seis palabras (es configurable) y se muestran ordenadas alfabéticamente en dos grupos de tres palabras cada uno. Además, mantiene ciertos códigos para obtener rápidamente palabras que no estén en la lista: por ejemplo, la palabra *that* puede ser siempre obtenida tecleando *t* y *2*. Los usuarios son capaces de aprenderse estos códigos para muchas palabras, acelerando la comunicación.

- *Expansión de abreviaturas:* el usuario escribe una serie de letras que se expanden en una palabra mucho mayor, una frase o un mensaje de voz.

- *Frases pregrabadas,* que dan espontaneidad y rapidez. Son especialmente importantes al principio de la comunicación. El usuario puede crear, almacenar (por categorías) y enviar a la salida de voz cualquiera de los mensajes con solamente 4 pulsaciones.
- *Síntesis de voz,* que permite leer documentos y mejorar la comunicación.
- *Control de entrada* con un emulador de ratón.

2.5.4 PAL

PAL (Predictive Adaptive Lexicon) es un sistema desarrollado en la Universidad de Dundee para ayudar a personas con discapacidad física y/o lingüística, como se describe en [Morr92], [Newe92], [Swif87a] y [Swif87b].

La aplicación muestra una ventana superpuesta a los programas Windows©, con una lista de cinco palabras como máximo. Dichas palabras dependen de las letras que se han escrito de la palabra en curso. Si una palabra no está en el diccionario es necesario escribirla por completo. A partir de ese momento se introduce en un diccionario del usuario y se predice en ocasiones posteriores. Se pueden generar diferentes diccionarios de usuario para distintos temas.

En la actualidad, según [PAL00], se está introduciendo información gramatical para incluir información de toda la frase escrita al predecir. La adaptación al estilo del usuario se realiza actualizando las frecuencias de las reglas más utilizadas. Además, es posible predecir frases cortas.

2.5.5 PredictAbility

Figura 2.5: Pantalla principal de PredictAbility.

PredictAbility (distribuido por Inclusive Technology Ltd.) es un programa de predicción de palabras diseñado para trabajar con cualquier aplicación que acepte texto como entrada, p. ej., editores o procesadores de texto. El usuario puede modificar su apariencia, tanto el tamaño de letra como el color y número de palabras predichas. Una de las posibles configuraciones se muestra en la figura 2.5.

Con respecto a la predicción de palabras, permite crear diccionarios a partir de ficheros, así como borrar palabras de diccionarios existentes, para adaptarse rápidamente al usuario, como se describe en [Pred99].

2.5.6 Predice

Predice es una aplicación de ayuda a la escritura y comunicación para personas con problemática muy diversa. Se ha implementado en el Laboratorio de Tecnología de Rehabilitación (LTR), habiéndose convertido en uno de los sistemas de ayuda a la escritura de más amplia difusión en nuestro país. Su desarrollo puede ser seguido, desde 1992 hasta la actualidad, en [Vall92], [Pala94], [Meri98], [Rodr98], [Sani99] y [Garc00]. El Centro Estatal de Autonomía Personal y Ayudas Técnicas (CEAPAT), perteneciente al Instituto de Migraciones y Servicios Sociales (IMSERSO) se encarga de su distribución.

Tiene varios modos de funcionamiento: por una parte, puede ser utilizado como editor de texto, teniendo las opciones usuales de manejo de ficheros, impresión de texto, etc., como puede verse en su pantalla principal, mostrada en la figura 2.6. En [Pala95] se describe su funcionamiento.

Por otra parte, se puede utilizar como comunicador, para lo cual incorpora una base de datos de mensajes pregrabados por el usuario, a los que se accede con un sistema de búsqueda por palabras clave, incluyendo además información probabilística. Con este sistema, el usuario almacena sus propios mensajes, que aparecen en pantalla cuando se escriben una o más palabras contenidas en ellos. Al mismo tiempo existe la posibilidad de escribir cualquier texto nuevo y de enviar a la salida de voz lo que esté escribiendo.

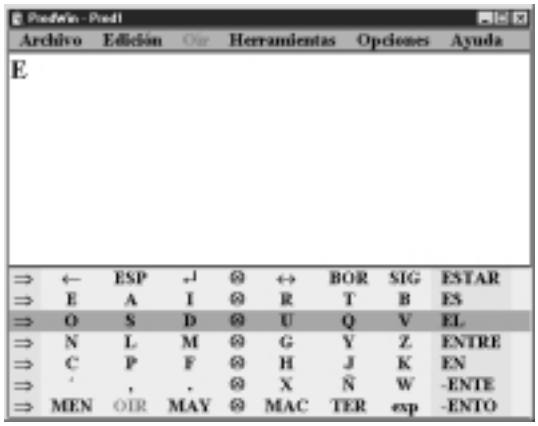


Figura 2.6: Pantalla principal de *Predice*.

Está preparado para ofrecer la máxima flexibilidad de entrada posible, de forma que pueda adaptarse a gran variedad de necesidades de usuario. Es posible utilizarlo con uno o dos pulsadores (con barrido de las distintas opciones, tanto automático como manual), ratón, palanca de control (*joystick*) o teclado.

Incluye predicción de las palabras, terminaciones o expresiones más probables. Al principio de cada sesión, la predicción se lleva a cabo utilizando información probabilística y gramatical. A medida que el usuario va escribiendo, el programa aprende del texto y la

predicción se adapta a él.

Si el usuario lo desea, el sistema también puede aprender a priori, creando una serie de diccionarios temáticos adaptados a distintos entornos: trabajo, casa, estudios, deporte, etc., con un vocabulario más específico y, por tanto, más eficaz y rápido en la predicción.

La predicción básica incluida en el programa es en castellano y la describimos en [Pala97]. El usuario, además, puede entrenar el sistema con textos en cualquier idioma.

Las personas con dificultades de visión también tienen acceso al programa, ya que es posible aumentar significativamente el tamaño de letra tanto del texto como de los menús. Si esto es insuficiente, es posible configurarlo para que todas las opciones de los menús sean leídas al mismo tiempo que se realiza el barrido automático. De esta forma el usuario no necesita ver la pantalla para saber qué opción se está barriendo y puede pulsar cuando oye la opción deseada.

En el apartado 3.5 (99) se puede encontrar una descripción detallada del diseño de su interfaz.

2.5.7 Prophet

Prophet (distribuido por ACE [3]) es un programa para Windows© que muestra y gestiona una ventana con la lista de palabras predichas que puede ser utilizada con cualquier procesador de textos. Su aspecto se puede observar en la figura 2.7. También está disponible una versión que incorpora un emulador de teclado en pantalla que se utiliza con barrido automático de sus opciones y pulsador o con acceso directo con ratón, como se puede ver en la figura 2.8.

Figura 2.7: Pantalla principal de Prophet para sueco.

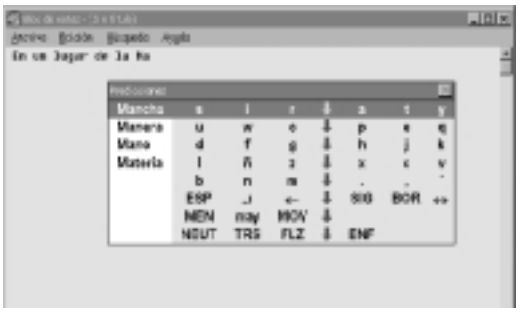


Figura 2.8: Pantalla principal de Profeta para castellano, con emulador de teclado.

Hay versiones de este programa para sueco, noruego, danés, inglés y castellano. En la versión para sueco, se utilizan como base para la predicción los trigramas de palabras [4], además de métodos heurísticos. Véase [Carl97a] para obtener información más detallada sobre los algoritmos de predicción. La información específica de cada idioma (excepto para el castellano) es muy reducida, puesto que es importante la posibilidad de utilizar los algoritmos de predicción para diferentes idiomas con un esfuerzo reducido, empleando

[3] La versión en castellano se distribuye bajo licencia del LTR.

[4] Este método se detalla en la sección 2.2.1

métodos para los cuales sea posible el entrenamiento automático de los modelos de lenguaje. En la versión para castellano se utiliza el mismo motor de predicción que se integró en la versión de *Predice* de 1995.

2.5.8 SAW (Switch Access to Windows)

SAW (Switch Access to Windows) es un programa distribuido por Oxford ACE Centre que permite al usuario controlar Windows© utilizando uno o dos pulsadores, un trackball o un licornio. SAW reemplaza el teclado y el ratón por una serie de menús de selección que sitúa sobre la pantalla, como puede ser observado en la figura 2.9.

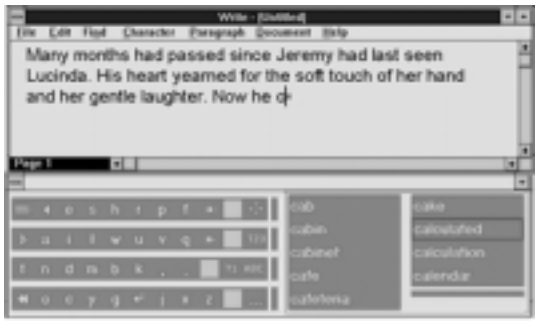


Figura 2.9: Pantalla principal de SAW.

En estos menús se sitúan las letras, palabras, números o símbolos, que son barridos elemento a elemento de forma manual o automática, y seleccionados utilizando los pulsadores. SAW proporciona acceso a todas las operaciones que se pueden realizar sobre Windows©: escribir, dibujar, mover el cursor por la pantalla, etc. En las últimas versiones incluye, además, técnicas de aceleración de escritura y acceso a CD-ROM.

Para ampliar información sobre SAW, véase [SAW00].

2.5.9 SoothSayer Word Prediction

Según [Soot00], SoothSayer es un programa para Windows©, consistente en una ventana en que se muestra una lista con las palabras predichas simultáneamente a la escritura del usuario. Incluye un diccionario de más de 11000 palabras en inglés y la posibilidad de añadir palabras a un diccionario de usuario.

Además de predicción de palabras, ofrece expansión de abreviaturas, predicción de frases y síntesis de voz para facilitar la comunicación. En la actualidad se está construyendo una librería de diccionarios en internet que permite el intercambio de diccionarios de usuario.

SoothSayer Word Prediction© es distribuido por Applied Human Factors©, Inc. (AHF).

2.5.10 Telepatic

Telepatic (distribuido por Madentec), es un programa para Windows© que gestiona una ventana con un listado de palabras predichas. Se puede utilizar con cualquier programa de edición y procesamiento de textos. Se accede a las palabras utilizando los números, teclas de función o cualquier combinación de teclas configurada por el usuario. Una posible configuración de la ventana puede observarse en la figura 2.10. La posición de la ventana de predicción y el número de palabras también son configurables.

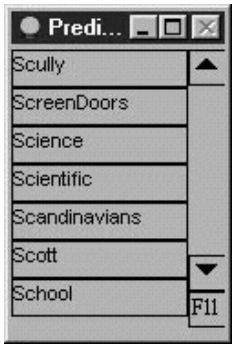


Figura 2.10: Pantalla principal de Telepatic 2000.

La predicción inicial se basa en un diccionario en inglés de 20000 palabras. Además, es posible entrenar la predicción de palabras con cualquier texto, creando así diccionarios adaptados a cualquier tema. La predicción tiene apoyo gramatical, por lo cual las personas que tienen problemas para construir frases gramaticalmente correctas escriben mejor si se apoyan en las predicciones.

Además de la predicción, incluye también expansión de abreviaturas, expandiendo expresiones frecuentes e incluso párrafos.

2.5.11 TPV

TPV es un emulador de teclado, distribuido por el CEAPAT, que incluye predicción de palabras para castellano. Los algoritmos de predicción han sido desarrollados en el LTR y son los mismos que se utilizan en *Predice* (véase sección 2.5.6). .

Esta herramienta es muy flexible, permitiendo al usuario configurar su propio teclado, tecla por tecla. Para cada tecla, es posible definir su contenido, aspecto (color del texto, del fondo, tamaño, etc.), comando ejecutado al pulsar (insertar un carácter o secuencia de caracteres en el texto, abrir una aplicación, dar paso a otro teclado), emisión de un sonido al pulsar (y cuál), etc. Funciona con el teclado convencional, con acceso directo (ratón) y con barrido automático controlado con pulsadores.

2.5.12 Write Away 2000

Information Services Inc. distribuye Write Away 2000, un editor de texto orientado a personas con discapacidad física. Su interfaz, que se muestra en la figura 2.11, se ha

intentado simplificar al máximo para evitar la excesiva cantidad de opciones normalmente existentes en los procesadores de texto habituales. Permite configurar los tamaños de las letras del texto, de los menús y de la ventana de predicción. El acceso a la ventana de predicción se puede realizar mediante ratón o teclado y las funciones más frecuentes están accesibles a través de iconos. La ventana de predicción se mueve con el cursor al insertar texto para minimizar el movimiento de ojo que el usuario debe hacer entre las palabras predichas y el texto que está escribiendo, aunque también es posible fijar su posición en la pantalla.

Figura 2.11: Pantalla principal de Write Away 2000.

Según [Writ00], la predicción de palabras se realiza considerando varias fuentes de información:

- *Listas de palabras más frecuentes*, que son dinámicamente actualizadas a medida que el usuario escribe. Además, es posible añadir nuevas entradas al diccionario, que pueden ser tanto palabras como expresiones compuestas por más de una palabra (frases, nombres propios, etc.).
- *Diccionario principal* de Write Away 2000, que tiene cuatro listas entre 600 y 8000 palabras frecuentes entre los cuales es fácil cambiar. Estas listas son modificables.
- *Listas específicas de palabras*, asociadas a un *tema* determinado, que se generan a partir de cualquier texto disponible sobre dicho tema.

La mayor parte de los sistemas descritos se han desarrollado para inglés[5]. Aquellos que permiten su utilización en varios idiomas están basados en métodos estadísticos, fácilmente reentrenables, pero carentes de la potencia que permite la incorporación de conocimiento lingüístico específico. Por esto, la predicción de estos sistemas es de baja calidad para el castellano, y está plenamente justificada la necesidad de realizar un esfuerzo tanto de investigación como de evaluación para optimizar la predicción en nuestro idioma. La incorporación de conocimiento específico sobre el mismo mejorará las prestaciones obtenidas, como ocurre en otras áreas, como, por ejemplo, el reconocimiento [Ferr96], o la comprensión de habla [Cola99].

[5]Excepto *Predice*, TPV y la versión de Profet para castellano

Capítulo 3

Predicción de palabras

Como se ha mostrado en las secciones anteriores, la predicción de palabras puede ofrecer diversos beneficios a personas con una problemática muy heterogénea. Estos beneficios se resumen, fundamentalmente, en un aumento de la calidad y/o cantidad de texto generado, dependiendo del tipo de problemas del usuario: las personas con dificultades lingüísticas (dislexia, faltas de ortografía frecuentes, etc.) valorarán el número de palabras predichas, mientras que los usuarios con algún tipo de discapacidad física, se beneficiarán del aumento de velocidad y de la reducción del esfuerzo necesario para escribir el texto.

Independientemente del objetivo buscado por el usuario, una predicción de palabras de mayor calidad beneficia a ambos grupos de personas, y esto justifica la introducción de métodos de predicción sofisticados, que consideren toda la información disponible a la hora de seleccionar las palabras predichas.

En esta tesis se ha realizado un estudio de la predicción de palabras para castellano, considerando las distintas fuentes de información que podrían ser utilizadas para aumentar su calidad y la forma de introducir y combinar la potencia de cada una de ellas. Además, se ha llegado a realizar un sistema completo con el que se ha verificado experimentalmente el resultado de cada uno de los métodos estudiados y sus combinaciones, permitiendo finalmente obtener la configuración óptima del sistema predictivo final.

En este capítulo se describe el estudio que ha dado lugar a los distintos elementos que componen la arquitectura y a la metodología de integración que esta tesis propone junto con las aportaciones y metodologías específicas que se han realizado en cada uno de los elementos. También se describe el sistema completo de ayuda a la escritura que se ha implementado para validar la arquitectura. Destacaremos por tanto las propuestas de la tesis para los diversos métodos de predicción, las fuentes de información utilizadas por dichos métodos, la interfaz con el usuario y el módulo que coordina todo el sistema.

3.1 Arquitectura general del sistema

En la figura 3.1 se muestra la arquitectura general del sistema de predicción de palabras.

La arquitectura propuesta es un modelo general y se irá refinando a lo largo de este

Figura 3.1: Arquitectura general de la predicción de palabras.

capítulo hasta llegar a la arquitectura final en la figura 3.6 (página 107). Como se puede apreciar en el diagrama, los métodos y los diccionarios son independientes entre sí, facilitando el diseño del sistema y la portabilidad a otros idiomas.

En el esquema pueden distinguirse los siguientes bloques fundamentales:

- *Interfaz de usuario (IU).* La interfaz del programa correspondiente. Proporciona al módulo de coordinación el texto del usuario y muestra las palabras predichas.

- Módulos que participan en el proceso de predicción:

  - *Módulo de coordinación.* Es el bloque que se encarga de procesar la entrada procedente de la interfaz de usuario (texto escrito por el usuario) y gestionar los flujos de información entre los distintos métodos de predicción (coordinando los datos que le proporciona y pide a cada uno) y las transacciones con los diccionarios. Además, obtiene el listado de palabras predichas propuesto por cada método y envía las más adecuadas a la interfaz de usuario.

  - *Gestores de diccionarios.* Encargados de seleccionar la lista de palabras que cumplen las restricciones impuestas por los métodos de predicción usados en cada caso. Reciben del módulo de coordinación las restricciones y devuelven las listas de palabras correspondientes, de entre las contenidas en los diccionarios del sistema.

- Los *diccionarios* son las fuentes de información generales del sistema (contienen las palabras pertenecientes al léxico disponible y la información frecuencial y gramatical básica necesaria para apoyar a todos los métodos de predicción: categoría de cada palabra, frecuencia, etc.). En este sistema se utilizan tres tipos diferentes de diccionarios: diccionario general, personal y diccionarios temáticos.
- *Métodos de predicción.* Son los métodos encargados de decidir, a partir de la información proporcionada por el módulo de coordinación y de las fuentes de información específica que cada uno requiere (conjunto de reglas, matrices de secuencias de categorías, etc.), las restricciones que deberán cumplir las palabras (conjunto de categorías, frecuencias, rasgos, palabras o lemas). Se han incluido métodos para considerar los factores que en el encuadre científico tecnológico se han mostrado como determinantes para aumentar la calidad de la predicción: información gramatical (categorías y rasgos), frecuencial, etc. Esta información se utiliza en etapas posteriores del proceso para obtener la lista de palabras predichas a partir de los diccionarios.

• Módulos que participan en el proceso de entrenamiento:

- *Módulos de entrenamiento de diccionarios.* Este bloque engloba los procedimientos automáticos y manuales necesarios para generar los diccionarios usados en el proceso de predicción a partir de las fuentes de entrenamiento de los mismos (normalmente, aunque no siempre, en una etapa previa a la predicción). En el proceso de generación se usan como fuente de conocimiento los textos disponibles, categorizados o no.
- *Módulos de entrenamiento de información específica.* Engloban los procedimientos automáticos y manuales necesarios para generar la información específica que necesita cada método de predicción a partir de las fuentes de entrenamiento disponibles. Dichas fuentes son tanto los textos, categorizados o no, como el conocimiento lingüístico experto necesario.

Una vez visto a grandes rasgos el esquema del sistema de predicción, se pasará a detallar los grandes bloques del sistema. El estudio, contenido, entrenamiento y gestión de los diccionarios se describe en el apartado 3.2. El apartado 3.3 está dedicado a los distintos métodos de predicción y la sección 3.4 al funcionamiento del módulo de coordinación. Las consideraciones sobre el diseño de la interfaz de usuario y la integración de la predicción de palabras se explican en el aparatado 3.5.

3.2 Diccionarios

Los diccionarios son las fuentes de información básicas utilizadas por todos los métodos de predicción, y contienen, tanto las palabras, como toda la información asociada a cada una de ellas que puede ser necesaria en cualquier punto del proceso de predicción.

A partir del estudio del comportamiento del sistema de predicción, se comprobó que es necesario disponer de varios tipos de diccionarios diferentes, cada uno de los cuales aporta

una información distinta y requiere sus propios mecanismos de entrenamiento y gestión. La división en distintos diccionarios puede ser encontrada también en algunos de los sistemas descritos en el encuadre científico-tecnológico, como, por ejemplo, PredicAbility o Prophet.

En esta sección se justifica la necesidad de cada tipo de diccionario y su aportación al sistema global. Además, se describe la estructura, los procesos de generación y mantenimiento, y las formas de uso de cada uno de ellos en la arquitectura predictiva final que hemos aportado en el marco de este trabajo.

3.2.1 Diccionario general

El diccionario general es el que proporciona la información frecuencial y gramatical necesaria para que los diversos métodos predigan palabras. Se utiliza en dos puntos del proceso de predicción:

- Para obtener información (categorías, rasgos, frecuencia, etc.) sobre las palabras escritas por el usuario.
- Para obtener la lista de palabras predichas.

Este diccionario es estático, es decir, no evoluciona con el uso del programa, ni es ampliable por el usuario.

El diagrama de bloques correspondiente al diccionario general es el que se muestra en la figura 3.2. Se puede apreciar que el gestor del diccionario ha sido dividido en dos módulos, cada uno de los cuales se encarga de una de las tareas indicadas anteriormente. En las secciones siguientes se explicará en detalle el contenido de cada bloque.

3.2.1.1 Estructura del diccionario general

El diccionario general contiene más de 160.000 entradas, considerando tanto palabras como signos de puntuación. Cada entrada del diccionario contiene la siguiente información:

- *Significante.* Fonema o secuencia de fonemas o letras que, asociados con un significado, constituyen un signo lingüístico[1]. Contiene la palabra. P. ej., *cosas*.
- *Lema.* Es la voz que se emplea como entrada de un artículo de diccionario y representa a todas las formas paradigmáticas que pueda tener[2]. Coincide en todas las formas de una palabra que tienen la misma raíz, categoría y significado, pero que cambian de género, número, tiempo verbal, etc. En los nombres y adjetivos es su forma en masculino singular (si existe), en los verbos es el infinitivo. P. ej.: en la entrada correspondiente a la palabra *cosas* como nombre, el lema es *cosa*, y en la entrada de *cosas* como verbo, el lema es *coser*.

[1] Definición según el Diccionario de la Real Academia de la Lengua [DRAE00].

[2] Definición del Diccionario General de la Lengua Española VOX [VOX00]. Según consulta realizada al Departamento de Español al día de la RAE, *lema* aparecerá definido en la próxima edición del DRAE como "entrada de un diccionario o enciclopedia" (pendiente de aprobación final).

Figura 3.2: Diagrama de bloques del diccionario general.

- *Frecuencia absoluta.* Es el número de veces que aparece la palabra en los textos de entrenamiento.
- *Categoría gramatical* a la que pertenece la palabra. Una categoría gramatical, según [Gram96], es cada una de las clases en las que se puede agrupar una palabra por su significado, constitución y funciones que puede desempeñar. En esta tesis, se añade un criterio adicional a la clasificación de palabras: el comportamiento sintáctico real de la palabra en las estructuras gramaticales en las que aparece. La clasificación final obtenida al añadir este nuevo criterio se explica en el apartado 3.2.1.4. Dos de las categorías utilizadas son *verbo* y *nombre*, por ejemplo.
- *Rasgos* asociados a la palabra. Coinciden, en parte, con los accidentes gramaticales explicados en [Gram96]: informaciones vinculadas a las terminaciones de las palabras que aportan significado gramatical, por ejemplo, *género*, *número*, *tiempo* o *persona*. Además de estos, en esta tesis se han añadido otros rasgos para marcar aquellas palabras que presentan comportamientos homogéneos. El conjunto total de rasgos se describe en el apartado 3.2.1.4.

Aproximadamente un 20% de las palabras del diccionario tienen *homógrafos*[3] con diferente información gramatical, por ejemplo, la palabra *llama* (fuego, animal o forma del verbo llamar). En estos casos, existe una entrada diferente por cada uno. Como no se considera información semántica, los homógrafos cuya información gramatical coincide están englobados en una sola entrada. En el ejemplo anterior, para la palabra *llama*, en

[3]Las palabras homógrafas son términos de orígenes distintos, cuyo significado no guarda ninguna relación entre sí, pero que coinciden en la forma [Gram96].

el diccionario general aparecerá una entrada como nombre femenino singular, para los dos primeros significados, y dos entradas diferentes para el verbo llamar, con rasgos diferentes, una para la tercera persona del presente y otra para el imperativo.

De las 160.000 entradas del diccionario, más de 130.000 tienen significantes diferentes.

3.2.1.2 Textos de entrenamiento del diccionario general

Gran parte de las palabras y la información frecuencial procede del corpus de entrenamiento MUNDO3-6[4], formado por textos de la edición electrónica del periódico "El Mundo" (de Enero a Junio de 1994). El resto fueron añadidas en sucesivas etapas, tanto automáticas como manuales, como se muestra en la sección siguiente, en la que se describe el proceso de generación del diccionario general.

3.2.1.3 Generación del diccionario general

A continuación se describen, en orden cronológico, las etapas que dieron lugar a la versión final del diccionario, hasta conseguir el conjunto de palabras con la información frecuencial y gramatical definitiva a partir de los corpora originales:

- Obtención, a partir de los textos originales, de la lista de palabras presentes y sus frecuencias absolutas en dicho corpus de entrenamiento. Estas frecuencias son las que se utilizan en el cálculo de probabilidades de las palabras predichas, combinadas con la información frecuencial procedente de las diversas técnicas de predicción.

- Etiquetado automático de las palabras de la lista con el sistema *SMORPH*, que es una herramienta para la construcción de diccionarios electrónicos, análisis morfológico y el etiquetado de textos reales, desarrollado en el *Groupe de Recherche dans les Industries de la Langue* (GRIL, Universidad de Blaise Pascal, Clermont Ferrand, Francia). SMORPH procesa los textos de entrada, asignando de forma automática a cada palabra su categoría gramatical y el conjunto de rasgos asociados a ella. Utiliza el siguiente conjunto de categorías:

Nombre	Adjetivo	Adverbio
Verbo	Determinante	Artículo
Posesivo	Numeral	Pronombre
Preposición	Conjunción	Signo de puntuación

  Un subconjunto de los rasgos utilizados por SMORPH es el siguiente:

Singular masculino	Singular femenino	Plural femenino
Plural masculino	1, 2 y 3 persona	Indicativo
Subjuntivo	Imperfecto	Indefinido
Pretérito indefinido	Infinitivo	Participio
Coma	Guión	Punto
Paréntesis (abertura)	Paréntesis (cierre)	

[4] Para ampliar información, véase el apéndice A sobre corpus de entrenamiento y evaluación.

Para palabras con más de una categoría diferente, SMORPH proporciona todas las posibilidades, sin discriminar cuál puede ser la correcta en un punto determinado de un texto. Por ejemplo, *sobre* puede ser preposición, nombre y, además, varias formas del verbo sobrar: cada vez que aparezca la palabra *sobre* en el texto que se categoriza, SMORPH entrega la información correspondiente a todas esas entradas, independientemente del uso particular que se esté haciendo de dicha palabra.

Para ampliar información sobre el funcionamiento del sistema SMORPH, véase [Aïtm95].

La categorización automática utilizando SMORPH, además, permitió un primer filtrado de las entradas erróneas, que no superaron esta etapa, porque no fueron categorizadas. También se eliminaron en este paso las palabras de otros idiomas, muy comunes en los textos periodísticos, especialmente en el caso de los nombres propios de personas y lugares.

- A continuación, se añadió a cada entrada la información frecuencial, según su frecuencia en los textos de entrenamiento. Algo más del 20% de las palabras del diccionario se corresponden con varias entradas diferentes, por lo cual, además, es necesario establecer un criterio para repartir la frecuencia entre ellas.

  La primera aproximación al problema fue dividir uniformemente la frecuencia entre todas las categorías que tenga cada palabra. Este proceso da lugar a datos erróneos, como, por ejemplo en el caso de *una*. Esta palabra se etiqueta como determinante, pronombre y varias formas del verbo unir (una en imperativo y dos personas distintas de presente de subjuntivo). Si se reparte uniformemente la frecuencia entre las cinco posibilidades, cada vez que se escriba la palabra *una* en un texto, los métodos de predicción que necesiten información gramatical asumirán que es un verbo, más frecuentemente que un determinante o pronombre.

  Para reducir el impacto negativo de esta estimación, se realizó un segundo paso, en el cual se rectificó manualmente, de forma aproximada, la frecuencia de las palabras más frecuentes, reduciendo la aportación de los significados menos probables, manteniendo la frecuencia absoluta total (por ejemplo, disminuyendo la frecuencia de *una* como verbo y aumentando su frecuencia como determinante y pronombre). Aunque este método no es exacto, el diccionario rectificado ofrece más garantías en cuanto a los resultados obtenidos, ya que la información estadística refleja mejor el comportamiento de las palabras.

  Este problema no habría existido si el categorizador hubiera proporcionado para cada palabra únicamente una categoría según su entorno, o si se hubiera realizado una categorización manual de los textos de entrenamiento, asignando a cada palabra exactamente la categoría y rasgos que le corresponden por la posición que ocupa en el texto. En este caso, se podría asignar a cada entrada del diccionario (a cada uno de los homógrafos) su información frecuencial exacta. Este proceso no se ha realizado por el enorme esfuerzo que supondría categorizar de forma manual o semiautomática los más de ocho millones y medio de palabras contenidos en los textos de entrenamiento.

- Los paradigmas de ciertas palabras, especialmente las pertenecientes a las categorías

cerradas[5], se completaron manualmente, ya que en los textos de entrenamiento no aparecían todas las formas.

- A continuación se obtuvieron todas las formas de los verbos que aparecen en los corpora de entrenamiento, para completar paradigmas. Para ello se utilizó un generador automático de formas verbales (parte del sistema SMORPH), a partir del listado de infinitivos, tanto regulares como no regulares, presentes en el diccionario.

  Esta lista de formas verbales también fue categorizada de forma automática utilizando el sistema SMORPH, obteniendo todas las posibles categorías de cada palabra. Por ejemplo, *fuera* puede ser del verbo *ser*, del verbo *ir* y, además, adverbio.

  Una vez categorizadas, se añadieron al diccionario las formas de los verbos irregulares y las formas de verbos regulares que tuvieran homógrafos pertenecientes a otras categorías gramaticales (si no estaban incluidas). Se asignó frecuencia uno (la mínima) a cada una de las entradas.

- En una etapa posterior, se eliminaron las formas flexivas de los verbos regulares que no tuvieran homógrafos pertenecientes a otras categorías. Este proceso reduce de forma importante el tamaño del diccionario, ya que hay gran cantidad de verbos regulares y muchas de sus formas apenas se usan (por ejemplo, algunas del subjuntivo).

  Esta disminución en el tamaño del diccionario tiene importantes ventajas en reducción de recursos necesarios y tiempo de procesamiento. La ausencia de estas formas verbales en el diccionario, que produciría problemas tanto a la hora de la categorización, como en la predicción, se compensa de dos maneras:

  - Incluyendo un módulo de conjugación de verbos regulares, que permita etiquetar cualquier forma de verbo regular que no esté incluida en el diccionario, asignándole tanto la categoría como el conjunto de rasgos adecuado.
  - Añadiendo el conjunto de desinencias verbales en la predicción de terminaciones (apartado 3.3.4.2), para ayudar en la escritura de las formas no incluidas en el diccionario. Así, en todas las formas verbales escritas está asegurado el ahorro de varias pulsaciones, aunque no estén en el diccionario.

  Además, como se explica en el apartado 3.2.2 correspondiente al diccionario personal, las formas verbales utilizadas por el usuario son incluidas en otros diccionarios (personal y temáticos) la primera vez que se utilizan, y, a partir de ese momento, son consideradas en los mecanismos de predicción. Con esto, la repercusión de la eliminación de formas verbales regulares queda limitada a la primera vez que se utiliza cada forma, que además está apoyada por la predicción de sufijos.

- Por último, tras estos procesos que permitieron introducir el conjunto definitivo de palabras en el diccionario, se modificó la clasificación por categorías: algunas de las categorías iniciales fueron modificadas manualmente, utilizando como nuevo criterio el comportamiento sintáctico observado de cada palabra. Además, parte de la carga informativa fue trasladada a un conjunto de rasgos asociado a cada palabra.

[5] Categorías con un conjunto reducido de palabras: preposiciones, determinantes, etc. en contraposición con los nombres, verbos y adjetivos.

3.2.1.4 Elección del conjunto de categorías y rasgos

El conjunto final de categorías y rasgos del diccionario general hemos decidido en esta tesis que no corresponda exactamente con la clasificación tradicional de los libros de lingüística, ni con, por ejemplo, dos de los esquemas de categorización con los que se había trabajado previamente, los del sistema SMORPH y los generados en el proyecto ESPRIT-860 (Linguistic Analysis of European Languages). Esta es una de las aportaciones conseguidas gracias al estudio realizado en esta tesis observando el comportamiento lingüístico real de las palabras y tomando la consecuente decisión sobre cómo recoger estos efectos con la potencia concreta que nos ofrece el modelo gramatical formal que presentaremos más adelante.

La categorización del sistema SMORPH fue comentada en la sección anterior sobre la generación del diccionario general.

En el proyecto ESPRIT-860 se estableció una clasificación lingüística para varios idiomas europeos, entre ellos el castellano. El análisis fue fundamentalmente morfológico, aunque también se consideraron algunos aspectos relacionados con el contenido semántico [Casa89]. En la versión final se definieron clases sintácticas principales (verbos, nombres, adjetivos, pronombres, artículos, adverbios, preposiciones, conjunciones, interjecciones y varios) y se refinó esta clasificación usando subcategorías de primer orden, tiempo/modo/grado, número, locución, género y terminación verbal [Kugl89].

En esta tesis se tomó como base la categorización del sistema SMORPH y se modificó según el siguiente criterio: el conjunto final de categorías y rasgos tiene que *facilitar la realización de un modelo del castellano (simplificado) como secuencia de categorías, en forma de gramática independiente de contexto, de la forma más sencilla posible.*

Según este criterio, mientras que no se manifieste que una palabra o conjunto de palabras tiene un comportamiento sintáctico distinto, en una lectura de izquierda a derecha, que el resto de las palabras de la misma categoría, no se justifica una nueva categoría (por muy tradicional que sea). Al contrario, cuando se manifieste un comportamiento diferente, está justificada la adopción de una nueva categoría, aunque en otras gramáticas no se considere diferente.

Por ejemplo, existe un conjunto reducido de palabras que se encuentran separadas en varias categorías tradicionales en lingüística (artículos determinados, indeterminados, posesivos, indeterminados o demostrativos, etc.). Tras una observación detenida del comportamiento sintáctico de las palabras incluidas en esas categorías, se verificó que en muchas ocasiones es el mismo, como en *mi amigo*, *el amigo* y *este amigo*, por lo cual se han unificado en una única categoría.

Sin embargo, otro conjunto de palabras, normalmente agrupadas como preposiciones, se han separado en tres grupos: por una parte *a* y *de*, a las cuales no pueden seguir la palabra *el*, por otra parte el resto de las preposiciones, y, por último, otra categoría para los artículos contractos, *al* y *del* que restringen el sintagma nominal que les sigue (debe ser masculino singular y sin artículo, porque ya está incluido en la contracción).

Parte de la categorización final del diccionario procede del sistema SMORPH, que ha aportado fundamentalmente nombres, adjetivos y verbos agrupados según la categorización tradicional. Otra parte, correspondiente a las demás categorías, ha sido modificada

manualmente siguiendo criterios sintácticos, que se indicarán al explicar cada una.

Nuestro estudio nos ha llevado a complementar la clasificación en categorías desplazando parte de la carga expresiva de las categorías a los rasgos con una metodología original. Estos rasgos permiten indicar para cada palabra su género, número, tiempo verbal, etc., expresando esa información sin necesidad de introducir nuevas categorías. Las ventajas de este tratamiento se podrán ver en el apartado 3.3.3 correspondiente a las gramáticas formales. Al igual que en el caso de las categorías, hemos decidido que no todos los rasgos se correspondan con accidentes gramaticales de la lingüística tradicional.

A continuación se explica el conjunto final de categorías y de rasgos. Se ha preferido introducir aquí la descripción de categorías por ser necesaria para la evolución lógica del documento, aunque su justificación última no se verá hasta el apartado 3.3.3, ya que el criterio utilizado para la selección de ambos conjuntos considera especialmente su comportamiento en la gramática de contexto libre que utiliza el método gramatical de predicción. Por esta razón, al justificar las categorías y los rasgos, se propondrán ejemplos en los cuales se vea su uso en las reglas en las que se utilizan.

3.2.1.4.1 Verbo

El conjunto de palabras incluidas en esta categoría coincide con la clasificación tradicional, agrupando a las palabras que indican acciones, estados o procesos. Por ejemplo, *leer*. Son los núcleos del sintagma verbal, que es el predicado de la oración.

Los rasgos que admite y sus valores son los siguientes:

- *Tipo de verbo*, que engloba una serie de características que determinan su comportamiento. Puede tomar los siguientes valores:
  - Regular. Está presente en todas las formas de los verbos regulares. Es especialmente importante a la hora de la categorización, ya que a partir de los infinitivos marcados como regulares se podrán reconocer todas sus formas flexivas, aunque no estén incluidas en el diccionario general.
  - Copulativo. Marca todas las formas de los verbos *ser*, *estar*, *resultar* y *parecer*. Con él se podrá discriminar este conjunto de verbos y permitirá, en los módulos gramaticales, escribir reglas específicas para formar frases como *las manzanas son buenas* y *las manzanas parecen buenas*.
  - Verbo que exige después de *que* un verbo en subjuntivo, como, por ejemplo, *poder* o *hacer*. Este rasgo permite modelar construcciones como *puede que llueva* o *hizo que volviera*, y no *puede que llueve* o *hizo que vuelve*. No son las únicas construcciones donde aparecen este tipo de verbos, ya que pueden utilizarse en reglas que aceptan verbos sin imponerles restricciones (p. ej., frases como *hace un pastel* o *puede venir*), pero cuando se escribe la palabra *que* después de ellos, deben ir seguida por un verbo en subjuntivo.
  - Verbo que exige después de *que* una forma flexiva de un verbo. Su comportamiento es similar al del rasgo anterior, exigiendo una forma flexiva de un verbo

con distinto modo dependiendo de la construcción concreta. Es el caso del verbo *creer* en expresiones del tipo *creía que vendrías, no creo que venga, creo que viene ya*.

  - Verbo que admite la construcción *de* más infinitivo, como *acabar*, en *acabar de llegar*.
  - Verbo que puede ir seguido por *a* más un verbo en infinitivo, como *venir* o *llegar*.
  - Verbo al que le puede seguir *que* más infinitivo, como *tener*.
  - Verbo que puede ir acompañado por un infinitivo, como, por ejemplo, *querer*.
  - Verbo al que puede seguir un participio, como *dejar*.
  - Verbo que puede ir seguido por un gerundio, como *seguir*.

    Cada verbo del diccionario está marcado con un número diferente de rasgos de este grupo, dependiendo de su comportamiento particular (o con ninguno, si no responde a ninguno de los fenómenos descritos por estos rasgos).

- *Tipo de terminación verbal*, que puede tener los siguientes valores:
  - Infinitivo (p. ej., *cantar*).
  - Participio (p. ej., *respirado*).
  - Gerundio (p. ej., *queriendo*).
  - Flexivo. Marcará el resto de formas, como p. ej. *podemos*.

    Los verbos marcados como flexivos, además, están subdivididos con otro rasgo según su *modo*, que admite los siguientes valores:
    * Indicativo, como *dormía*.
    * Subjuntivo, p. ej., *sea*.
    * Condicional, como *podría*.
    * Imperativo, que marca palabras como *sigue*.

    Estos dos últimos rasgos permiten especificar ciertas condiciones en las reglas, por ejemplo, correspondiendo a restricciones impuestas por verbos anteriores.

- *Tiempo verbal*, con los valores:
  - Presente (p. ej., *amo*).
  - Pretérito imperfecto (p. ej., *amaba*).
  - Pretérito indefinido (p. ej., *amé*).
  - Futuro (p. ej., *amaré*).

- *Persona*, cuyos posibles valores son primera, segunda, tercera y cualquiera de ellas (se marcan todas ellas), por ejemplo, para las formas verbales no flexivas.

- *Número*, que puede ser singular y plural. En las formas flexivas sirve para distinguir entre las personas del singular y las del plural (p. ej., *amo*, correspondiente a la primera persona del singular y *amamos*, primera persona del plural). En el participio, el número (junto con el género) se utilizan en las construcciones que exijan concordancia, como en *Las casas son pintadas*. Las demás formas no flexivas tendrán ambos valores del número.

- *Género*, puede ser femenino y masculino. Se utilizará en los participios en las construcciones que exigen concordancia, como las pasivas: *La puerta es abierta*. Las demás formas verbales están marcadas con ambos valores del género.

3.2.1.4.2 Nombre

Coincide con la clasificación tradicional, agrupando a las palabras que designan personas, animales, ideas, etc. Por ejemplo, *mesa*. El conjunto de rasgos que admite, y sus posibles valores son:

- *Género*. Puede ser masculino, femenino y, además, se ha añadido un valor específico para los nombres singulares que comienzan por *a tónica* (con o sin *h*) como, por ejemplo, *alma* o *hacha*, que tienen un comportamiento peculiar con la concordancia[6]: concuerdan con el artículo masculino, como en *el alma* o *un alma*, pero, sin embargo, si después se escribe un adjetivo, este debe ser femenino *el alma buena*. En plural se comportan como palabras femeninas, están categorizadas como tales y siguen las reglas de concordancia habituales: *las almas buenas*.

  Este comportamiento particular hace que no puedan utilizarse las reglas de concordancia habituales y justifica la inclusión del nuevo valor para el género, que no concuerde con los demás, y la elaboración de reglas específicas para tratarlo.

  Las palabras que tengan género neutro, como, por ejemplo, *periodista*, están marcadas como femeninas y masculinas, y concuerdan con ambos valores.

- *Número*. Puede tomar los valores singular y plural. Las palabras de número neutro, como *tesis* toman ambos valores.

- *Tipo de nombre*. Divide a los nombres en comunes (*casa*) y propios (*Javier* o *Alcalá*). Los nombres propios están marcados como tales en el diccionario, aunque están introducidos en minúsculas (incluida la primera letra), ya que el procesamiento de las palabras se realiza siempre en minúsculas. Para considerar la información que pueden aportar las mayúsculas en un texto, sería necesario al menos considerar la ambigüedad que se produce después de un punto (donde se escribe la primera letra en mayúscula sea nombre propio o no), y cuando se escribe una palabra en mayúsculas por completo. También sería necesario considerar que el usuario puede no escribir la primera letra del nombre propio en mayúscula, especialmente si el proceso es costoso, como en las interfaces por barrido de los sistemas orientados a personas con discapacidad física.

- *Nombre seguido por subjuntivo*. Este rasgo marca a los nombres que admiten la construcción *que* + *verbo en subjuntivo*, como *suerte* o *error* en *parece una suerte que sea de ese modo*, *es un error que vengas*, y no *parece una suerte que es de ese modo* ni *es un error que vienes*.

[6] Según [Gram96], *concordancia* es la correspondencia entre las desinencias de dos o más palabras. En esta tesis se considera que la concordancia de dos palabras en un determinado rasgo depende de la relación existente entre el valor de dicho rasgo asignado a ambas palabras.

3.2.1.4.3 Adjetivo

Coincide con la clasificación tradicional, incluyendo a las palabras que expresan cualidades, estados u otras circunstancias que se refieren a un sustantivo. Por ejemplo, *curioso*. Los rasgos que admite son:

- *Género.* Puede ser masculino, femenino y neutro. Este último es para adjetivos como, por ejemplo, *puntual* que concuerdan con palabras femeninas y masculinas y se marcan con ambos valores del género.
- *Número.* Puede tomar los valores singular y plural. Al igual que en el caso de los nombres, los adjetivos de número neutro toman ambos valores.
- *Tipo de adjetivo.* Especifica si puede formar parte de construcciones de tipo comparativo o no. Por ejemplo, los adjetivos marcados como comparativos, como *mejor*, *mayor* o *peor*, pueden utilizarse como adjetivos aislados (*esa manzana es mejor*), o en expresiones comparativas (*esa manzana es mejor que aquella*).
- *Adjetivo superlativo.* Marca los adjetivos en grado superlativo, como *óptimo* o *bellísima*.
- Tipo de adjetivo según la *posición* que puede tomar con respecto al nombre. Se consideran tres tipos:
  - Antepuestos: los que solamente pueden estar antes del nombre, como *buen* que admite construcciones como *un buen amigo*, pero no *un amigo buen*.
  - Pospuestos: los que solamente pueden estar después del nombre, como *grande*: se puede decir *la mano grande*, pero no *la grande mano*.
  - Los que pueden estar en ambas posiciones, como en *mi nuevo coche* y *mi coche nuevo*.
- *Adjetivo que impone modo verbal.* Para los adjetivos que cuando son utilizados en construcciones de tipo *adjetivo que verbo*, imponen el modo subjuntivo del verbo, como *bueno* en *es bueno que vengas* o indicativo como *cierto* en *es cierto que vienes*.

3.2.1.4.4 Determinante

Se engloba en una sola categoría a todos los tipos de determinantes tradicionales: las palabras que suelen preceder a los sustantivos para determinarlos, es decir, para precisar o limitar la extensión de su significado. Se han agrupado los artículos definidos e indefinidos, determinantes posesivos, demostrativos, cardinales, etc., ya que tienen un comportamiento homogéneo en muchas estructuras.

Las diferencias entre ellos se expresan con los rasgos que se explican a continuación, que se utilizan fundamentalmente en las estructuras formadas por varios determinantes seguidos, para expresar las restricciones que cada uno impone a los que le siguen: por ejemplo, se puede escribir *las casas*, *dos casas* y *las dos casas*, pero no *dos las casas*.

El conjunto de rasgos que se aplican a los determinantes es el siguiente:

- *Género*, que puede tomar como valores masculino *el*, femenino *estas* o neutro, cuando pueden acompañar a nombres de ambos géneros, como *mis*.
- *Número*, singular, plural y neutro.
- *Admite construcciones comparativas*, que marca a los determinantes cuyo lema es *tanto*, para permitir modelar construcciones como *tanto café como chocolate*.
- *Tipo de determinante*. Los determinantes se dividen en definidos o indefinidos, y cada uno de ellos, a su vez tiene sus propios rasgos.

  Los *determinantes definidos* se dividen en:

  - Artículo, que agrupa el paradigma de *el*.
  - Demostrativo, con los paradigmas de *este, ese, y aquel*, menos las formas de género neutro, como *esto*.
  - Relativo, con los paradigmas de *cuanto y cuyo*, que siempre necesitan un nombre en sus construcciones, por lo cual no están duplicados como pronombres.
  - Posesivo, incluyendo todas las formas de *mi, tu, su, nuestro y vuestro*. Las formas *mío, tuyo, suyo* están incluidas en el diccionario como adjetivos pospuestos, ya que se comportan como ellos y no como los pronombres/determinantes (p. ej., *los dos coches suyos*).

  Los *determinantes no definidos* se dividen en:

  - Indeterminado, agrupando, entre otras, a *un, cualquier, todo, tanto, mucho, algunos, medio, varios, cierto, etc.*.
  - Cardinal, con los números.
  - Interrogativo, con *qué, cuál y cuánto* y sus formas.

3.2.1.4.5 Pronombre

Agrupa a las palabras que pueden ser núcleos de sintagma nominal por sí mismas, por ejemplo, *tú*. Mayoritariamente coinciden con la clasificación tradicional, aunque se han eliminado ciertas palabras cuyo comportamiento no es homogéneo con el del resto de la categoría.

Parte de las entradas están duplicadas, como pronombres y como determinantes, porque pueden aparecer de forma correcta con ambas funciones en diferentes construcciones: por ejemplo, *ninguna*, que puede ser pronombre, y actuar por sí misma como sujeto, en frases como *ninguna quería ir*, o hacer de determinante en *ninguna chica quería ir*. Sin embargo, otras solamente pueden ser pronombres, como *ninguno*, que no admiten otra palabra para hacer de núcleo del sintagma nominal. Otro grupo de palabras, como *cierta*, siempre necesitan otra palabra para formar el sintagma nominal, por lo cual sólo aparecen como determinantes.

A continuación se muestran los rasgos que pueden acompañar a los pronombres:

- *Género*, que puede tomar como valores masculino para palabras como *él*, femenino para, por ejemplo, *ellas* o neutro para *yo*.
- *Número*, singular, plural y neutro.
- *Tipo de pronombre*, los pronombres pueden ser *personales* o *no personales*.
  - Están marcados como *personales* los pronombres *yo, tú, ellos, etc.*, que designan a personas. Además, tienen flexión de persona (que se manifiesta en la concordancia con el verbo).
    Algunos de estos *pronombres personales* están marcados con el rasgo *no puede seguir a una preposición*. Son, por ejemplo, *tú*, que no puede estar en construcciones del tipo *a tú*. La única construcción en la que se permite que sigan a una preposición es del tipo *entre tú y yo*.
  - Los pronombres *no personales*, como, por ejemplo, *ninguna, cuatro, etc.*, no tienen flexión de persona, y admiten modificadores especificativos (mientras que los *personales* no pueden llevarlos nunca). P. ej.: no se puede decir *tú de las chicas*, pero sí *ninguna de las chicas*, *cuatro de las chicas* o *cuatro de ellas*.
    Estos pronombres de *tipo no personal* se clasifican en *definidos* o *no definidos*. A su vez, cada uno de ellos tiene su propia subclasificación.
    - Los *pronombres definidos* se dividen en:
      - Determinado (paradigma completo de *el*, para construcciones del tipo *el de azul*).
      - Demostrativo (paradigmas de *este, ese, aquel*, con y sin tilde[7]). Su diferencia con los determinantes demostrativos, además de las formas acentuadas, son las formas *esto, eso y aquello* que solamente aparecen como pronombres, ya que nunca admiten otra palabra como núcleo de su sintagma nominal.
      - Relativo que agrupa, a *que, como, cuando, quien, etc.*.
    - Los *pronombres no definidos* se dividen en:
      - Indeterminado que agrupa a palabras como *ninguna, tanta, nada, mucho o poco* entre otras.
      - Cardinal, que incluye algunos números.
      - Interrogativo con *qué, cómo, dónde, etc.*.

3.2.1.4.6 Pronombre que sigue a preposición

Incluye solamente a *mí, ti, sí*, que son formas de los pronombres personales que siempre deben seguir a una preposición, en expresiones como *a mí* o *para ti*, y no pueden aparecer aislados. Tampoco pueden aparecer tras la preposición *con*, porque existen las formas *conmigo, contigo* y *consigo*.

[7] Según [RAE00], "los demostrativos este, ese, aquel, con sus femeninos y plurales, no deben llevar tilde cuando funcionan como adjetivos. Sí pueden llevar tilde cuando funcionan como pronombres, pero esta tilde SÓLO ES OBLIGATORIA cuando existe riesgo de anfibología o posibilidad de confusión entre el uso adjetivo y el pronominal, riesgo que se produce únicamente en casos excepcionales."

3.2.1.4.7 Pronombres con preposición

Agrupa a las palabras *conmigo, contigo* y *consigo*, que tienen un comportamiento particular, como se justifica en el párrafo anterior.

3.2.1.4.8 Clíticos

Contiene los clíticos *los, le, me, etc.*. Estas palabras pueden aparecer solas (*lo come*) o en secuencias de dos (*me lo comí*). Cualquiera de los clíticos puede aparecer sólo, pero cuando aparecen de dos en dos deben mantener cierto orden entre ellos. Hay algunos que solamente pueden ocupar la primera posición (*se*), otros que únicamente pueden aparecer en la segunda (*las*), y un tercer grupo que puede ocupar ambas (*me*). Esta información se ha incluido en el diccionario en forma de rasgo, para poder modelar construcciones como *se las come*, y no admitir otras *las se come*.

3.2.1.4.9 Numeral ordinal

Contiene los ordinales: palabras que indican orden de una forma exacta: *segundo, último, etc.*

3.2.1.4.10 Preposición

Las preposiciones no coinciden con la clasificación tradicional (palabras que hacen de nexo de unión entre palabras, expresando relaciones de dirección, medio, procedencia, etc.) ya que, por una parte, se han añadido algunas palabras cuyo comportamiento es similar al de las preposiciones, y, por otra, se ha segregado alguna de ellas porque tienen un comportamiento particular, como se verá a continuación.

El conjunto de palabras agrupadas en la categoría preposición es: *en, por, con, para, sobre, entre, sin, desde, según, hasta, contra, ante, tras, bajo, hacia, cabe, so, incluso, mediante, excepto y durante*. Como se puede comprobar, se han incluido ciertas palabras en esta categoría, como *mediante o excepto*, que no son preposiciones en la clasificación tradicional, porque tienen un comportamiento similar en ciertos casos al de las demás palabras de esta categoría, esperando un sintagma nominal después: *excepto esos niños* o *mediante una incisión*.

Los comportamientos particulares de estas palabras en otras construcciones pueden ser modelados con reglas específicas para cada una, como se verá más adelante.

No se ha añadido ningún rasgo a estas palabras.

3.2.1.4.11 Artículo contracto

Esta categoría agrupa a las palabras *al* y *del*. Se han separado de las demás preposiciones ya que al tener unido un artículo determinado, restringen el tipo de sintagma nominal que puede seguirlas: será en masculino singular, y considerando las construcciones que pueden seguir a este tipo de artículos.

No se ha añadido ningún rasgo a estas palabras. Las restricciones de género y número se introducen en las reglas que los tratan.

3.2.1.4.12 Preposición a y de

Sólo contiene esas dos preposiciones, y se han separado porque los sintagmas que les siguen no pueden comenzar por la palabra *el*.

No se ha añadido ningún rasgo a estas preposiciones.

3.2.1.4.13 Adverbio

Coincide con la clasificación tradicional, agrupando a las palabras que modifiquen al verbo, como *estupendamente* o *sí*. Se incluyen los adverbios de los diferentes tipos (tiempo, modo, cantidad, etc.).

Admite un rasgo que indica si pueden ser utilizados en construcciones de tipo comparativo (*más, menos, mejor, etc.*) o gradaciones (*nada, poco, algo, etc.*).

3.2.1.4.14 Conjunción

Agrupan a un conjunto de palabras que unen palabras, sintagmas o proposiciones del mismo o distinto nivel sintáctico. Admiten el rasgo *tipo de conjunción* que las divide en conjunciones copulativas como *y, e, o, u, ni, mas* y subordinativas como *que, aunque, como, pero, si, ya, bien, ora, según, pues, porque, etc.*. Se ha añadido *según* a la lista de conjunciones subordinativas, ya que puede encabezar frases subordinadas.

3.2.1.4.15 Interjección

Son frases compuestas por una sola palabra, que sirven para expresar brevemente y de forma repentina impresiones y sentimientos, como dolor, rabia, alegría sorpresa, etc. Se ha introducido un número reducido de ellas, ya que no es un tipo de palabra común en lenguaje escrito. El conjunto de interjecciones presente en el diccionario es: *ay, eh, oh, venga, vaya y ojalá.*

3.2.1.4.16 Cajón de sastre

Esta categoría es especial, ya que las palabras que pertenecen a ella no tienen un comportamiento homogéneo: contiene un conjunto de palabras de difícil clasificación, que requieren un tratamiento individual.

Aunque tradicionalmente cada una de ellas esté incluida en alguna categoría, su comportamiento sintáctico no se corresponde con el de ninguna otra palabra, por lo cual en el módulo de predicción gramatical será necesario elaborar reglas específicas para cada una de ellas: cuando esta categoría aparezca en las reglas, siempre deberá estar especificado el significante al que se refiere.

Se han separado en una categoría específica para que no sean predichas cuando una regla determinada exija cualquiera de las otras categorías.

El conjunto de palabras clasificadas como cajón de sastre es: *sino, embargo, respecto, priori, sabiendas, obstante, san, don, doña.*

Como el tratamiento se hace individualmente para cada palabra, no es necesario asignarles rasgos ya que se pueden exigir los valores adecuados en cada regla, imponiendo las restricciones necesarias.

3.2.1.4.17 Signo

Incluye los signos de puntuación menos el punto.

Están introducidos en las reglas como elementos separadores, marcadores de incisos, en las frases subordinadas, etc. Algunos de los signos contenidos en el diccionario son: *,, :, :, -, , ¡, (,), %, ?, ¿ y !.* En el módulo de predicción por pares y tríos de categorías, también tienen sus estadísticas específicas.

3.2.1.4.18 Punto

Sólo contiene el punto, elemento que se considera la separación entre frases. Cuando aparece un punto en el análisis, éste recomenzará, lanzando desde el principio las reglas que expanden la oración. En los modelos estadísticos categoriales tras un punto se predicen las categorías que más frecuentemente comienzan una frase.

3.2.1.5 Mecanismos de gestión del diccionario general

Como se mencionó anteriormente, el diccionario general se utiliza en dos puntos del proceso de predicción:

- La *categorización* de palabras escritas por el usuario, para averiguar la información gramatical asociada a la misma y utilizarla en los métodos de predicción correspondientes.

- La *predicción* propiamente dicha, en la que se obtienen las palabras predichas con mayor probabilidad, de entre todas las entradas del diccionario.

Cualquiera de los dos procesos puede implementarse realizando un recorrido del diccionario hasta encontrar la o las palabras adecuadas. El inconveniente de esta estrategia es que, dado el tamaño del diccionario (más de 130.000 palabras), un recorrido lineal implica demasiado tiempo y se ralentiza la predicción. Por esta razón, se hace necesario realizar una gestión más eficiente de ambos procesos. Además, los modos de acceso al diccionario necesarios para realizar cada uno de los dos procesos son muy distintos, por lo cual habrá que utilizar técnicas de aceleración diferentes en cada caso.

3.2.1.5.1 Gestión del acceso al diccionario para categorización

La *categorización* de una palabra implica localizarla en el diccionario para obtener la información asociada a ella. Independientemente de la ordenación del diccionario, cualquier solución que implique recorrerlo supone que para categorizar ciertas palabras sea necesario realizar decenas de miles de comparaciones. Además, intentar categorizar una palabra que no está incluida en el diccionario implica recorrerlo por completo (incluso varias veces, si se utilizan los métodos de categorización que se ven en la sección 3.3.2.1.5 sobre palabras desconocidas), aumentando el tiempo de procesamiento.

Para acelerar el proceso, ya que al categorizar se conoce la palabra, se ha modificado el acceso al diccionario para hacerlo a través de una *tabla Hash*.

La categorización de las palabras implica aplicar una sencilla función sobre el código ASCII de las letras, ponderándolo por la posición que ocupan en la palabra para obtener el índice en la tabla Hash. En la tabla puede ocurrir que el índice dado apunte a una posición imposible del diccionario general (en cuyo caso no será necesario recorrer ninguna posición, ya que esto indica que la palabra no está incluida en él), o puede colisionar con otra u otras palabras contenidas en el diccionario, y será necesario recorrer, como máximo, el número máximo de colisiones (menos de 20) hasta encontrar la palabra o darnos cuenta de que estamos en el caso de una palabra desconocida. Este proceso habrá que repetirlo varias veces si se utilizan las técnicas descritas en la sección 3.3.2.1.5 sobre categorización de palabras desconocidas, ahorrando en cada repetición las más de 130.000 comparaciones necesarias para recorrer el diccionario completo.

3.2.1.5.2 Gestión del acceso al diccionario para predicción

El otro punto del sistema en el que se accede al diccionario es la *predicción* propiamente dicha, cuando se genera la lista de palabras predichas. En este caso, es necesario recorrer el diccionario completo para calcular la probabilidad de cada palabra.

Por la naturaleza del problema, es inevitable recorrer el diccionario, aunque se pueden incluir mecanismos que reduzcan el número de palabras a procesar. Para ello se ha implementado un *sistema de índices* paralelo al diccionario. Este sistema consta de varios listados de índices, cada uno de los cuales da acceso a un conjunto limitado de palabras

que cumplen determinadas restricciones. De esta forma se evita recorrer el diccionario completo, reduciendo la búsqueda al subconjunto determinado por el listado de índices correspondiente.

Estos índices dan acceso a las palabras pertenecientes a cada categoría gramatical por separado, y, además, seleccionan las palabras por la primera o las dos primeras letras, para el caso en que esta información esté disponible: p. ej., las preposiciones que empiecen por *a* o los nombres que empiecen por *ni*. Las palabras que tienen varias categorías aparecen en tantos listados de índices como sea necesario.

Además, el acceso a las palabras se realiza por orden frecuencial, según su probabilidad para cada categoría. De esta forma se puede parar el procesamiento cuando la frecuencia de la palabra para una determinada categoría sea lo suficientemente pequeña como para no poder ser candidata ni ella ni ninguna de las siguientes (que tendrán una frecuencia inferior para la categoría que esté siendo procesada).

Estos sistemas de índices consiguen reducir la lista de palabras a procesar de 130.000 al conjunto de las que cumplen las restricciones buscadas, disminuyendo drásticamente el tiempo necesario para procesar la lista. Esta optimización es especialmente beneficiosa en los casos en que las listas de palabras son muy reducidas, o incluso nulas (preposiciones que empiecen por *z*, por ejemplo), porque se pasa de procesar todas las palabras del diccionario, comprobando si cumplen las condiciones buscadas, a recorrer una lista mínima o, incluso, vacía.

3.2.1.6 Limitaciones del diccionario general

El diccionario general es la fuente de información básica para el sistema de predicción. Aporta las frecuencias iniciales y las categorías y rasgos de todas sus palabras. Se deben tener en cuenta, sin embargo, dos factores que hacen que pueda no ser suficiente para conseguir un sistema potente de predicción:

- El vocabulario, tanto en su contenido como en su distribución frecuencial, está adaptado a los textos utilizados para generar el diccionario.
- Carece de capacidad de adaptación.

Estas dos características hacen que el diccionario esté mejor adaptado para predecir textos periodísticos (especialmente sobre política y economía), mientras que el usuario puede desear escribir sobre otros temas, con su vocabulario específico y una distribución frecuencial de las palabras muy diferente. Parte del vocabulario es común y mantiene sus frecuencias altas (como las palabras función: determinantes, preposiciones, etc.), sin embargo, las palabras que aportan el significado (los nombres, verbos y adjetivos) pueden sufrir grandes modificaciones (cuando se escribe de fútbol pueden ser palabras frecuentes *jugador, árbitro o gol*, que no aparecerán, en cambio, en los textos de música o jardinería). Además del tema, hay otros factores que pueden condicionar el vocabulario, como puede ser el propio estilo del usuario, si está escribiendo texto o conversando, etc.

Estas limitaciones del diccionario general justifican la necesidad de introducir módulos *dinámicos*, que adapten su predicción al texto del usuario, con un entrenamiento previo

o, incluso, de forma simultánea al proceso de escritura.

3.2.2 Diccionario personal y diccionarios temáticos

El diccionario personal se genera a la vez que el usuario introduce el texto. Como se verá a continuación, almacena el vocabulario que está utilizando, la frecuencia con que usa cada palabra, etc. De esta forma, permite que la predicción se adapte al estilo del usuario y al tema acerca del cual está escribiendo.

El diccionario personal se genera en cada sesión, empezando vacío al principio, y borrándose cuando acaba. El rendimiento de este diccionario mejora a medida que la sesión avanza, ya que se va introduciendo el vocabulario nuevo y se van ajustando las frecuencias a las reales del tema concreto. El momento en el cual se obtienen los mejores resultados es el final de la sesión, por lo cual, lo más conveniente sería mantenerlo para sesiones sucesivas en las que se escribiera del mismo tema.

Esta posibilidad existe, permitiéndose incluso generar distintos diccionarios sobre temas concretos a partir de textos externos, que pueden no haber sido generados por el usuario. De esta forma, si al usuario le gusta escribir sobre ciertos temas (deportes, coches, trabajo, música), puede entrenar los *diccionarios temáticos* correspondientes a estos temas antes de empezar a escribir, a partir de textos escritos por otra persona, copiados de Internet, etc. En sesiones posteriores, cuando desee escribir sobre un tema concreto, activará el diccionario adecuado y dispondrá, desde el principio de la sesión, de la potencia que el diccionario personal le proporciona al final de un texto, es decir, tras su entrenamiento.

El diagrama de bloques de los módulos encargados de gestionar ambos diccionarios es el mostrado en la figura 3.3.

Figura 3.3: Diagrama de bloques del diccionario personal y los temáticos.

En las secciones siguientes se explica cada módulo, las relaciones entre ellos y las fuentes de entrenamiento usadas en el proceso de generación.

3.2.2.1 Textos de entrenamiento del diccionario personal y de los temáticos

El diccionario personal se genera a partir del texto escrito en la sesión en curso: al principio de cada sesión el diccionario está vacío, porque no hay texto, y se completa a medida que éste se introduce.

Los diccionarios temáticos se pueden generar a partir de cualquier texto. Si el usuario desea entrenar un diccionario para escribir sobre un tema determinado, los textos deben tratar de dicho tema, en caso contrario, los resultados serán peores que si no se utiliza diccionario temático, como se comprobará en el capítulo de evaluación.

3.2.2.2 Contenido del diccionario personal y de los temáticos

El contenido de ambos tipos de diccionarios es el siguiente:

- Cada palabra utilizada en el texto origen, denominada *palabra principal*.
- Su *frecuencia absoluta* en dicho texto.
- Las palabras que hayan seguido a la palabra principal en el texto (*palabras siguientes*, para la construcción de tablas de *pares de palabras*) y la frecuencia absoluta de aparición de las mismas. Estas palabras se almacenan en orden *frecuencial* y, dentro de las palabras de la misma frecuencia, por orden *temporal*, dando mayor preferencia a las últimas palabras escritas.
- *Información gramatical* asociada a la palabra principal. Se incluye el lema, categoría, rasgos y frecuencia del diccionario general (si están disponibles), de todos los homógrafos con diferente información gramatical de la palabra. En caso contrario se aplican las rutinas de categorización de palabras desconocidas, que estimarán el valor de cada uno de los campos, como se explica en la sección 3.3.2.1.5 (página 65).

Con esta información se consigue adaptar el vocabulario al tema específico del texto que está siendo escrito (ya que se incluye cada palabra y su frecuencia) y, además, la información de las palabras siguientes permite utilizar bigramas de palabras[8] en la predicción.

3.2.2.3 Generación del diccionario personal y de los temáticos

El diccionario personal, como se comentó anteriormente, se genera simultáneamente a la escritura del texto. Hemos seguido la siguiente metodología de consignación de información cada vez que se completa la escritura de una palabra:

[8] Véase la sección 2.2.1, en la que se discuten los modelos de lenguaje estadísticos.

- Si la palabra no está ya en el diccionario personal, se incluye, y se añade la información gramatical del diccionario general, o se categoriza automáticamente si es desconocida.

- Se incrementa su frecuencia absoluta en el diccionario personal.

- Se reordenan los índices de acceso al diccionario por orden frecuencial (su necesidad se mostrará en la sección siguiente, sobre los mecanismos de gestión del diccionario).

- Se introduce la palabra en la lista de palabras siguientes de la anterior a ella si es necesario.

- Se actualiza su frecuencia como palabra siguiente, en la entrada correspondiente a su palabra anterior.

- Se reordenan las palabras siguientes para que la última introducida sea la primera entre las de su misma frecuencia (información temporal).

Este proceso se repetirá cada vez que se complete una palabra o un signo (también los signos son incluidos en el diccionario con su conjunto de palabras siguientes y su frecuencia).

Tanto la longitud del diccionario como el número de palabras que pueden seguir a la palabra principal son variables, por lo cual, es necesario incluir en el algoritmo de generación mecanismos que permitan adaptar dichas longitudes a las características del texto que está siendo escrito, de forma óptima.

Si se desea *generar* un diccionario temático a partir de un texto dado, se crea el diccionario personal de la misma forma que si el usuario estuviera escribiendo dicho texto, y se almacena al final con un formato determinado, que permite posteriormente recuperar toda la información. A cada diccionario temático se le asigna un nombre que permite distinguir entre los diferentes diccionarios que puede crear el usuario.

Además, los diccionarios temáticos se pueden *actualizar* añadiendo nuevos textos. Para ello se lee el temático inicial (creado anteriormente, a partir de otros textos), se añade el contenido del nuevo texto de entrenamiento al diccionario (con el mismo proceso anterior) y se almacena el resultado con el mismo nombre.

3.2.2.4 Gestión del diccionario personal y de los temáticos

Tanto para el diccionario personal como para los temáticos se ha implementado un sistema de índices similar al del diccionario general, para reducir el número de palabras a procesar y acelerar así el proceso de predicción. Los beneficios producidos por esta aceleración serán mayores en los diccionarios temáticos, por el mayor tamaño que pueden alcanzar, ya que pueden ser entrenados a partir de gran cantidad de texto. Considerando la velocidad de escritura de los usuarios con discapacidad física y el vocabulario medio que se utiliza generalmente para escribir, es muy difícil que la longitud del diccionario personal sea superior a las mil entradas (es decir, mil palabras diferentes escritas en una sola sesión).

3.2.2.5 Problemática del diccionario personal y de los temáticos

Aunque el diccionario personal tiene ventajas sobre el diccionario general por su capacidad de adaptación, esta misma flexibilidad resulta problemática de cara a determinados usuarios.

La capacidad de aprendizaje hace que el diccionario almacene todas las palabras escritas por el usuario. Esto es especialmente beneficioso cuando el usuario escribe sobre temas técnicos, con gran cantidad de vocabulario específico, o cuando lo utiliza para comunicarse, usando motes, nombres de personas o sitios, diminutivos, etc. Estas palabras, que no están incluidas en el diccionario general, se predecirán desde la primera vez que se escriban.

El problema reside en que, de la misma forma que se aprende las palabras nuevas, también almacena las faltas de ortografía, y las predice posteriormente. Este comportamiento hace que las personas que comentan numerosas faltas (niños aprendiendo a escribir, estudiantes de segunda lengua, etc.) vean fomentadas las palabras escritas incorrectamente. En el caso de las personas que utilicen interfaces controladas por pulsador, también pueden escribir las palabras de forma incorrecta por problemas con el barrido, y no corregirlas para evitar la pérdida de tiempo. Estas palabras también se incluyen en el diccionario, dando lugar a predicciones no deseadas.

Por esta razón, aunque los resultados (como se muestra en el apartado correspondiente) de la predicción del diccionario personal sean muy satisfactorios, hemos visto necesario el estudio de la inclusión de mecanismos de control en la arquitectura que eviten este efecto negativo.

Hemos introducido dos métodos de control en la arquitectura:

- Eliminación total de la capacidad de aprendizaje de palabras nuevas: el diccionario personal adaptará las frecuencias al vocabulario que está siendo escrito, *siempre que la palabra esté incluida en el diccionario general.* En caso contrario, el diccionario personal ignorará dicha palabra. De esta forma se asegura que ninguna falta ortográfica es incluida en el diccionario ni se predice. Aunque las palabras nuevas no se predigan, y haya que escribirlas por completo cada vez que se utilicen, la adaptación de la frecuencia de las palabras conocidas justifica la existencia del diccionario, como se demuestra en el capítulo de resultados.

  Este mecanismo de control está especialmente indicado para los casos en que las faltas de ortografía sean muy frecuentes.

- Para las personas que escriben correctamente, cometiendo faltas esporádicamente, se puede incluir un método de control menos drástico, que les permita beneficiarse de la predicción de las palabras nuevas. Para ello, se introducen en el diccionario todas las palabras, pero no se utilizan en la predicción hasta que no han sido introducidas por el usuario al menos dos veces (o cualquier otro límite inferior que sea razonable, dependiendo del grado de confianza del usuario).

  Si la palabra es nueva, y está bien escrita, las primeras veces que se introduzca no se utilizará para predecir, pero se acumulará su frecuencia hasta que alcance el límite. A partir de ese momento se utilizará como si fuera una palabra conocida. Si el

usuario escribe mal una palabra por un error de precisión en el uso del pulsador, por ejemplo, es difícil que cometa exactamente el mismo error cada vez que la escriba, por lo cual su frecuencia no se incrementará, y no se fomentará el error mostrando la palabra en el menú. De esta forma, el sistema aprenderá las palabras nuevas, usándolas con cierto retardo, y se evitará, en parte, el ruido introducido por las palabras incorrectamente escritas.

Como se acaba de explicar, entre el diccionario general, personal y temáticos se incluye información probabilística, gramatical, temporal y de estilo del usuario:

- Probabilística: información de la frecuencia de la palabra (considerando tanto la frecuencia en el diccionario general como la particular del usuario).
- Gramatical: el diccionario general incluye información tanto de la o las categorías a las que pertenece cada palabra, como de los rasgos asociados a las mismas.
- Estilo del usuario: el sistema predictor consigue adaptarse al estilo del usuario al aprender su vocabulario, y la frecuencia de uso de cada palabra.
- Temporal: ya que a igualdad de frecuencia, prevalecen las últimas palabras que el usuario ha escrito (el uso de una palabra hace que sea más probable volver a utilizarla en el mismo texto). Tienen así prioridad en la predicción las más recientemente utilizadas.

En las secciones siguientes se verá cómo utilizan esta información los diferentes métodos de predicción incluidos en el sistema.

3.3 Métodos de predicción

En la sección anterior se han descrito las fuentes de información en las que se basa el sistema de predicción de palabras propuesto en esta tesis, pero existen varias maneras de utilizar dicha información para obtener las palabras predichas.

El funcionamiento general de todos los métodos de predicción de este sistema es el siguiente:

1. Cada método sugiere una lista de palabras (todos pueden predecir una lista incluso al principio del texto), para lo que utilizan distintas fuentes de información obtenidas a partir del texto en curso:
    - Los caracteres escritos de la palabra en curso, dado que las palabras predichas tienen que comenzar, al menos, por lo escrito por el usuario. Esta información es la única utilizada por todos los métodos.
    - La palabra o palabras anteriores, que son utilizadas, por ejemplo, por los métodos basados en bigramas de palabras.

- La categoría de la o las palabras anteriores, utilizadas en los métodos basados en pares o tríos de categorías (que no necesitan conocer el significante).
- La frase completa que se está escribiendo, tanto como secuencia de categorías, como sus lemas y significantes, para los métodos basados en gramáticas formales.

2. Si alguna palabra de la lista de predicciones es la deseada, el usuario la elige y se inserta en el texto que está siendo escrito, como si hubiera sido escrita por el procedimiento habitual.

3. Si ninguna de las palabras es la adecuada, el usuario escribirá una nueva letra, y el proceso comenzará de nuevo.

A continuación se muestran los diferentes métodos que se han estudiado e implementado, sus necesidades en cuanto a fuentes de información, sus exigencias y su modo de funcionamiento. Se explican por orden de complejidad, indicando en cada caso sus puntos fuertes y débiles y justificando el uso de métodos más complicados para superar las deficiencias de métodos más sencillos y aumentar progresivamente la calidad de la predicción.

Se comienza por el método basado simplemente en información frecuencial, para añadir posteriormente información gramatical básica, e irla ampliando hasta llegar a los sistemas más complejos, basados en el uso de gramáticas formales y su combinación con estrategias probabilísticas.

3.3.1 Métodos estadísticos básicos

El método más sencillo de predicción planteado se basa en el uso de *unigramas*[9], es decir, la frecuencia absoluta de cada palabra. La única información que necesita es la parte ya escrita de la palabra en curso, y la lista de palabras que genera es el conjunto de palabras más frecuentes que comienzan por dicha cadena de letras.

Cuando se aplica este método sobre el diccionario general, las palabras predichas tras un espacio siempre son las mismas (las más probables en los textos de entrenamiento: *de*, *la*, *el*, *que* y *en*). De igual forma, tras una determinada letra siempre se predice lo mismo, independientemente del usuario, del tema y de la posición dentro de la frase que se esté escribiendo: por ejemplo, tras escribir *los l*, se predicen como palabras más probables *la*, *los*, *las*, *lo* y *le*, ya que son las palabras más frecuentes del diccionario que comienzan por *l*.

Cuando se aplica este método sobre el diccionario personal, se predicen las palabras más probables que el usuario ha utilizado, mejorando su rendimiento, como se comprobará en la sección de evaluación. Al contrario de lo que ocurre con el diccionario general, como el diccionario personal es dinámico, la lista de palabras más frecuentes en cada punto se modifica a medida que se escribe el texto.

Al aplicar este método de predicción sobre los distintos diccionarios se obtienen hasta tres listas de palabras predichas: una procedente del diccionario general, otra del personal,

[9] *Unigram*, en la literatura de habla inglesa.

y otra del diccionario temático activo (si el usuario está utilizando alguno). El módulo de coordinación[10] se encarga de combinarlas, como se verá en secciones posteriores.

Aunque el método basado en *unigramas* fue el primero en implementarse, su escasa potencia predictiva hace que en la actualidad solamente se utilice cuando los demás métodos no predicen el número suficiente de palabras.

Con la misma filosofía del modelo *unigram*, se puede ampliar la cantidad de información utilizada en la predicción, considerando secuencias de palabras: pares (*bigramas*), tríos (*trigramas*), etc. A la hora de predecir se considera, no sólo la secuencia de letras de la palabra en curso, sino también la o las palabras precedentes. Utilizando el ejemplo anterior, tras escribir *los l*, se predicen las palabras que comiencen por *l* que más frecuentemente hayan seguido a la palabra *los* en los textos de entrenamiento.

El problema que tienen estos métodos es la cantidad de texto necesario para entrenar el sistema, ya que debe ser suficiente para que cada secuencia válida de palabras aparezca un número de veces relevante, estadísticamente hablando. Necesitan, además, gran cantidad de recursos del sistema, tanto para generar los diccionarios como para manejarlos, especialmente si el tamaño de los mismos es grande. Estos problemas aumentan si la cantidad de palabras a considerar en la secuencia crece.

Por las razones anteriormente expuestas, en la predicción de palabras sólo se han incluido bigramas (no se contemplan secuencias más largas), y exclusivamente en el diccionario personal y los temáticos. En el diccionario general no se consideran por los excesivos recursos necesarios para gestionar las secuencias posibles en las más 130.000 palabras distintas que contiene.

3.3.1.1 Limitaciones de los métodos estadísticos básicos

La baja eficacia del método basado en unigramas en la predicción se debe fundamentalmente a la escasa información que considera: no tiene en cuenta las relaciones de cada palabra con las palabras anteriores, por lo cual realiza predicciones que carecen de sentido en el contexto de la frase. Además, cuando se aplica al diccionario general, la lista de palabras predichas depende de información frecuencial no adaptada a los textos del usuario.

Al intentar mejorar el sistema (considerando mayor cantidad de información), realizando las estadísticas sobre secuencias de palabras, aumenta la cantidad de textos y recursos necesarios para gestionar el sistema, especialmente si se contempla gran cantidad de palabras (como en el diccionario general), o en secuencias de mayor longitud.

Si no se dispone de la cantidad suficiente de textos de entrenamiento, el rendimiento del sistema disminuye, por falta de datos, y es necesario incluir mecanismos que reduzcan dichos efectos negativos. En apartados posteriores se explica cómo, en este sistema, se ha optado por agrupar las palabras en categorías, según su comportamiento, dando lugar a los modelos que se explican a continuación.

[10] Véase la figura 3.1 en la página 32.

3.3.2 Modelos estadísticos categoriales

Estos modelos solucionan parte del problema de los métodos estadísticos básicos, reduciendo la cantidad de texto necesario para ser entrenados. Además, introducen información gramatical de corto alcance en el proceso de predicción. Como es bien conocido, todas las frases tienen una estructura y las categorías gramaticales de las palabras consecutivas están relacionadas: no todas las categorías pueden aparecer en cualquier punto de la frase, ni siguiendo a cualquier otra. Los modelos estadísticos categoriales se basan en calcular las probabilidades de cada secuencia de categorías, e introducir esa información en el proceso de predicción.

El principal problema de estos métodos es la calidad de la información adicional que necesitan:

- Hay que añadir a cada palabra toda la información gramatical que sea necesaria.
- Es necesario introducir las relaciones entre las categorías gramaticales (la probabilidad de cada secuencia).
- Hay que modificar el predictor, para utilizar dicha información.

La información gramatical de las palabras que se va a utilizar es la contenida en el diccionario general. Posiblemente este método conseguiría mejores resultados con un conjunto de categorías diferente, pero se ha dado prioridad a la clasificación que está mejor adaptada al método basado en gramáticas formales.

Una vez conocido el conjunto de categorías asignadas a las palabras, se deben expresar las relaciones entre dichas categorías. El modelo más sencillo considera solamente las relaciones entre pares de categorías consecutivas. Este método se entrena realizando estadísticas sobre los pares de categorías (*bipos*[11]) que aparecen consecutivamente en un texto de entrenamiento determinado (corpus etiquetado según dicho sistema de categorías). La información se codifica en forma de matriz, que contiene, para cada par de categorías, la probabilidad de que cada una siga a la anterior:

$$p(C_t/C_{t-1}) = \frac{Frec(C_{t-1}\ C_t)}{Frec(C_{t-1})} \tag{3.1}$$

Donde $Frec(C_{t-1}\ C_t)$ es el número de veces que aparece la secuencia $C_{t-1}\ C_t$ en el texto de entrenamiento, y $Frec(C_{t-1})$ es el número de veces que aparece la categoría C_{t-1}.

En la predicción del sistema actual se utilizan las estadísticas obtenidas a partir del texto de entrenamiento etiquetado PAIS, que se describe en el apéndice A (página 179) sobre los corpora de entrenamiento.

Una vez conseguida la información sobre los pares de categorías, es necesario implementar un sistema para utilizar dicha información en la predicción, dando preferencia a las palabras pertenecientes a las categorías que más frecuentemente siguen a la categoría de la palabra anterior.

[11] Del inglés *bi-Part-Of-Speech*, acrónimo que usaremos en el resto del documento por comodidad

El procedimiento que se utiliza es calcular la probabilidad de cada palabra, condicionada por la información sobre la o las categorías a las que pertenece y la o las categorías de la palabra anterior, y presentar como palabras predichas las más probables. El proceso para calcular dicha probabilidad a partir de la información frecuencial básica, se basa en las ideas que se pueden encontrar en [Jeli91], [Ricc96] y [Alle94], adaptadas a este trabajo. La fórmula resultante, suponiendo que la relación entre las palabras de la secuencia depende solamente de sus categorías gramaticales, es:

$$p(w_t/w_{t-1}) = \sum_{\forall C_{t-1}^j \in \mathcal{C}_{t-1}} \sum_{\forall C_t^i \in \mathcal{C}_t} p(w_t/C_t^i) \cdot p(C_t^i/C_{t-1}^j) \cdot p(C_{t-1}^j/w_{t-1}) \qquad (3.2)$$

Donde:

- $p(w_t/w_{t-1})$ es la probabilidad de que la palabra w_t siga a w_{t-1}.
- $p(C_t^i/C_{t-1}^j)$ es la probabilidad de que la categoría C_t^i siga a la categoría C_{t-1}^j. Se obtiene de la matriz de *bipos*.
- $p(w_t/C_t^i)$ es la probabilidad de predecir la palabra w_t dada la categoría C_t^i. Esta probabilidad depende, entre otros factores, de la frecuencia absoluta de la palabra en el diccionario que se considere, y, por lo tanto, será diferente si la información se obtienen del diccionario general, personal o de los temáticos. Este método se aplicará tanto a unos diccionarios como a otros y, posteriormente, el módulo de coordinación considerará ambas probabilidades a la hora de generar la lista definitiva de palabras predichas.
- $p(C_{t-1}^j/w_{t-1})$ es la probabilidad de que la palabra w_{t-1} esté categorizada como C_{t-1}^j. Se obtiene del diccionario general.
- $\mathcal{C}_{t-1}$ y $\mathcal{C}_t$ son los conjuntos de categorías a las que pertenecen, respectivamente, w_{t-1} y w_t, ya que ambas pueden pertenecer a más de una categoría, y esto debe ser considerado al calcular la probabilidad de las palabras predichas.

La fórmula 3.2 permite calcular la probabilidad de cada palabra y debe aplicarse a todas las del diccionario. Posteriormente se seleccionan las más probables, y se presentan al usuario como palabras predichas.

El proceso expuesto utiliza la probabilidad de que cada categoría siga a la o las de la palabra anterior y, al incluir esa información en la fórmula, modifica la lista de palabras predichas en cada punto de la oración, dependiendo de la palabra anterior.

Al contrario que en el método basado en frecuencia simple descrito en el apartado anterior, tras un espacio en blanco, la lista de palabras mostrada dependerá de la palabra anterior. Por ejemplo, al aplicar este método al diccionario general, al principio de la frase se predicen palabras de las categorías que más frecuentemente comienzan una oración, en este caso pronombres, preposiciones y determinantes: *la*, *el*, *de* y *los*. Cuando el usuario escribe *las c*, tendrían preferencia en la lista de palabras predichas (a partir de dicho diccionario general), los nombres: *columna*, *congreso*, *comisión* y *caso*.

Al igual que los métodos estadísticos, éste se puede aplicar a todos los diccionarios, y el módulo de coordinación se encargará de decidir cuál es la lista de palabras predichas definitiva, a partir de las que se predicen según cada diccionario.

Este método es extensible. Para pasar a considerar un número mayor de categorías en la predicción, en lugar de basarse solamente en una categoría anterior, se pueden utilizar más categorías en los cálculos y predecir las que siguen más frecuentemente a las tres, cuatro o, en general, n anteriores. Las modificaciones que se deben realizar en el sistema son:

- Durante el entrenamiento: en lugar de contabilizar los pares de categorías que aparecen de forma consecutiva en el texto, se amplia el rango, y se buscan los grupos de tres, cuatro o más categorías. El número de categorías a considerar está limitado, fundamentalmente, por el volumen de texto categorizado del que se dispone. En este trabajo se han utilizado estadísticas sobre tríos de categorías gramaticales (*tripos*[12]), y la ecuación 3.1 se modifica para considerar la categoría de las dos palabras anteriores:

$$p(C_t/C_{t-2}C_{t-1}) = \frac{Frec(C_{t-2}C_{t-1}C_t)}{Frec(C_{t-2}C_{t-1})} \tag{3.3}$$

  Esta información se almacena en el sistema en forma de matriz, denominada *matriz de tripos*.

- Durante la predicción: es necesario modificar la expresión 3.2 de la probabilidad de cada palabra para incluir la información de la penúltima escrita (w_{t-2}). Para ello:

  - Se añade el término correspondiente a la probabilidad de que dicha palabra (w_{t-2}) pertenezca a cada una de sus categorías posibles, y se incluye un nuevo sumatorio que las recorra todas.
  - La probabilidad de la categoría predicha pasa a estar condicionada por la categoría de las dos palabras anteriores.

  Con la inclusión de estos dos términos, la expresión pasa a ser:

$$p(w_t/w_{t-2}w_{t-1}) = \sum_{C^k_{t-2}\in \mathcal{C}_{t-2}} \sum_{C^j_{t-1}\in \mathcal{C}_{t-1}} \sum_{C^i_t\in \mathcal{C}_t} p(w_t/C^i_t)\cdot p(C^i_t/C^k_{t-2}C^j_{t-1}) \cdot p(C^j_{t-1}/w_{t-1})\cdot p(C^k_{t-2}/w_{t-2}) \tag{3.4}$$

Al igual que con los *bipos*, una vez se haya calculado la probabilidad de cada palabra, se eligen las más probables y se presentan al usuario.

El proceso de generalización que se ha realizado para pasar de *bipos* a *tripos* puede ser aplicado tantas veces como categorías se desee considerar en la predicción, aumentando así la cantidad de información que se considera en el proceso.

[12] Del ingles, *tri-Part-Of-Speech*, acrónimo que usaremos en el resto del documento por comodidad

Con este método, una alta frecuencia de palabra puede ser reducida por una baja probabilidad de categoría gramatical y viceversa. El resultado es que las palabras más frecuentes de las categorías más probables son elegidas, consiguiendo así una mayor precisión en la predicción.

3.3.2.1 Mejoras introducidas en los modelos categoriales básicos

3.3.2.1.1 Tratamiento de rasgos

Los métodos categoriales descritos (*bipos* y *tripos*), proporcionan información que permite calcular la probabilidad de la categoría de las palabras siguientes. Sin embargo, como se ha visto en la sección correspondiente al diccionario general, una gran parte de la información sobre cada palabra se expresa utilizando rasgos.

Por el gran número de rasgos incluidos y su complejidad, no es factible introducir la gestión de todos ellos en paralelo con los *bipos* y *tripos*. Además, los valores de parte de los rasgos dependen de relaciones de larga distancia, fuera del alcance de los *tripos*. Por ejemplo, en *el niño es muy bueno*, los rasgos de *bueno*, dependen de *niño*, pero los tripos sólo consideran *es* y *muy* al predecir *bueno*, por lo cual no tienen información suficiente como para asignar a los rasgos el valor adecuado.

Algunos rasgos, sin embargo, dependen de la o las palabras inmediatamente anteriores, como el género y el número de los adjetivos y nombres del sintagma nominal, que deben concordar entre sí. Es posible introducir un mecanismo de control que permita asignar el valor adecuado a los rasgos en determinadas condiciones. Para ello se tendrá en cuenta que:

- Dos palabras *concuerdan* si el género y el número de ambas coinciden (p. ej.: *la* y *casa* concuerdan porque ambas son palabras femeninas y singulares).
- Dos palabras *concuerdan* si alguna de las dos tiene género y/o número neutros. Por ejemplo, *tesis* es una palabra neutra en número, porque tiene la misma forma para el singular que para el plural, y concuerda tanto con *la* como con *las*. *Valiente* es una palabra de género neutro y puede concordar tanto con palabras femeninas como masculinas.

En *la casa de los espíritus*, *la* y *casa* concuerdan, y lo mismo *los* y *espíritus*. La palabra *de* no transmite ni aporta información de concordancia.

No se contempla la concordancia para el caso de los verbos, ya que en castellano no siempre el sintagma nominal que precede a un verbo es su sujeto y no siempre deben concordar. Siguiendo con el ejemplo anterior, si el usuario escribe *La casa de los espíritus está en la colina*, *está* no concuerda con *los espíritus*, sino con *la casa*, por lo que sería erróneo predecir verbos en plural obligándolos a concordar con el último nombre (*espíritus*): se restringirían los rasgos de forma incorrecta, y sería contraproducente. Como los *bipos* y *tripos* no disponen de información suficiente como para predecir los rasgos en casos como éste (por no tener información de largo alcance, sino solamente de la o las dos últimas categorías), se ha decidido no asignarle ningún valor al número en los verbos.

Por lo tanto, la gestión de rasgos que se realiza en paralelo con los *bipos* y *tripos* consiste en comprobar si concuerdan en género y número los *artículos contractos (*al *y* del*, que exigen masculino singular), determinantes, adjetivos y nombres* con los *adjetivos y nombres* que les siguen. Independientemente de si las categorías se predicen considerando una o dos palabras anteriores, en la concordancia de rasgos sólo se considera la palabra anterior.

Esta información se ha introducido en la arquitectura de predicción en forma de filtro que rechaza las palabras que no concuerdan. La metodología concreta que hemos propuesto es la siguiente:

- Se predicen las categorías siguientes con los *bipos* y *tripos* tal y como se ha explicado anteriormente, y se asignan las probabilidades adecuadas a cada una.
- Se recorre el diccionario y, para cada palabra:
  - Se comprueba si la palabra es determinante, adjetivo o nombre, y si la palabra anterior es determinante, adjetivo, nombre o artículo contracto.
  - En caso afirmativo, se verifica si ambas palabras concuerdan en género y número.
  - Si la palabra no concuerda, se rechaza (filtra) y se vuelve a comenzar el proceso con la siguiente palabra.
  - Si la palabra concuerda, o no es necesario verificar su concordancia, se le asigna su probabilidad y se sigue el proceso habitual con los *bipos/tripos* descrito anteriormente.

Como ejemplo de los resultados de los *bipos* con filtrado de rasgos, al aplicarlos al diccionario general, tras escribir *un d*, se predicen las palabras *día*, *director*, *domingo*, y *dinero*, todas ellas concordantes en género y número con *un*.

En el apartado 4.1.3.3.4 (página 133) del capítulo de evaluación se realizan experimentos con y sin tratamiento de rasgos, pudiéndose apreciar la mejora que producen.

3.3.2.1.2 Técnicas de suavizado de matrices

Uno de los problemas fundamentales de los modelos de lenguaje estadísticos se debe a la falta de suficientes datos para calcular de forma fiable las probabilidades a usar. Por muy grandes que sean los textos de entrenamiento, no contienen todas las combinaciones válidas de categorías, lo que se materializa en la aparición de valores nulos en las matrices de *bipos* y *tripos*, que prohíben las secuencias de palabras pertenecientes a las categorías implicadas. Para evitar el efecto negativo de los ceros en estas matrices, es necesario usar algún método que asegure que todas las probabilidades son distintas de cero.

En la literatura se presentan varias alternativas, como, por ejemplo, el suavizado umbral (*floor smoothing*) [Rabi89], el descuento lineal (*linear discounting*), el retroceso a modelos más sencillos (*back off*) [Katz87] y [Jeli91], la interpolación lineal [Nada84], etc. En [Mart99] se puede encontrar una comparación entre varias de estas técnicas en función

de la reducción en perplejidad que proporcionan y sus necesidades computacionales. Dos de las técnicas más sencillas se explican a continuación:

- El *retroceso a estadísticos sobre secuencias de categorías de longitud inferior*: si falla el valor del *n-pos* buscado[13], se usará en su lugar el valor del *(n-1)-pos*. Es decir, si el valor en la matriz de *tripos* es cero, se utilizará el valor de la matriz de *bipos*, considerando únicamente la categoría precedente, y, si también falla el *bipos*, se usa la frecuencia de cada categoría (*unipos*).

  En el caso de los *tripos*, se tiene:

$$p_r(C_i/C_kC_j) = \begin{cases} p(C_i/C_kC_j) & si\ p(C_i/C_kC_j) \neq 0 \\ p(C_i/C_j) & si\ p(C_i/C_kC_j) = 0\ y\ p(C_i/C_j) \neq 0 \\ p(C_i) & si\ p(C_i/C_kC_j) = 0\ y\ p(C_i/C_j) = 0 \end{cases} \tag{3.5}$$

  y para el caso de los *bipos*:

$$p_r(C_i/C_j) = \begin{cases} p(C_i/C_j) & si\ p(C_i/C_j) \neq 0 \\ p(C_i) & si\ p(C_i/C_j) = 0 \end{cases} \tag{3.6}$$

  En ambos casos, $p_r(x)$ se refiere a la probabilidad que se introduce en las fórmulas 3.4 y 3.2 para el cálculo de la probabilidad de la palabra.

- Otro posible método para eliminar los ceros en las matrices de *bipos* y *tripos* es el *suavizado umbral*, que consiste en sumar un cierto valor ϵ a los elementos de la matriz. A continuación se normaliza, para que la suma de las probabilidades sea la unidad. El umbral considerado, en nuestro caso, es el valor mínimo de la matriz (que sea distinto de cero), y se suma a todos sus elementos y no sólo a los ceros, para evitar la penalización innecesaria de los valores no nulos.

$$\epsilon_{bi} = min\{\ p(C_i/C_j)\ /\ p(C_i/C_j) \neq 0,\ \forall\ i,\ j\} \tag{3.7}$$

$$p'(C_i/C_j) = p(C_i/C_j) + \epsilon_{bi},\ \forall\ i,\ j \tag{3.8}$$

$$p_s(C_i/C_j) = \frac{p'(C_i/C_j)}{\sum_{\forall\ j} p'(C_i/C_j)},\ \forall\ i,\ j \tag{3.9}$$

  Donde ϵ_{bi} es el umbral elegido y $p_s(C_i/C_j)$ es la probabilidad de la matriz de *bipos* suavizada para la secuencia de categorías $C_j\ C_i$.

  De la misma forma, para el caso de los *tripos*, tenemos:

$$\epsilon_{tri} = min\{\ p(C_i/C_kC_j)\ /\ p(C_i/C_kC_j) \neq 0,\ \forall\ i,\ j,\ k\} \tag{3.10}$$

$$p'(C_i/C_kC_j) = p(C_i/C_kC_j) + \epsilon_{tri},\ \forall\ i,\ j,\ k \tag{3.11}$$

[13] *Fallo*: en este contexto se refiere a que no se ha encontrado ningún ejemplo del estadístico considerado en el corpus de entrenamiento, es decir, su probabilidad es 0.

$$p_s(C_i/C_kC_j) = \frac{p'(C_i/C_kC_j)}{\sum_{\forall\ j,\ k} p'(C_i/C_kC_j)},\ \forall\ i,\ j,\ k \tag{3.12}$$

De nuevo, $p_s(x)$ se refiere a la probabilidad que se introduce en las fórmulas 3.4 y 3.2 para el cálculo de la probabilidad de la palabra.

Igualmente se ha verificado el comportamiento de técnicas más refinadas de suavizado: las disponibles en las herramientas de modelado de lenguaje de CMU y Cambridge [Clar97] (descuento lineal y absoluto [Ney94], Good-Turing [Katz87] y Witten-Bell [Witt91]). Dichas técnicas mejoran ligeramente los resultados obtenidos con la de suavizado umbral, pero siguen siendo peores que las que usan tripos sin suavizar con retroceso. Esta es la razón de que se haya optado por abandonar la línea de técnicas de suavizado.

3.3.2.1.3 Entrenamiento con textos categorizados con ambigüedad

Como se comentó anteriormente, uno de los mayores problemas de los métodos categoriales está relacionado con la necesidad de tener una gran cantidad de textos categorizados exactamente según el sistema de categorías contenido en el diccionario. Este problema es mayor cuanto más larga sea la secuencia de categorías que se desee considerar y cuanto mayor sea el número de categorías del sistema.

Debido a la ambigüedad del lenguaje, la categorización de los textos debe hacerse de forma manual (un proceso muy costoso si se persigue una precisión absoluta, especialmente en casos como el que nos ocupa, en que el sistema de categorías y rasgos es complejo), que además, debe ser repetido si se modifica el esquema de categorías. Una vez categorizado el texto concreto, el entrenamiento queda limitado a dicho texto, es decir, adaptado a su estilo, y, con la calidad que permita obtener el número de palabras que contenga.

En el sistema actual, las matrices de bipos y tripos han sido entrenadas a partir de un texto de más de 50000 palabras etiquetado sin ambigüedad. Además, se han realizado pruebas entrenando con textos de mayor longitud, categorizados de forma automática, a partir de la información gramatical del diccionario general, y, por tanto, conteniendo todas las posibles categorías de cada palabra y la probabilidad de cada una.

En el entrenamiento sin ambigüedad, cada secuencia de palabras se corresponde con una secuencia de categorías determinada y, cada vez que aparece, aumenta en una unidad la cuenta de dicha secuencia, para posteriormente aplicar la fórmula 3.1 o 3.3. Al entrenar con ambigüedad, cada lista de palabras se puede corresponder con varias secuencias diferentes de categorías, y debe aumentar la cuenta de cada una de ellas considerando la probabilidad de que cada palabra pertenezca a cada categoría, es decir, en lugar de aumentar en una unidad la frecuencia de la secuencia $C_j\ C_i$, el incremento es $p(C^j_{t-1}/w_{t-1}) \cdot p(C^i_t/w_t)$. La misma formulación es extensible al caso de trigramas entrenados con ambigüedad.

Cabe esperar que los resultados no sean tan buenos como los de los modelos entrenados a partir de textos categorizados sin ambigüedad, pero si las diferencias no son grandes, permitirían generar las matrices de forma sencilla, con un esfuerzo mínimo (ya que la categorización con ambigüedad es un proceso automático), y se podrían adaptar fácilmente

a textos de usuarios o estilos diferentes, es decir, el proceso dejaría de estar limitado a la disponibilidad de un texto determinado que es necesario categorizar de forma manual.

En el apartado 4.1.3.3.3 (página 131) del capítulo de evaluación se pueden ver los resultados obtenidos por los modelos entrenados con ambigüedad.

3.3.2.1.4 Mecanismos de aceleración

Cualquier modelo categorial implica recorrer el diccionario por completo calculando la probabilidad de cada palabra, para obtener la lista de palabras más probables. Este proceso exige demasiado tiempo, y es necesario acelerarlo. Para ello se utilizan los sistemas de índices explicados en la sección 3.2.1.5 al hablar de los mecanismos de gestión del diccionario general.

Con estos sistemas de índices se reducen las palabras a procesar a aquellas pertenecientes a las categorías predichas (y que comienzan por las letras adecuadas). Además, al estar ordenadas por orden de frecuencia descendente, no será necesario recorrerlas todas, sino solamente las de frecuencia superior a un determinado valor, fácilmente calculable en cada caso. De esta forma se eliminará la necesidad de procesar las menos frecuentes, acelerando el proceso de predicción.

3.3.2.1.5 Gestión de las palabras desconocidas

Uno de los problemas de los métodos categoriales, que comparte con cualquier sistema que precise información gramatical, son las *palabras desconocidas*[14], es decir, las que no se encuentran en el diccionario general. Al no estar contempladas en el sistema, no existe información sobre ellas, lo que impide un funcionamiento correcto de las técnicas de predicción utilizadas.

En el caso de los *bipos* y los *tripos*, la escritura de una palabra desconocida implica la aparición de una categoría desconocida, por lo que no se sabrá qué pares o tríos de categorías utilizar. La repercusión negativa de este tipo de palabras en los resultados de los diversos métodos de predicción, hace necesario gestionar este problema.

Antes de afrontar la gestión de las palabras desconocidas, cabe destacar que, por muy completo que sea el diccionario, nunca contendrá todas las palabras que aparezcan en los textos de prueba o del usuario, por varias causas, entre las cuales se encuentran:

- El *volumen de palabras del idioma*: aunque el diccionario sea de gran tamaño, es muy difícil que contenga todas las palabras.
- Las *formas flexivas de los verbos*: que, como se ha comentado anteriormente, fueron eliminadas en la generación del diccionario general para reducir su tamaño.
- Los *neologismos* o *palabras técnicas* sobre temas concretos.

[14] *Out of Vocabulary word* (OOV), en la nomenclatura inglesa

- Los *nombres propios*: especialmente problemáticos en el caso del uso de predicción de palabras en comunicadores, donde es más común el uso de nombres de personas, calles, ciudades, etc.

- El *vocabulario personal* del usuario: apelativos coloquiales, tanto para personas como para sitios, etc.

- Las *faltas de ortografía*: por una parte, los errores habituales (confusiones entre letras, como *b* y *v*, cambios de orden, etc.). Por otra parte, los errores derivados de la falta de precisión en el uso del barrido, en los casos en los que la predicción de palabras esté integrada en este tipo de interfaces. A la dificultad de controlar el barrido con el pulsador, se añade tiempo necesario para subsanar el error, que hace que muchas veces el usuario decida ignorarlo, sabiendo que el interlocutor va a comprender la palabra igualmente.

- Las palabras de otros idiomas.

Por las causas anteriormente expuestas, no es realista considerar que el diccionario vaya a contener todas las palabras que aparezcan en los textos y, debido a ello, hemos decidido introducir ciertas estrategias en la rutina de categorización para asignar información gramatical a las palabras desconocidas. La metodología aportada en nuestra arquitectura, tras el estudio que hemos llevado a cabo en esta tesis, implica la comprobación de:

- *Prefijos.* La rutina de categorización dispone de la siguiente lista de prefijos, cada uno de los cuales se puede aplicar a ciertas categorías:

  - *Extra-*, que se aplica a adverbios, adjetivos o verbos. Por ejemplo: *extralimitar*, *extradeportivo*.
  - *In-*, que se aplica a adverbios, adjetivos o verbos.
  - *Re-*, que se aplica a nombres y verbos.

  La rutina de categorización comprueba si la palabra comienza por una de las cadenas anteriores, elimina dicha cadena, y verifica si la palabra resultante está en el diccionario. Si lo está, comprueba si su categoría admite el prefijo eliminado, y en caso afirmativo, asigna a la palabra buscada la categoría adecuada.

- *Terminaciones distinguidas.* La rutina de categorización dispone de la siguiente lista de terminaciones características, cada una de las cuales corresponde siempre a una categoría gramatical determinada:

  - *-mente*. A las palabras terminadas por *-mente* que no estén en el diccionario, se le asigna automáticamente la categoría *adverbio*. Se debe tener en cuenta que hay palabras que acaban en *-mente* que no son adverbios, como *mente* (nombre), o *aumente* (verbo) que están en el diccionario, por lo cual serán encontradas y categorizadas correctamente.
  - *-ísimo*, *-ísima*, *-ísimos*, *-ísimas*, *-érrimo*, *-érrima*, *-érrimos*, *-érrimas*. A todas las palabras que terminen por una de estas cadenas de letras se les asigna la categoría adjetivo, el género y número adecuados y el rasgo superlativo.

  - *-dad*, *-dades*, *-tud*, *-tudes*, *-ción*, *-sión*. Estas terminaciones corresponden con nombres femeninos, en singular o plural. No se consideran distinguidas *-siones* y *-ciones*, ya que pueden ser terminaciones ciertas formas de verbos, como *ilusiones* (del verbo *ilusionar*), que no estén presentes en el diccionario general.
  - *-ismo* e *-ismos*. Se corresponden con nombres masculinos, en singular y plural, respectivamente. Puede haber otras palabras, como *mismo* con idéntica terminación, pero estarán contenidas en el diccionario.
  - *-ible* e *-ibles*. A las palabras con esta terminación se les asignará la categoría adjetivo, género neutro, y número singular y plural, respectivamente. Puede haber palabras con estas terminaciones que no sean adjetivos, pero están contenidas en el diccionario, como *combustible*.

  Las terminaciones distinguidas se pueden añadir a ciertas palabras transformándolas en otras nuevas, de la misma o de diferente categoría. En este caso, al contrario que en el caso de los prefijos, no se realiza la comprobación de si la palabra *original* existe, por ejemplo, dada la palabra *correctamente*, no se verifica si existe la palabra *correcta* por varias razones:

  - La categoría que se le asigna a la palabra es la misma, independientemente de la categoría de la palabra original, al contrario de lo que ocurre con los prefijos, que pueden aplicarse a palabras de categorías diferentes, y por lo tanto, es necesario conocer la categoría de la palabra original para asignársela a la palabra desconocida.
  - La palabra original puede sufrir modificaciones adicionales cuando se le añade el sufijo. Así, por ejemplo, la palabra *paupérrimo* es un adjetivo en grado superlativo, pero su correspondiente adjetivo en grado positivo es *pobre*. Por esta razón no es posible en todos los casos averiguar mediante transformaciones de qué palabra original proviene la que se está intentando categorizar.

- *Formas de verbos regulares.* Como no todas las formas de verbos regulares están incluidas en el diccionario, ante una palabra desconocida se comprueba si su terminación es un sufijo de verbo regular, en cuyo caso se transforma hasta obtener el infinitivo correspondiente, y se verifica si éste está en el diccionario como infinitivo regular. Este proceso se realizará varias veces, hasta encontrar todos los homógrafos, ya que es muy común que varias formas coincidan (p. ej.: comía puede ser primera y tercera persona), y, en algunos métodos de predicción, es imprescindible conocer todas las formas, con sus rasgos particulares.

  El conjunto de rasgos que se asigna a la palabra se obtienen combinando parte de los rasgos asociados al infinitivo (p. ej. el que indica que es un verbo regular), con los correspondientes al sufijo.

- *Clíticos.* Al igual que las formas de verbos regulares, no todos los verbos con clíticos están en el diccionario. El proceso para decidir si una palabra se corresponde con un verbo con uno o más clíticos es más complejo, por las modificaciones adicionales que pueden producirse en la palabra:

  - Se comprueba si la terminación de la palabra es un clítico. Se consideran las cadenas de uno y dos clíticos.

  - Se elimina la terminación y se verifica si la palabra resultante es verbo. Se restringe la comprobación a las formas no personales y al imperativo, porque es mucho más frecuente que los clíticos se añadan a estas formas. También se consideran aquellas formas no personales e imperativos que no están en el diccionario, pero que se reconocen automáticamente como formas de verbos regulares.
  - Cuando el clítico es *-nos* y está unido a una forma acabada en *-mo*, se debe añadir una *-s* a la forma antes de buscarla en el diccionario, ya que dicha *-s* ha sido eliminada del verbo al añadir el clítico (p. ej. en *vayámonos*).
  - Si no se ha reconocido la forma de verbo y ésta tiene tilde, se elimina y se repite el proceso.
  - A esto se debe añadir que las cadenas de clíticos pueden tener longitud diferente (p. ej.: cóme*lo*, cóme*telo*), haciendo necesario realizar estas comprobaciones en diferentes puntos. En las cadenas de dos clíticos, no todos pueden ocupar la primera o la segunda posición, y en cada punto del proceso se debe verificar si el clítico que se elimina puede aparecer en la posición que se está trabajando.
  - Los rasgos que se asignan a la palabra son los del verbo encontrado, añadiéndole el rasgo *clítico*.

- *Números.* Los números escritos como cadena de cifras, no están incluidos como tales en el diccionario, pero si se utilizan, serán reconocidos en la rutina de categorización y no interrumpirán el correcto funcionamiento de los métodos de predicción.

- *Signos.* Se categoriza como signo todo carácter aislado que no sea letra ni número.

En nuestra propuesta, estos procesos se aplican de forma combinada y recursiva, ya que a las palabras se les pueden añadir varios sufijos o prefijos. Por ejemplo, al intentar categorizar *extralimitándose*, se eliminará *extra-*, obteniendo *limitándose*, que no está presente en el diccionario, pero admite que se elimine el clítico *-se*, obteniéndose *limitándo*. Dado que esta palabra sigue siendo desconocida, el proceso continúa quitando la tilde y generando como resultado *limitando*, que tampoco está en el diccionario, pero que es reconocida como forma del verbo regular *limitar*. Esta propuesta la presentamos como una aportación importante del trabajo de tesis ya que la consideramos una característica relevante de nuestra arquitectura que consigue ampliar la cobertura, extendiendo la validez del conocimiento almacenado en el diccionario general a significantes no explicitados en el mismo, pero relacionados a través de esta metodología con otros que sí aparecen.

Con el proceso visto anteriormente, se consigue categorizar algunas de las palabras que no están incluidas en el diccionario, pero será inevitable que en los textos aparezcan palabras que el sistema no sea capaz de clasificar. Incluso a estas últimas es necesario asignarle una categoría y, para dar una solución a este problema, hemos incluido en nuestra arquitectura la estrategia que se describe a continuación.

El conjunto de categorías que se asigna a estas palabras desconocidas ha sido entrenado. Se partió del texto etiquetado PAIS y se contabilizaron las categorías a las que pertenecían las palabras no incluidas en el diccionario. El resultado fue que mayoritariamente corresponden a nombres y adjetivos, y nunca son números, signos, ni pertenecen a las categorías cerradas (ya que están completamente caracterizadas en el diccionario). El

número de verbos también era muy reducido, ya que el conjunto de ellos que está incluido en el sistema SMORPH es una representación prácticamente exhaustiva del sistema verbal castellano.

Los valores que se asignan a los rasgos son aquellos que permiten el comportamiento más flexible posible de las palabras, sin imponer ninguna restricción adicional. Por ejemplo, a los nombres y adjetivos se les asocian género y número neutros.

Según los criterios expuestos, a las palabras desconocidas se les asigna la siguiente estructura de categorías y rasgos:

nombre	74%	género y número neutros
adjetivo	23%	género y número neutros
adverbio	1.5%	sin rasgos
verbo	1.5%	

Esta distribución en categorías de las palabras desconocidas es característica de los textos periodísticos, donde el porcentaje de palabras nuevas suele ser muy alto, especialmente nombres propios de personas y lugares.

Evidentemente, este proceso de categorización de palabras desconocidas, además de clasificar las palabras correctas no incluidas en el diccionario, le asigna una categoría a palabras que no existen en el castellano, tanto a las palabras con faltas de ortografía, como a las de otros idiomas, etc. De todas formas, aunque en ciertos casos este método asigne las categorías de forma errónea, el efecto global de introducirlo en el proceso de predicción es beneficioso, como puede comprobarse en el apartado 4.1.3.3.4 (página 133) del capítulo de resultados.

3.3.2.2 Limitaciones de los métodos categoriales

Los métodos categoriales basados en secuencias fijas de categorías, solamente consideran información sobre las n últimas palabras escritas. El número n no suele ser muy grande, debido a la cantidad de texto de entrenamiento categorizado que se requiere, por lo cual, normalmente la predicción se realiza en base a las 2 o 3 categorías anteriores.

Este número puede ser insuficiente, ya que, tanto la categoría como los rasgos de las palabras que se predicen, pueden depender de palabras cuya distancia con la actual es mayor. Piénsese por ejemplo en la frase *La casa de los espíritus, que está en la colina, es blanca*: las palabras *está*, *es* y *blanca* dependen de *casa*. Estas *relaciones de larga distancia* no pueden ser modeladas con *bipos* y *tripos*, que consideran una información de contexto limitada y se deben desarrollar métodos más sofisticados que permitan expresarlas. A estos modelos dedicaremos el siguiente apartado.

3.3.3 Gramáticas formales

Hasta ahora, en los métodos de predicción vistos, la información gramatical que se utiliza para predecir es la correspondiente, como mucho, a las dos palabras anteriores, perdiéndose

información contextual de más largo alcance. Buscando tener en cuenta la información gramatical de la frase en curso desde el principio, se decide introducir gramáticas formales en la predicción de palabras.

Se plantearon las exigencias de los modelos vistos en la sección 2.2.3.1 del encuadre científico tecnológico y su posible empleo para la predicción, y se eligió el más adecuado, tomando como criterios su modo de funcionamiento, la potencia del modelo (flexibilidad), la complejidad algorítmica (velocidad), la robustez, la disponibilidad de información, etc.

Como compromiso entre la dificultad del analizador y la sencillez y potencia expresiva de la gramática, se toman como punto de partida las *gramáticas independientes del contexto* [Alle94]. A modo de recordatorio, cabe indicar que son aquellas cuyas reglas de reescritura son de la forma:

Símbolo no terminal → secuencia de símbolos terminales y no terminales

Es decir, los símbolos no terminales se descomponen siempre de la misma forma, independientemente del contexto, y están compuestos por cualquier secuencia de símbolos terminales y no terminales.

Una vez se decide utilizar este tipo de gramáticas, es necesario generar un sistema completo capaz de predecir basándose en él. Los pasos a seguir son:

- Elección del conjunto de categorías y de sus rasgos.
- Clasificación de las palabras según dicho sistema de categorías y asignación de los rasgos correspondientes.
- Generación de un conjunto de reglas para modelar un subconjunto lo más amplio posible del castellano, también según dichas categorías y dicho esquema de rasgos.
- Desarrollo de un sistema de predicción de palabras capaz de *analizar* el texto escrito hasta el momento, verificar qué reglas está cumpliendo y *predecir* las categorías y rasgos gramaticalmente correctos en ese punto de la frase.

El conjunto de categorías y rasgos que se utilizan en este sistema se describe en la sección 3.2.1.4 del diccionario general. El criterio seguido en la selección, como se comentó anteriormente, fue el de buscar un conjunto que facilitara *la realización de un modelo del castellano (simplificado) como secuencia de categorías, en forma de gramática independiente de contexto, de la forma más sencilla posible.* En el diccionario general las palabras están etiquetadas según este sistema de rasgos y categorías.

3.3.3.1 Selección del analizador

Una vez decidida la utilización de una gramática independiente del contexto, se eligió un analizador que pudiera verificar la validez del fragmento de frase escrito y que, además, se pudiera modificar para predecir las categorías gramaticalmente correctas.

Algunos de los factores que se consideraron en la selección del analizador fueron:

- Las restricciones propias del problema, como, por ejemplo, que el analizador debe recorrer la frase de izquierda a derecha, ya que sólo está disponible el principio de la frase.
- La capacidad predictiva.
- La posibilidad de mantener en paralelo varias ramas con las distintas interpretaciones válidas para la frase en curso, permitiendo tratar fácilmente el problema de la ambigüedad en el análisis.
- La necesidad o no de recomenzar todo el análisis cada vez que una rama no sea válida: normalmente existen varias posibilidades que se corresponden con el principio de la frase, pero muchas fallan tras las primeras palabras, dejando de ser válidas. En algunos algoritmos, en caso de fallo es necesario retroceder y volver a empezar el análisis desde el principio, siguiendo un camino diferente. Se elegirá un analizador que no necesite retroceder.
- La posibilidad de incluir información probabilística, gestión de rasgos, etc., es decir, de ampliar la potencia del analizador.

Una vez estudiados todos los elementos, se eligió el *analizador de Earley (*top down chart parser*)*, que es un analizador de gramáticas independientes del contexto con las siguientes características:

- Analiza el texto de izquierda a derecha.
- Para un determinado fragmento de texto, mantiene simultáneamente todos los posibles análisis válidos. De esta forma:
  - Se gestiona la *ambigüedad* de las frases, tanto por los diferentes significados de las palabras que la componen, como por las distintas formas de segmentar las frases en sintagmas.
  - En caso de *fallo* de alguna regla, no es necesario retroceder al principio del análisis, ya que continuarán activas todas interpretaciones que no hayan fallado. Así se elimina la duplicidad de esfuerzo.
- Es posible añadirle en paralelo la gestión de frecuencias y de rasgos.

Por estas razones, el *analizador de Earley* parece suficientemente potente y flexible como para cubrir los requisitos necesarios. En [Stol95] y [Earl70] se puede encontrar una descripción del funcionamiento básico del analizador.

3.3.3.2 Funcionamiento del analizador modificado

A continuación, para ilustrar el proceso, se mostrará el funcionamiento del *analizador modificado para actuar como predictor*, al escribir la frase *La casa está en la colina*.

Notación: se escribirán con minúsculas los *símbolos terminales* (categorías, que se obtienen del diccionario, a partir de cada palabra), y como secuencia de letras mayúsculas y minúsculas los *símbolos no terminales*, que se obtienen a partir de las reglas.

El conjunto de reglas es:

(R1) O → SN verbo SPrep

(R2) SN → nombre

(R3) SN → artículo nombre

(R4) SN → artículo adjetivo nombre

(R5) SPrep → preposición SN

El analizador, al ser de estrategia descendente, tiene una primera fase predictiva, en la cual se activan todas las reglas que son gramaticalmente correctas en el punto de la frase que se está analizando. Se marca con un ∘ la posición de la regla hasta donde ya se ha analizado. Al principio de la frase, las reglas que se activan son las correspondientes al *elemento inicial*, que en este caso es *O* (oración), y se marca como posición del análisis la anterior al primer elemento.

De esta forma, la regla 1 queda:

(R1) O → ∘ SN verbo SPrep

Como la marca está en la posición anterior a un símbolo no terminal, éste no coincide con la categoría de ninguna palabra del diccionario, sino que se genera a partir de las reglas correspondientes a dicho símbolo. En estos casos, se activan las reglas que lo derivan, situando la marca al principio de las reglas correspondientes. En este ejemplo, será necesario añadir las reglas que derivan el sintagma nominal (SN):

(R2) SN → ∘ nombre

(R3) SN → ∘ artículo nombre

(R4) SN → ∘ artículo adjetivo nombre

Las reglas que esperan un símbolo no terminal (como la regla R1, que espera un SN), detienen su análisis hasta que se completa alguna de las reglas que lo derivan.

En cada punto del análisis se predicen todas las categorías gramaticales esperadas por las reglas activas, es decir, las que aparecen después de las marcas. Las reglas que esperan un símbolo no terminal no predicen categorías, pero sí las predicen las reglas que expanden dicho símbolo no terminal.

Utilizando la información probabilística asociada a cada regla (como se verá en próximas secciones), se averigua la probabilidad de que cada categoría gramatical sea la siguiente, y se calcula la probabilidad de cada palabra. Posteriormente se seleccionan las palabras más probables, y se muestran al usuario como palabras predichas.

En este momento se escribe la primera palabra de la frase (*la*) seleccionándola de la lista de palabras predichas, si está en ella, o completándola letra a letra. Se categoriza la palabra, obteniendo su información gramatical del diccionario general: *la* aparece en tres entradas del diccionario, como *artículo, pronombre y nombre* (nota musical, aunque es muy poco frecuente, excepto en los textos de música). Por ello, avanzan la posición de la marca tanto las reglas que esperan un artículo (R3 y R4), como R2 que espera un nombre:

(R2) SN → nombre ∘

(R3) SN → artículo ∘ nombre

(R4) SN → artículo ∘ adjetivo nombre

Se puede ver que la marca de (R2) está en la última posición de la regla, indicando que ha terminado su análisis completando un sintagma nominal. Con dicho sintagma nominal, avanzan todas las reglas que lo esperen, de la misma forma que si fuera un símbolo terminal.

De todas formas, se mantiene una copia de R1 con la marca en la primera posición, porque aún faltan dos reglas que derivan el sintagma nominal y se deben mantener todas las posibilidades, para soportar la ambigüedad.

Dados estos pasos, las reglas activas en este punto son:

(R3) SN → artículo ∘ nombre

(R4) SN → artículo ∘ adjetivo nombre

(R1.1) O → SN ∘ verbo SPrep

(R1.2) O → ∘ SN verbo SPrep

Según este conjunto de reglas, se predicen nombres, adjetivos y verbos.

La siguiente palabra de la frase, en el recorrido de izquierda a derecha, es *casa*, que se corresponde con tres entradas del diccionario: nombre y dos formas del verbo *casar*. Al incluir esta información en las reglas se interrumpe R4, que queda eliminada del análisis. El verbo avanza R1.1 a la siguiente posición, donde espera el símbolo no terminal SPrep, por lo cual se deben lanzar las reglas que lo derivan. La entrada *nombre* avanza R3 y finaliza la regla, dando lugar a un SN, que avanza R1.2. Después de todo el proceso, las reglas activas son:

(R5) SPrep → ∘ preposición SN

(R1.1) O → SN verbo ∘ SPrep

(R1.2) O → SN ∘ verbo SPrep

Con este conjunto de reglas se predicen verbos y preposiciones. Cuando se completa la siguiente palabra, *está* (verbo), se avanza R1.2 y se interrumpe R5, y con ella R1.1. Como R1.2 pide un SPrep, se expanden las reglas correspondientes.

El análisis continúa de forma similar hasta que se completa la frase de entrada. La oración es gramaticalmente correcta según el conjunto de reglas definido, si se acaban las palabras en el mismo punto en el que se acaba el análisis de la regla correspondiente al símbolo inicial (O). Es decir, si cuando se acaba la oración de entrada, alguno de los análisis tiene la marca situada en el último punto de la regla que expande el símbolo "O":

(R1) O → SN verbo SPrep ∘.

En este ejemplo, la frase sí es gramaticalmente correcta.

Este es el funcionamiento del analizador modificado. Se ha podido ver que se avanza de izquierda a derecha, que se mantienen en paralelo todas las posibles interpretaciones

de la frase y que, en caso de interrupción de alguna regla, no es necesario recomenzar el análisis, ya que se puede seguir con las reglas que continúen activas.

En el desarrollo de este predictor de categorías se ha utilizado como base el sistema de tablas generado por el programa *Bison*, que es un generador de analizadores: si se le proporciona una gramática en el formato adecuado, es capaz de generar el código C de un analizador para dicha gramática. En [Donn95] y [Donn99] se puede encontrar información adicional sobre este sistema. El código fuente de Bison ha sido modificado para cambiar la información generada y añadir nuevos elementos. Posteriormente se ha implementado otro sistema diferente, que utiliza estas tablas para predecir categorías en lugar de analizar, y se han añadido nuevas características que aumentan su potencia, como se explica en las secciones siguientes.

3.3.3.3 Aumento de la potencia del analizador

A continuación se describen un conjunto de aportaciones a la organización y descripción de las reglas de la gramática formal y a la metodología de su uso que constituyen diversas ampliaciones que se han realizado al analizador para incrementar su potencia. Estas contribuciones han sido fruto del esfuerzo realizado para dotar a nuestro analizador sintáctico de la potencia expresiva adecuada para describir los efectos que hemos considerado relevantes, tras el estudio de los fenómenos lingüísticos que hemos observado en la práctica. Para cada una de las ampliaciones se ofrece, si procede:

- Su descripción y la justificación de su necesidad.
- Un ejemplo.
- La mejora que produce, como reducción del número de reglas y de categorías necesarias para describir los mismos fenómenos lingüísticos.
- Forma de introducir la mejora en el proceso de predicción.

En los ejemplos se utiliza el formato definitivo de las reglas del sistema, por lo que, a continuación, se proporciona una breve descripción de éste, que ayude a la comprensión de los ejemplos.

Se parte del formato básico utilizado por el programa *Bison* (mencionado en la sección anterior), con potencia suficiente como para permitir formalizar una gramática de contexto libre convencional. Como ya se ha explicado, en este tipo de gramáticas cada regla de reescritura define la secuencia de símbolos terminales y no terminales que expanden un determinado símbolo no terminal. Los símbolos terminales son el conjunto de categorías explicado en el apartado 3.2.1.4, y se escriben en minúsculas: *pron, v*. Los símbolos no terminales se escriben como combinación de letras mayúsculas y minúsculas, comenzando por mayúscula, como *SN, SPrep*. Un ejemplo de regla básica es:

```
SN:
    det adj n adj
    ;
```

El SN es el símbolo no terminal expandido por la regla, separado del resto por dos puntos, después la secuencia de símbolos que lo forman, y, por último, un punto y coma para cerrar la regla.

Cada símbolo no terminal puede ser expandido por más de una regla. En ese caso, dichas reglas pueden escribirse como listado de reglas completas (como la anterior), o unificarse todas tras el mismo símbolo, separándolas por una barra vertical, como en el ejemplo siguiente:

```
SN:
    det n
  | n adj
  | pron
  ;
```

En cada uno de los apartados se explica el aumento de potencia expresiva necesario para que las reglas soporten la información nueva que precisa cada una de las mejoras introducidas en el analizador.

3.3.3.3.1 Tratamiento de la información frecuencial

Es muy común que en una gramática existan varias formas de derivar cada símbolo no terminal. Las distintas reglas pueden tener probabilidades diferentes (como se explica, por ejemplo, en [Mage91]), y es necesario incluir esta información en el proceso de predicción, dando prioridad a las categorías procedentes de las reglas más probables. Además, por la ambigüedad de las palabras, que pueden pertenecer a varias categorías, una palabra puede avanzar más de una regla. Los distintos significados pueden tener diferente probabilidad, y sería recomendable aumentar la prioridad de las ramas seguidas por las categorías más probables, para robustecer el proceso de predicción.

Los datos necesarios para considerar estos factores en el cálculo de la probabilidad de la categoría predicha por cada rama son:

- La probabilidad de que cada regla expanda su símbolo no terminal correspondiente. Esta información se añade al final de cada regla, en tanto por mil, después del signo &, como puede verse en el siguiente ejemplo:

  ```
  SN:
      det n        & 750
    | n adj        & 100
    | pron         & 150
    ;
  ```

  En este caso, el sintagma nominal está compuesto por un determinante seguido por un nombre (un 75% de las veces), por un nombre seguido por un adjetivo (10% de las veces) o por un pronombre (el 15% restante).

  Esta probabilidad se puede entrenar, contabilizando la cantidad de veces que se utiliza cada símbolo no terminal y cada regla que lo genera. Si la regla R_r deriva el

símbolo no terminal S_s, la probabilidad de dicha regla será:

$$p(R_r) = \frac{\#R_r}{\#S_s} \tag{3.13}$$

Donde $\#S_s$ es el número de veces que se utiliza el símbolo no terminal S_s en el texto de entrenamiento, y $\#R_r$ es el número de veces que es la regla R_r la que deriva dicho símbolo no terminal.

Como se entrena a partir de texto categorizado con ambigüedad, es necesario considerar la probabilidad de que las palabras que conforman la frase pertenezcan a las categorías requeridas por la regla cuya probabilidad se está entrenando, de forma similar a la que se utiliza en el entrenamiento de los *bipos/tripos* a partir de textos ambiguos (véase el apartado 3.3.2.1.3 en la página 64). Este entrenamiento proporciona datos aproximados, debido a que no se dispone de textos etiquetados con la información de las reglas utilizadas, aunque se han realizado comprobaciones, comparando sus resultados con cálculos manuales realizados sobre un subconjunto del mismo corpus, obteniéndose valores similares a los de la aproximación. El número de datos a estimar en la versión actual de las reglas es aproximadamente 260.

En versiones futuras del sistema se adaptarán otros métodos de entrenamiento de la probabilidad de las reglas (como el *Input-Output algorithm*, explicado en [Pere91], cuya versión modificada se explica en [Sanc99]) para que consideren las características especiales del sistema, como la imposición y prohibición de significantes, lemas, o el tratamiento de rasgos.

- La probabilidad de que las palabras que componen el fragmento de frase escrito pertenezcan a la categoría precisada por dicha rama. Esta información está contenida en el diccionario general.

  P. ej., para el conjunto de reglas del ejemplo anterior, la palabra *la* avanzaría las tres reglas, ya que en el diccionario general consta que puede ser determinante (con una probabilidad del 50%), pronombre (el 49% de las veces) y nombre (de nota musical, un 1% de las veces).

Basándonos en estos datos, se pretende calcular la probabilidad de la categoría de la palabra que el usuario va a escribir, tomando como base la secuencia de palabras anteriores del texto, las reglas y toda la información probabilística asociada a las reglas y las palabras. Si cada palabra tuviera solamente una categoría posible, las ramas activas del analizador serían aquellas cuya secuencia de categorías hasta el momento coincida con la secuencia de categorías de las palabras escritas por el usuario. Esa secuencia sería única, y la probabilidad de cada rama sería el producto de las probabilidades de las reglas implicadas en su generación.

Según esto, la probabilidad de que la siguiente palabra perteneciera a una determinada categoría sería la suma de las probabilidades de todas las ramas que esperaran dicha categoría. A su vez, en este caso, en el que no hay ambigüedad, la probabilidad de cada rama sería el producto de la probabilidad de las reglas que hayan dado lugar a ella, quedando la expresión de la probabilidad de la categoría como sigue:

$$p(C_t^i) = \sum_{R_{C_t^i}} \left(\prod_{RR_r} p(R_r) \right) \tag{3.14}$$

Donde $R_{C_t^i}$ son las ramas que esperan la categoría C_t^i, y RR_r es el conjunto de todas las reglas R_r que se han explorado en la rama r.

Como se ha visto, las palabras pueden tener varios significados y pertenecer a distintas categorías gramaticales, por lo que la probabilidad de seguir por cada rama debe ser ponderada por la probabilidad de que cada palabra de la frase (w_j) pertenezca a la categoría necesaria para seguir por esa rama (C_{t-j}^k).

Por lo tanto, la probabilidad de cada rama, si hay ambigüedad, se calcula con la expresión:

$$p(Rama_{amb}) = \prod_{RR_r} p(R_r) \prod_{C_{t-j}^k} p(C_{t-j}^k / w_j) \tag{3.15}$$

La probabilidad, entonces de que cada categoría sea la predicha es:

$$p(C_t^i) = \sum_{R_{C_t^i}} p(Rama_{amb}) \tag{3.16}$$

Una vez obtenida la probabilidad de que cada categoría sea la siguiente, se predicen las palabras. Para ello se calcula la probabilidad de cada palabra, dependiendo de la o las categorías a las que puede pertenecer, según la expresión:

$$p(w) = \sum_{C_t^i} p(w/C_t^i) \cdot p(C_t^i) \tag{3.17}$$

Posteriormente, se eligen las más probables. De esta forma se consigue dar prioridad a las palabras predichas por las reglas más frecuentes, que hayan sido seguidas por las categorías más comunes de las palabras.

3.3.3.3.2 Posibilidad de introducir elementos opcionales

En ocasiones existen elementos que pueden estar en alguna posición de la regla, pero no son imprescindibles, es decir, serán gramaticalmente correctas tanto las frases que contengan ese elemento como las que no. Es posible introducir ambas reglas en el modelo de lenguaje, pero se puede ampliar la potencia del analizador para que pueda admitir elementos opcionales, y, así, reducir el número de reglas.

Cualquier símbolo de la regla, tanto terminal como no terminal, puede ser opcional. Por ejemplo, el adjetivo en el sintagma nominal es opcional, y también se puede añadir un sintagma preposicional opcional (símbolo no terminal) como complemento. Cada elemento

opcional llevará asociada su probabilidad de aparición, que permite realizar el tratamiento estadístico adecuado.

En las reglas del sistema, los elementos opcionales llevan a su derecha un símbolo de interrogación, seguido por su probabilidad de aparición, en tantos por mil. Esto permite, por ejemplo, escribir reglas como la que sigue:

```
SN:
    det    adj ? 400    n    & 1000
    ;
```

Esta regla admite como válidos los sintagmas nominales formados por un determinante más un nombre, pudiendo insertarse un adjetivo entre ambos, con una probabilidad del 40%.

En la versión actual del sistema de predicción, la probabilidad de los elementos opcionales es entrenada, contabilizando el número de veces que se utiliza cada regla, y el número de veces que el elemento opcional es elegido. Esta contabilidad se pondera por la probabilidad léxica, es decir, por la probabilidad de que el conjunto de palabras que forma la frase se corresponda con el conjunto de categorías requeridas por la regla.

En la predicción, el comportamiento del analizador hasta el elemento opcional es idéntico a los casos anteriores. Sin embargo, cuando se llega a un elemento opcional, la regla R se duplica, en R_1 y R_2, heredando todos los datos que contenga (probabilidad de las palabras, de la regla, etc.).

R_1 continúa en el mismo punto y predice el elemento opcional. Su probabilidad de regla se multiplica por Pr_{Opc} (la que acompaña al símbolo opcional), y, con ella, la probabilidad de las predicciones que realice. R_2 avanza una posición más, pasando a predecir en el punto siguiente al elemento opcional y su probabilidad de regla se multiplica por $(1 - Pr_{Opc})$. A partir de ese momento, las dos reglas continuarán en el sistema de forma independiente, como si se hubieran desplegado ambas desde el principio. Cuando el usuario escriba la palabra siguiente, ésta se corresponderá con el elemento opcional o con el siguiente y fallará el análisis de una de las dos reglas. También puede ocurrir que la palabra tenga diferentes homógrafos, que sean aceptados por ambas reglas, y se avancen las dos, cada una con su probabilidad. En este caso, serán las palabras siguientes las que eliminen la ambigüedad. Por ejemplo, en la regla:

```
SN:
    ∘  det    adj ? 400    n    & 1000
    ;
```

Cuando se escribe un determinante, se avanza el análisis, llegando al elemento opcional. En ese momento, la regla se desdobla en dos y se actualiza la información frecuencial como se explicó en el párrafo anterior, obteniéndose:

```
SN:
      det    ∘    adj    n    & 400
    | det    adj    ∘    n    & 600
    ;
```

En esta situación se predicen adjetivos (con probabilidad 0.4) y nombres (con proba-

bilidad 0.6). Si a continuación se escribe un adjetivo, la segunda regla falla y la primera avanza, prediciendo nombres.

En una misma regla pueden encadenarse varios elementos opcionales. Para gestionarlo, la regla se replicará tantas veces como elementos opcionales consecutivos haya, avanzando cada vez una posición más y multiplicando las probabilidades de las reglas por el factor adecuado, hasta encontrar un símbolo no opcional, o el final de la regla. Por ejemplo, dada la regla:

```
SN:
    ∘  det   n   adj ? 400   SPrep ? 300   & 1000
    ;
```

Tras escribir un determinante y un nombre, se avanza la regla, y se desdobla en:

```
SN:
      det   n   ∘   adj   SPrep ? 300   & 400
    | det   n   adj   ∘   SPrep         & 180
    | det   n   adj   SPrep ∘           & 420
    ;
```

La primera regla predice adjetivos, con la probabilidad de aparición de los adjetivos (0.4), la segunda despliega las reglas del sintagma preposicional, con la probabilidad de que no aparezca el adjetivo anterior, pero sí el sintagma preposicional ($(1 - 0.4) \cdot 0.3 = 0.18$). La tercera regla tiene la marca en la última posición, y considera finalizado el análisis del SN. Su probabilidad será la de que no aparezcan ni el adjetivo ni el sintagma preposicional: $((1 - 0.4)(1 - 0.3) = 0.42)$.

3.3.3.3.3 Tratamiento de significantes y lemas

Existen ciertas relaciones entre las palabras que no pueden ser expresadas como secuencia de categorías (a menos que se separen las palabras implicadas en una categoría especial para ellas), ya que exigen en un punto determinado un *significante* o un *lema* concretos, por ejemplo, si se desea introducir ciertas expresiones en forma de regla, como *en el caso de que* o modelar las perífrasis verbales o los verbos compuestos: *haber + participio*, o *ir a + infinitivo*.

Por otra parte, algunas palabras siguen normalmente las reglas convencionales de su categoría, salvo en ciertas estructuras concretas, por ejemplo, secuencias como *preposición + infinitivo*, en las que no todas las preposiciones pueden preceder a un infinitivo, siendo necesario prohibir *según*, *bajo*, *cabe*, *so* y *hacia*.

Por estas razones se ha ampliado la potencia del analizador, para que no solamente exprese relaciones entre categorías gramaticales, sino también permita imponer y/o prohibir ciertos lemas y/o significantes en determinadas posiciones de las reglas, sin necesidad de separarlos en una categoría específica.

Imposición de significantes y lemas

Como se ha explicado anteriormente, es posible imponer en cualquier punto de las reglas un determinado lema o significante acompañando a la categoría, permitiéndose así modelar ciertas expresiones y perífrasis verbales mediante reglas. Por ejemplo, la regla necesaria para describir *en el caso de que* es:

```
Expresion1:
  prep [sgte {en}]
  det  [sgte {el}] ? 60
  n    [sgte {caso}]
  prep [sgte {de}]
  conj [sgte {que}] & 1000
  ;
```

En ella se puede ver el formato utilizado para imponer los significantes, escribiendo a la derecha de la categoría correspondiente [sgte{*significante*}] (para imponer el lema se cambiaría la palabra clave *sgte* por *lema*). Se puede observar que en todos los símbolos se impone el significante, y que la segunda palabra (*el*) es opcional, modelando con una sola regla las expresiones *en el caso de que* y *en caso de que*.

Esta regla, en cada punto del análisis, solamente predice y considera válidas las palabras que coinciden exactamente con el significante impuesto. Sólo se puede imponer un significante (ya que es exactamente la palabra predicha por la regla), y únicamente en los símbolos terminales.

Para modelar esta expresión con una gramática independiente del contexto, sería necesario separar los significantes que se desea imponer en una categoría específica para cada uno, y utilizar dicha categoría en las reglas que modelen las expresiones: en el ejemplo anterior, se debería crear una categoría específica para *en, el, caso, de* y *que*, y usar dichas categorías en esta regla. Por otra parte, como se separan de la categoría a la que pertenecen, sería necesario replicar las reglas de cada categoría inicial, añadiendo, por cada regla que requiere una preposición, otras para *en* y *de*, y lo mismo para *caso, el* y *que* con sus categorías respectivas.

La imposición de lemas es un proceso similar, que permite modelar expresiones en las que puede tomar parte cualquier forma de una determinada palabra, como, por ejemplo, el verbo *haber* en los verbos compuestos. Al igual que en el caso de los significantes, para modelar la imposición de lemas con un analizador convencional sería necesario crear una nueva categoría que contenga solamente las palabras cuyo lema sea el que se quiere imponer. Esta categoría se utilizaría en las expresiones en las que se impone el lema, y, por otra parte, sería necesario replicar las reglas de la categoría original para la nueva.

Como este procedimiento no es eficiente, el analizador ha sido modificado para permitir la introducción de lemas y significantes en las reglas sin tener que segregarlos a una categoría específica. En este caso, el analizador sólo considera válidas las palabras con dicho lema o significante, y, cuando predice, muestra exactamente ese significante, o un conjunto de palabras con dicho lema (esa información es extraída del diccionario general). De esta forma se reduce el número de categorías y reglas necesarias para modelar la lengua.

El tratamiento estadístico que conlleva la imposición de lemas y significantes hace necesario modificar las ecuaciones utilizadas para hallar la probabilidad de las palabras. Cuando las ramas predicen una categoría, la probabilidad de la palabra se calcula multiplicando la probabilidad de la categoría por la probabilidad de cada palabra fijada la categoría (probabilidad léxica), según la ecuación 3.17 (página 77).

Sin embargo, por cada rama que impone un significante, toda la probabilidad de la regla debe ser aplicada a dicha palabra, es decir, esa rama predice exclusivamente una palabra, que será toda su aportación a la lista de palabras predichas (si su probabilidad es suficientemente grande). Esto modifica la expresión de la probabilidad final de cada palabra, para añadir el término con la aportación de las ramas que imponen dicha palabra como significante. La nueva expresión es:

$$p(w) = \sum_{C_t^i} p(w/C_t^i) \cdot \left(\sum_{R_{C_t^i}} p(Rama_{amb_{C_t^i}}) \right) + \sum_{R_w} p(Rama_{amb_w}) \tag{3.18}$$

Donde R_w son las ramas que esperan la palabra w.

Las ramas en las que se impone un lema aportan un nuevo término, ya que su comportamiento no se puede modelar con ninguno de los dos anteriores: toda la probabilidad de estas ramas se aplica al mismo lema, pero es necesario repartirlo en sus diferentes significantes. Para ello se pondera la probabilidad de la rama por la probabilidad de cada significante fijado el lema, apareciendo el siguiente término:

$$p(w/Lema_w) = \frac{\#(w \cap Lema_w)}{\#Lema_w} \tag{3.19}$$

Es decir, la probabilidad de que fijado el $Lema_w$, la palabra sea w. La información necesaria para realizar este cálculo se obtiene del diccionario general: $\#(w \cap Lema_w)$ es la suma de las frecuencias absolutas de todas las entradas cuyo significante es w y lema $Lema_w$ (puede haber varias entradas que coincidan, por ejemplo, es muy común que varias entradas correspondan con distintas formas verbales y tengan el mismo significante y lema, pero distintos rasgos). $\#Lema_w$ se calculará sumando las frecuencias absolutas de todas las entradas cuyo lema sea $Lema_w$.

En la expresión para hallar la probabilidad de la palabra w es necesario añadir otro término, para considerar la aportación de las ramas que imponen los lemas, quedando entonces:

$$\begin{aligned} p(w) = \sum_{C_t^i} p(w/C_t^i) \cdot \left(\sum_{R_{C_t^i}} p(Rama_{amb_{C_t^i}}) \right) + \sum_{R_w} p(Rama_{amb_w}) \\ + p(w/Lema_w) \cdot \sum_{R_{L_w}} p(Rama_{amb_{L_w}}) \end{aligned} \tag{3.20}$$

Donde R_{L_w} son las ramas que imponen $Lema_w$.

Prohibición de significantes y lemas

Hay expresiones que son seguidas por todas las palabras pertenecientes a una determinada categoría gramatical *excepto* por algunas de ellas. Por ejemplo, secuencias como *preposición + infinitivo*, en las que no todas las preposiciones pueden preceder a un infinitivo, siendo necesario prohibir *según*, *bajo*, *cabe*, *so* y *hacia*.

Para modelar la expresión anterior con una gramática independiente de contexto convencional se tendrían que separar las palabras prohibidas en otra categoría, y, al igual que en la imposición, habría que replicar las reglas que se cumplen tanto para dichas palabras como para el caso general.

De nuevo, no es necesario, ya que se ha modificado el analizador para soportar la posibilidad de *prohibir* uno o más lemas o significantes de una determinada categoría. Cuando el análisis llegue al punto en el que se prohíben los lemas o significantes, no considerará válida ninguna palabra que coincida con ellos. Al predecir, mostrará palabras pertenecientes a la categoría adecuada, pero que no se correspondan con ninguno de los lemas y/o significantes prohibidos. La regla del ejemplo siguiente:

```
OracionCompuesta:
  O
  prep  [sgte {~bajo ~cabe ~hacia ~so ~tras ~ante }]
  conj  [sgte {que}]
  O     &1000
  ;
```

permite expresar oraciones compuestas del tipo *Me he enterado de que has venido*. En el segundo símbolo se prohíben una serie de preposiciones, que nunca pueden utilizarse en esa posición, y en el tercer símbolo se impone la palabra *que*.

Como se puede observar en el ejemplo, la prohibición de significantes utiliza el mismo formato que la imposición, añadiendo un ~ a la izquierda de la palabra prohibida. De forma similar se prohíben lemas, sustituyendo la palabra clave *sgte* por *lema*.

La gestión de las prohibiciones se añade en forma de filtro en el proceso de predicción: en un primer paso se predicen las palabras correspondientes a la categoría propuesta por la regla, y posteriormente se rechazan las palabras que coincidan con los lemas o significantes prohibidos. En cada elemento de la regla se pueden prohibir tantos lemas y/o significantes como se desee.

3.3.3.3.4 Tratamiento de rasgos

Como se ha visto en la sección 3.2.1.4, relativa a la información gramatical asociada a cada palabra, gran parte de dicha información está introducida en el diccionario en forma de rasgos y no de categorías. La ventaja fundamental de esta opción es la reducción en el número de categorías y reglas del modelo de lenguaje, con respecto a una gramática con la misma cobertura basada solamente en categorías.

A continuación se compara el número de categorías y reglas necesarias para expresar

ciertas relaciones entre los rasgos de las palabras, utilizando las dos posibles maneras de formalizarlo: por una parte, las categorías usadas en este sistema (*nombre, adjetivo, verbo,* etc.), a las que se denominará *categorías*, y a las que se asocian *rasgos*, que pueden ser tratados de forma independiente en las reglas (*masculino, femenino, singular, etc.*). Por otra parte, las $categorías_{rasgos}$, en las que se incluye en la categoría la información del rasgo, por lo que se consideran categorías diferentes aquellas que tienen distinto valor del rasgo, como $nombre_{mascsing}$ o $nombre_{femplu}$.

Con respecto al *número de categorías*, cada una de las del sistema actual equivale a tantas $categorías_{rasgos}$ como combinación de valores de rasgos admita. Por ejemplo, un *nombre* que admita como rasgos el género (masculino, femenino, neutro o que comience por a tónica) y el número (singular, plural o neutro), equivale a:

$$\begin{array}{lll} nombre_{masc\ sing} & nombre_{masc\ plu} & nombre_{masc\ neut} \\ nombre_{fem\ sing} & nombre_{fem\ plu} & nombre_{fem\ neut} \\ nombre_{neut\ sing} & nombre_{neut\ plu} & nombre_{neut\ neut} \\ nombre_{a-tónica\ sing} & & \end{array}$$

Es decir, cada *categoría* equivale a 10 $categorías_{rasgos}$.

Con respecto al *número de reglas*, el caso más simple, considerando un sintagma nominal formado únicamente por un nombre, supone, con el sistema actual, una sola regla, que indica la categoría *nombre*, mientras que con $categorías_{rasgos}$ serían necesarias 10 reglas, una por categoría.

En estos ejemplos se ha comprobado que el uso de rasgos disminuye el número de reglas y categorías necesarias para modelar expresiones formadas por una sola palabra (una sola categoría). Es común que los rasgos de una palabra influyan en los de otras (por ejemplo, tras la palabra *el*, se deben escribir nombres con género masculino y número singular), y es necesario incluir estas restricciones en las reglas que relacionan dichas palabras.

Las modificaciones realizadas en el analizador permiten modelar las relaciones de concordancia entre rasgos de distintas palabras, e imposición o prohibición de un valor de un rasgo en una palabra.

Concordancia entre rasgos

La *concordancia* consiste en asegurar que los valores de determinados rasgos de varias palabras coinciden. Para cada regla se permite definir qué rasgo debe concordar y entre cuales de las categorías, sin ninguna restricción, de forma que una palabra puede restringir los valores de determinados rasgos de palabras que se encuentren a cualquier distancia de ella. Dado el ejemplo: *La casa de mis tías, las que fueron a Santander, es azul*, con la potencia incluida en el analizador, es posible expresar las siguientes concordancias:

- *la* y *casa* concuerdan en género y número.
- *mis*, *tías* y *las* concuerdan en género y número.
- *tías* y *fueron* concuerdan en número.

- *casa* y *es* en número, independientemente de lo que se haya insertando entre las dos palabras.
- *casa* y *azul* concuerdan en género y número, aunque disten 8 palabras. *Casa* y *azul*, concuerdan siendo casa una palabra femenina y azul neutra (concuerda con palabras femeninas y masculinas).

A continuación se muestra el formato utilizado para determinar exactamente cuáles son los rasgos que deben concordar y en qué categorías de la regla.

A la derecha de cada símbolo de la regla, se indica el nombre del rasgo cuya concordancia se exige y uno de los cuatro signos auxiliares (*X, Y, Z y M*). Se asociará el mismo signo a todos los símbolos que deban concordar entre sí. Esto limita a 4 el conjunto concordancias se puede exigir en cada regla. En la práctica este número es suficiente, ya que en la mayoría de las reglas no es necesario utilizarlos todos. El número de elementos de la regla a los cuales se les puede exigir concordancia no está limitado.

Por ejemplo, en la siguiente regla se modela un tipo particular de sintagma nominal con un sintagma preposicional asociado:

```
SN:
    det  [gen{X} num{Y}] ? 175
    adj  [gen{X} num{Y}] ? 500
    n    [gen{X} num{Y}]
    adj  [gen{X} num{Y}] ? 250
    prep
    det  [gen{Z} num{M}] ? 75
    n    [gen{Z} num{M}]
    ;
```

En ella se han considerado las siguientes concordancias: de género y número de los cuatro primeros elementos, utilizando los signos X e Y, respectivamente, y de género y número de los últimos elementos entre sí, pero no con los primeros, por lo que se utilizan los otros signos, Z y M. La preposición no tiene ninguna restricción.

Se puede observar que los determinantes y adjetivos son opcionales, pero, si aparecen, deben concordar con las palabras adecuadas.

Para comprobar la concordancia, a la hora de analizar o de predecir, el analizador modificado averigua el valor del rasgo correspondiente de la palabra (p. ej., en la concordancia en *número*, el valor sería *singular o plural*), y lo compara con los valores correspondientes de las demás palabras que deban concordar con ella, buscando coincidencias (p. ej., dos palabras en singular coinciden en número).

Una vez determinados los rasgos con los que deben concordar las palabras siguientes, se filtran las palabras que se predicen, para eliminar las que no concuerdan. Además, no se consideran válidas las palabras escritas por el usuario que no concuerdan con los rasgos dados, interrumpiendo el análisis de la regla que exige la concordancia.

A las palabras neutras en género se les han asignado (en el diccionario general) los valores femenino y masculino, de esta forma, concuerdan tanto con las femeninas como con las masculinas (pero no con las palabras que empiezan con a tónica). El mismo tratamiento

se realiza con las palabras neutras en número, asignándoles los rasgos singular y plural. Si en un sintagma nominal se escribe *crisis*, el adjetivo posterior puede ser singular o plural. Sin embargo, si se escribe *la crisis* los adjetivos deben estar en singular. Es decir, los rasgos que se utilizan para predecir son una *combinación* elaborada de los rasgos de todas las palabras implicadas.

La posibilidad de exigir concordancias a los rasgos aporta gran flexibilidad al tratamiento. En realidad, la potencia expresiva de la gramática es la misma (la de una gramática independiente del contexto), pero se reduce drásticamente la cantidad de reglas necesarias para modelar las diferentes estructuras del lenguaje.

Con respecto al número de reglas necesario para modelar la *concordancia* de género y número entre dos categorías, por ejemplo, en *determinante + nombre*, utilizando las *categorías* del sistema actual son necesarias dos reglas: una que exige concordancia entre género (masculino, femenino y neutro) y número (singular, plural y neutro), y otra para los *nombres* singulares que comienzan por a tónica. Si se expresa con $categorías_{rasgos}$ habría que dar la lista de todas las posibles combinaciones de rasgos compatibles:

$$
\begin{array}{ll}
determ_{fem\ sing} + nombre_{fem\ sing}, & determ_{fem\ sing} + nombre_{neut\ sing}, \\
determ_{neut\ sing} + nombre_{fem\ sing}, & determ_{neut\ sing} + nombre_{neut\ sing}, \\
determ_{masc\ sing} + nombre_{masc\ sing}, & determ_{masc\ sing} + nombre_{neut\ sing}, \\
. & \\
. & \\
determ_{masc\ sing} + nombre_{a-tónica\ sing}, & determ_{neut\ sing} + nombre_{a-tónica\ sing},
\end{array}
$$

obteniéndose aproximadamente 50 reglas.

Para modelar la concordancia de tres palabras (por ejemplo, *determinante + nombre + adjetivo*), el número de reglas necesario, utilizando $categorías_{rasgos}$, se multiplica por 2 por cada rasgo que se contemple, ya que, para cada uno de ellos se debe expresar la concordancia con su valor exacto y con el neutro, es decir, son necesarias aproximadamente 200 reglas. Si se utiliza el sistema actual, el modelado se resuelve con dos reglas, una en la que se exige concordancia de género y número entre las tres palabras, para los nombres femeninos, masculinos o neutros, y otra para los $nombres_{sing\ a-tónica}$.

Como se puede observar, hay una gran diferencia entre el número de categorías y de reglas necesario para modelar las mismas estructuras con *categorías*, exigiendo concordancia de rasgos, y con $categorías_{rasgos}$, siendo este último método mucho menos eficiente.

Imposición de un valor determinado a un rasgo

En algunas ocasiones, las palabras no requieren que un rasgo concuerde con el suyo, sino que exigen que tenga un valor concreto, por ejemplo, después del verbo *haber*, si se escribe un verbo, debe estar en *participio*.

También existen ciertas construcciones particulares como *el águila blanca*, en que el artículo debe ser masculino, *águila* está categorizado con género *palabra que comienza por*

a tónica, y cualquier adjetivo posterior es femenino. Además, todas estas palabras son singulares.

El analizador ha sido modificado para permitir la imposición de tantos rasgos como se requiera en cualquiera de las palabras.

El formato utilizado para imponer rasgos tiene una estructura similar a la empleada para los lemas y significantes: dentro de los corchetes, a la derecha del símbolo al que afectan, se indica el nombre del rasgo, y el valor que se impone. Véanse, por ejemplo, las reglas de la siguiente lista:

```
FormaVerbal:
      v & 700
    | v [lema {ir}] prep [sgte{a}] v [tipo-tv{inf}] & 100
    | v [lema {haber}] v [tipo-tv{par} gen{m} num{s} ] & 100
    | v [lema {estar}] v [tipo-tv{ger}] & 100
    ;
```

Estas reglas modelan las formas verbales simples y compuestas y dos perífrasis verbales del castellano. La primera de ellas, indica que puede utilizarse un verbo sin ninguna restricción, como por ejemplo, *comer*, *voy*, *estamos*. La segunda regla modela las perífrasis verbales, compuestas por cualquier forma del verbo *ir* (solamente se impone el lema), más *a*, más cualquier verbo en infinitivo, para lo cual se impone que el rasgo *tipo-tv(tipo verbal)* tenga el valor *inf(initivo)*, como en *ir a cenar*, o *vamos a salir*.

La tercera regla es para las formas compuestas por *haber* (en cualquier tiempo, sin ninguna restricción), seguido por un verbo (no se imponen lema ni significante), en *participio masculino singular*, lo que se consigue imponiendo en *tipo-tv* el valor *par(ticipio)*, en *gen(ero) m(asculino)*, y en *num(ero) s(ingular)*: *he estado*, *habéis venido*.

Por último, la regla para la construcción estar + gerundio, como en *estamos acabando*. Al igual que en los casos anteriores, se impone el lema *estar* a la primera palabra, y en el segundo verbo se impone *ger(undio)* al rasgo *tipo-tv*.

A continuación, se explicará la potencia que le aporta al sistema la posibilidad de prohibir e imponer rasgos, de forma similar a como se hizo con la concordancia.

Para *imponer* que un sintagma nominal de cualquier número de elementos sea singular (exigiendo, además, concordancia de género), con el sistema actual se necesitan dos reglas: una para los *nombres* singulares que comienzan por a tónica, *gen(a)*, y otra para los demás, que son las siguientes:

```
SN:
      det  [gen{m} num{s}] ? 175
      n    [gen{a} num{s}]
      adj  [gen{f} num{s}] ? 250 & 10
|     det  [gen{X} num{s}] ? 175
      n    [gen{X} num{s}]
      adj  [gen{X} num{s}] ? 250 & 990
```

Para imponer que un sintagma sea masculino (gestionando en paralelo la concordancia

de número), sería suficiente con la siguiente regla:

```
SN:
    det  [gen{m} num{Y}] ? 175
    n    [gen{m} num{Y}]
    adj  [gen{m} num{Y}] ? 250 & 1000
```

Sin embargo, la imposición de un rasgo, cuando se trata con $categorías_{rasgos}$, equivale a escribir una regla para todas las combinaciones válidas de los demás rasgos. Por ejemplo, para imponer número singular en un sintagma nominal que sólo consta de determinante más nombre son necesarias 9 reglas para el singular y 7 para el neutro (porque no hay palabras neutras que comiencen por a tónica). Es decir, 16 reglas en total. Si el sintagma nominal es de 3 palabras (añadiendo un adjetivo), se necesitarán aproximadamente 32 reglas, ya que habría que añadirle a cada regla la posibilidad de concordar con el tercer elemento tanto con el valor impuesto como con el neutro.

La imposición de rasgos supone añadir un filtro más a la predicción de palabras, que elimina de la lista de palabras predichas a aquellas que no contienen el rasgo impuesto. Además, si el usuario escribe una palabra sin dicho valor en el rasgo correspondiente, se interrumpirá el análisis de esa regla.

Prohibición de ciertos valores de los rasgos

Algunas veces en un punto de la frase, una palabra puede tener cualquier rasgo *excepto* un valor determinado. Un ejemplo puede ser, en el caso de los adjetivos, la necesidad de prohibir los adjetivos pospuestos en la posición anterior al nombre, es decir, asegurar que *grande* sólo aparece después del nombre, y no se admite como gramaticalmente correcto *el grande niño*. De forma similar se debería evitar *el niño gran*.

De nuevo, esto podría resolverse separando las palabras en diferentes categorías y replicando las reglas cumplidas por ambos grupos de palabras, pero no es necesario porque el analizador ha sido modificado para permitir la prohibición de palabras cuyos rasgos tenga ciertos valores en cualquier punto de las reglas.

Un ejemplo de imposición y prohibición de valores en los rasgos se muestra en la siguiente regla:

```
SintagmaLo:
  pron        [sgte {lo}]
  adj         [gen{m} num{s} tipo-a{~a1}]
  SPrep       ? 500 & 1000
  ;
```

donde se modelan expresiones formadas por *lo* seguido por un *adjetivo* más, opcionalmente, un sintagma preposicional, como *lo malo*, o *lo bueno de los niños*. En este tipo de expresiones está restringido el uso de cierto conjunto de adjetivos, como *buen* o *gran*. Estos adjetivos están etiquetados en el diccionario general con el rasgo tipo-a{a1} (adjetivos antepuestos: que únicamente puede escribirse *antes* del nombre), y en esta regla están prohibidos.

Del mismo modo, hay adjetivos que solamente pueden escribirse *después* del nombre, información que aparece en el diccionario como tipo-a{a3}. Considerando estas restricciones, la regla que modela el sintagma nominal con un adjetivo antes y otro después del nombre, queda de la siguiente forma:

```
SN:
    det    adj [tipo-a{~a3}] ? 100    n    adj [tipo-a{~a1}] ? 100    & 1000
    ;
```

En esta regla, antes del nombre se prohíben los adjetivos pospuestos, y después del nombre se restringen los antepuestos, eliminando la posibilidad de predecir, o aceptar como válidas, estructuras como *el grande niño buen*. Los adjetivos que puedan aparecer en ambas posiciones, no están etiquetados como *a1* ni como *a3*, por lo que no están prohibidos por la regla en ninguno de los casos, y se permite tanto *las valientes mujeres* como *las mujeres valientes*.

A continuación, se explica la potencia que le aporta al sistema la posibilidad de prohibir rasgos, de forma similar a como se hizo con la concordancia e imposición. La *prohibición* de un rasgo en el sistema actual únicamente requiere una regla, mientras que, si se utilizan $categorías_{rasgos}$, es necesario especificar el listado de todas las reglas con las combinaciones no prohibidas de los rasgos: es decir, si se desea prohibir en algún punto los $nombres_{sing\ a-tónica}$, será necesario proporcionar toda la lista de sintagmas femeninos, masculinos y neutros, con género singular, plural y neutro, es decir, listar todas las reglas excepto las relativas a los $nombres_{sing\ a-tónica}$.

La prohibición de rasgos se incluye en forma de filtro que elimina de la lista de predicciones las palabras que contengan el rasgo prohibido, e interrumpe el análisis cuando la palabra escrita por el usuario lo contiene.

Tratamiento de los rasgos en los símbolos no terminales

En algunas estructuras hay rasgos que deben transmitirse de un sintagma a otro, como, por ejemplo, en la frase *la casa de mi amigo es nueva*, el género y número de *la casa* debe guardarse hasta que aparezca *nueva*, sin afectar al género y número de ninguna de las palabras intermedias, que, a su vez, deben cumplir sus propias concordancias entre sí. Esto quiere decir, entre otras cosas, que el primero de los sintagmas tiene libertad de género y número, pero, una vez escrito el primero, éste restringe los valores que pueden tomar los rasgos de las palabras siguientes para que concuerden con él.

De nuevo se ha ampliado la potencia del analizador para soportar la concordancia con símbolos no terminales, que permite que las reglas puedan ser subdivididas en fragmentos reutilizables en otros puntos, y, además, que cumplan las reglas de concordancia de las diferentes estructuras en las que puedan ser insertadas. Se ha añadido la posibilidad de transmitir rasgos *a* la regla y *desde* la regla.

Para tratar los rasgos de los símbolos no terminales, es necesario definir el conjunto de rasgos característicos de cada uno, que serán los que se necesiten en las reglas en las que se utilice cada símbolo no terminal (para imponer valores o concordancias con otros rasgos de la regla). Una vez definido el conjunto de rasgos, se debe indicar, en cada regla,

la forma de obtener los valores correspondientes a partir de las palabras del sintagma, y la manera de transmitir a la regla los valores de dichos rasgos si se imponen desde fuera o deben concordar.

Por ejemplo, con la regla siguiente se pueden modelar frases como *la casa es bonita*.

```
O :
    SN   [gen{X} num{Y}] ? 175
    v    [lema{ser} num{Y}]
    adj  [gen{X} num{Y}]
    ;
```

En este tipo de frases, el sintagma nominal inicial, además de la concordancia interna de género y número, debe concordar con los rasgos del adjetivo, y sólo en número con el verbo. Los rasgos que se transmiten al exterior del sintagma nominal se escriben a la derecha del símbolo no terminal, utilizando los mismos signos auxiliares indicados anteriormente. De esta manera, la regla del sintagma nominal quedaría como sigue:

```
SN    [gen{X} num{Y}]:
det   [gen{X} num{Y}] ? 175
n     [gen{X} num{Y}]
;
```

Este formato indica que el sintagma nominal tiene como rasgos los del determinante y nombre que lo forman. Se elegirán los más restrictivos, es decir, si uno de ellos es neutro con respecto a un rasgo y el otro no, el rasgo del símbolo no terminal es el más restrictivo. Además, este formato también se utiliza para transmitir valores de rasgos al interior de la regla: si el símbolo no terminal debe concordar con rasgos de palabras anteriores a él, X e Y ya tendrán los valores adecuados cuando se invoque la regla correspondiente, y se predecirá con los valores adecuados desde el primer momento.

Por ejemplo, dada la regla:

```
O :
    SN   [gen{X} num{Y}] ? 175
    v    [lema{ser} num{Y}]
    SN   [gen{X} num{Y}]
    ;
```

para modelar frases como *mis primos son periodistas*. El primer sintagma nominal consta de un determinante con número plural y género neutro, y un nombre masculino plural, y devuelve como valores de sus rasgos X=*masculino* e Y=*plural*. El siguiente elemento de la regla, *son*, es una forma del verbo *ser* que concuerda en número con el sintagma nominal. Cuando se llega a la posición anterior al último sintagma nominal, X e Y ya tienen valores predefinidos con los que concordar, y sólo admitirán como válidos sintagmas nominales masculinos plurales. *Periodistas* es una palabra plural de género neutro, que, por tanto, cumple las condiciones de concordancia y la frase se considera gramaticalmente correcta.

Con la misma regla se podría modelar, por ejemplo, *es profesora*. En esta frase el primer sintagma nominal (opcional) no aparece, por lo que no se impone ninguna restricción al número de la forma del verbo *ser* que se espera. Al escribir *es*, se impone el número de

la siguiente palabra, pero no el género, es decir, se podrían admitir como continuaciones válidas tanto *profesor* como *profesora*, y ambas serían gramaticalmente correctas.

A continuación se muestra la manera de seleccionar qué rasgos y de qué elementos determinan el valor de los rasgos del símbolo no terminal. Por ejemplo:

```
SN      [gen{X} num{Y}]:
   det  [gen{X} num{Y}] ? 175
   n    [gen{X} num{Y}]
   prep
   det  [gen{Z} num{M}] ? 75
   n    [gen{Z} num{M}]
   ;
```

La regla anterior modela un sintagma nominal que consta de un determinante y un nombre, seguidos por un sintagma preposicional (formado por la preposición, determinante y nombre siguientes). El determinante y el nombre iniciales deben concordar entre sí, y posiblemente con otros elementos de las reglas que le invocan, y, por eso, los signos que se utilizan para indicar la concordancia entre ellos son X e Y, los mismos que aparecen a la derecha del símbolo no terminal SN. Sin embargo el determinante y nombre finales (tras la preposición), utilizan signos distintos, Z y M, expresando que deben concordar entre sí, pero el valor de sus rasgos no se transmite al exterior, ni puede estar condicionado previamente, ya que ninguno de estos signos acompaña al SN.

Hay reglas que tienen un comportamiento peculiar con respecto a los rasgos, ya que contemplan casos particulares, en los cuales los valores de los rasgos del símbolo no terminal no se obtienen de los valores de los rasgos de los elementos de la regla. En estos casos, cuando se definen los rasgos del símbolo no terminal, se les asocia directamente un valor, en lugar de los signos $X, \ldots M$. Este es el caso de la regla que trata los sintagmas nominales compuestos por un determinante más un nombre que comienza por a tónica en singular:

```
SN      [gen{f} num{s}]:
   det  [gen{m} num{s}] ? 175
   n    [gen{a} num{s}]
   ;
```

Esta regla impone un determinante masculino singular, más un nombre que comienza por a tónica en singular, para modelar casos como *el águila*. Sin embargo, el género devuelto por la regla, para su uso y concordancia con el resto de los símbolos, es femenino (*el águila parece blanca*), y, por eso, se impone dicho valor a la derecha de SN.

Estas reglas no se utilizan en los momentos en los que se imponen otros rasgos, ni cuando se requiere concordancia con rasgos incompatibles con ellos: si en un punto se exige un sintagma nominal que concuerde con masculino, no se expande la regla anterior.

Es posible también imponer un valor de un rasgo, y su concordancia con otro. Esto puede ser interesante para modelar relaciones en las que el rasgo debe concordar con otros rasgos exteriores, y, en algunos puntos (en ciertas reglas), además, deben tener un determinado valor. En este caso, el valor del rasgo de las palabras predichas será el impuesto *si concuerda con el otro valor*. Si el valor que se impone es incompatible con el valor con el que debe concordar, se interrumpe el análisis. Un ejemplo de este

comportamiento, es la siguiente regla:

```
SN      [gen{X} num{Y} st-prondet{Z} per{3}]:
   pron [gen{X} num{Y} st-prondet{Z} tipo-pron{~pers} st-prondet{~ard}]
   & 1000
   ;
```

Esta regla modela los sintagmas nominales formados únicamente por un *pronombre no personal*, que no sea *determinado*. El tipo de pronombre o determinante puede venir impuesto desde el exterior (st-prondet{Z}), y, si el valor que se impone es *pronombre personal o determinado*, esta regla no se expande.

De la misma forma, es posible utilizar la recursión, permitiendo que una regla, en su proceso de expansión durante el análisis, se llame a sí misma[15]. Un ejemplo de su uso son las siguientes reglas, utilizadas para modelar desde las formas verbales más sencillas (un solo verbo) hasta las perífrasis verbales de cualquier longitud:

```
SFormaVerbal        [tipo-tv{M} tv-fl-md{X} num{Y} per{Z}]:
     v      [tipo-tv{M} tipo-tv{fl} tv-fl-md{X} num{Y} per{Z}] & 700
   | v      [tipo-tv{M} tv-fl-md{X} lema{ir} num{Y} per{Z}]
     prep   [lema{a}]
     SFormaVerbal [tipo-tv{inf}] & 100
   | v      [tipo-tv{M} tv-fl-md{X} lema{haber} num{Y} per{Z}]
     SFormaVerbal [tipo-tv{par}] & 200
     ;
```

Estas son tres de las reglas utilizadas para modelar las formas verbales. La primera de ellas modela las formas verbales compuestas por una sola palabra.

La segunda regla genera las perífrasis verbales de la forma *ir + a + infinitivo*. El verbo *ir* debe concordar con el modo de verbo flexivo (*tv-fl-md*) que se le imponga desde el exterior, después se impone la preposición *a*, y por último, una forma verbal en infinitivo: este infinitivo no es necesariamente un verbo, sino que llama recursivamente a las reglas que expanden SFormaVerbal.

La tercera regla modela los tiempos compuestos, formados por *haber + participio*. En ella se puede observar que la primera palabra es una forma del verbo *haber* que debe concordar con los rasgos impuestos desde el exterior, y el segundo elemento es un participio masculino singular: como es una forma no flexiva, no se le pasa información sobre el modo, ni debe concordar con ninguna información tomada del exterior. Como este elemento es también una SFormaVerbal, vuelve a desplegar este mismo conjunto de reglas:

- La primera regla, que en este caso predice verbos en participio.
- La segunda regla, con la perífrasis verbal, pero, esta vez, el verbo *ir* en participio. De esta forma se predicen formas como *haber ido a comprar*.
- La tercera regla, que predice *haber + participio*, pero esta vez *haber* está en participio.

[15] No se permite la *recursión por la izquierda*, es decir, si un símbolo no terminal forma parte de alguna de las reglas que lo expande, no puede ocupar la primera posición.

Con la segunda regla el proceso es similar, permitiendo estructuras de tipo *voy a haber hecho* (aunque semánticamente no tenga mucho sentido). Se puede, además, recursivamente aplicar ambas, llegando a construcciones de tipo *he ido a haber hecho*.

Si al conjunto de reglas anterior se le añaden las que modelan *estar + gerundio* y *pensar + en + infinitivo*, al aplicarse recursivamente el proceso anterior se pueden analizar estructuras como *hemos estado pensando en ir a cenar*: *hemos* impone un participio, en este ejemplo, *estado*, que al ser del verbo *estar* permite que continúe la regla que modela *estar + gerundio*. Al gerundio *pensando*, como forma del verbo *pensar*, le puede seguir *en* y un infinitivo, y a *ir* le sigue *a* más un infinitivo. Por supuesto, en cualquier punto se podría detener el proceso de formar perífrasis, ya que, por ejemplo, *pensando*, avanza la regla *pensar + en +* infinitivo, pero además puede acabar la SFormaVerbal si consideramos la primera regla, que modela las formas compuestas por una sola palabra.

Con este método, en el que se exigen concordancias e imponen rasgos, es posible escribir conjuntos de reglas, condicionando su paso por unas u otras al valor concreto de los rasgos en cada punto de la frase. Esto permite, entre otras cosas, hacer reglas recursivas, controlando con los rasgos el paso por las reglas adecuadas. En conclusión, toda esta metodología, basada además en las diversas contribuciones de mecanismos de manejo de información que hemos incluido en el analizador de la gramática, han conseguido aumentar sensiblemente la potencia descriptiva con que contamos para recoger los fenómenos lingüísticos que hemos considerado más relevantes cara a la utilización de la gramática formal como método de predicción.

3.3.3.3.5 Técnicas de poda de caminos

En cada punto del análisis, debido al número de reglas del sistema, a los elementos opcionales y a la ambigüedad de las palabras, se despliegan gran cantidad de reglas, pudiendo llegar a superar las 50.000 o incluso 100.000. Procesar tal cantidad de reglas puede llevar demasiado tiempo, haciendo el sistema inviable para su uso en tiempo real.

Para mejorar su tiempo de respuesta se ha implementado un mecanismo de poda de caminos[16] usando búsqueda en haz[17].

Los mecanismos de poda evitan la explosión combinatoria, expandiendo únicamente las ramas más prometedoras del proceso en cada nivel. La decisión de qué ramas eliminar se basa en el uso de un heurístico que selecciona los más próximos al objetivo, como se explica en [Lowe80]. Existen distintas variantes de estas técnicas, desde la que utiliza un umbral sobre la probabilidad más alta para eliminar las hipótesis menos probables, con una adecuada estructuración del espacio de búsqueda (explicada en [Ney92] y [Coll96]), hasta las que usan histogramas para la limitación del mismo ([Ortm96]). En [Stei94] pueden encontrarse detalles adicionales para refinar la estrategia de poda.

En el analizador modificado, se utiliza una variante de la búsqueda en haz basada en umbral descrita en [Coll96], en la que la aplicación de la poda de caminos sólo se

[16] *Pruning* en la nomenclatura inglesa.

[17] *Beam search* en la nomenclatura inglesa.

lleva a cabo si el número de ramas activas en la búsqueda supera un cierto número. Los parámetros que se deben definir previamente son el número de ramas activas a partir del cual se aplica el algoritmo de poda (δ_a) y el umbral de recorte a usar (β_{bs}). Los valores de estos parámetros se fijan dependiendo de cuánto se desee aumentar la velocidad del análisis. Una vez fijados los valores, cada vez que se avanza el análisis (es decir, cada vez que se completa una palabra), se aplica el siguiente algoritmo:

- Se calcula el número de ramas activas, $\mathcal{N}$.
- Si $\mathcal{N} \leq \delta_a$ se continúa el análisis sin modificaciones.
- Si $\mathcal{N} > \delta_a$ se calcula la probabilidad de la mejor rama activa p_{mejor} y se eliminan aquellas ramas activas cuya probabilidad sea menor que el umbral de recorte aplicado a la mejor probabilidad, es decir, todas aquellas cuya

$$p_{rama} < \frac{p_{mejor}}{\beta_{bs}} \tag{3.21}$$

Los parámetros de control han sido seleccionados experimentalmente, de forma que reduzcan el espacio de búsqueda hasta que el tratamiento computacional del árbol resultante pueda realizarse en un tiempo razonable, pero controlando a su vez que no se produce una pérdida apreciable de rendimiento en el sistema.

3.3.3.4 Reglas de reescritura

Las reglas, junto con el diccionario general, constituyen las fuentes de información básicas de la predicción gramatical de palabras cuando se utiliza el analizador. Son las que expresan formalmente las estructuras consideradas gramaticalmente correctas, basándose en el conjunto de categorías y rasgos del diccionario general.

Se han obtenido a partir del estudio de corpora de texto hasta determinar los patrones repetitivos más importantes, definiendo de modo progresivo las reglas declarativas, dando cuenta del mayor número de observaciones lingüísticas posibles que sean relevantes sintácticamente, e introduciendo asimismo información frecuencial, y las relaciones entre los rasgos, lemas y significantes. El listado final de reglas se muestra en el apéndice C (página 195).

3.3.3.5 Gestión de errores

El método de predicción basado en la gramática formal procesa de izquierda a derecha el fragmento escrito de la frase en curso, y predice las categorías y las restricciones sobre los rasgos, lemas o el significante de la palabra esperada, en base a las reglas que estén activas.

En cada punto del análisis están activas todas las reglas que pueden ser seguidas por el conjunto de palabras de la frase, y, al continuar escribiendo, si no se aplican métodos de recuperación de errores, se van eliminando las reglas que no son seguidas por las nuevas palabras, prediciendo según las reglas que continúen activas.

En este escenario es posible que, en un punto determinado, la palabra escrita no sea considerada válida por ninguna de las reglas activas (p. ej., si es una falta de ortografía) y, entonces, se eliminarían todas las reglas, el análisis se interrumpiría y este método de predicción no sería capaz de proponer ninguna categoría.

La tipología de errores con los que se enfrenta el analizador modificado es más amplia que la de un analizador tradicional, ya que sus nuevas características producen posibilidades de error adicionales. Las reglas pueden ser interrumpidas por:

1. La escritura en el texto de una palabra desconocida (falta de ortografía, tecnicismo, etc., según la tipología vista en el apartado 3.3.2.1.5).

2. Falta de cobertura de la gramática: la regla que sigue la frase que está siendo escrita por el usuario no está incluida en la gramática.

3. Inserción de una palabra en una posición no debida (inserción).

4. Eliminación de una palabra (borrado).

5. Escritura de una palabra perteneciente una categoría no esperada (sustitución).

6. Incoherencia de los rasgos de la palabra con respecto a los rasgos impuestos, concordantes o prohibidos.

7. No coincidencia del significante o lema de la palabra con el lema o palabra impuesta.

8. Coincidencia del significante o lema con la palabra o el lema prohibido.

Para reducir los efectos negativos de estos errores gramaticales sobre la predicción de palabras es necesario incluir mecanismos de recuperación que aumenten la flexibilidad y robustez sistema.

En el modelo actual sólo está tratada la inserción de palabras desconocidas, proponiéndose el uso de la rutina de categorización de palabras desconocidas explicada en el apartado 3.3.2.1.5, que le asigna las categorías nombre, adjetivo, verbo y adverbio, con sus frecuencias asociadas. Esta categorización puede utilizarse como si hubiera sido extraída del diccionario general, y la palabra hará avanzar las reglas que esperan dichas categorías.

Si, a medida que avanza el análisis, aparece cualquiera de los demás tipos de error y se llega a una situación en la que no hay ninguna regla activa, se retrocede a modelos gramaticales menos restrictivos, los basados en *tripos* o *bipos* de categorías, que solamente consideran información de la o las dos categorías anteriores al predecir, o incluso a los modelos basados en unigramas. Cuando aparece el siguiente punto del texto, que se considera el comienzo de la siguiente oración, se sitúa el analizador en el punto inicial del análisis.

En el apartado de líneas futuras se esbozan posibles planteamientos para gestionar los tipos de error aún no tratados en el analizador.

3.3.4 Métodos adicionales a la predicción de palabras

En esta sección se explican diversos métodos que no generan por sí mismos una lista de palabras predichas, pero pueden mejorar el rendimiento de la predicción o complementarla.

3.3.4.1 Filtro de las palabras rechazadas

En numerosas ocasiones ocurre que hay palabras en el diccionario que comienzan por la misma cadena de letras que la que se está escribiendo, y que aparecen en la lista de candidatas desde la primera letra hasta que el usuario prácticamente acaba de escribir la palabra que desea.

Piénsese, por ejemplo, en el proceso para escribir *correspondientemente*: dependiendo de los métodos de predicción gramaticales empleados y del contenido de los diccionarios, es posible que las palabras *correspondiente*, *correspondientes*, *corresponder* y *correspondencia* aparezcan en la lista de palabras predichas desde el momento en que se escribe la *c*, y ocupen la lista al menos hasta que se llegue a *correspondi*, momento en el cual sólo desaparecen las dos últimas.

Cabe suponer que el usuario que está utilizando el sistema verifica la lista cada vez que escribe una letra y ve las palabras en el momento en el que son mostradas. Si el usuario no elige ninguna de las palabras de la lista, es porque ninguna de ellas es la que desea escribir, y no es necesario volver a mostrarlas aunque al seguir escribiendo letras sigan siendo predicciones válidas.

En el ejemplo anterior, se le enseñaría dicha lista al escribir la *c*, y, cuando el usuario escribiera la *o*, se considerarían todas las palabras rechazadas, y no se volverían a mostrar, dejando libre la lista para otras palabras, acelerando el proceso de aparición de la palabra deseada.

Los resultados de este filtro de palabras rechazadas son buenos, como se muestra en el capítulo correspondiente, especialmente si se considera que las palabras solamente aparecen una vez en la lista de palabras predichas. De cara al uso del sistema, sin embargo, es preferible modificarlo ligeramente, permitiendo que las palabras se muestren más de una vez, ya que puede ocurrir que el usuario no vea la palabra deseada a la primera y, por no predecirse ésta en sucesivas ocasiones, tenga que completarla letra a letra.

Para evitar esta penalización, el sistema debe contemplar la posibilidad de configurar el número máximo de veces que las palabras son mostradas al usuario. Este valor dependerá de diversos factores como el grado de atención/cansancio de la persona en un momento determinado o su capacidad lecto-escritora (que puede hacer que ignore parte de las palabras de la lista, porque para leerla necesite demasiado tiempo), entre otros.

3.3.4.2 Predicción de terminaciones

Como se ha explicado en secciones previas, puede que el usuario desee escribir palabras que no estén contenidas en los diccionarios (especialmente si está escribiendo sobre un tema técnico o especializado), y entonces no se beneficie de la predicción de palabras.

En castellano hay una serie de terminaciones, bastante comunes, que pueden ayudar a completar palabras, acelerando la escritura de aquellas que no estén contenidas en el diccionario. Son, por ejemplo, las terminaciones de verbos, para ayudar a escribir los tiempos no incluidos en el diccionario, o algunas cadenas (correspondientes a sufijos o simplemente frecuentes estadísticamente) como *-mente*, *-ación*. Estas terminaciones pueden resultar muy útiles, ahorrando al usuario la escritura de varias letras y el espacio posterior, es decir, produciendo los mismos beneficios que la predicción de palabras .

Por esta razón se ha incluido en el sistema, en paralelo con la predicción de palabras, la de terminaciones, y al usuario se le mostrarán ambas listas.

Las terminaciones no se corresponden exactamente con desinencias, sino con las cadenas de letras con las que, en los textos de entrenamiento, han acabado las palabras más frecuentemente. La lista que se le muestra al usuario está compuesta por las terminaciones más probables que comiencen por la última letra que ha escrito, por ejemplo, tras escribir una *m*, se muestra *-mente*, *-mos* y *-mas*, y tras *p* enseña *-porte*, *-plica* y *-parte*.

3.3.4.3 Escritura automática de caracteres

Además de la predicción de palabras, hay ciertos caracteres que se pueden añadir en ciertas condiciones, ahorrando pulsaciones al usuario, como, por ejemplo, los espacios después de los signos (punto, dos puntos, coma o punto y coma).

Un paso más en este sistema es la eliminación del espacio anterior al signo: si se escoge una palabra de la lista de predicciones, el sistema añade automáticamente un espacio después. Si el usuario escribiera un signo a continuación, debería borrar dicho espacio, pero no es necesario, porque éste se eliminará automáticamente, se escribirá el signo, y se añadirá el espacio posterior, ahorrando dos pulsaciones.

Si el signo es un punto, el siguiente carácter se escribirá automáticamente en mayúsculas, ahorrado otra pulsación.

El ahorro producido por este algoritmo depende de la cantidad de signos que el usuario emplee en el texto, y es bastante alto, como puede comprobarse en el apartado 4.1.3.6 (página 158) del capítulo de evaluación.

3.4 Módulo de coordinación

El módulo de coordinación es el bloque que nos ha permitido la realización y validación de las diversas metodologías de integración de las fuentes de información y los métodos de predicción que hemos probado y con el que hemos encontrado la opción más conveniente. Este módulo se encarga de:

- Obtener y procesar la entrada procedente de la interfaz de usuario (texto escrito por el usuario).
- Gestionar los flujos de información con los distintos métodos de predicción, coordinando los datos que le proporciona y solicita a cada uno.

- Gestionar las transacciones con los diccionarios.
- Obtener el listado de palabras predichas propuesto por cada método y diccionario, y seleccionar el conjunto definitivo de palabras.
- Gestionar los métodos adicionales a la predicción de palabras.
- Enviar las palabras y sufijos predichos a la interfaz de usuario.

La primera tarea y la última realizan la comunicación entre el sistema de predicción de palabras y el usuario, obteniendo la entrada (el texto escrito) y mostrando la salida (palabras predichas). Su complejidad depende de la interfaz de la aplicación en la cual esté integrada la predicción, ya que su diseño será diferente para cada tipo de sistema final: la predicción puede estar incluida en un sistema controlado por barrido o en una ventana que se utiliza con el teclado convencional, el procesador de texto en el que el usuario escribe puede estar incluido en la aplicación o ser externo, etc.

La segunda tarea consiste en gestionar los flujos de información con los distintos métodos de predicción. Como se ha visto en los apartados correspondientes a la descripción de dichos métodos, cada uno exige una determinada información de entrada (procedente del texto escrito por el usuario, procesado en mayor o menor medida), y proporciona una información de salida diferente (un conjunto de restricciones sobre las palabras siguientes). A continuación se muestran los datos de entrada que el módulo de coordinación proporciona a cada uno y la salida que recibe de ellos:

- El método basado en el analizador gramatical necesita tanto el listado de palabras de la frase en curso como la información gramatical asociada a cada una (el conjunto de sus posibles lemas, categorías y rasgos, obtenidos al categorizar). A la salida proporciona un listado con las categorías gramaticales a las que puede pertenecer la palabra siguiente, y, asociada a cada una, las restricciones impuestas por la regla que la ha predicho: valores de los rasgos impuestos, prohibidos y con los que debe concordar, y lemas y/o significantes prohibidos o impuestos. Además también obtiene la probabilidad con la cual cada una de las categorías puede ser la siguiente.
- El método basado en *tripos* necesita conocer la categoría a la que pertenecen las dos últimas palabras escritas y el género y número de la última de ellas. Con ello obtiene las categorías a las que puede pertenecer la palabra siguiente, su género y número (si la secuencia de categorías que se genera requiere concordancia) y la probabilidad con la cual cada una de las categorías puede ser la categoría siguiente.
- El método basado en *bipos* es idéntico al anterior, exceptuando que no requiere la categoría gramatical de la penúltima palabra.

La tercera tarea, la gestión de los diccionarios, está dividida en varias partes, ya que cada uno de ellos es capaz de proporcionar diferente información, a partir de distintos datos, como se muestra a continuación:

- El diccionario general se utiliza en dos puntos del proceso de predicción:

  - Al categorizar. El módulo de coordinación le proporciona la palabra y el diccionario entrega a la salida toda la información que contiene sobre ella: el conjunto de categorías gramaticales a las que puede pertenecer, sus rasgos y los lemas correspondientes.
  - Al predecir. A partir del conjunto de restricciones impuestas por los módulos de predicción, el diccionario general proporciona el listado de palabras que las cumplen y toda la información asociada a cada una.

- El diccionario personal y los temáticos solamente se utilizan en la etapa de predicción, generando dos listados de palabras predichas diferentes cada uno:
  - Un primer listado conteniendo los bigramas correspondientes a una determinada palabra: tomando como entrada dicha palabra, los diccionarios proporcionan aquellas que la han seguido en el texto actual/de entrenamiento más frecuentemente, lo que se denomina sus *palabras siguientes*.
  - Un segundo listado obtenido tomando como entrada el conjunto de restricciones gramaticales, donde los diccionarios dan como salida el conjunto de palabras pertenecientes a ellos que las cumplen y su información frecuencial asociada.

El proceso que lleva a cabo el módulo de coordinación para obtener el listado de palabras predichas, a partir de todos los diccionarios y métodos de predicción, se muestra a continuación. En todos los pasos se considera que las palabras propuestas se filtran de forma que, además de las restricciones impuestas por cada método, solamente pasan las palabras que comienzan por las letras ya escritas de la palabra en curso.

- En primer lugar, se procesa el texto del usuario para obtener el listado de palabras de la frase en curso.

- Posteriormente se aplica la predicción basada en bigramas de palabras, que solamente está disponible para el diccionario temático y el personal. Si hay algún diccionario temático activo, el módulo de coordinación le proporciona la última palabra escrita por el usuario y obtiene su lista de *palabras siguientes* para ese tema. Si el conjunto anterior de palabras es inferior al número máximo de palabras predichas, se obtiene la lista de *palabras siguientes* del diccionario en curso.

- Si aún no se ha alcanzado el número máximo de palabras, se pasa a la predicción gramatical. El módulo de coordinación intenta obtener, a partir del método de predicción activo más potente, el conjunto de restricciones que deben cumplir las palabras predichas. Si dicho módulo no proporciona una salida adecuada (por ejemplo, si la frase no es gramaticalmente correcta según el conjunto de reglas del sistema y el método basado en el analizador no predice ninguna categoría), el módulo de coordinación solicitará información a los métodos basados en *tripos* y *bipos*, por ese orden, hasta que uno de ellos le proporcione el conjunto de restricciones buscado.

- Si se ha activado algún diccionario temático, el módulo de coordinación averigua la probabilidad de cada palabra de dicho diccionario, según la fórmula 3.20 (página 81). Las palabras que cumplen las restricciones gramaticales son añadidas a la lista por orden de probabilidad.

- Si la lista aún admite más palabras, se realiza una combinación entre las procedentes del diccionario personal y general que cumplen las restricciones del módulo gramatical. Se averigua la probabilidad de la palabra en cada diccionario, aplicando la fórmula 3.20, y se combinan según la siguiente ecuación:

$$p(w) = Q_{curso} \cdot p_{curso}(w) + Q_{gral} \cdot p_{gral}(w) \tag{3.22}$$

  Donde:

  - $p_{curso}(w)$ y $p_{gral}(w)$ son el resultado de aplicar la fórmula 3.20 para el diccionario general y en curso.
  - $Q_{curso} = 0.35$ y $Q_{gral} = 0.65$ son los pesos que controlan la aportación de cada diccionario. Han sido entrenados para maximizar el número de pulsaciones ahorradas en textos de longitud entre 1000 y 2500 palabras (longitud aproximada de los textos que pueden ser generados en una sesión). Si se aumenta Q_{curso} con respecto a Q_{gral}, disminuye la eficiencia en textos de pequeña longitud, cuando el diccionario en curso no está aún suficientemente entrenado. Si se disminuye la aportación del diccionario en curso, se reduce la capacidad de adaptación, y, por tanto, empeoran los resultados de la predicción, especialmente en textos de mayor longitud.

- Si aún no se ha completado la lista de palabras predichas, se añaden las palabras más probables del diccionario temático, en curso y general (por ese orden) que comiencen por las letras escritas de la palabra en curso, aunque no cumplan las restricciones gramaticales.

- En cada uno de los pasos anteriores, antes de añadir una palabra a la lista de predicciones, se comprueba:

  - Si ha sido rechazada por el usuario (que ya se le haya mostrado durante la escritura de la palabra en curso y no haya sido elegida).
  - Si la palabra pertenece o no al diccionario general. Si está activado el control que impide la predicción de faltas de ortografía, no se añaden a la lista de predicciones las palabras del diccionario temático o del personal que no están incluidas en el diccionario general.

- A las palabras predichas por el diccionario se añade la lista de sufijos más probables que comiencen por la última letra escrita.

Tras esta serie de pasos, el módulo de coordinación dispone de la lista de palabras y sufijos a predecir, que envía a la interfaz del programa correspondiente, y espera la actuación del usuario, para, después, volver a comenzar el ciclo.

3.5 Interfaz de usuario

La *interfaz de usuario* es una de las partes críticas de una aplicación, ya que determina el acceso a su funcionalidad. La *interfaz de entrada* es la parte mediante la cual el usuario

introduce datos, indica al programa qué tarea realizar, etc., y en la *interfaz de salida* se muestran los resultados.

La facilidad o dificultad de la interacción del usuario con cada programa dependen básicamente de su interfaz, tanto de entrada como de salida, por lo que es necesario aplicar estrategias de diseño que faciliten el uso de la aplicación, haciéndola más *amigable*. En [Smit96] y [Mart96] se pueden encontrar recomendaciones para diseñar adecuadamente interfaces de usuario.

Los programas orientados a personas con alguna discapacidad deben considerar que estos usuarios pueden encontrar problemas con las interfaces de entrada y/o salida convencionales, por lo que éstas deben ser adaptadas.

A continuación, se explicarán brevemente los mecanismos incluidos en la interfaz de la aplicación *Predice* (Editor Predictivo para Personas con Discapacidades Motrices), que será la que se utilizará en la evaluación de usuario. Una breve descripción de *Predice* puede ser encontrada en el apartado 2.5.6 (página 25) del encuadre científico-tecnológico, ya que es una aplicación de ayuda a la escritura muy utilizada entre los usuarios de pulsador. En [Pala00] se puede encontrar una descripción de la última versión de *Predice*, con funcionalidades añadidas (el módulo de matemáticas y el comunicador), con el diseño de la interfaz optimizado para cubrir las necesidades de cada función.

3.5.1 Adaptaciones de la interfaz de *Predice* para personas con distintas discapacidades

Predice es un editor de texto que incluye predicción de palabras, preparado para poder ser utilizado por personas con problemas de índole muy diferente: auditivos, visuales y, muy especialmente, físicos. Su pantalla principal se muestra en la figura 3.4.

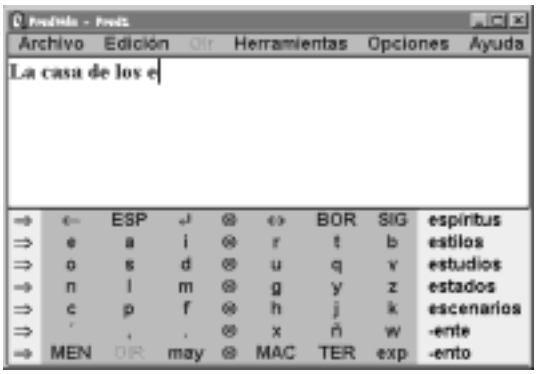


Figura 3.4: Pantalla principal de *Predice*.

Entre las adaptaciones incluidas en *Predice*, para que pueda ser utilizado por personas con distintos tipos de discapacidad, están las siguientes:

- Para personas con *discapacidad auditivas*, se reproducen por escrito todos los mensajes sonoros del sistema.
- Para los usuarios con *discapacidades visuales* se han realizado distintas adaptaciones,

dependiendo de la severidad del problema. Para casos leves, es suficiente con la posibilidad de controlar tamaños y colores de las letras y fondos de los menús. Cuando esta solución no es suficiente, se toman otras medidas, pudiéndose llegar incluso a la lectura de todas las opciones de los menús secuencialmente.

- Para personas con *discapacidades físicas* se han incluido en el sistema una serie de adaptaciones que dependen de la movilidad del usuario, ya que no pueden utilizar el sistema de la misma manera una persona que tenga un ligero problema de precisión en sus movimientos y otra con espasmos, que únicamente pueda controlar adecuadamente el cuello o una rodilla.

  - Para las personas con *problemas motrices ligeros*, que puedan utilizar el teclado y/o el ratón, aunque con dificultades, se han incluido en *Predice* mecanismos para acelerar el acceso, permitiendo la interacción con el programa utilizando el método más cómodo en cada caso concreto. *Predice* les ayudará, disminuyendo el número de pulsaciones necesarias para escribir el texto por el apoyo de la predicción de palabras y por los mecanismos auxiliares (inserción automática de espacios y mayúsculas).
  - El mayor esfuerzo en el diseño de la interfaz ha sido realizado para optimizar el acceso de las personas con *discapacidades físicas severas* (p. ej., parálisis cerebral). Para estos usuarios el acceso al ordenador con los métodos habituales (teclado o ratón) está totalmente vetado, y, además, suelen tener otros problemas (no se pueden comunicar con normalidad, ni desplazarse de forma independiente). Las personas con problemas especialmente graves dependen del ordenador para realizar tareas básicas de su vida diaria (p. ej., comunicación o control de entorno), por lo que es particularmente importante realizar un diseño óptimo de la interfaz. Las adaptaciones de *Predice* para que pueda ser utilizado por estas personas se describen en el siguiente apartado.

3.5.2 Adaptación para personas con discapacidades físicas graves

Los usuarios con problemas físicos graves no pueden utilizar el teclado ni el ratón para acceder a las aplicaciones, por lo que es necesario adaptar éstas para que se pueda acceder a todas sus opciones con uno o dos *pulsadores*. Para ello se implementa un sistema de *barrido*, en el que un cursor resalta secuencialmente todas las opciones del programa, y el usuario presiona el pulsador cuando se resalte la deseada. Este método es muy lento y un diseño inadecuado de la interfaz puede ralentizarlo aún más.

Predice incluye barrido de sus opciones, y puede ser controlado con uno o dos pulsadores. Para aumentar su flexibilidad, se han incluido varios parámetros a configurar en el barrido. Cada usuario debe elegir el más conveniente según sus características, ya que, lo que para una persona resulta beneficioso, para otra puede ser contraproducente. Se permite controlar:

- La *velocidad* del barrido, ajustándola al máximo que el usuario pueda controlar, ya que determinará la velocidad de escritura.

- La *duración mínima* de la pulsación, para ser considerada válida. Este parámetro es importante para los usuarios con espasmos o temblores, que, de forma involuntaria, puedan generar pulsaciones espúreas. El sistema no procesará ninguna pulsación de duración menor a la que se configure.

- La *forma de uso* del pulsador. Puede pulsarse cuando se desea seleccionar una opción, o utilizarse de forma inversa: normalmente pulsado, y soltarlo para seleccionar. Su configuración dependerá de cuál sea el movimiento que el usuario realice con mayor precisión y produzca menos cansancio.

- El *modo* de barrido, es decir, la forma en el que se recorren las distintas opciones en el barrido. Existen varios métodos, cada uno de los cuales requiere un número de pulsaciones y un tiempo diferente de acceso a las opciones. En [Cher94] se puede encontrar un estudio sobre las interfaces de barrido. En dicho artículo se estudian diferentes métodos de acceso, con uno o más pulsadores, y diferentes modos de barrido, por ejemplo, barrido por filas y columnas, división binaria[18], una división y filas y columnas[19], y, por último, dos divisiones y filas y columnas[20]. Para cada uno de los métodos proporciona el tiempo medio necesario para seleccionar una opción y el número de pulsaciones requerido. Para menús de menos de 6 filas y columnas, el método más rápido es el barrido por filas y columnas. Para menús más grandes, los dos últimos métodos presentan reducción de tiempo significativa, y a partir de 8*8 el último método es el más rápido. El método por división binaria es siempre el método más lento, con el agravante de ser el que más pulsaciones requiere para seleccionar cualquier opción. Además también se estudia la posibilidad de utilizar dos o más pulsadores, aunque en este caso deben considerarse las limitaciones físicas y cognitivas del usuario, que pueden impedirle manejar de forma fácil más de un pulsador.

  En [Lesh98] se muestran también los resultados de un estudio de diversas configuraciones de la matriz, introduciendo, no solamente modificaciones en la posición de las opciones, sino también predicción de palabras y de caracteres.

  En *Predice* se han incluido:

  - *Barrido automático lineal controlado por un pulsador.* Es el método más lento y el que requiere menor cantidad de pulsaciones para elegir cada opción, por lo que es el método que menos cansancio físico produce.
  - *Barrido automático por filas y columnas controlado por un pulsador.* Es más rápido que el anterior, aunque requiere dos pulsaciones para seleccionar cada opción.
  - *Barrido manual controlado por dos pulsadores* (con uno de ellos se controla el barrido y con el otro se selecciona la opción deseada). Deja total libertad al usuario para controlar la velocidad a la que realiza el barrido, aunque el número

[18] Se divide por dos la matriz, y se barren ambas mitades. Cuando una de ellas es seleccionada, se divide a su vez por dos y se repite el proceso hasta que se elige la opción deseada.

[19] Divide la matriz por dos, y la mitad seleccionada es barrida por filas y columnas.

[20] Aplica dos veces la división binaria, y en el conjunto final barre por filas y columnas.

de pulsaciones necesarias para seleccionar una opción es mucho mayor que con los otros métodos.

Dependiendo de cuál sea el criterio primordial para el usuario, velocidad o esfuerzo, elegirá uno u otro método. El más habitual es el barrido automático por filas y columnas, por su compromiso entre velocidad y esfuerzo.

Independientemente del método de barrido utilizado, el acceso a las opciones del programa es siempre muy lento, y es necesario incluir todos los mecanismos disponibles para mejorarlo. En *Predice*, la tarea crítica es la escritura de textos. El usuario estará la mayor parte de tiempo en la matriz de letras (parte inferior de la pantalla), por lo que es el primer elemento a optimizar.

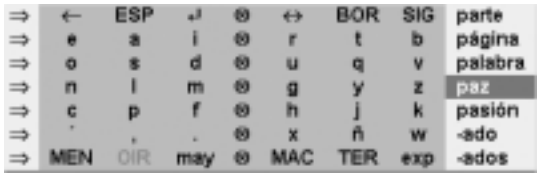


Figura 3.5: Matriz de letras de *Predice*.

El diseño final de la matriz cuando se utiliza barrido por filas y columnas, se muestra en la figura 3.5, en la que se puede apreciar:

- Contiene, no sólo las letras y las palabras y sufijos predichos, sino, también, una serie de símbolos muy útiles en la escritura: espacio, retorno de carro, signos de puntuación frecuentes, y acceso a algunos menús útiles: movimiento de cursor, acceso a expresiones, etc.

- Las palabras están situadas a la derecha, para que no ocupen una posición en el barrido cuando no son útiles (si el usuario lo desea, se pueden situar a la izquierda). La columna izquierda contiene un acelerador (⇒) que, cuando se selecciona, hace saltar el cursor a la columna de las palabras, para que no sea necesario esperar el barrido de todas las letras de la fila correspondiente.

- En cada línea hay un símbolo (⊗), que permitirá que se interrumpa el barrido de dicha fila, si se ha entrado en ella por error, sin necesidad de seleccionar ninguna de las otras opciones. De esta forma, se disminuye el coste en tiempo y número de pulsaciones necesarias para corregir los errores de barrido.

- El resto de las opciones están situadas siguiendo criterios estadísticos, para minimizar el tiempo necesario para escribir textos en castellano. Se han realizado estadísticas sobre el uso de los diferentes caracteres en los textos (letras, signos de puntuación, mayúscula, etc.). Una vez obtenidos los datos, se sitúan los caracteres más probables en las posiciones que requieren menos tiempo de acceso.

  En una matriz barrida por filas y columnas, las posiciones que menor tiempo de *acceso* requieren son las más cercanas a la esquina superior izquierda: la opción situada en la primera fila y primera columna, requiere menos de dos tiempos de barrido para ser seleccionada (dependiendo del tiempo que el usuario tarde en seleccionar

cada opción, una vez esté resaltada). La opción situada en la primera fila y segunda columna, requiere entre uno y tres tiempos de barrido. El tiempo requerido por cada posición aumenta en diagonal, hacia la esquina inferior derecha.

Según esta distribución de tiempos de acceso, se sitúan las letras en las posiciones no ocupadas por los caracteres explicados anteriormente. Se comienza por la esquina superior izquierda, para el carácter más probable, y se continúa con los siguientes caracteres, por orden de frecuencia descendente, ocupando las diagonales hacia la derecha.

La posición más rápida se reserva para el carácter de borrado, por ser muy utilizado en las interfaces con barrido. Se puede observar la concentración de las vocales (letras más frecuentes) en la parte superior izquierda de la matriz, mientras que en la inferior derecha se sitúan las letras menos utilizadas, como la w y la $\tilde{n}$.

Esta matriz es personalizable, de manera que si un usuario prefiere situar las opciones de otra forma (por ejemplo, por orden alfabético, o como en un teclado qwerty) puede hacerlo, aunque las pruebas realizadas con diferentes configuraciones muestran que el tiempo necesario para escribir los textos aumenta significativamente.

3.5.3 Inclusión de mecanismos adaptativos en una interfaz de barrido

La *adaptabilidad* consiste en introducir un cierto grado de inteligencia en la interfaz, que analiza las interacciones del usuario. Tras este análisis, el sistema propone modificaciones a la interfaz para que los accesos sean óptimos: que las opciones más utilizadas sean más cómoda y rápidamente accesibles y que el usuario necesite realizar un esfuerzo menor para llevar a cabo las acciones habituales.

El concepto de interfaz adaptativa no está ligado a programas para personas con discapacidad, aunque puede resultar especialmente importante en estos casos. Dentro del trabajo de esta tesis se han ampliado los conceptos de adaptabilidad para mejorar una interfaz con barrido, incluyendo en el sistema:

1. La *adaptabilidad de posición*, que consiste en promocionar las opciones más utilizadas por el usuario a posiciones del menú de acceso más rápido. El sistema, simultáneamente a su utilización, realiza estadísticas sobre el uso de las diferentes opciones del mismo. Posteriormente, en base a las estadísticas, se modifica la posición de las opciones dentro de cada menú, e, incluso, el menú al que pertenecen: p. ej., si una de las opciones más utilizadas está en un submenú, al adaptarse promocionará al menú principal.

   Así se conseguirá, tras un tiempo, que la interfaz esté optimizada para su uso habitual, reduciendo al máximo el tiempo y esfuerzo necesarios para utilizar el programa por parte del usuario, ya que las opciones más utilizadas estarán en posiciones de acceso rápido del menú principal.

2. La *adaptabilidad de estado*, que consiste en deshabilitar las opciones que no se utilizan nunca, o casi nunca. Se utilizan las estadísticas sobre uso de cada opción, para detectar aquellas que el usuario no utiliza. Dichas opciones son deshabilitadas, y, a partir de ese momento, no serán barridas, acelerando el acceso de las opciones que

se barrían a continuación. Se reduce así el número de posiciones efectivas que se barren en el menú adaptado.

Las opciones deshabilitadas no se anulan permanentemente, sino que existe otra opción que permite al usuario recuperar la funcionalidad plena del sistema cuando sea necesario.

En este tipo de adaptabilidad hay que considerar que hay opciones que no se pueden deshabilitar, aunque nunca se hayan utilizado, como, por ejemplo, las letras, la opción de salida de cada menú, o la opción de escape de barrido en caso de error.

3. La *adaptación automática de la velocidad de barrido*, en la línea de la propuesta descrita en [Simp99], tomando como criterio el porcentaje de errores cometidos al escribir. Los errores se contabilizan considerando el número de veces que se utiliza la opción de borrado y el carácter de escape de barrido de cada fila. El usuario establece un umbral a este número de errores, y si dicho umbral se supera (por ejemplo, por fatiga del usuario, o por una inadecuada configuración inicial), la aplicación sugerirá una disminución en la velocidad de barrido.

   Así mismo, se establecerá otro umbral mínimo, y, si el número de errores es menor (por ejemplo, cuando el usuario adquiere destreza en el manejo del programa), la aplicación pregunta si se desea aumentar la velocidad.

   En la versión actual, el valor de los umbrales se configura manualmente, aunque se podría estudiar la manera de calcularlos en base al valor medio del porcentaje de pulsaciones erróneas para cada usuario.

   Cualquier cambio en el programa debe ser validado por el usuario antes de realizarse. Además, tras cada modificación, hay un tiempo de adaptación para el usuario, en el que puede que cometa un número mayor de errores, o realice un uso sub-óptimo de la interfaz, hasta que se adapte a sus nuevas características. La aplicación considerará ese tiempo como un transitorio en el cual no recogerá información sobre la que tomar decisiones de adaptabilidad.

   En el marco de esta tesis se dirigió el proyecto fin de carrera [Garc00] en el que se puede encontrar los detalles sobre la implementación de los diferentes mecanismos adaptativos y los resultados de la evaluación automática del sistema. La adaptabilidad no está disponible en la versión actual de *Predice* del mercado, por lo que aún no se ha podido realizar una evaluación de usuario a largo plazo del sistema. La evaluación automática proporciona resultados positivos, obteniendo que la adaptabilidad produce una reducción tanto el número de pulsaciones como el tiempo necesario para ejecutar cada acción.

3.6 Arquitectura detallada del sistema

En la figura 3.6 se muestra el diagrama en el que se incluyen el conjunto de métodos de predicción, diccionarios, gestores y entrenadores que componen el sistema completo, considerando la estructura de cada módulo y las relaciones entre ellos, tal y como se ha detallado en los apartados correspondientes.

A modo de resumen, podemos destacar las características fundamentales de la arquitectura propuesta:

- Modularidad, ya que la separación estricta entre los módulos que gestionan los diccionarios y el resto (los que constituyen los métodos de predicción) permite incorporar fácilmente nuevos métodos de predicción y fuentes de conocimiento, sin necesidad de invertir esfuerzo adicional de rediseño de la arquitectura.

- Flexibilidad, por una parte, dada la independencia de la arquitectura de la tarea concreta para la que se emplea la predicción, y por otra, del idioma objetivo, ya que los módulos procedimentales se mantendrían, siendo necesario únicamente incluir la información lingüística específica.

- Potencia, dados los excelentes resultados obtenidos en la evaluación, por su capacidad integradora de múltiples métodos y fuentes de conocimiento (diccionario general, diccionario personalizado de la sesión en curso, diccionarios temáticos, modelos estocásticos con unigramas, bigramas, bipos y tripos, analizador sintáctico que sigue las reglas de la gramática formal, filtro de palabras rechazadas y predictor de sufijos). El elemento clave es la propuesta de un gestor de flujos de información (módulo de coordinación) que combina de forma óptima la potencia desplegada por cada uno de ellos.

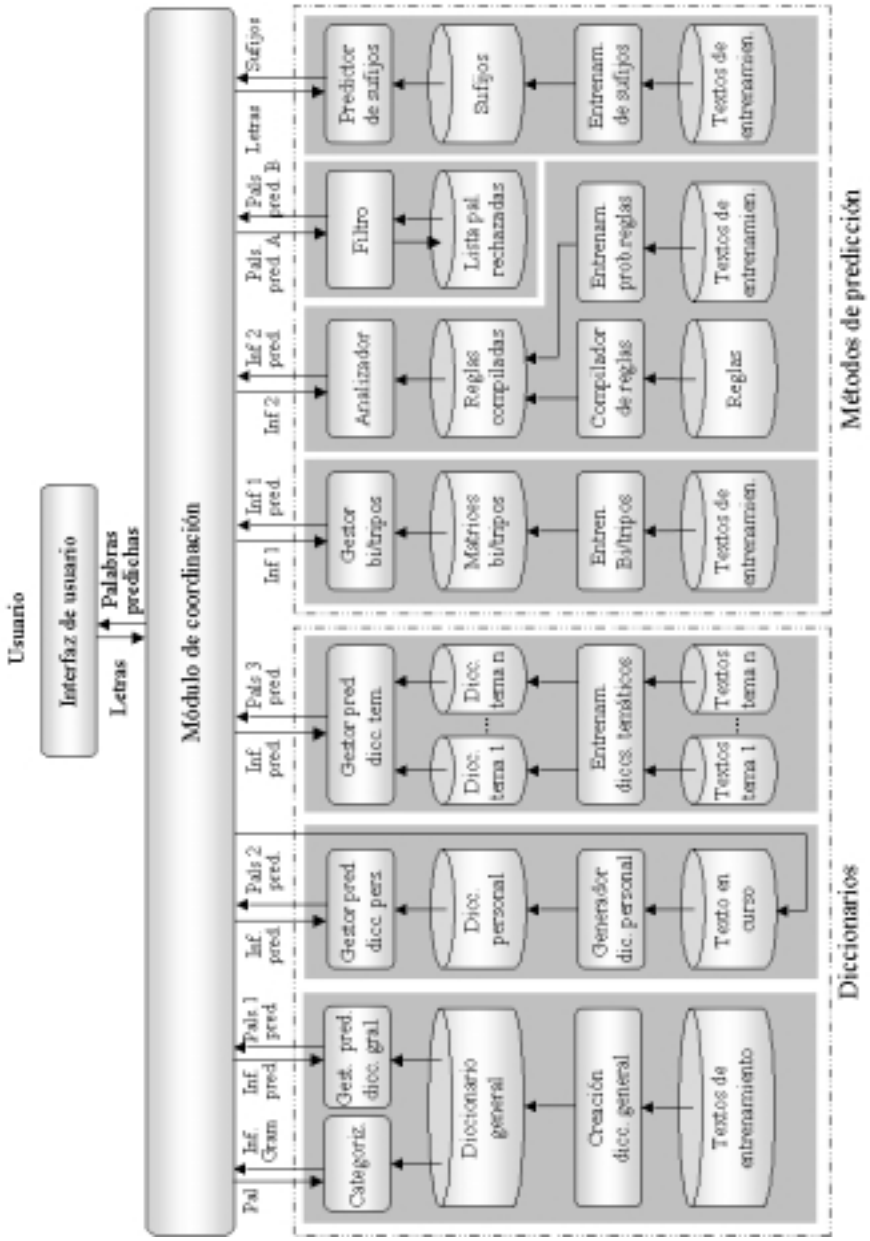


Figura 3.6: Diagrama de bloques detallado del sistema de predicción de palabras.

Capítulo 4

Evaluación de la predicción de palabras

En el presente capítulo se evaluará la eficiencia de todos los métodos de predicción vistos anteriormente, lo que permitirá realizar una comparación entre ellos, para seleccionar el método o combinación de métodos más adecuada. Está dividido en dos partes, una primera con los resultados de la evaluación cuantitativa, llevada a cabo mediante procedimientos automáticos, y la segunda, con la evaluación subjetiva de los usuarios sobre el sistema completo, y ejemplos de los beneficios producidos por su uso con respecto al aumento de la calidad de la escritura.

4.1 Evaluación automática

Antes de comenzar con la descripción de las pruebas y sus resultados, se justifican las decisiones tomadas acerca de la configuración de la predicción, las métricas y el conjunto de pruebas a realizar, tomando como base los criterios habituales en la evaluación automática de cualquier diseño y, además, añadiendo un conjunto de restricciones derivadas del uso del sistema por usuarios reales.

4.1.1 Consideraciones acerca de la evaluación automática

En la evaluación automática de cualquier sistema que implique procesamiento de lenguaje natural (PLN), existen una serie de factores que pueden influenciar en gran medida los resultados, independientemente de la calidad del sistema de PLN en sí. Para hacer comparaciones válidas entre métodos es necesario controlar dichos factores o al menos tener en cuenta sus efectos, como se explica en [Pala96].

4.1.1.1 Textos de entrenamiento y de prueba

La selección de los textos de entrenamiento y prueba a utilizar en el sistema es uno de los aspectos clave en la etapa de evaluación, ya que pueden condicionar de forma importante

los resultados obtenidos.

En el diseño de los corpora de entrenamiento y prueba se deben tener en cuenta, al menos, los siguientes factores: su tamaño, si son representativos o no del dominio que se desee evaluar, y el grado de acuerdo entre ellos.

Con respecto al *tamaño* de los corpora, como en cualquier aplicación de PLN, se deben utilizar textos de la mayor longitud posible:

- En la etapa de entrenamiento, para obtener modelos fiables.
- En la evaluación, para obtener resultados estadísticamente significativos.

Esta suele ser una premisa habitual en la evaluación de los sistemas que tratan lenguaje natural. Sin embargo, el uso de textos muy grandes en la evaluación produce resultados que pueden no ser realistas en algunos casos. Por ejemplo, si se utilizan textos muy largos para evaluar los diccionarios personales, se obtienen unos niveles de adaptación muy altos, que producen buenos resultados. Sin embargo, los usuarios nunca escribirán textos tan largos, así que nunca obtendrán el nivel de adaptación conseguido en la evaluación, y el rendimiento en los casos reales será inferior.

Al evaluar los métodos de predicción, en general se realizarán pruebas automáticas con tanto texto como sea posible, y, en los métodos de predicción que dependan de la cantidad de texto escrito (como los diccionarios personales), se repetirán ciertas pruebas con textos más pequeños para observar su comportamiento en condiciones similares a las encontradas por el usuario.

El problema de utilizar corpora pequeños es que pueden producir resultados muy dependientes del texto, aunque este efecto puede reducirse si se seleccionan cuidadosamente. Si solamente se dispone de un corpus pequeño, se puede dividir en varias partes, y utilizar el método *cross-validation*, que consiste en entrenar con todas las partes menos una, y utilizar esa para probar el sistema. Este proceso se repite varias veces, cada una de ellas dejando fuera uno de los fragmentos, y finalmente se promedian los resultados. De esta forma se reduce la probabilidad de que el texto elegido produzca resultados particularmente buenos o malos, como se explica en [Alle94].

En general, son preferibles las estrategias que implican probar con varios textos de tamaño medio, promediando los resultados.

Los textos utilizados deben ser *representativos* de los dominios que se desea evaluar, ya que tanto el vocabulario como la información frecuencial y el estilo de la escritura varían de unos entornos a otros, pudiendo influir en los resultados. Para evaluar el comportamiento del sistema en un dominio concreto debe obtenerse la mayor cantidad posible de textos pertenecientes a él, que nos permitan entrenar y probar el sistema de forma fiable.

Si se desea realizar la evaluación en las *condiciones del usuario*, se deberían generar los corpora de entrenamiento y prueba a partir de textos introducidos por el usuario (datos grabados), totalmente adaptados a su estilo de escritura. Estos textos son muy difíciles de obtener, especialmente si se necesitan corpora grandes, y, además, hay que considerar la confidencialidad de los textos. En [Cope98] se puede encontrar un estudio sobre la evaluación de la predicción de palabras a partir de textos de usuario grabados.

Otro factor clave es el *grado de acuerdo* entre los textos de entrenamiento y prueba: éstos deben ser totalmente disjuntos. En caso contrario, las medidas obtenidas con textos que se solapen no deben ser consideradas válidas.

4.1.1.2 Configuración de la predicción

En el proceso de predicción hay numerosas variables que configurar, que pueden ser decisivas en los resultados obtenidos, por lo cual, a la hora de comparar diferentes métodos es necesario asegurar que sus valores se han mantenido coherentes en todo el proceso.

Uno de los datos que más influencia tiene en los resultados es el *número de palabras predichas*. Obviamente, cuanto mayor sea la longitud de la lista de palabras que se muestra al usuario, más probable es que la palabra buscada se encuentre en dicha lista, y menos pulsaciones serán necesarias, mejorando los resultados.

Por supuesto, para las pruebas automáticas se puede utilizar cualquier número, habiéndose decidido realizar las pruebas con un número reducido de palabras, en condiciones realistas para el usuario. En los estudios publicados en [Mill56] se indica que el número máximo de opciones que el usuario puede procesar de un vistazo es 7. Este número mantiene el equilibrio entre el aumento en la carga cognitiva necesaria para procesar la lista y el ahorro de tiempo: si se muestran más palabras, el usuario emplea más tiempo y esfuerzo para leerlas, buscando la palabra adecuada, por lo cual la predicción resulta contraproducente y puede decidir ignorarla.

Las mismas restricciones se deben considerar cuando se trata de evaluar la predicción de terminaciones.

Otro factor a tener en cuenta es que, normalmente, en los sistemas de ayuda a la escritura, además de predicción de palabras se incluyen otras técnicas auxiliares, como la inserción automática de espacios tras los signos, o las mayúsculas tras los puntos. El ahorro de pulsaciones producido por estas técnicas tiene una repercusión en los resultados que debe ser considerada.

4.1.1.3 Métricas

Para que los resultados sean significativos se debe seleccionar el conjunto más adecuado de parámetros que representen el rendimiento del sistema. A continuación se muestran ciertas medidas que pueden resultar relevantes en la evaluación de la predicción de palabras. Cada uno de los parámetros recoge cierta información sobre el proceso de predicción, de forma que la métrica a seleccionar depende del interés del evaluador.

Los parámetros que se explicarán son el ahorro de pulsaciones, la cobertura de la predicción y la capacidad de aprendizaje. Se evita el uso de medidas de tiempo y velocidad porque son muy dependientes tanto del diseño de la interfaz como del propio usuario, centrando la descripción en medidas que puedan ser obtenidas automáticamente.

4.1.1.3.1 Medidas basadas en el número de pulsaciones

El número de pulsaciones es uno de los puntos clave a evaluar en un sistema de predicción de palabras, ya que es una medida directamente relacionada con el esfuerzo físico necesario para escribir el texto, especialmente importante para usuarios con discapacidad física.

Sobre este parámetro pueden realizarse dos tipos de medidas: el *número de pulsaciones* necesarias para escribir el texto o el *ahorro* en el número de pulsaciones que produce el método de predicción considerado.

La primera medida consiste en contabilizar el número de pulsaciones que el usuario debe teclear antes de que se le ofrezca la palabra deseada. Dependiendo del nivel de detalle requerido, se pueden realizar las siguientes medidas:

- Número de palabras en función de las pulsaciones necesarias para escribirlas, es decir, del número de pulsaciones necesarias para que se muestren en la lista, o de la longitud de la palabra (medida en pulsaciones, por consistencia), si la palabra no aparece.
- Si se necesita una medida más compacta, se puede calcular el número medio de pulsaciones necesarias para escribir una palabra. Esta cifra puede ponerse en relación con la longitud media de las palabras del corpus (medida en pulsaciones), obteniéndose el porcentaje de pulsaciones tecleadas por palabra.

La segunda medida es complementaria de la anterior. La idea es contabilizar las pulsaciones que el usuario *no* tiene que realizar. Esta medida puede ser considerada más relevante desde el punto de vista del usuario, porque muestra la reducción de esfuerzo que se consigue con el sistema de predicción. La fórmula que se aplica para calcular el ahorro de pulsaciones es:

$$Ahorro(\%) = \frac{\#pulsaciones_{sin\ predicción} - \#pulsaciones_{con\ predicción}}{\#pulsaciones_{sin\ predicción}} \cdot 100 \qquad (4.1)$$

Donde:

- $\#pulsaciones_{sin\ predicción}$ es el número de pulsaciones necesarias para escribir el texto sin la ayuda de la predicción de palabras.
- $\#pulsaciones_{con\ predicción}$ es el número de pulsaciones asistido por predicción.

Es importante tener en cuenta que, aunque las medidas están basadas en pulsaciones, los textos están compuestos por caracteres. Es necesario definir una relación entre los caracteres y el número de pulsaciones necesarias para escribir cada uno, para que las medidas que se obtengan permitan realizar comparaciones objetivas.

Esta relación depende de la interfaz de usuario y no tiene por qué coincidir en diferentes sistemas, así que, si se van a realizar comparaciones, es necesario definir exactamente el

número de pulsaciones que se considera por cada carácter, y contabilizar de la misma forma en todos los sistemas. Por ejemplo, en interfaces con barrido por filas y columnas, para escribir cada letra se necesitan dos pulsaciones, mientras que si el barrido es lineal, o si se utiliza el teclado convencional, la mayoría de los caracteres precisan una sola pulsación.

En los experimentos de la sección siguiente, se utiliza una relación entre caracteres y pulsaciones similar a la que existe en el teclado convencional, es decir:

- 1 pulsación por cada letra minúscula, excepto las acentuadas.
- 2 pulsaciones por cada vocal minúscula acentuada.
- 2 pulsaciones por cada letra mayúscula no acentuada.
- 3 pulsaciones por cada vocal mayúscula acentuada.
- 1 pulsación por número.
- 1 pulsación por cada signo que no precise el uso de la tecla *mayúscula*, que son:

  , . ; -] # = / [¡ '+ <

- 2 pulsaciones por cada signo que precise pulsar la tecla de *mayúsculas*:

  Ł & * ( ) [ ] { } @ : ∧ | _ % ∼ \ $

- 4 pulsaciones por cada signo que no pertenezca a los grupos anteriores (se supone que se precisa pulsar la tecla ALT, más los tres dígitos de su código ASCII).
- 1 pulsación por cada palabra o sufijo seleccionado.

4.1.1.3.2 Medidas relativas a la cobertura de la predicción

En general, la cobertura de la predicción puede ser definida como el número de palabras correctamente predichas, es decir, que el sistema ha mostrado en la lista de predicciones antes de que el usuario haya completado su escritura letra a letra. Este parámetro permite evaluar hasta qué punto el corpus de prueba está cubierto por las capacidades predictivas del sistema.

Estas medidas son más importantes que las anteriores para las personas que necesitan el apoyo gramatical del sistema, como por ejemplo, los usuarios disléxicos, o personas aprendiendo una segunda lengua. Estos usuarios pueden no ser capaces de completar las palabras correctamente, pero sí de reconocerlas cuando se les presentan en el menú, y, así escribirlas de forma correcta. Por esto, es importante que escriban con la ayuda de la predicción la mayor cantidad posible de palabras.

Se considera que una palabra ha sido correctamente predicha si al menos ha ahorrado un carácter (y probablemente un espacio el blanco posterior, a no ser que se escriba un signo).

En la cobertura de la predicción influyen los siguientes factores:

- El número de palabras *diferentes* en el corpus de entrenamiento.
- El número de palabras *diferentes* en el corpus de prueba.
- El grado de *acoplamiento* entre los textos de prueba y los diccionarios del sistema, que puede medirse como el número o porcentaje de palabras de los corpora de prueba que aparecen en los diccionarios. Este número de palabras puede ser diferente al número de palabras que se predicen, fundamentalmente por dos razones:
  - Si el sistema aprende, cada palabra nueva debe ser escrita una sola vez, y la próxima vez será predicha.
  - Las palabras cortas puede que sean escritas por el usuario antes que predichas, incluso aunque estén en el diccionario.
- La *entropía o perplejidad* del lenguaje, ya que en idiomas con mayor perplejidad, la complejidad de la predicción es mayor (por ejemplo, en los lenguajes en los cuales las palabras pueden tener muchas formas diferentes).
- El grado de *concordancia gramatical* entre el texto de prueba y el de entrenamiento, puesto que si se utiliza información gramatical, se obtienen resultados muy diferentes al evaluar con textos que siguen en mayor o menor medida las reglas del modelo.

4.1.1.3.3 Velocidad de aprendizaje

Otro parámetro que se puede cuantificar es la capacidad de aprendizaje. Para ello se evalúa la velocidad a la que el sistema aprende, como indicación sobre la adaptabilidad o flexibilidad del sistema de predicción. Una posibilidad es contabilizar la reducción de pulsaciones en intervalos regulares a lo largo del texto (por ejemplo, cada 100 palabras se calcula el ahorro en las últimas 25), y se representa en una gráfica el ahorro de pulsaciones en función del número de palabras escritas desde el principio del texto, como se indica en [Clay98].

En los sistemas con capacidad de aprendizaje, en la gráfica se observa la mejora del rendimiento a lo largo del texto.

4.1.1.3.4 Validación estadística de los resultados

En cualquier sistema del cual se obtengan resultados sobre su rendimiento a partir de un determinado texto de prueba, se pueden plantear cuestiones sobre la posible generalización de las conclusiones obtenidas.

En la literatura estadística pueden encontrarse referencias que permiten abordar el problema de la estimación de fiabilidad estadística, como en [Weis93], además de referencias específicas para sistemas de reconocimiento estadístico de patrones [Raud91], reconocimiento de habla [Gill89], lenguaje natural [Gibb98], etc. De los diversos métodos de validación estadística de resultados, se va a utilizar el cálculo de bandas de probabilidad, más ampliamente aplicado que otros como, por ejemplo, el test de McNemar [Gill89][Hunt90].

Para calcular las bandas de fiabilidad se utiliza la siguiente fórmula:

$$Banda/2 = \alpha \cdot \sqrt{\frac{Resultado_{experimento} \cdot (1 - Resultado_{experimento})}{N}} \tag{4.2}$$

Donde:

- $Banda/2$ es la semiamplitud de la banda de fiabilidad en tanto por ciento.
- $Resultado_{experimento}$ es el resultado del experimento en tanto por ciento, por ejemplo, porcentaje de palabras predichas, o de pulsaciones ahorradas.
- N es el número total de elementos del texto de prueba. En el ejemplo anterior serían el número total de palabras del texto o el número total de pulsaciones necesarias para escribir el texto sin predicción.
- α es un parámetro que se fija dependiendo del valor concreto de confianza requerida. P. ej., para una confianza del 95%, el valor de α es 1.96.

Una vez obtenido el valor de la banda de fiabilidad, el resultado del experimento se puede generalizar para todos los textos del mismo dominio[1], asegurando que el resultado *global* está dentro del siguiente intervalo:

$$Resultado_{global} = Resultado_{experimento} \pm Banda/2 \tag{4.3}$$

con la fiabilidad utilizada en el cálculo de la banda.

Para comparar diferentes métodos de predicción se realizan pruebas con cada uno sobre un texto de prueba determinado y se calculan sus bandas de fiabilidad. Se considerará que uno es mejor que otro solamente si *su resultado es mejor, y, además, las bandas de fiabilidad de ambos métodos no se solapan.* Si las bandas se solapan, se concluirá que las diferencias no son estadísticamente significativas con el grado de confianza elegido.

Como la amplitud de las bandas depende inversamente del valor de N (longitud del texto de prueba), dicha amplitud puede estrecharse aumentando el número de palabras de dicho texto. Es decir, al obtener resultados con textos pequeños puede que las bandas se solapen y no se pueda concluir que las diferencias en los resultados sean estadísticamente significativas. Si se aumenta la longitud de los textos de prueba, los resultados serán más fiables, y eso se refleja como una reducción en la amplitud de las bandas de fiabilidad. Si se dispone de suficiente texto, se podrá aumentar el número de palabras utilizadas en las pruebas hasta que las bandas no se solapen, y se pueda afirmar que un método es mejor que el otro (a no ser que las tasas converjan al aumentar el número de palabras).

Si los textos de prueba no son suficientes para conseguir resultados con diferencias estadísticamente significativas, se realizarán varias pruebas con los textos de los que se disponga. Si uno de los métodos siempre obtiene mejor resultado que el otro, aunque las

[1] Aquellos para los cuales el texto de prueba sea representativo: textos periodísticos de un periodo y periódico concretos, novelas de un autor determinado sobre un tema y época determinados, etc.

bandas se solapen, la conclusión sería que se aprecia una mejora consistente, aunque no se pueda afirmar que las diferencias sean estadísticamente significativas.

Evidentemente, es imposible garantizar que una determinada tasa de predicción se mantendrá si se modifican las condiciones del experimento (por ejemplo, los resultados obtenidos con textos de *El Mundo* no serán iguales a los obtenidos con textos de cartas personales), pero tampoco es ese el objetivo que se persigue. La función principal de las estimaciones de fiabilidad es determinar la relevancia estadística de las diferencias en tasa obtenidas por distintos algoritmos, métodos o sistemas en el dominio sobre el que se está trabajando.

4.1.2 Descripción del sistema de evaluación automática

La predicción de palabras se ha integrado en un entorno de evaluación, que facilita la configuración de todos sus parámetros y la realización de las pruebas de forma automática.

Esta nueva herramienta es similar al *entorno de pruebas parametrizable* [2] sugerido en [Thom94], aunque considerando las particularidades de la predicción de palabras. Todavía no se ha definido ningún estándar sobre la evaluación de la predicción de palabras, al contrario que para otros aspectos del procesamiento de lenguaje natural, en los que sí hay iniciativas sobre normalización, como las que pueden encontrarse en [Lehm96] y [Eagl98].

La herramienta de evaluación lleva a cabo todo el proceso de forma automática. Para cada experimento lleva a cabo los siguientes pasos, que en [Gibb98] se describen como los necesarios en la evaluación de los sistemas de PLN:

- Configuración del experimento: en esta fase se establecen los valores de todos los parámetros implicados en la evaluación: corpus de prueba y, en su caso, de entrenamiento, número máximo de palabras predichas, tipo de predicción deseada (utilizando o no información gramatical, etc.), uso de diccionario personal o no, etc.
- Entrenamiento de alguna parte del sistema si es necesario, por ejemplo, de algún diccionario temático.
- Realización del experimento bajo las condiciones elegidas, contabilizando los parámetros deseados.
- Obtención de los resultados, a partir de las medidas realizadas.
- Análisis de dichos resultados, comparación entre resultados de diferentes experimentos o con los límites teóricos.
- Generación del informe correspondiente, que deberá ser suficientemente completo como para permitir reproducir el experimento y extenso como para permitir realizar la evaluación. Debe incluir, al menos:
  - Valores de los parámetros de entrada: texto utilizado, diccionario, longitud de las listas de palabras y sufijos predichos, método de predicción evaluado, etc.

[2] Parameterisable Test Bed (PTBs)

- Parámetros medidos en el texto: longitud, longitud media de sus palabras, grado de solapamiento con el diccionario general, etc.
- Si se desea realizar una depuración exhaustiva se proporciona información del proceso, paso a paso.
- Resultados del experimento.

Además, se contempla la posibilidad de realizar experimentos *especiales*, en los cuales no se evalúa un método de predicción determinado, sino se averiguan los límites teóricos de la predicción: permiten calcular, para cada texto, el número de pulsaciones necesarias para escribirlo *sin ninguna ayuda* (ni predicción, ni inserción de espacios tras signos, etc.), y con *predicción de palabras perfecta* (cada palabra es predicha en primera opción y sólo es necesario seleccionarla). Estas dos medidas permiten situar cada método dentro de los límites teóricos de rendimiento de la predicción.

Una vez integrada la predicción en la herramienta de evaluación, se debe definir un conjunto de experimentos bajo diferentes condiciones, con diferentes textos, con diccionario personal o no, incluyendo los métodos complementarios (como inclusión automática de espacios tras signos, etc.), suficientemente exhaustivo como para permitir una evaluación objetiva de los diferentes métodos y sus combinaciones.

4.1.3 Resultados

En esta sección se presentan los resultados de los distintos experimentos realizados sobre cada método para estudiar su comportamiento, el efecto de distintas configuraciones, su sensibilidad a los datos de entrenamiento, etc.

La mayor parte de los experimentos están estructurados en *series*, que permitirán, no sólo evaluar los resultados de cada uno de forma individual, sino compararlos. Para cada serie de experimentos se proporciona (si es relevante y no ha aparecido con anterioridad):

- Descripción de sus objetivos.
- Nombre del texto de prueba utilizado, número de palabras que contiene y número de pulsaciones necesario para escribirlo sin ayuda de predicción de palabras. Una descripción más detallada de cada texto puede ser encontrada en el apéndice A (página 179).
- Resultados del experimento de referencia con el que se comparan los experimentos de la serie. Normalmente es el resultado obtenido por el mejor experimento de alguna de las series previas.
- Resultados de cada uno de los experimentos de la serie, en valor absoluto y en comparación con los del experimento de referencia.
- Gráfica resumen, que muestra la evolución de los parámetros fundamentales a lo largo de los experimentos de la serie.
- Análisis y discusión de los resultados.

Los resultados que se aportarán sobre cada experimento son los siguientes (elegidos entre las posibles métricas explicadas en la sección 4.1.1.3 en la página 111):

- Las pulsaciones ahorradas, por su relación con la disminución del esfuerzo físico que tendría que realizar un usuario para escribir el texto.
- El número de palabras predichas, porque proporciona información sobre la cobertura de la predicción.
- Las palabras predichas en función del número de letras que ha sido necesario escribir antes de que se muestren al usuario. Este dato aporta información adicional que permite matizar el parámetro anterior y se mostrará en los experimentos en los que sea relevante.

Dichos datos serán presentados, para cada experimento, en forma de:

- Tabla que contiene los siguientes datos sobre las pulsaciones ahorradas y el número total de palabras predichas:
  - El resultado del experimento de referencia, en valor absoluto y porcentaje sobre el total.
  - El resultado del experimento que se evalúa, en valor absoluto y porcentaje sobre el total.
  - A los porcentajes se les añaden las bandas de fiabilidad, según lo explicado en la sección 4.1.1.3.4 (página 114) sobre validación estadística de resultados. Las bandas proporcionadas son para una fiabilidad del 95%, por lo que la fórmula que se aplica a los resultados es:

    $$Banda_{95\%}/2 = \pm 1.96 \cdot \sqrt{\frac{t \cdot (1-t)}{N}} \tag{4.4}$$

    Donde t es el resultado correspondiente en tanto por ciento, y N es el número de palabras del texto de prueba.
  - Mejora relativa del resultado del experimento sobre el de referencia, que se calcula aplicando la siguiente ecuación:

    $$Mejora\ relativa\ \% = 100 \cdot \left(\frac{Resultado_{experimento} - Resultado_{referencia}}{Resultado_{referencia}} \right) \tag{4.5}$$

- Gráfica con la información detallada sobre el número de palabras predichas en función del número de letras que ha sido necesario escribir antes de que sean mostradas al usuario. Se proporciona el porcentaje sobre el número total de palabras del texto. Es decir: si en la posición 0 de la gráfica aparece un 27%, y en la posición 1 un 23%, se debe interpretar que el 27% de las palabras del texto han sido predichas sin que el usuario escriba ninguna letra (en la *letra cero*), y el 23% cuando se ha tecleado una sola letra (en la *letra uno*).

Por último mencionar que las operaciones se han realizado con una precisión mayor a la que se ha utilizado en la presentación de resultados (en la que se muestran únicamente uno o dos decimales). Esto puede dar lugar a la aparición de algún cero en posiciones de las gráficas que realmente no son nulas, debido al redondeo de la cifra original. La misma razón puede provocar que al repetir las operaciones con los datos de las tablas el resultado no coincida de forma exacta con el que se muestra. Se debe considerar como válido este último, ya que ha sido obtenido a partir de datos con mucha mayor precisión.

4.1.3.1 Límite máximo teórico de la predicción

A continuación se muestra el resultado del experimento en el que se trata de evaluar el *método de predicción perfecto*: aquel en el que se predice cada palabra en la letra cero, es decir, en el que el usuario solamente tiene que elegir cada palabra, sin escribir ninguna letra.

Este método predice el 100% de palabras, pero sin embargo *no* se ahorra el 100% de las pulsaciones, ya que al menos se necesita una por palabra (para seleccionarla), y aquellas correspondientes a la escritura de los signos de puntuación, las mayúsculas, etc.

Por ejemplo, dado el texto de prueba TESIS, de 37496 palabras, y que precisa 253932 pulsaciones para ser escrito sin predicción[3], el resultado del método de predicción de palabras perfecto sobre él es:

Número de palabras predichas: 37496 (100%).

Número de pulsaciones necesarias: 62656 (24.67%).

Número de pulsaciones ahorradas: 191276 (75.33%).

Si, además de una predicción perfecta, se añaden los mecanismos auxiliares, como adición automática de espacios y mayúsculas después de punto, el número de pulsaciones necesarias para escribir el texto disminuye, quedando como sigue:

Número de pulsaciones necesarias: 57637 (22.70%)

Número de pulsaciones ahorradas: 196295 (77.30%).

Este máximo es teórico, y los métodos que se presentan están bastante alejados de este porcentaje de ahorro de pulsaciones, pero es importante considerar, a la hora de interpretar los resultados de los experimentos, que el máximo ahorro posible, aplicando sólo predicción de palabras, es aproximadamente un 75% y no el 100%.

[3]En la sección 4.1.1.3.1 (página 112) se muestra la relación entre los caracteres escritos y el número de pulsaciones que requiere cada uno.

4.1.3.2 Evaluación de los métodos estadísticos básicos

En esta sección se evalúa el más sencillo de los métodos de predicción utilizados, basado exclusivamente en la información frecuencial de las palabras, sin ningún tipo de filtrado gramatical.

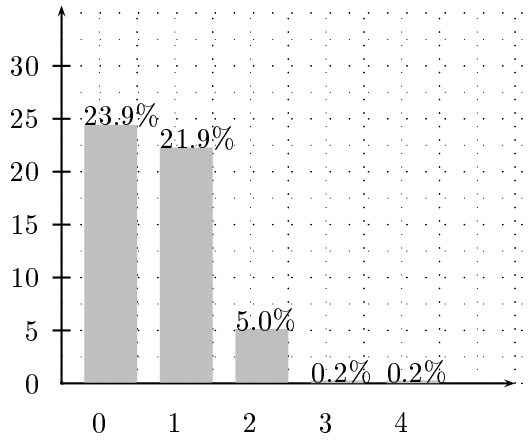
Como se explicó en la sección 3.3.1 (página 56) sobre los métodos estadísticos básicos, el modelo *unigram* solamente depende del diccionario general, por lo que en esta serie de experimentos se muestra la influencia en los resultados de la cantidad de palabras que éste contiene. Para ello se ha diseñado una serie de experimentos en los que, partiendo del diccionario general descrito en el apartado 3.2.1 (página 34), se limita el número de palabras que se utilizan en la predicción a 200, 400, 800, 1200, 5000, 12000, 40000 y 132543 palabras (tamaño total), respectivamente, considerando siempre las más probables.

Las características del texto de prueba son:

Nombre del texto de prueba	TESIS
Número de palabras	37496
Pulsaciones para escribirlo sin predicción	253932

Como no hay ningún método de predicción anterior con el que comparar los de esta serie, el primer experimento, considerando solamente las 200 palabras más probables del diccionario general, será utilizado como referencia para los siguientes. Sus resultados son:

Número de palabras predichas	19211 (51.23%±0.51)
Número de pulsaciones ahorradas	36803 (14.49%±0.14)



Resultado 2.1: Resultados del experimento de referencia: predicción utilizando unigramas con un diccionario general de 200 palabras.

En este experimento se parte de un diccionario de sólo 200 palabras, sin ningún tipo de filtrado gramatical. Cabe destacar el elevado porcentaje de palabras que se predicen en la *letra cero* y uno. Con este método de predicción, las palabras que se predicen en la *letra cero* son siempre la mismas, las más frecuentes en castellano *de, la, el, que* y *en.* Se incluyen entre las palabras predichas en la *letra cero* las de una sola letra, *a, y, e, etc.*, que no necesitan ser predichas, ya que se realiza el mismo esfuerzo seleccionándolas directamente como letras que como palabras. Es decir, con las cinco primeras palabras

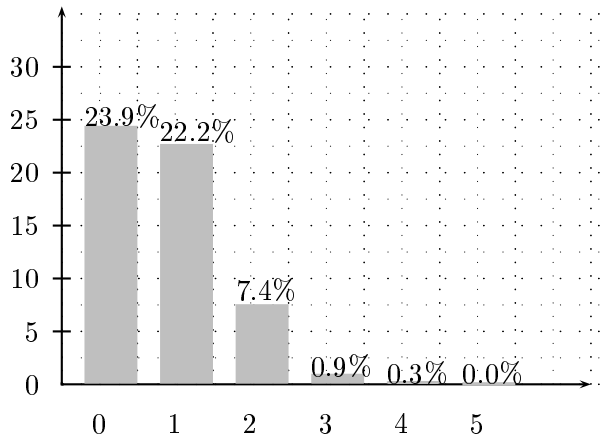
del diccionario y las palabras de una letra se cubre el 23.9% del texto de prueba.

El porcentaje de palabras predichas en la *letra uno* también es bastante elevado, sin embargo, a partir de la tercera letra decrece drásticamente, ya que con un diccionario de 200 palabras, tras escribir tres letras, si la palabra estaba en él, ya ha sido predicha, y, si no, el usuario tendrá que escribirla letra a letra.

El porcentaje de pulsaciones ahorradas, sin embargo, es pequeño (solamente un 14.49%), y se debe fundamentalmente a que las palabras que se predicen son muy cortas: las treinta primeras palabras del diccionario general miden dos o tres letras, por lo que el ahorro de pulsaciones que se produce al predecir correctamente cada una es mínimo.

En el siguiente experimento se duplica el tamaño del diccionario, hasta considerar las 400 palabras más probables. Sus resultados son:

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	19211 (51.23%±0.51)	20574 (54.87%±0.50)	7.09%
Pulsacs. ahorradas	36803 (14.49%±0.14)	42889 (16.89%±0.15)	16.54%



Resultado 2.2: Resultados de la predicción utilizando unigramas con un diccionario general de 400 palabras.

En él se puede apreciar la mejora en los resultados al duplicar la longitud del diccionario general: el número de palabras predichas en la *letra cero* coincide con el del experimento anterior (ya que éstas son siempre las mismas), pero aumentan las palabras predichas a partir de la *letra uno*, y con ellas el porcentaje de pulsaciones ahorradas.

De forma similar, en los siguientes experimentos se puede apreciar la mejora, tanto en número de palabras predichas como en pulsaciones, que se produce al seguir aumentando el tamaño del diccionario general.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	19211 (51.23%±0.51)	22870 (60.99%±0.49)	19.05%
Pulsacs. ahorradas	36803 (14.49%±0.14)	52018 (20.49%±0.16)	41.34%

Resultado 2.3: Resultados de la predicción utilizando unigramas con un diccionario general de 800 palabras.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	19211 (51.23%±0.51)	23989 (63.98%±0.49)	24.87%
Pulsacs. ahorradas	36803 (14.49%±0.14)	56527 (22.26%±0.16)	53.59%

Resultado 2.4: Resultados de la predicción utilizando unigramas con un diccionario general de 1200 palabras.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	19211 (51.23%±0.51)	28408 (75.76%±0.43)	47.87%
Pulsacs. ahorradas	36803 (14.49%±0.14)	74133 (29.19%±0.18)	101.43%

Resultado 2.5: Resultados de la predicción utilizando unigramas con un diccionario general de 5000 palabras.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	19211 (51.23%±0.51)	31194 (83.19%±0.38)	62.38%
Pulsacs. ahorradas	36803 (14.49%±0.14)	85127 (33.52%±0.18)	131.30%

Resultado 2.6: Resultados de la predicción utilizando unigramas con un diccionario general de 12000 palabras.

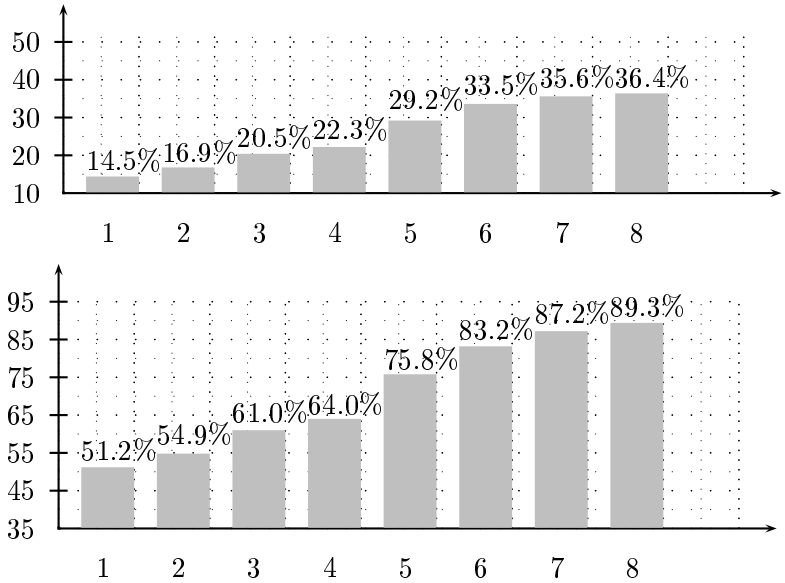
Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	19211 (51.23%±0.51)	32682 (87.16%±0.34)	70.12%
Pulsacs. ahorradas	36803 (14.49%±0.14)	90503 (35.64%±0.19)	145.91%

Resultado 2.7: Resultados de la predicción utilizando unigramas con un diccionario general de 40000 palabras.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	19211 (51.23%±0.51)	33504 (89.35%±0.31)	74.38%
Pulsacs. ahorradas	36803 (14.49%±0.14)	92367 (36.37%±0.19)	150.88%

Resultado 2.8: Resultados de la predicción utilizando unigramas con el diccionario general completo: 132543 palabras.

En las siguientes tablas se puede ver un resumen con la evolución de los resultados de toda la serie, tanto en porcentaje de palabras predichas como de pulsaciones ahorradas.



Resultado 2.9: Evolución del porcentaje de pulsaciones ahorradas y palabras predichas al variar el tamaño del diccionario general.

Donde:

1: Resultados del experimento de referencia, utilizando un diccionario general de 200 palabras.

2: Resultados utilizando un diccionario general de 400 palabras.

3: Resultados utilizando un diccionario general de 800 palabras.

4: Resultados utilizando un diccionario general de 1200 palabras.

5: Resultados utilizando un diccionario general de 5000 palabras.

6: Resultados utilizando un diccionario general de 12000 palabras.

7: Resultados utilizando un diccionario general de 40000 palabras.

8: Resultados utilizando un diccionario general completo, con 132543 palabras.

La conclusión fundamental de esta serie de experimentos es que la predicción mejora al aumentar el tamaño del diccionario considerando siempre las palabras más probables.

Esta mejora es más rápida en los diccionarios más pequeños, tendiendo a saturar las curvas en tamaños mayores. Por ejemplo, entre los tres últimos experimentos se introducen 90.000 palabras en el diccionario, pero solamente produce un 2% de mejora en palabras predichas y un 1% en pulsaciones ahorradas.

Sin embargo, la tendencia es ascendente, por lo que sería previsible que siguiera mejorando si se aumenta aún más el tamaño del diccionario.

4.1.3.3 Evaluación de los modelos estadísticos categoriales

A continuación se evalúan los modelos estadísticos categoriales. Se mostrará el efecto sobre la predicción del uso de los *bipos* y *tripos* sencillos, así como de las mejoras introducidas en ellos. Además, se incluyen los resultados de los estadísticos entrenados a partir de textos con ambigüedad.

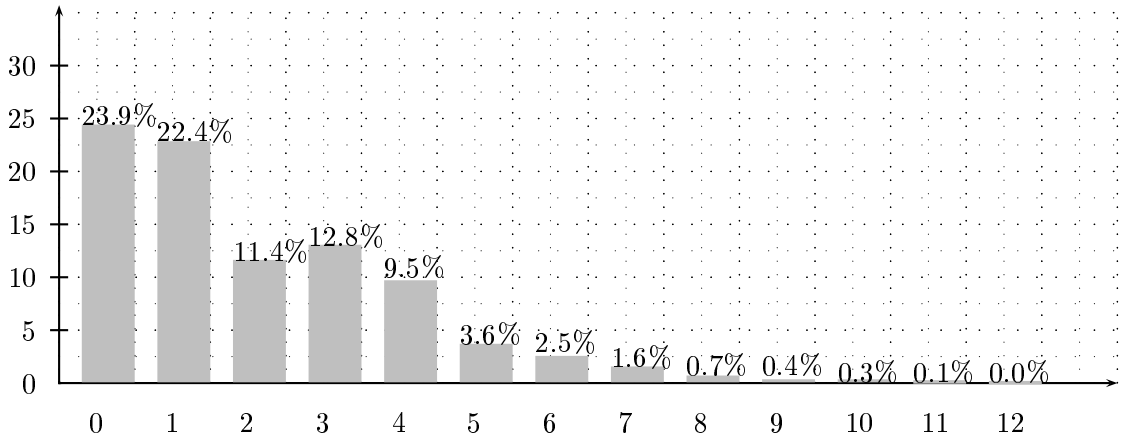
4.1.3.3.1 Evaluación de los *bipos* y la influencia del suavizado y del retroceso

En esta serie de experimentos se evalúa el efecto de la introducción de información gramatical básica en el proceso de predicción, dando prioridad a las palabras pertenecientes a las categorías gramaticales que más frecuentemente pueden seguir a la categoría de la palabra anterior.

Se utilizará el mismo texto de prueba que en la serie de experimentos anterior.

El experimento que se tomará como referencia es el mejor de la serie anterior, que sólo considera la información frecuencial del diccionario general completo (130000 palabras). Sus resultados son:

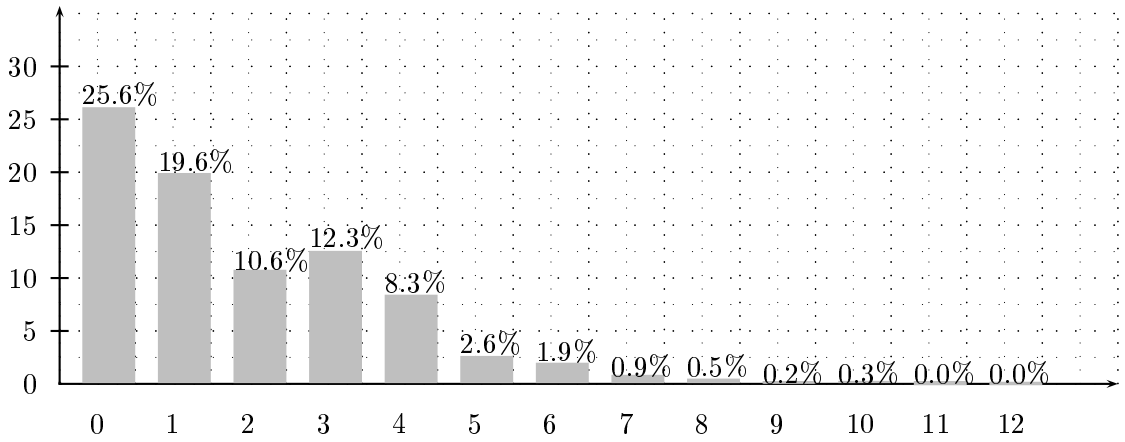
Número de palabras predichas	33504 (89.35%±0.31)
Número de pulsaciones ahorradas	92367 (36.37%±0.19)



Resultado 3.1: Resultados del experimento de referencia: predicción utilizando unigramas con el diccionario general completo.

En el siguiente experimento se utilizan las matrices de *bipos* entrenadas a partir de PAIS (texto de 55000 palabras categorizado sin ambigüedad) sin suavizado y sin retroceso, es decir, si, por ejemplo, la posición de la matriz correspondiente al par C_1 C_2 es cero, después de C_1 no se predicen nunca palabras pertenecientes a C_2. Los resultados de este experimento son:

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33504 (89.35%±0.31)	31050 (82.81%±0.38)	-7.32%
Pulsacs. ahorradas	92367 (36.37%±0.19)	90437 (35.61%±0.19)	-2.09%



Resultado 3.2: Resultados del experimento que utiliza *bipos* entrenados a partir de un texto de 55000 palabras categorizado sin ambigüedad, sin suavizado y sin retroceso.

En estas gráficas se puede observar que disminuye tanto el número de palabras predichas como el de pulsaciones ahorradas. Este empeoramiento se debe fundamentalmente al *número de ceros en la matriz*, es decir, al número de posiciones de la matriz de *bipos* que son nulas, que se corresponden con secuencias de categorías que no aparecen en los textos de entrenamiento. La matriz de *bipos* utilizada contiene un 45.99% de posiciones nulas. Este porcentaje de ceros tiene efectos negativos, que será necesario compensar, como se verá en los siguientes experimentos.

Sin embargo, el número de palabras encontradas en la *letra cero* se ha incrementado de 23.9% a 25.6%, es decir, la información aportada por la matriz de *bipos* aumenta la calidad de la lista de palabras predichas en la *letra cero*, por lo que, a continuación, se buscan métodos alternativos que nos permitan aprovechar esta mejora en la calidad de las palabras predichas sin empeorar el rendimiento global.

En el siguiente experimento se aplica a la matriz de *bipos* un suavizado umbral (explicado en la sección 3.3.2.1.2 (página 62) sobre suavizado de matrices). Los resultados obtenidos son:

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33504 (89.35%±0.31)	31053 (82.82%±0.38)	-7.32%
Pulsacs. ahorradas	92367 (36.37%±0.19)	90533 (35.65%±0.19)	-1.99%

Resultado 3.3: Resultados del experimento aplicando suavizado umbral a la matriz de *bipos*.

Como se puede comprobar, la eliminación de los ceros de la matriz produce una mejora no significativa en el porcentaje de pulsaciones ahorradas con respecto al método anterior, pero no se consigue mejorar los resultados del diccionario general en porcentaje total de

palabras predichas ni pulsaciones ahorradas, por lo que se continúa modificando el modelo, introduciendo mayor cantidad de información.

Se han realizado experimentos con otras formas de suavizado, como son el descuento lineal y absoluto [Ney94], Good-Turing [Katz87] y Witten-Bell [Witt91]). Dichas técnicas mejoran ligeramente los resultados obtenidos con respecto al uso del suavizado umbral, pero siguen sin alcanzar la mejora de rendimiento que producen los modelos con retroceso que se muestran a continuación.

En el siguiente experimento se comprueba el rendimiento del modelo *bipos* con retroceso a *unipos*: se utiliza la matriz sin suavizar, y cuando el método de predicción basado en *bipos* no predice el número suficiente de palabras, se muestran predicciones procedentes del *unipos*.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33504 (89.35%±0.31)	32366 (86.32%±0.35)	-3.40%
Pulsacs. ahorradas	92367 (36.37%±0.19)	93808 (36.94%±0.19)	1.56%

Resultado 3.4: Resultados del experimento utilizando con *bipos* sin suavizar con retroceso a *unipos*.

Este método de predicción, con retroceso al modelo anterior, produce mejores resultados en porcentaje de pulsaciones ahorradas que el suavizado, consiguiendo, además, una mejora significativa con respecto al experimento de referencia.

Se puede observar que el número de palabras predichas es menor que en el experimento de referencia, y, a pesar de todo, disminuye el número de pulsaciones necesarias para escribir el texto. Esto es así porque se predicen menos palabras, pero éstas son más largas, y, además, se predicen antes (en la *letra cero* se predice un mayor número de palabras, 26.8%).

En el último experimento se comprueba la eficiencia del retroceso a unigramas si el método basado en *bipos* sin suavizar no predice el número suficiente de palabras. Sus resultados se muestran a continuación:

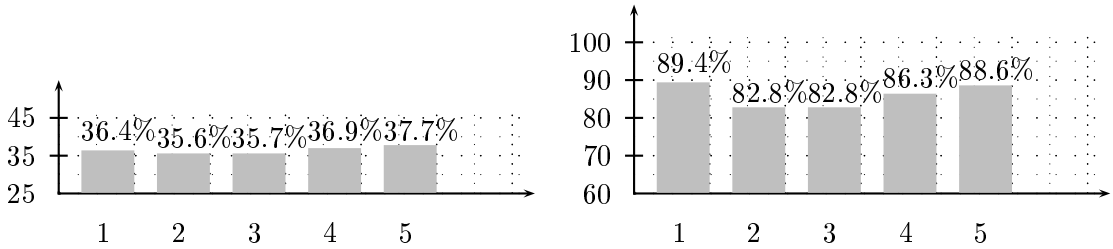
Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33504 (89.35%±0.31)	33232 (88.63%±0.32)	-0.81%
Pulsacs. ahorradas	92367 (36.37%±0.19)	95742 (37.70%±0.19)	3.65%

Resultado 3.5: Resultados del experimento con *bipos* sin suavizar, retrocediendo a unigramas.

Como se puede observar, este experimento es el que produce mejores resultados en número de pulsaciones ahorradas, consiguiendo una mejora relativa del 3.65% con respecto al experimento de referencia. El número de palabras predichas sigue siendo ligeramente menor que la predicción basada exclusivamente en unigramas, por las mismas razones que se explicaron para el experimento anterior: se encuentran palabras diferentes, más largas, en detrimento de las palabras más cortas, que producen un ahorro menor en el número de pulsaciones. Como las palabras cortas sólo se muestran en la *letra cero* o *letra uno*, el usuario acaba de escribirlas antes de que se predigan, disminuyendo en número total de

palabras predichas.

A continuación se muestran las gráficas resumen de los resultados más significativos de esta serie de experimentos:



Resultado 3.6: Evolución del porcentaje de pulsaciones ahorradas y palabras predichas en los experimentos de esta serie.

Donde:

1: Resultados del experimento de referencia: predicción utilizando unigramas con el diccionario general completo.

2: Resultados de la predicción utilizando *bipos* entrenados a partir de un texto de 55000 palabras categorizado sin ambigüedad, sin suavizado y sin retroceso.

3: Resultados del experimento anterior aplicándole a la matriz de *bipos* un suavizado umbral.

4: Resultados del experimento utilizando *bipos* sin suavizar y retroceso a *unipos*.

5: Resultados del experimento con *bipos* sin suavizar y retroceso a unigramas.

En estas gráficas se puede apreciar la evolución del número de pulsaciones ahorradas y palabras predichas al introducir el modelo basado en *bipos* en la predicción de palabras, con las sucesivas mejoras que se han añadido.

El modelo sencillo produce un empeoramiento en ambos parámetros, que se corrige al introducir las técnicas de retroceso a unigramas, que proporcionan mejores resultados que otros métodos probados, como diversas técnicas de suavizado de matrices y el retroceso a *unipos*.

Con respecto a los resultados obtenidos por el modelo basado en *bipos* con retroceso a unigramas, éste produce una mejora relativa del 3.65% en el porcentaje de pulsaciones ahorradas con respecto al experimento de referencia. Como se ha podido comprobar, en número total de palabras predichas disminuye, porque se encuentra un menor número de palabras cortas, que son también las más frecuentes. En el modelo basado exclusivamente en unigramas se encuentran siempre en la *letra cero* las palabras *de, la, el, que* y *en*, lo que, por ser las más frecuentes del idioma, es muy productivo en términos de porcentaje de palabras predichas, pero poco eficiente en pulsaciones ahorradas por la reducida longitud de las palabras. El aumento en la calidad de la predicción se manifiesta especialmente en las predicciones realizadas en la *letra cero*, que, al no ser constantes sino dependientes de la palabra anterior, aumentan su porcentaje de aciertos, y como, además, las palabras predichas son más largas, mejora el porcentaje de pulsaciones ahorradas.

4.1.3.3.2 Evaluación de los *tripos* y de la influencia del suavizado y del retroceso

En esta serie de experimentos se aumenta la cantidad de información gramatical considerada en el sistema, incluyendo estadísticas sobre las secuencias de tres categorías gramaticales, en lugar de dos.

Se utilizará el mismo texto de prueba y experimento de referencia que en la serie de experimentos anterior.

En el siguiente experimento se introduce el modelo de predicción basado en *tripos* entrenados a partir de un texto de 55000 palabras categorizado sin ambigüedad, sin suavizado de la matriz y sin retroceso a otros modelos de predicción. Sus resultados son:

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33504 (89.35%±0.31)	28436 (75.84%±0.43)	-15.13%
Pulsacs. ahorradas	92367 (36.37%±0.19)	83374 (32.83%±0.18)	-9.74%

Resultado 3.7: Resultados del experimento utilizando un modelo de predicción basado en *tripos* entrenados a partir de un texto de 55000 palabras categorizado sin ambigüedad, sin suavizado y sin retroceso.

Se puede observar un empeoramiento significativo en los resultados del sistema, debidos, al igual que ocurría con los *bipos*, al número de ceros de la matriz de *tripos*. Sus efectos son más acusados porque el porcentaje de ceros es mayor (78.91%).

Se han realizado experimentos con matrices de *tripos* entrenadas a partir de textos de diferentes longitudes, entre 1000 y 55000 palabras. Las matrices contienen porcentajes de ceros entre el 95.32% y el 78.91%, y se ha verificado que los resultados del sistema de predicción son peores con las matrices con mayor número de ceros, como era de esperar.

De la misma forma que en el modelo basado en *bipos*, se han incluido métodos que disminuyen el efecto negativo de estos ceros. A continuación se muestran los resultados del experimento en el que se utiliza suavizado umbral:

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33504 (89.35%±0.31)	29556 (78.82%±0.41)	-11.78%
Pulsacs. ahorradas	92367 (36.37%±0.19)	85859 (33.81%±0.18)	-7.05%

Resultado 3.8: Experimento anterior con la matriz de *tripos* suavizada.

Se puede observar, que los resultados mejoran con respecto a los del sistema basado en *tripos* sin suavizar, pero no con respecto al experimento de referencia, por lo cual se comprueban otros métodos.

En el siguiente experimento se utiliza la matriz de *tripos* sin suavizar, y, si el método categorial no predice el suficiente número de palabras, se utiliza el módulo basado en *bipos*, sin más retrocesos adicionales. Sus resultados son:

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33504 (89.35%±0.31)	32289 (86.11%±0.35)	-3.63%
Pulsacs. ahorradas	92367 (36.37%±0.19)	94149 (37.08%±0.19)	1.93%

Resultado 3.9: Experimento utilizando con *tripos* sin suavizar, con retroceso a *bipos*.

Este método produce una mejora con respecto al experimento de referencia en número de pulsaciones ahorradas, pero no en número de palabras predichas. En el siguiente experimento se retrocede a unigramas si el modelo basado en *tripos* no proporciona el número de palabras suficiente.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33504 (89.35%±0.31)	33277 (88.75%±0.32)	-0.68%
Pulsacs. ahorradas	92367 (36.37%±0.19)	95947 (37.78%±0.19)	3.88%

Resultado 3.10: Experimento utilizando *tripos* sin suavizar, con retroceso a unigramas.

El método que produce mejores resultados es el que se muestra a continuación, en que se combinan el retroceso a bipos y a unigramas, secuencialmente.

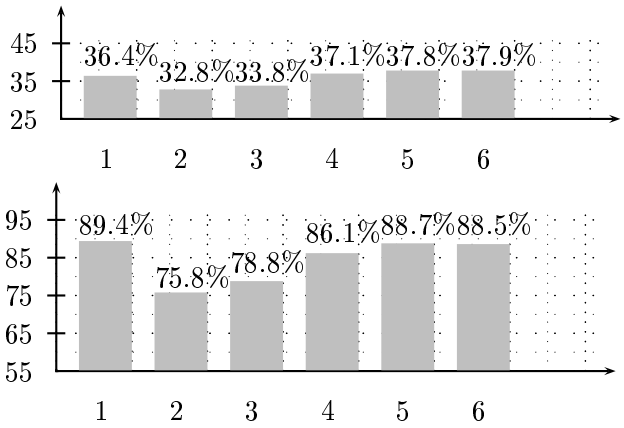
Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33504 (89.35%±0.31)	33187 (88.51%±0.32)	-0.95%
Pulsacs. ahorradas	92367 (36.37%±0.19)	96157 (37.87%±0.19)	4.10%

Resultado 3.11: Experimento con tripos sin suavizar, retrocediendo a *bipos* y a unigramas.

En estos resultados podemos ver una mejora en el número de pulsaciones ahorradas mayor incluso que la que produce el mejor de los métodos basado en *bipos*, aunque la diferencia no sea significativa.

El número de palabras predichas se reduce ligeramente, por las mismas razones que se explicaron en la sección anterior: se predice menor número de palabras cortas, aumentando el número de palabras largas correctamente predichas, y, con ellas, el número de pulsaciones ahorradas.

A continuación se muestran las gráficas con la evolución del número de pulsaciones ahorradas y palabras predichas en los modelos de predicción utilizados en esta serie de experimentos.



Resultado 3.12: Evolución del porcentaje de pulsaciones ahorradas y palabras predichas en los experimentos de esta serie.

Donde:

1: Experimento de referencia: predicción utilizando unigramas con el diccionario general completo.

2: Predicción utilizando *tripos* entrenados a partir de un texto de 55000 palabras categorizado sin ambigüedad, sin suavizado y sin retroceso.

3: Experimento anterior con la matriz suavizada.

4: Experimento con *tripos* sin suavizar, retrocediendo a *bipos*.

5: Experimento con *tripos* sin suavizar, retrocediendo a unigramas.

6: Experimento con *tripos* sin suavizar, retrocediendo a *bipos* y a unigramas, secuencialmente.

Las conclusiones de este modelo son similares a las obtenidas para el método de predicción basado en *bipos*. Por una parte, no puede utilizarse un modelo basado exclusivamente en las estadísticas entrenadas, por los efectos negativos que producen los ceros de la matriz de *tripos*. Es necesario introducir mecanismos adicionales para mejorar los resultados. Se han probado diferentes métodos de suavizado, de los cuales el retroceso a métodos más sencillos, *bipos* y unigramas, secuencialmente, es el que ha proporcionado resultados más satisfactorios.

Por otra parte, el número total de palabras predichas empeora con respecto al experimento de referencia, porque disminuye el número de palabras de dos o tres letras predichas en la *letra cero* por el modelo basado en unigramas. Sin embargo, las predicciones son de mayor calidad, consiguiéndose una mejora significativa en el ahorro de las pulsaciones necesarias para escribir el texto.

4.1.3.3.3 Evaluación de los *bipos* y *tripos* entrenados a partir de textos con ambigüedad

En esta serie de experimentos se han utilizado *bipos* y *tripos* generados a partir de textos categorizados con ambigüedad.

Se han realizado experimentos análogos a los anteriores, para comprobar su rendimiento con y sin suavizado y con y sin retroceso a modelos anteriores, obteniendo resultados similares, es decir, que el modelo basado en *bipos* mejora si se añade retroceso a unigramas y los *tripos* producen mejores resultados con retroceso a *bipos* (también entrenados con ambigüedad) y a unigramas, por lo que solamente se muestran los resultados de los mejores experimentos de cada modelo.

Se utilizará el mismo texto de prueba y experimento de referencia que en la serie de experimentos anterior.

A continuación se muestran los resultados de la predicción utilizando el método basado en *bipos* con retroceso a unigramas. La matriz de *bipos* ha sido entrenada a partir del texto PAIS (el mismo con el que se han entrenado los *bipos* originales), etiquetado con ambigüedad.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33504 (89.35%±0.31)	33244 (88.66%±0.32)	-0.78%
Pulsacs. ahorradas	92367 (36.37%±0.19)	95210 (37.49%±0.19)	3.08%

Resultado 3.13: Experimento con *bipos* entrenados con ambigüedad a partir de un texto de 55000 palabras con retroceso a unigramas.

Como se puede observar, aumenta el número de pulsaciones ahorradas. No se obtienen resultados tan buenos como con los *bipos* entrenados a partir del mismo texto sin ambigüedad, pero las diferencias no son significativas. En parte, puede ser debido a que la ambigüedad en la categoría de las palabras hace que cada par de palabras aumente la cuenta de varios posibles pares de categorías, generando una matriz con menor número de ceros (39.20%), reduciéndose así los efectos negativos de dichos ceros.

A continuación se muestran los resultados del experimento en el que se utilizan *tripos* entrenados a partir del texto PAIS, de 55000 palabras, etiquetado con ambigüedad.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33504 (89.35%±0.31)	33235 (88.64%±0.32)	-0.80%
Pulsacs. ahorradas	92367 (36.37%±0.19)	95918 (37.77%±0.19)	3.84%

Resultado 3.14: Experimento con *tripos* entrenados a partir de un texto de 55000 palabras etiquetado con ambigüedad, con retroceso a *bipos* y unigramas.

De nuevo se puede observar que también mejora la calidad de la predicción y, aunque estos resultados no alcanzan los de los *tripos* generados a partir de textos categorizados sin ambigüedad, las diferencias no son significativas. De nuevo, parte de la mejora producida puede ser debida a la disminución en el porcentaje de ceros de la matriz, que pasa de ser

un 78.91% a un 69.62%, entrenando a partir del mismo texto.

En los experimentos anteriores, los *bipos* y *tripos* han sido entrenados a partir de un texto de 55000 palabras porque también estaba disponible categorizado sin ambigüedad, y se podían comparar los resultados de los sistemas con las matrices generadas a partir del mismo texto con y sin ambigüedad.

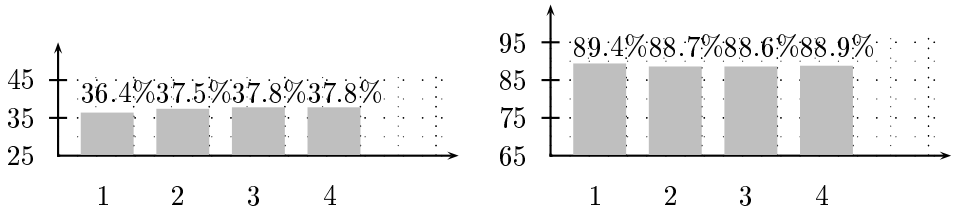
Habiendo visto que los resultados obtenidos por los modelos no son significativamente diferentes, a continuación se muestran los resultados del experimento en el que se entrenan los *tripos* a partir de un texto de un millón de palabras (MUNDO1M) categorizado con ambigüedad.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33504 (89.35%±0.31)	33319 (88.86%±0.32)	-0.55%
Pulsacs. ahorradas	92367 (36.37%±0.19)	95951 (37.79%±0.19)	3.88%

Resultado 3.15: Experimento con *tripos* entrenados con ambigüedad a partir de un texto de 1 millón palabras, con retroceso a *bipos* y unigramas.

Se puede observar una ligerísima mejora con respecto al experimento anterior, es decir, el aumento en la cantidad de textos de entrenamiento produce un aumento en la calidad de la predicción, aunque el texto esté categorizado con ambigüedad. La mejora, sin embargo, no es significativa, ni consigue superar los resultados obtenidos por los modelos entrenados sin ambigüedad.

A continuación se muestran los resultados conjuntos de esta serie de experimentos.



Resultado 3.16: Evolución del porcentaje de pulsaciones ahorradas y palabras predichas en los experimentos de esta serie.

Donde:

1: Experimento de referencia: predicción utilizando unigramas con el diccionario general completo.

2: Experimento con *bipos* entrenados con ambigüedad a partir de un texto de 55000 palabras, con retroceso a unigramas.

3: Experimento con *tripos* entrenados con ambigüedad a partir de un texto de 55000 palabras, con retroceso a *bipos* y unigramas.

4: Experimento con *tripos* entrenados con ambigüedad a partir de un texto de 1 millón palabras, con retroceso a *bipos* y unigramas.

En estas tablas se puede observar la mejora en el porcentaje de pulsaciones ahorradas

producido por los métodos, tal y como ha sido comentado en cada uno de ellos.

Los modelos generados a partir de textos categorizados con ambigüedad producen un ahorro en el número de pulsaciones ligeramente inferior al producido por los métodos entrenados con textos categorizados sin ambigüedad, por lo que si la cantidad de texto etiquetado manualmente es insuficiente, pueden utilizarse el texto categorizado con ambigüedad para entrenar el sistema, y se obtendrán resultados satisfactorios. Si se dispone de texto etiquetado de forma manual se obtendrán resultados mejores.

Sería necesario estudiar el comportamiento de estos modelos categorizados con ambigüedad para otros esquemas de categorización antes de extrapolar estas conclusiones a sistemas con mayor número de categorías.

El porcentaje de palabras predichas disminuye con respecto al experimento de referencia por las razones explicadas anteriormente.

4.1.3.3.4 Evaluación del efecto de la categorización de las palabras desconocidas y del tratamiento de los rasgos

En esta serie de experimentos se evalúa el efecto de dos de las mejoras introducidas en los modelos categoriales básicos. Por una parte, la categorización de las palabras desconocidas, que se explica en el apartado 3.3.2.1.5 (página 65), y, por otra, el tratamiento básico de rasgos que se realiza en paralelo con la predicción de los *bipos* y *tripos*, que se explica en la sección 3.3.2.1.1 (página 61).

El texto de prueba utilizado en esta secuencia de experimentos es el mismo que se utiliza en los experimentos anteriores.

Se ha elegido como experimento de referencia el que mejores resultados ha proporcionado de los que se han visto hasta ahora, es decir, el que basa la predicción en *tripos* entrenados a partir de un texto categorizado sin ambigüedad, con retroceso a *bipos* y unigramas.

A continuación se realiza el mismo experimento introduciendo la gestión de palabras desconocidas. Sus resultados son:

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33187 (88.51%±0.32)	33165 (88.45%±0.32)	-0.07%
Pulsacs. ahorradas	96157 (37.87%±0.19)	96252 (37.90%±0.19)	0.10%

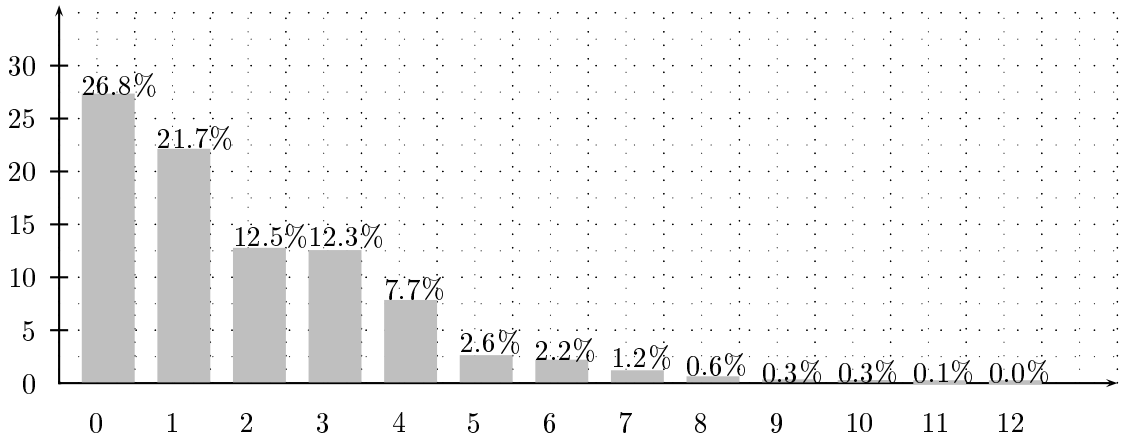
Resultado 3.17: Resultados del experimento de referencia, añadiendo categorización de palabras desconocidas.

Como se puede observar, se produce una ligerísima mejora en el porcentaje de pulsaciones ahorradas, que, de hecho, no es significativa. El número de palabras desconocidas del texto es 3658 (el 9.76%). De ellas, 700 (1.87%) son categorizadas utilizando los prefijos, terminaciones distinguidas, etc., y al resto 2958 (7.89%) se les asigna el conjunto de categorías que se indica en la sección 3.3.2.1.5 (página 65).

En el siguiente experimento se añade la gestión de rasgos paralela a los *tripos*, reali-

zando predicciones que concuerden en género y número con la última palabra escrita. Los resultados obtenidos son:

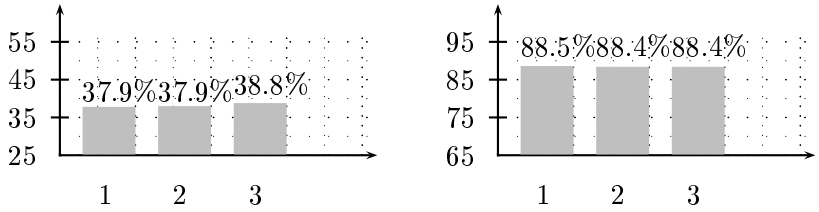
Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33187 (88.51%±0.32)	33176 (88.48%±0.32)	-0.03%
Pulsacs. ahorradas	96154 (37.87%±0.19)	98665 (38.85%±0.19)	2.59%



Resultado 3.18: Experimento anterior añadiendo gestión de rasgos.

Se puede observar que se produce una mejora significativa en el ahorro de pulsaciones al añadir la gestión de rasgos al sistema, ya que aumenta la cantidad de información que se utiliza al predecir después de algunas de las categorías.

A continuación se muestran los resultados conjuntos de los experimentos.



Resultado 3.19: Evolución del porcentaje de pulsaciones ahorradas y palabras predichas en los experimentos de esta serie.

Donde:

1: Experimento de referencia: predicción utilizando *tripos* con retroceso a *bipos* y unigramas.

2: Efecto de la categorización de palabras desconocidas.

3: Efecto de la gestión de rasgos junto a la categorización de palabras desconocidas.

Se puede observar la mejora que se produce en el porcentaje de pulsaciones ahorradas al añadirle potencia al modelo basado en *tripos*. Al igual que ocurrió en los experimentos anteriores, el número total de palabras predichas disminuye, pero éstas son de mayor

calidad (ahorran un mayor número de pulsaciones).

4.1.3.3.5 Evaluación de los métodos categoriales con textos de dominio periodístico

En esta serie de experimentos se evalúan los mejores métodos de predicción de cada una de las series anteriores, aplicándolos a un texto periodístico de más de 100000 palabras. Se podrá comprobar que hay tendencias que se mantienen con respecto a los resultados vistos anteriormente, mientras que en otros experimentos se observará su dependencia con el estilo del texto.

Las características del texto utilizado para las pruebas son:

Nombre del texto de prueba	FEB100K
Número de palabras	103583
Pulsaciones para escribirlo sin predicción	751191

Como se indica en el apéndice A (página 179) correspondiente a la descripción de los textos, FEB100K procede del mes de Febrero de la edición electrónica del periódico *El Mundo*, es decir, mantiene el mismo estilo que los textos a partir de los cuales se ha entrenado el diccionario (aunque no se solapa con ellos, ya que en la generación del diccionario se han utilizado textos de otros meses).

El experimento de referencia de esta serie utiliza predicción de palabras basada en el modelo unigram, es decir, solamente considera la información frecuencial del diccionario general, y sus resultados son:

Número de palabras predichas	92012 (88.83%±0.19)
Número de pulsaciones ahorradas	266824 (35.52%±0.11)

Resultado 3.20: Experimento de referencia: predicción utilizando unigramas con el diccionario general completo.

A continuación se muestran los resultados al incluir el modelo *bipos* (entrenados a partir de un texto de 55000 palabras categorizado sin ambigüedad), con retroceso a unigramas.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	92012 (88.83%±0.19)	91673 (88.50%±0.19)	-0.37%
Pulsacs. ahorradas	266824 (35.52%±0.11)	273812 (36.45%±0.11)	2.62%

Resultado 3.21: Resultados de la predicción utilizando *bipos* con retroceso a unigram.

Como se puede comprobar, el resultado es similar al obtenido con el texto TESIS, de carácter científico, con el que se realizaron las pruebas anteriores, es decir, mejora el porcentaje de pulsaciones ahorradas de forma significativa, mientras que el porcentaje de palabras predichas es inferior, por la disminución en el número de palabras cortas predichas. El número de palabras predichas en la *letra cero* mejora de un 25.5% a un

28.3%, por el aumento de la calidad de la predicción.

A continuación se realiza el experimento con predicción basada en *tripos* entrenados a partir de un texto de 1 millón de palabras con ambigüedad, con retroceso a *bipos* y unigram. El texto a partir del cual se entrenan los *tripos* también procede del periódico *El Mundo*, aunque no del mismo mes, por lo cual no se solapa con el texto de prueba.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	92012 (88.83%±0.19)	91820 (88.64%±0.19)	-0.21%
Pulsacs. ahorradas	266824 (35.52%±0.11)	275479 (36.67%±0.11)	3.24%

Resultado 3.22: Resultados de la predicción utilizando *tripos* entrenados a partir de un texto de 1 millón de palabras con ambigüedad con retroceso a *bipos* y unigram.

Se puede observar que la predicción es mejor que la del experimento anterior basado en *bipos*, tanto en número de palabras predichas como de pulsaciones ahorradas.

A continuación se realiza el experimento utilizando predicción basada en *tripos* entrenados a partir de un texto de 55000 palabras etiquetado sin ambigüedad, con retroceso a *bipos* y unigramas. Se debe recordar que en los experimentos de las series anteriores, aplicado a texto de dominio científico, los *tripos* entrenados sin ambigüedad proporcionaban mejores resultados que los entrenados con ambigüedad.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	92012 (88.83%±0.19)	91514 (88.35%±0.20)	-0.54%
Pulsacs. ahorradas	266824 (35.52%±0.11)	274192 (36.50%±0.11)	2.76%

Resultado 3.23: Predicción utilizando *tripos* entrenados sin ambigüedad con retroceso a *bipos* y unigramas.

Como se puede observar, los resultados de este experimento son peores que los del experimento anterior, al contrario a lo que ocurría con los textos de ámbito científico (aunque la diferencia no es significativa). Es decir, los modelos basados en *tripos* entrenados a partir de un texto del mismo dominio etiquetado con ambigüedad proporcionan resultados ligeramente mejores, posiblemente porque, los efectos negativos de la ambigüedad se compensan con la adaptación al estilo de escritura.

A continuación se añaden las otras características que habían mejorado los *tripos* en experimentos anteriores. En primer lugar, la categorización de palabras desconocidas.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	92012 (88.83%±0.19)	91473 (88.31%±0.20)	-0.59%
Pulsacs. ahorradas	266824 (35.52%±0.11)	274195 (36.50%±0.11)	2.76%

Resultado 3.24: Experimento anterior con categorización de palabras desconocidas.

La categorización de palabras desconocidas no produce una variación apreciable en la eficiencia de la predicción basada en *tripos*.

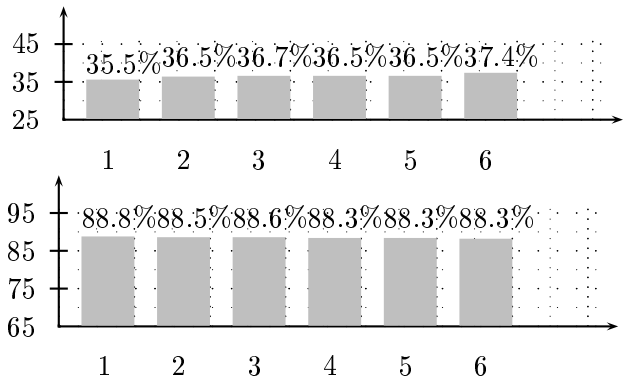
A continuación se añade tratamiento de rasgos.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	92012 (88.83%±0.19)	91419 (88.26%±0.20)	-0.64%
Pulsacs. ahorradas	266824 (35.52%±0.11)	280855 (37.39%±0.11)	5.26%

Resultado 3.25: Experimento anterior añadiendo tratamiento de rasgos.

Se puede observar que el tratamiento de rasgos proporciona mejores resultados independientemente del dominio del texto de prueba, consiguiendo incluso mejorar la predicción de los *tripos* generados a partir del texto de *El Mundo*.

A continuación se muestran los resultados conjuntos de la serie de experimentos:



Resultado 3.26: Evolución del porcentaje de pulsaciones ahorradas y palabras predichas en los experimentos de esta serie.

Donde:

1: Experimento de referencia: predicción utilizando unigramas con el diccionario general completo.

2: Predicción utilizando *bipos* entrenados a partir de un texto de 55000 palabras categorizado sin ambigüedad con retroceso a unigram.

3: Predicción utilizando *tripos* entrenados a partir de un texto de 1 millón de palabras con ambigüedad con retroceso a *bipos* y unigram.

4: Predicción utilizando *tripos* entrenados sin ambigüedad con retroceso a *bipos* y unigramas.

5: Experimento anterior con categorización de palabras desconocidas.

6: Experimento anterior añadiendo tratamiento de rasgos.

Se puede observar la evolución de las pulsaciones ahorradas y las palabras predichas. La única diferencia que se puede apreciar con respecto al comportamiento visto en experimentos anteriores es el aumento en número de pulsaciones ahorradas correspondiente a los *tripos* entrenados a partir de un texto de un millón de palabras categorizadas con ambigüedad del mismo dominio que el texto de prueba.

Se puede decir, entonces, que el entrenamiento utilizando gran cantidad de texto categorizado con ambigüedad, es una opción razonable, ya que consiguen resultados que no difieren esencialmente de los proporcionados por los modelos entrenados sin ambigüedad, cuando no se dispone de gran cantidad de material etiquetado manualmente. Esta conclusión es especialmente relevante debido al gran coste necesario para conseguir dicho material.

4.1.3.3.6 Evaluación con textos pequeños de los métodos estadísticos categoriales

A continuación se muestra el comportamiento del sistema para una serie de textos de pequeña longitud, como los que podría escribir un usuario en una o varias sesiones.

Los textos pertenecen a dominios diferentes. Los primeros son cuentos cortos, después dos fragmentos de *El Quijote*, un texto periodístico y un apartado de esta tesis. Con cada uno de ellos se han realizado dos experimentos, uno con el diccionario general, utilizando solamente información frecuencial, que se ha tomado como referencia y otro utilizando toda la potencia asociada a los *tripos*: retroceso a *bipos* y unigramas, tratamiento de rasgos y categorización de palabras desconocidas.

A continuación se muestran los resultados obtenidos para cada uno de ellos.

Experimento 1

Nombre del texto de prueba	CUENTO1
Número de palabras	986
Pulsaciones para escribirlo sin predicción	6470

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	866 (87.83%±2.04)	848 (86.00%±2.17)	-2.08%
Pulsacs. ahorradas	2006 (31.00%±1.12)	2121 (32.78%±1.14)	5.73%

Experimento 2

Nombre del texto de prueba	CUENTO2
Número de palabras	2672
Pulsaciones para escribirlo sin predicción	14878

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	2334 (87.35%±1.26)	2341 (87.61%±1.25)	0.30%
Pulsacs. ahorradas	4843 (32.55%±0.75)	5107 (34.33%±0.76)	5.45%

Experimento 3

Nombre del texto de prueba	CUENTO3
Número de palabras	1416
Pulsaciones para escribirlo sin predicción	8682

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	1253 (88.49%±1.66)	1231 (86.94%±1.76)	-1.76%
Pulsacs. ahorradas	2984 (34.37%±1.00)	3134 (36.10%±1.01)	5.03%

Experimento 4

Nombre del texto de prueba	MOLINOS
Número de palabras	2980
Pulsaciones para escribirlo sin predicción	17715

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	2582 (86.64%±1.22)	2545 (85.40%±1.27)	-1.43%
Pulsacs. ahorradas	5507 (31.09%±0.68)	5765 (32.54%±0.69)	4.68%

Experimento 5

Nombre del texto de prueba	QUIJOTE(2500)
Número de palabras	2446
Pulsaciones para escribirlo sin predicción	14660

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	2081 (85.08%±1.41)	2072 (84.71%±1.43)	-0.43%
Pulsacs. ahorradas	4545 (31.00%±0.75)	4776 (32.58%±0.76)	5.08%

Experimento 6

Nombre del texto de prueba	FEB1K
Número de palabras	1286
Pulsaciones para escribirlo sin predicción	8536

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	1167 (90.75%±1.58)	1160 (90.20%±1.62)	-0.60%
Pulsacs. ahorradas	3181 (37.27%±1.03)	3368 (39.46%±1.04)	5.88%

Experimento 7

Nombre del texto de prueba	TESIS(2000)
Número de palabras	2059
Pulsaciones para escribirlo sin predicción	14267

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	1884 (91.50%±1.20)	1879 (91.26%±1.22)	-0.27%
Pulsacs. ahorradas	5555 (38.94%±0.80)	5923 (41.52%±0.81)	6.62%

La tendencia que se manifiesta en los resultados es una mejora relativa en el ahorro de pulsaciones comprendida entre un 4.68% y un 6.62% en todos los casos, entre los métodos basados en *tripos* y aquellos que utilizan solamente información frecuencial.

Sin embargo, si nos fijamos en los porcentajes de pulsaciones ahorradas, es destacable su gran varianza entre los distintos textos, llegándose incluso a observar que el valor absoluto de los resultados obtenidos por los unigramas en ciertos textos supera al obtenido por los modelos basados en *tripos* para textos diferentes.

El porcentaje de palabras predichas empeora en todos los experimentos excepto en dos de ellos, por las razones que se comentaron al explicar los resultados de los bipos.

4.1.3.4 Evaluación del modelo formal

En esta sección se evalúa el resultado del módulo basado en una gramática de contexto libre.

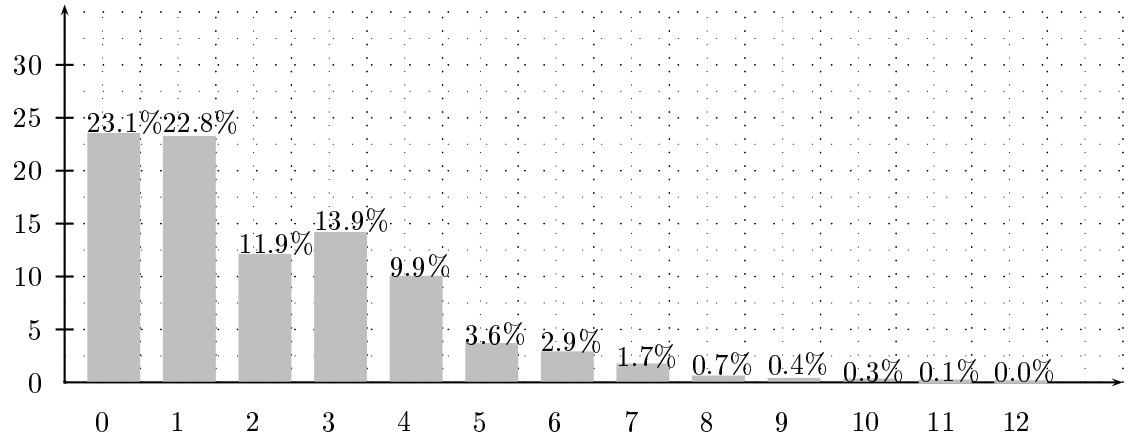
En la primera serie de experimentos se compara el mejor método categorial estadístico (*tripos* con gestión de rasgos), con el modelo basado en el analizador, para las primeras 17500 palabras del texto TESIS.

Las características del texto de prueba son:

Nombre del texto de prueba	TESIS(17500)
Número de palabras	17500
Pulsaciones para escribirlo sin predicción	122425

En el experimento de referencia se utiliza solamente la información frecuencial del diccionario general (modelo *unigram*), cuyos resultados son:

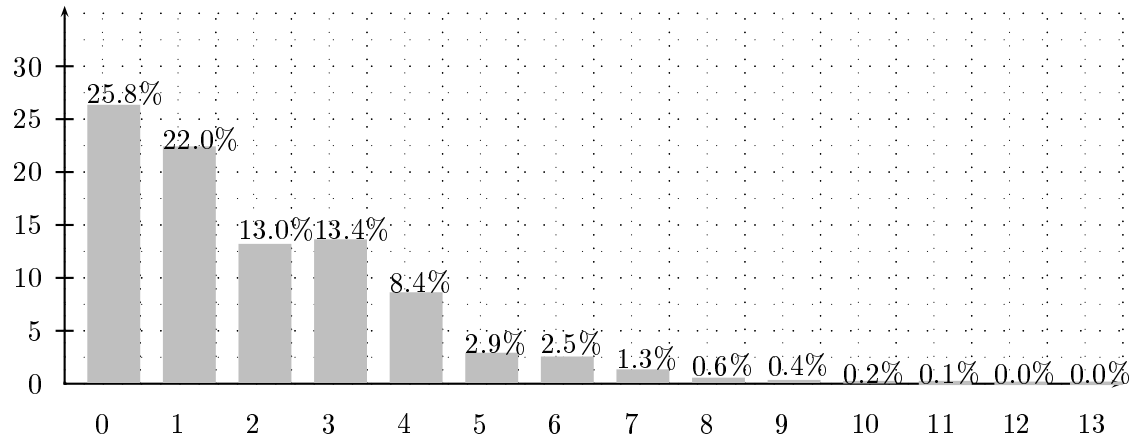
Número de palabras predichas	15990 (91.37%±0.42)
Número de pulsaciones ahorradas	46310 (37.83%±0.27)



Resultado 4.1: Experimento de referencia, que utiliza solamente la información frecuencial del diccionario general.

A continuación se muestran los resultados del experimento aplicando *tripos*, para poder realizar la comparación de ambos métodos gramaticales.

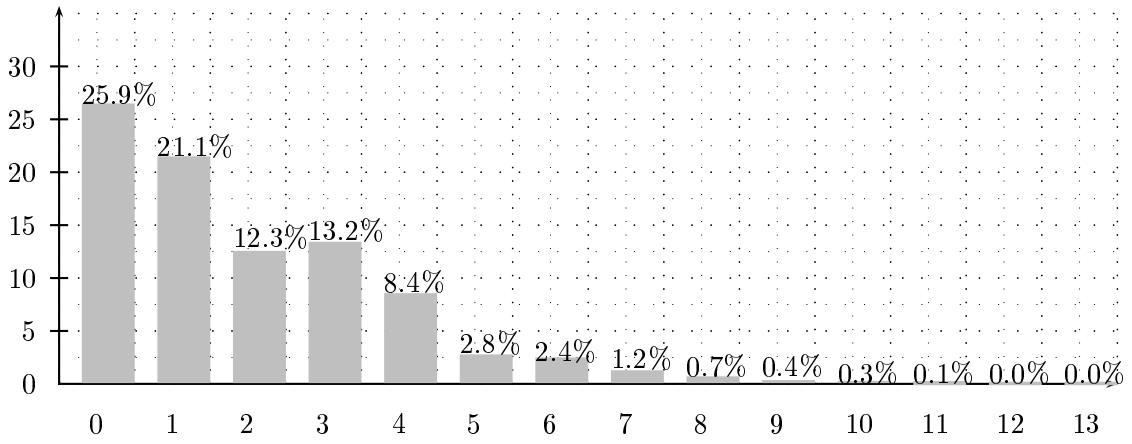
Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	15990 (91.37%±0.42)	15854 (90.59%±0.43)	-0.85%
Pulsacs. ahorradas	46310 (37.83%±0.27)	49195 (40.18%±0.27)	6.23%



Resultado 4.2: Experimento anterior añadiendo *tripos* con gestión de rasgos sobre el diccionario general.

Por último, los resultados del módulo gramatical formal, aplicado sobre el mismo texto. Se utiliza retroceso al método basado en *tripos* cuando la longitud de la lista de palabras predichas no sea suficientemente larga.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	15990 (91.37%±0.42)	15537 (88.78%±0.47)	-2.83%
Pulsacs. ahorradas	46310 (37.83%±0.27)	48883 (39.93%±0.27)	5.56%



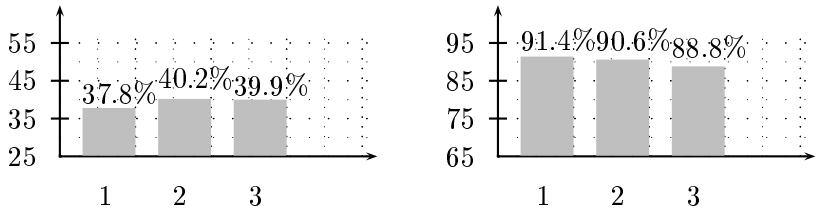
Resultado 4.3: Experimento inicial con el filtrado gramatical basado en un modelo formal aplicado al diccionario general.

Como se puede comprobar, el método formal produce una disminución en el porcentaje de pulsaciones ahorradas, aunque las bandas se solapan, por lo que no se puede afirmar que la diferencia sea significativa. También se puede observar una disminución en el número de palabras predichas.

Este empeoramiento es debido a que la gramática no está perfectamente adaptada al texto que se está escribiendo. Se debe considerar que el texto TESIS contiene gran cantidad de referencias bibliográficas (que, al no ser palabras, no están incluidas en el diccionario), incisos y frases de longitud muy elevada que pueden interrumpir el funcionamiento del analizador, ya que estos fenómenos aún no están modelados en la gramática.

A pesar de ello, se puede observar que en la *letra cero* se produce una ligera mejora con respecto al método basado en *tripos*.

Los resultados conjuntos de los experimentos de esta serie son:



Resultado 4.4: Evolución del porcentaje de pulsaciones ahorradas y palabras predichas en los experimentos de esta serie.

Donde:

1: Experimento de referencia, que utiliza solamente la información frecuencial del

diccionario general.

2: Experimento anterior añadiendo *tripos* con gestión de rasgos sobre el diccionario general.

3: Experimento inicial con el filtrado gramatical basado en un modelo formal aplicado al diccionario general.

4.1.3.4.1 Evaluación del modelo formal sobre textos de diferentes estilos

A continuación muestran los resultados de repetir la serie de experimentos anterior con otros dos textos de diferentes estilos.

El primero de los textos es:

Nombre del texto de prueba	CUENTO1
Número de palabras	986
Pulsaciones para escribirlo sin predicción	6470

El experimento de referencia, como en el caso anterior, es el modelo basado exclusivamente en la información frecuencial del diccionario general.

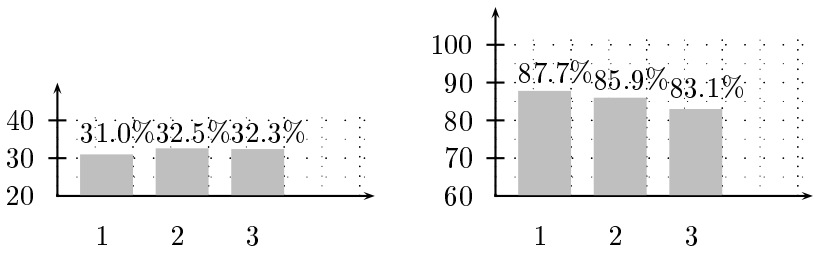
Cuando se añade el modelo basado en *tripos* con gestión de rasgos sobre el diccionario general, se obtiene:

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	865 (87.73%±2.04)	847 (85.90%±2.17)	-2.08%
Pulsacs. ahorradas	2006 (31.00%±1.12)	2104 (32.52%±1.14)	4.89%

Al incluir el modelo formal, los resultados empeoran ligeramente con respecto al último experimento:

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	865 (87.73%±2.04)	819 (83.06%±2.34)	-5.32%
Pulsacs. ahorradas	2006 (31.00%±1.12)	2091 (32.32%±1.14)	4.24%

Las gráficas que resumen los resultados de esta serie de experimentos son:



Resultado 4.5: Evolución del porcentaje de pulsaciones ahorradas y palabras predichas en los experimentos de esta serie.

Donde:

1: Experimento de referencia con predicción basada en la información frecuencial del diccionario en curso.

2: Experimento con predicción basada en *tripos* con gestión de rasgos aplicada al diccionario general.

3: Experimento con predicción basada en la gramática independiente del contexto.

En esta segunda serie se utiliza como texto de prueba el formado por las primeras 1400 palabras de "El Quijote":

Nombre del texto de prueba	QUIJOTE(1400)
Número de palabras	1390
Pulsaciones para escribirlo sin predicción	8223

El experimento de referencia se basa únicamente en la información frecuencial del diccionario general (unigramas).

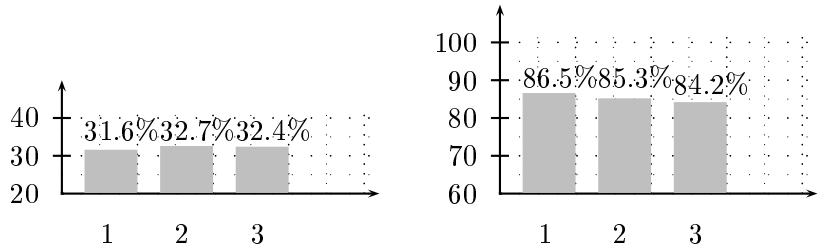
Al añadir el modelo gramatical estadístico, los resultados son:

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	1203 (86.55%±1.79)	1185 (85.25%±1.86)	-1.50%
Pulsacs. ahorradas	2602 (31.64%±1.01)	2687 (32.68%±1.01)	3.27%

Al igual que para el texto anterior, el modelo basado en el analizador mejora con respecto al modelo sin información gramatical, pero no en la misma medida que el modelo basado en *tripos*:

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	1203 (86.55%±1.79)	1170 (84.17%±1.92)	-2.74%
Pulsacs. ahorradas	2602 (31.64%±1.01)	2668 (32.45%±1.01)	2.54%

Los resultados conjuntos para esta serie son:



Resultado 4.6: Evolución del porcentaje de pulsaciones ahorradas y palabras predichas en los experimentos de esta serie.

Donde:

1: Experimento de referencia con predicción basada en la información frecuencial del diccionario en curso.

2: Experimento con predicción basada en *tripos* con gestión de rasgos aplicada al diccionario general.

3: Experimento con predicción basada en la gramática independiente del contexto.

Como se puede observar, el modelo basado en la gramática independiente del contexto no mejora con respecto a los modelos estadísticos categoriales en estos textos, ya que la gramática ha sido entrenada a partir de textos de estilo diferente. De todas formas, se obtienen resultados similares, es decir, aunque el texto no se ajuste a la gramática, se realiza un guiado gramatical que mejora los resultados con respecto al modelo sin filtrado gramatical.

Estos resultados sobre textos que no están relacionados con la gramática inducen a pensar que si se adapta la gramática al estilo del texto se producirían beneficios.

4.1.3.4.2 Efecto del analizador en un texto adaptado a la gramática

Como se ha podido comprobar en los experimentos anteriores, el guiado gramatical que realiza el analizador cuando el texto no sigue el estilo de la gramática no produce tan buenos resultados como el modelo basado en *tripos* (aunque mejora con respecto al método sin información gramatical).

A continuación se realiza un experimento en el que se prueba con un texto más adaptado a la gramática. El texto, denominado TESIS2, es una versión modificada de las primeras 2000 palabras de TESIS. Los cambios han consistido en:

- Eliminación de referencias bibliográficas insertadas en las frases, ya que son cadenas no incluidas en el diccionario, que, por lo tanto, interrumpen en análisis.
- Eliminación de la palabra "etc." insertada dentro de una frase, ya que el algoritmo de segmentación de oraciones considera el punto como fin de frase, y el analizador vuelve a situarse en el comienzo de la análisis.
- Eliminación o reescritura de algunos incisos escritos entre paréntesis.
- Eliminación de números en las enumeraciones.
- Las frases excesivamente largas, con concatenaciones complejas, han sido segmentadas en oraciones de menor longitud. La longitud media de las frases de TESIS2 es de 18.56 palabras.

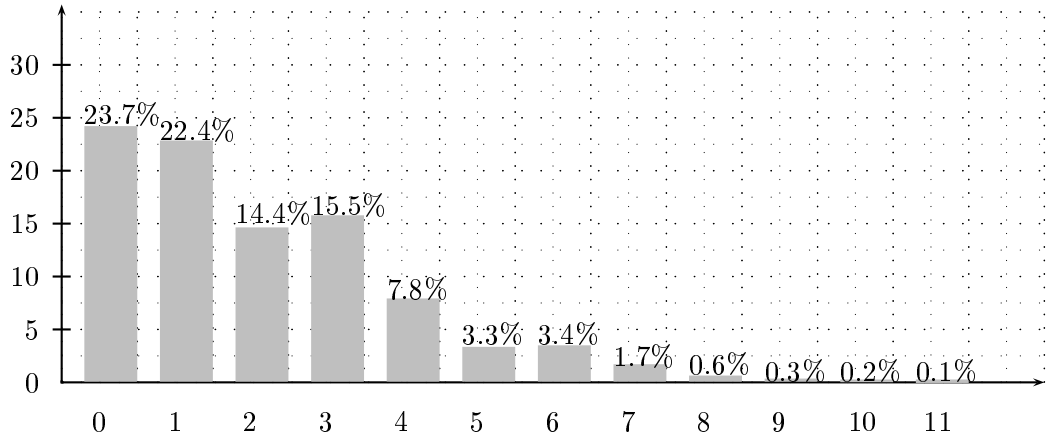
Para poder verificar los cambios que se han realizado, se pueden contrastar las dos versiones del texto en el capítulo B (página 187).

Las características del texto de prueba son:

Nombre del texto de prueba	TESIS2
Número de palabras	2058
Pulsaciones para escribirlo sin predicción	13957

Se toma como experimento de referencia el modelo basado en la información frecuencial del diccionario general (unigramas), cuyos resultados son:

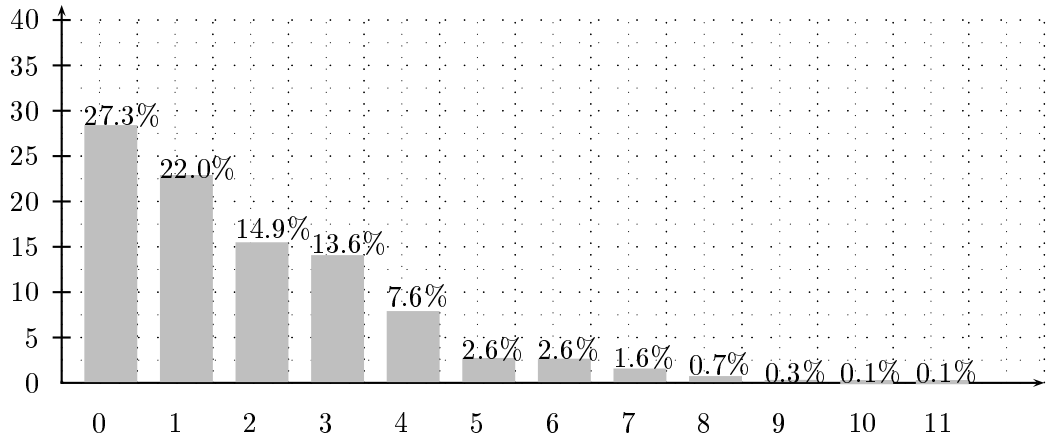
Número de palabras predichas	1924 (93.49%±1.07)
Número de pulsaciones ahorradas	5732 (41.07%±0.82)



Resultado 4.7: Experimento de referencia: predicción basada en la información frecuencial del diccionario general.

Si se añade el modelo basado en *tripos* con gestión de rasgos y retroceso a *bipos* y unigramas, se obtiene:

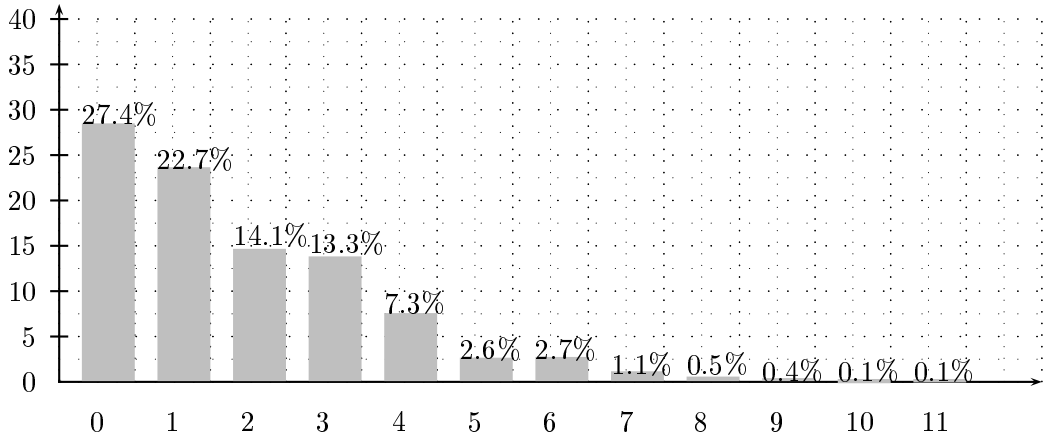
Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	1924 (93.49%±1.07)	1919 (93.25%±1.08)	-0.26%
Pulsacs. ahorradas	5732 (41.07%±0.82)	6127 (43.90%±0.82)	6.89%



Resultado 4.8: Experimento anterior añadiendo el modelo gramatical basado en *tripos* y retroceso a *bipos* y unigramas, con gestión de rasgos sobre el diccionario general.

Cuando se introduce el modelo basado en el analizador (con retroceso a *tripos*, *bipos* y unigramas) al ser aplicado a este texto, que está más adaptado a la gramática, se consiguen mejores resultados, como puede apreciarse a continuación:

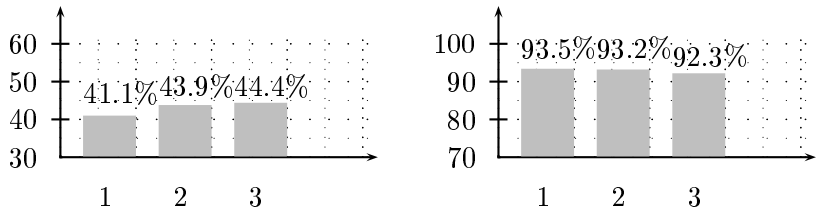
Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	1924 (93.49%±1.07)	1899 (92.27%±1.15)	-1.30%
Pulsacs. ahorradas	5732 (41.07%±0.82)	6193 (44.37%±0.82)	8.04%



Resultado 4.9: Experimento incluyendo el modelo basado en la gramática independiente del contexto (con retroceso a *tripos*, *bipos* y unigramas).

Se puede apreciar que se ha aumentado el ahorro en número de pulsaciones con respecto al modelo basado en *tripos*, y se ha mejorado el porcentaje de palabras predichas en las *letras cero y uno*, aunque por la longitud del texto utilizado las bandas se solapen, es decir, el analizador muestra una tendencia a mejorar el rendimiento proporcionado por los *tripos*, si la gramática está adaptada al estilo del texto que se está escribiendo.

Los resultados conjuntos de los experimentos de esta serie son:



Resultado 4.10: Evolución del porcentaje de pulsaciones ahorradas y palabras predichas en los experimentos de esta serie.

Donde:

1: Experimento de referencia, que utiliza solamente la información frecuencial del diccionario general.

2: Experimento anterior añadiendo *tripos* con gestión de rasgos, con retroceso a *bipos* y unigramas, sobre el diccionario general.

3: Experimento inicial con el filtrado gramatical basado en un modelo formal, con retroceso a *tripos*, *bipos* y unigramas, aplicado al diccionario general.

4.1.3.4.3 Efecto del retroceso al modelo basado en *tripos* cuando se interrumpe el análisis

A continuación se comprobará el efecto del retroceso a los modelos basados en *tripos*, *bipos* y unigramas, como mecanismo de recuperación cuando se interrumpe el análisis, ante una palabra desconocida o una frase cuya estructura no está contemplada en la gramática. Este método de gestión de errores ha sido utilizado en los experimentos anteriores relativos al analizador, y en este experimento queremos mostrar cual es su impacto, comparándolo con retroceso exclusivamente a unigramas.

El texto de prueba es el mismo que se utilizó en la serie anterior, y se toma como referencia el modelo basado en la información frecuencial del diccionario general (unigramas).

En el experimento cuyos resultados se muestran a continuación se utiliza predicción de palabras basada en la gramática independiente del contexto, con retroceso exclusivamente a unigramas cuando el análisis se interrumpe.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	1924 (93.49%±1.07)	1902 (92.42%±1.14)	-1.14%
Pulsacs. ahorradas	5732 (41.07%±0.82)	6022 (43.15%±0.82)	5.06%

Se puede observar que el modelo consigue una mejora con respecto a la predicción sin información gramatical, pero sus resultados son inferiores a los del modelo basado en *tripos* con retroceso a *bipos* y unigramas mostrado en la serie de experimentos anterior.

A continuación recordamos el resultado del último experimento de la serie anterior, basado en modelo gramatical en que cuando se interrumpe el análisis, se utiliza la predicción procedente del modelo basado en *tripos*, *bipos* y unigramas si es necesario.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	1924 (93.49%±1.07)	1899 (92.27%±1.15)	-1.30%
Pulsacs. ahorradas	5732 (41.07%±0.82)	6193 (44.37%±0.82)	8.04%

Como se puede observar, este mecanismo de recuperación de errores es más potente que el anterior, superando esta vez (con el texto adaptado a la gramática) los resultados obtenidos con los modelos basados en *tripos* con retroceso a *bipos* y unigramas.

4.1.3.5 Evaluación del diccionario personal y de los temáticos

En esta sección se evalúa la influencia de introducir en el proceso de predicción información sobre el tema y el estilo de escritura del usuario, generando, de forma simultánea al proceso de escritura, un diccionario sobre la sesión en curso y añadiendo la posibilidad de crear y utilizar diccionarios sobre los diferentes temas de interés del usuario.

4.1.3.5.1 Efecto del diccionario personal sobre textos de diferente longitud

Como se explicó en el apartado 3.2.2 (página 51) correspondiente al diccionario personal, éste se genera a la vez que se escribe el texto, por lo que está mejor adaptado al final

de la sesión de escritura.

Al realizar experimentos con textos de decenas de miles de palabras (como los utilizados para evaluar los métodos de predicción estadísticos), el diccionario personal consigue un grado de adaptación al estilo del texto muy alto, proporcionando, como se verá a continuación, unas tasas elevadas, tanto de pulsaciones ahorradas como de palabras predichas. Estos resultados pueden servir para evaluar el diccionario de forma teórica, pero no son representativos del grado de ayuda que le va a proporcionar al usuario, ya que los textos producidos en las sesiones de escritura habituales serán de varios cientos o pocos miles de palabras como máximo.

Por esta razón, en los experimentos de este apartado se han utilizando textos de diferente longitud: con los más pequeños se podrá apreciar el efecto del diccionario al principio de la sesión de escritura, mientras que los mayores mostrarán la potencia del método cuando alcanza un grado alto de adaptación al texto.

El sistema de predicción que se está evaluando en este apartado utiliza el diccionario personal, dando prioridad a los bigramas de palabras, y, si aún se necesitan más palabras, se aplican los métodos estadísticos categoriales (*tripos* y *bipos* y unigramas con gestión de rasgos) al diccionario personal, y, si la lista generada no es suficientemente larga, se añaden palabras procedentes del diccionario general, siguiendo los mecanismos habituales. Excepto en el experimento en el que se especifica lo contrario, las palabras nuevas son predichas desde la primera vez que aparecen en el texto.

En esta secuencia de experimentos se utilizan fragmentos de distinta longitud del texto TESIS, que se corresponden con el principio del texto, desde el origen hasta alcanzar el número de palabras indicado en cada uno.

Se toma como referencia en cada experimento los resultados del modelo basado en *tripos*, con retroceso a *bipos* y unigramas, con tratamiento de rasgos y categorización de palabras desconocidas, aplicados solamente al diccionario general.

Experimento 1

Nombre del texto de prueba	TESIS(100)
Número de palabras	100
Pulsaciones para escribirlo sin predicción	707

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	87 (87.00%±6.59)	86 (86.00%±6.80)	-1.15%
Pulsacs. ahorradas	273 (38.61%±3.59)	278 (39.32%±3.60)	1.83%

Resultado 5.1: Mejora producida por el diccionario personal sobre un texto de 100 palabras.

En este experimento se puede observar que en un texto de 100 palabras ya se consigue una mejora en el porcentaje de pulsaciones ahorradas, aunque el porcentaje de palabras predichas empeore ligeramente. Obviamente, con textos de este tamaño las bandas se solapan, por lo cual no se puede concluir que uno sea estrictamente mejor que otro, aunque

al repetir el experimento con diferentes textos de esta longitud se observa que el diccionario personal produce siempre una mejora (de distinta magnitud).

Experimento 2

Nombre del texto de prueba	TESIS(200)
Número de palabras	200
Pulsaciones para escribirlo sin predicción	1425

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	179 (89.50%±4.25)	178 (89.00%±4.34)	-0.56%
Pulsacs. ahorradas	576 (40.42%±2.55)	608 (42.67%±2.57)	5.56%

Resultado 5.2: Mejora producida por el diccionario personal sobre un texto de 200 palabras.

En un texto de 200 palabras puede observarse una mejora mayor en el porcentaje de palabras ahorradas. Además, el número de palabras predichas en la *letra cero* aumenta con respecto al experimento de referencia de un 22.5% hasta un 28.5%. Esto se debe, en gran medida, a la eficacia de los bigramas de palabras del diccionario personal.

En los experimentos siguientes se muestra el efecto de la adaptación del diccionario personal a medida que aumenta el tamaño del texto de prueba. Se puede observar, la mejora creciente en el porcentaje de pulsaciones ahorradas y de palabras predichas. Aunque no se muestren las gráficas correspondientes, en la *letra cero* se mejora en todos los casos al menos un 5%, y en la *letra uno* entre un 5% y un 12%.

Experimento 3

Nombre del texto de prueba	TESIS(500)
Número de palabras	500
Pulsaciones para escribirlo sin predicción	3483

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	443 (88.60%±2.78)	450 (90.00%±2.63)	1.58%
Pulsacs. ahorradas	1443 (41.43%±1.63)	1587 (45.56%±1.65)	9.98%

Resultado 5.3: Mejora producida por el diccionario personal sobre un texto de 500 palabras.

Experimento 4

Nombre del texto de prueba	TESIS(1000)
Número de palabras	1000
Pulsaciones para escribirlo sin predicción	6764

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	895 (89.50%±1.90)	914 (91.40%±1.74)	2.12%
Pulsacs. ahorradas	2793 (41.29%±1.17)	3135 (46.35%±1.19)	12.24%

Resultado 5.4: Mejora producida por el diccionario personal sobre un texto de 1000 palabras.

Experimento 5

Nombre del texto de prueba	TESIS(5000)
Número de palabras	5000
Pulsaciones para escribirlo sin predicción	34264

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	4459 (89.18%±0.86)	4613 (92.26%±0.74)	3.45%
Pulsacs. ahorradas	13773 (40.20%±0.52)	16956 (49.49%±0.53)	23.11%

Resultado 5.5: Mejora producida por el diccionario personal sobre un texto de 5000 palabras.

Experimento 6

Nombre del texto de prueba	TESIS(10000)
Número de palabras	10000
Pulsaciones para escribirlo sin predicción	68805

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	8867 (88.67%±0.62)	9232 (92.32%±0.52)	4.12%
Pulsacs. ahorradas	27311 (39.69%±0.36)	34784 (50.55%±0.37)	27.36%

Resultado 5.6: Mejora producida por el diccionario personal sobre un texto de 10000 palabras.

Experimento 7

Nombre del texto de prueba	TESIS
Número de palabras	37496
Pulsaciones para escribirlo sin predicción	253932

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33176 (88.48%±0.32)	34617 (92.32%±0.27)	4.34%
Pulsacs. ahorradas	98665 (38.85%±0.19)	130834 (51.52%±0.19)	32.61%

Resultado 5.7: Mejora producida por el diccionario personal sobre un texto de 37496 palabras.

Como se ha podido comprobar, desde el primer texto de prueba, con 100 palabras, el

uso del diccionario personal produce un aumento en el número de pulsaciones ahorradas, tendencia que se mantiene cuando se utilizan textos de mayor tamaño.

Con textos más grandes también crece el número de palabras predichas, especialmente porque las palabras nuevas sólo deben ser escritas la primera vez, y a partir de ese momento se predecirán, aunque no estén en el diccionario general. Además, el número de palabras acertadas en las *letras cero y uno* aumenta con respecto al experimento de referencia, especialmente por el efecto de los bigramas de palabras, que hacen que, para textos grandes, más del 50% de las palabras del texto se muestren al usuario habiéndose escrito cero o una letra.

En todos los parámetros se puede observar una mejora al aumentar la longitud del texto de prueba, ya que éste es también el texto del entrenamiento.

4.1.3.5.2 Efecto de la predicción o no de las palabras nuevas del diccionario personal

En estos experimentos se muestra el efecto de predecir o no las palabras nuevas[4] del diccionario personal. Como se explicó en la sección 3.2.2.5 (página 54), en los usuarios que tienen dificultades para escribir, estas palabras pueden corresponder con faltas de ortografía, por lo que se han incluido mecanismos de control que eviten su predicción, para que no se refuerce la escritura de palabras que no estén correctamente escritas.

El texto utilizado en esta secuencia de experimentos es TESIS:

Nombre del texto de prueba	TESIS
Número de palabras	37496
Pulsaciones para escribirlo sin predicción	253932

En el experimento de referencia se utiliza el modelo basado en *tripos* con retroceso a *bipos* y unigramas, con gestión de rasgos sobre el diccionario general.

A continuación se realizan experimentos añadiendo predicciones procedentes del diccionario personal, con el sistema de control que impide la predicción de las palabras nuevas.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33176 (88.48%±0.32)	32554 (86.82%±0.34)	-1.87%
Pulsacs. ahorradas	98665 (38.85%±0.19)	118305 (46.59%±0.19)	19.91%

Resultado 5.8: Experimento con el diccionario personal sin predecir ninguna palabra nueva.

Se puede observar que, aunque no se predigan las palabras nuevas, se produce un aumento en las pulsaciones ahorradas, debido a la adaptación de la información frecuencial al texto que se está escribiendo.

Si el usuario escribe más de una vez una palabra que no esté en el diccionario general, se puede considerar que está bien escrita, y predecirse, mejorando los resultados. Se

[4] No contenidas en el diccionario general.

puede configurar el número de veces que sea necesario escribir las palabras nuevas antes de incluirlas en el proceso de predicción. Los resultados obtenidos dependerán de este número, como se puede comprobar en los experimentos siguientes:

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33176 (88.48%±0.32)	33814 (90.18%±0.30)	1.92%
Pulsacs. ahorradas	98665 (38.85%±0.19)	126001 (49.62%±0.19)	27.71%

Resultado 5.9: Experimento con el diccionario personal, prediciendo las palabras cuando el usuario las escribe más de 5 veces.

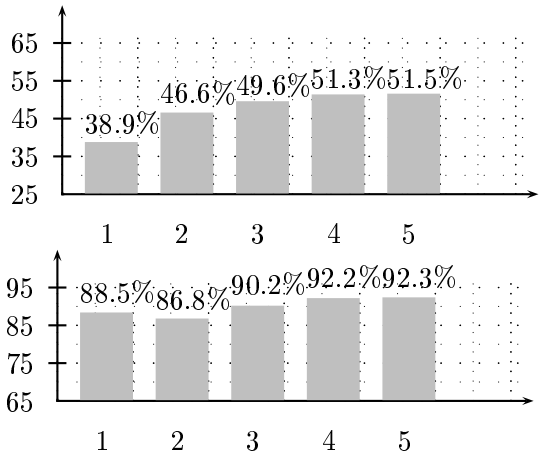
Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33176 (88.48%±0.32)	34581 (92.23%±0.27)	4.23%
Pulsacs. ahorradas	98665 (38.85%±0.19)	130300 (51.31%±0.19)	32.06%

Resultado 5.10: Experimento con el diccionario personal, prediciendo las palabras cuando el usuario las escribe 2 o más veces.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	33176 (88.48%±0.32)	34617 (92.32%±0.27)	4.34%
Pulsacs. ahorradas	98665 (38.85%±0.19)	130834 (51.52%±0.19)	32.61%

Resultado 5.11: Experimento prediciendo las palabras desde la primera vez que se escriben.

Como se puede observar, el rendimiento del sistema mejora si las palabras nuevas se predicen desde la primera vez que se escriben, aunque este método puede ser contraproducente para usuarios que cometan un excesivo número de faltas de ortografía.



Resultado 5.12: Evolución del porcentaje de pulsaciones ahorradas y palabras predichas en los experimentos de esta serie.

Donde:

1: Experimento de referencia, utilizando *tripos* sobre el diccionario general.

2: Experimento con el diccionario personal sin predecir nunca las palabras nuevas.

3: Experimento con el diccionario personal, prediciendo las palabras cuando el usuario las escribe más de 5 veces.

4: Experimento con el diccionario personal, prediciendo las palabras cuando el usuario las escribe 2 o más veces.

5: Experimento prediciendo las palabras desde la primera vez que se escriben.

En estas gráficas se muestra la evolución de los resultados al añadir el diccionario personal y permitir que se predigan las palabras nuevas, pudiéndose observar la mejoría cuando se predicen desde la primera vez que se escriben. En la evaluación de laboratorio se considerará siempre la predicción desde la primera vez que la palabra aparezca en el texto, aunque cuando el sistema sea utilizado para la escritura, debe ser configurado dependiendo de la capacidad de escritura de cada usuario concreto (cantidad de faltas de ortografía que cometa, etc.).

4.1.3.5.3 Evaluación del diccionario personal con textos pequeños

A continuación se muestra el comportamiento de la predicción utilizando el diccionario personal, para los mismos textos con los que se evaluó el diccionario general en el apartado 4.1.3.3.6 (página 138). Con cada uno de ellos se han realizado dos experimentos, uno utilizando el modelo basado en *tripos*, con retroceso a *bipos* y unigramas, y gestión de rasgos, que se ha tomado como referencia, y otro añadiendo el diccionario personal, para poder apreciar la mejora que este último produce.

A continuación se muestran los resultados obtenidos para cada uno de los textos.

Experimento 1

Nombre del texto de prueba	CUENTO1
Número de palabras	986
Pulsaciones para escribirlo sin predicción	6470

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	848 (86.00%±2.17)	859 (87.12%±2.09)	1.30%
Pulsacs. ahorradas	2121 (32.78%±1.14)	2380 (36.79%±1.18)	12.21%

Experimento 2

Nombre del texto de prueba	CUENTO2
Número de palabras	2672
Pulsaciones para escribirlo sin predicción	14878

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	2341 (87.61%±1.25)	2416 (90.42%±1.12)	3.20%
Pulsacs. ahorradas	5107 (34.33%±0.76)	6230 (41.87%±0.79)	21.99%

Experimento 3

Nombre del texto de prueba	CUENTO3
Número de palabras	1416
Pulsaciones para escribirlo sin predicción	8682

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	1231 (86.94%±1.76)	1265 (89.34%±1.61)	2.76%
Pulsacs. ahorradas	3134 (36.10%±1.01)	3474 (40.01%±1.03)	10.85%

Experimento 4

Nombre del texto de prueba	MOLINOS
Número de palabras	2980
Pulsaciones para escribirlo sin predicción	17715

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	2545 (85.40%±1.27)	2655 (89.09%±1.12)	4.32%
Pulsacs. ahorradas	5765 (32.54%±0.69)	6867 (38.76%±0.72)	19.12%

Experimento 5

Nombre del texto de prueba	QUIJOTE(2500)
Número de palabras	2446
Pulsaciones para escribirlo sin predicción	14660

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	2072 (84.71%±1.43)	2139 (87.45%±1.31)	3.23%
Pulsacs. ahorradas	4776 (32.58%±0.76)	5503 (37.54%±0.78)	15.22%

Experimento 6

Nombre del texto de prueba	FEB1K
Número de palabras	1286
Pulsaciones para escribirlo sin predicción	8536

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	1160 (90.20%±1.62)	1179 (91.68%±1.51)	1.64%
Pulsacs. ahorradas	3368 (39.46%±1.04)	3515 (41.18%±1.04)	4.36%

Experimento 7

Nombre del texto de prueba	TESIS(2000)
Número de palabras	2059
Pulsaciones para escribirlo sin predicción	14267

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	1879 (91.26%±1.22)	1920 (93.25%±1.08)	2.18%
Pulsacs. ahorradas	5923 (41.52%±0.81)	6904 (48.39%±0.82)	16.56%

Se puede apreciar que el uso del diccionario personal mejora los resultados del diccionario general en todos los experimentos, independientemente del ámbito de los textos. Como ocurría en la serie de experimentos del apartado 4.1.3.3.6 (página 138), los resultados concretos son diferentes en cada una de las pruebas, observándose que en los textos más largos la mayor adaptación del diccionario personal produce una mejora relativa mayor.

Esta tendencia queda alterada, por ejemplo, en el experimento con el texto FEB1K, en el que la mejora que produce el diccionario personal es apreciablemente menor que en los demás. Esto se debe a que este texto es de ámbito periodístico, por lo que el diccionario general ya está adaptado a su vocabulario y la aportación del diccionario personal se reduce a la información extra proporcionada por los bigramas de palabras.

Se puede concluir que el diccionario personal produce mejores resultados que el diccionario general en todos los experimentos, tanto en ahorro de pulsaciones como en porcentaje de palabras predichas.

4.1.3.5.4 Efecto del diccionario temático

Como se ha podido comprobar, el mejor rendimiento del diccionario personal se produce tras la escritura de la mayor cantidad posible de texto. Por ello este diccionario se almacena, de forma que pueda ser utilizado en sesiones posteriores en las que se escriba sobre el mismo tema, generándose *diccionarios temáticos*. En los siguientes experimentos se muestra el efecto del uso de los mismos.

Efecto sobre un texto del mismo dominio

En este experimento se muestra el resultado de utilizar un diccionario temático sobre un texto del mismo tema.

El texto de prueba utilizado en este experimento es:

Nombre del texto de prueba	QUIJOTE(8000)
Número de palabras	7813
Pulsaciones para escribirlo sin predicción	46010

En el experimento de referencia se utiliza el método de predicción basado en *tripos*, con retroceso a *bipos* y unigramas, y gestión de rasgos, aplicados sobre el diccionario general y el personal. A partir de esa configuración se utiliza, además, el diccionario temático generado a partir de QUIJOTE15-13K, que no contiene a QUIJOTE(8000), obteniéndose los siguiente resultados.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	6933 (88.74%±0.70)	7223 (92.45%±0.59)	4.18%
Pulsacs. ahorradas	17943 (39.00%±0.45)	19381 (42.12%±0.45)	8%

Resultado 5.13: Experimento anterior añadiendo un diccionario temático entrenado a partir de un texto de 13000 palabras sobre el mismo tema.

Se puede observar una mejora importante, tanto en el porcentaje de pulsaciones ahorradas, como en el número de palabras predichas. El número de palabras que se predicen en la *letra cero* también aumenta (desde un 28% hasta un 33.6%), fundamentalmente por el uso de los bigramas contenidos en el diccionario temático y la información frecuencial adaptada al tema, que se aplican desde el inicio de la sesión.

Lógicamente las ventajas del diccionario temático sobre el personal son mayores cuanto menor sea la longitud del texto de prueba (p. ej., para un texto de 200 palabras, donde el diccionario personal muestra una mejora relativa del 0.27%, el uso del diccionario temático incrementa esa mejora hasta casi el 34%).

Efecto sobre un texto de un dominio diferente

En este experimento se muestra la importancia de seleccionar el diccionario temático adecuado a la hora de escribir un texto sobre un tema concreto, ya que una elección

equivocada puede tener efectos contraproducentes. Para ello se realiza un experimento sobre un fragmento de "El Quijote", utilizando un diccionario temático generado a partir del texto TESIS.

El texto de prueba es QUIJOTE(8000). En el experimento de referencia se utiliza el método de predicción basado en *tripos*, con retroceso a *bipos* y unigramas, y gestión de rasgos, aplicado sobre los diccionarios general y personal.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	6933 (88.74%±0.70)	6923 (88.61%±0.70)	-0.14%
Pulsacs. ahorradas	17943 (39.00%±0.45)	16473 (35.80%±0.44)	-8.19%

Resultado 5.14: Resultados del experimento en el que se utiliza un diccionario temático entrenado a partir de un texto sobre un tema diferente al del texto de prueba.

Aunque los bigramas de palabras del diccionario temático sean un método de predicción muy potente, también son los más sensibles al grado de acuerdo (estilo, vocabulario utilizado, etc.) entre los textos de prueba y entrenamiento, pudiendo llegar, como se verifica en el experimento realizado, a empeorar significativamente los resultados cuando ambos textos no pertenecen al mismo dominio.

4.1.3.6 Efecto de los mecanismos auxiliares de aceleración

En esta serie de experimentos se muestra el efecto de los mecanismos auxiliares de aceleración de la escritura. Como se explicó en el apartado 3.3.4 (página 95), no son métodos de predicción en sí mismos, aunque también ahorran pulsaciones y/o aumentan el número de palabras predichas.

El texto utilizado en esta serie de experimentos es QUIJOTE(8000), y en el experimento de referencia se utiliza el método de predicción basado en *tripos*, con retroceso a *bipos* y unigramas, con gestión de rasgos, aplicado sobre los diccionarios general, personal y temático (generado a partir del texto QUIJOTE15-13K).

En el experimento siguiente se incorpora un mecanismo que automáticamente inserta un espacio tras los signos de puntuación, y, si el signo es un punto, escribe en mayúsculas la siguiente letra, obteniéndose como resultado:

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	7223 (92.45%±0.59)	7223 (92.45%±0.59)	0.00%
Pulsacs. ahorradas	19381 (42.12%±0.45)	20773 (45.15%±0.45)	7.18%

Resultado 6.1: Experimento anterior añadiendo la escritura automática de espacios tras los signos y mayúscula después de punto.

Como se puede observar, evidentemente, este mecanismo no afecta al número de palabras predichas, pero sí aumenta el número de pulsaciones ahorradas.

En el siguiente experimento se añaden a la lista de palabras predichas 3 terminaciones que comiencen con la última letra escrita.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	7223 (92.45%±0.59)	7333 (93.86%±0.53)	1.52%
Pulsacs. ahorradas	19381 (42.12%±0.45)	21106 (45.87%±0.46)	8.90%

Resultado 6.2: Experimento anterior añadiendo predicción de terminaciones.

Este mecanismo sí produce un aumento en el número de palabras predichas (realmente terminaciones predichas), además de en el de pulsaciones ahorradas, ya que la predicción de terminaciones ayuda a escribir palabras que no están incluidas en ninguno de los diccionarios (las palabras nuevas, la primera vez que se escriben), ahorrando una media de 3 o 4 pulsaciones cada vez que una de ellas es válida.

A continuación se añade al módulo de coordinación el filtro que rechaza las palabras que ya hayan sido mostradas al usuario. En este experimento se muestra cada palabra dos veces antes de rechazarla.

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	7223 (92.45%±0.59)	7352 (94.10%±0.52)	1.79%
Pulsacs. ahorradas	19381 (42.12%±0.45)	21249 (46.18%±0.46)	9.64%

Resultado 6.3: Experimento anterior rechazando las palabras cuando hayan sido mostradas dos veces al usuario.

Se puede observar que este filtro produce una mejora, tanto en el porcentaje de pulsaciones ahorradas, porque las palabras se muestran antes al usuario, como en el número de palabras predichas, ya que al necesitarse menos pulsaciones para que la palabra se prediga, se muestran algunas palabras que sin este filtro acabarían de escribirse antes de ser predichas.

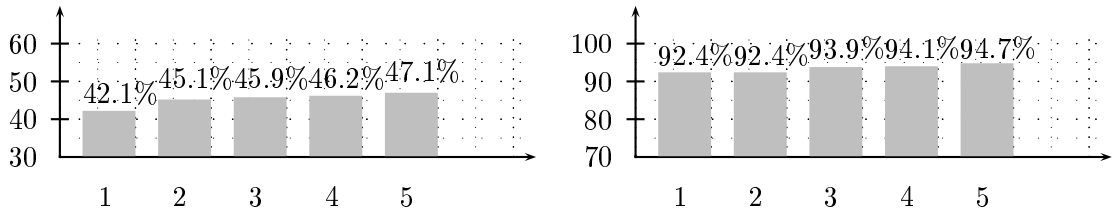
Se obtienen mejores resultados con la configuración del experimento siguiente, en el que solamente se muestra cada palabra una vez al usuario:

Medida	Referencia	Resultado	Mejora relativa
Palabras predichas	7223 (92.45%±0.59)	7401 (94.73%±0.50)	2.46%
Pulsacs. ahorradas	19381 (42.12%±0.45)	21665 (47.09%±0.46)	11.78%

Resultado 6.4: Experimento anterior rechazando las palabras cuando hayan sido mostradas una sola vez al usuario.

Los resultados, como se puede comprobar, son mejores, mostrándole únicamente una vez cada palabra al usuario. Sin embargo, como se indica en el apartado 3.3.4.1 (página 95) se recomienda no utilizarlo como configuración del sistema, para evitar la penalización que se produce si el usuario no lee la palabra en el momento que aparece.

A continuación se muestran los resultados conjuntos, en los que se aprecia el efecto de añadir los diversos mecanismos auxiliares a la predicción.



Resultado 6.5: Evolución del porcentaje de pulsaciones ahorradas y palabras predichas en los experimentos de esta serie.

Donde:

1: Experimento de referencia: predicción del método basado en *tripos*, aplicado a los diccionarios general, personal y temático.

2: Experimento anterior añadiendo la escritura automática de espacios tras los signos y mayúscula después de punto.

3: Experimento anterior añadiendo predicción de terminaciones.

4: Experimento anterior rechazando las palabras cuando hayan sido mostradas dos veces al usuario.

5: Experimento anterior rechazando las palabras cuando hayan sido mostradas una sola vez al usuario.

4.2 Evaluación subjetiva

En el presente apartado se muestran los resultados obtenidos al someter el sistema *Predice* (que incorpora predicción de palabras), a evaluación por parte de personas con discapacidad, para obtener datos sobre la influencia de la predicción en la cantidad y calidad de los textos que generan, así como su valoración subjetiva.

No se dispone de datos que nos permitan realizar una valoración de la predicción de palabras en sí misma, sino de dicha predicción incorporada en el sistema *Predice* completo, porque se deseaba realizar la evaluación con usuarios con discapacidades físicas, para verificar si la predicción producía los efectos esperados. Estas personas necesitan que los sistemas dispongan de una interfaz que pueda ser manejada con un pulsador, como la del *Predice* (descrita en el apartado 3.5, en la página 99), que es el que se ha utilizado en el proceso.

No se conoce ningún estándar a nivel nacional ni internacional que establezca ninguna métrica que permita "contabilizar" el efecto de la predicción sobre la calidad de los textos generados, por lo que se presentarán informes elaborados a partir de los que nos han proporcionado los expertos que trabajan directamente con los usuarios del sistema.

4.2.1 Evaluación a corto plazo

En este apartado se muestran las dificultades, los resultados y las impresiones de un usuario en la primera sesión de uso de *Predice*.

El usuario elegido es Javier, afectado de tetraplejia debida a un accidente de tráfico que le impide mover las manos con normalidad. Su capacidad intelectual no se ha visto afectada, y tiene capacidad lecto-escritora de alto nivel. Está habituado al uso del ordenador, que normalmente utiliza con el teclado convencional, pulsando las teclas con los nudillos, ya que conserva movimiento normal de los brazos.

La primera parte de la sesión se dedicó a que Javier se familiarizara con el sistema, explicándole el significado de las opciones, el funcionamiento del sistema de barrido, etc. Posteriormente escribió varios textos cortos con diferentes configuraciones, hasta encontrar la más adecuada: barrido por filas y columnas, con un tiempo de espera en cada opción de 0.7 segundos. Utilizó como pulsador el botón izquierdo del ratón.

A continuación comenzó la sesión propiamente dicha, en la que se le sugirió que escribiera un texto sobre un tema de su interés, para que fuera representativo de su estilo de escritura. Eligió un cuento sobre animales, por lo cual el vocabulario no estaba especialmente adaptado al contenido del diccionario general, ni se disponía de un texto previo sobre el mismo tema a partir del cual generar un diccionario temático que se pudiera utilizar.

La escritura fue dividida en varias partes, de una duración aproximada de media hora, entre las que paraba para descansar. Estos descansos eran necesarios, ya que la primera vez que utiliza el sistema hay una serie de factores a los que hay que adaptarse, y que requieren un esfuerzo cognitivo mayor que el de la escritura convencional:

- El barrido: hasta que el usuario se adapta al sistema, comete errores por imprecisión en el uso del pulsador.
- Las posiciones de las letras: las letras no están situadas en el mismo orden que en el teclado convencional, sino situadas de forma que se minimice el número medio de opciones que se barren para escribir un texto (para reducir el tiempo necesario para escribirlo). Esto hace que al principio sea necesario buscar cada letra antes de escribirla, hasta que se memoriza la posición de cada una.
- El uso de las palabras y sufijos predichos: antes de escribir cada nueva letra, es necesario realizar un esfuerzo adicional para comprobar si la palabra o sufijo deseado está en el menú.

Cabe destacar que el usuario realizó un gran esfuerzo por utilizar todas las opciones disponibles en el sistema (comprobando antes de escribir cada letra si la palabra que buscaba estaba en la lista de palabras predichas, etc.), y que el texto generado fue de una calidad perfecta, ya que corrigió todos los errores debidos al barrido (muy comunes en las primeras sesiones), y escribió todos los acentos y mayúsculas presentes en el texto[5].

Las tres primeras partes de la escritura duraron entre 26 y 33 minutos, y en cada una de ellas se escribió un párrafo que contenía entre 60 y 70 palabras.

[5]Es importante resaltar la calidad final del texto, por la gran cantidad de tiempo extra que requiere corregir los errores y añadir los acentos y las mayúsculas. Parte de los usuarios de sistemas controlados por barrido optan por escribir textos de forma más rápida, especialmente en la comunicación, sabiendo que, aunque el texto sea de menor calidad, será inteligible.

A continuación, Javier copió los mismos párrafos, esta vez con el diccionario personal ya entrenado, reduciéndose el tiempo necesario para escribirlos a su tercera parte, entre 6 y 10 minutos cada uno, por los beneficios que aporta el diccionario adaptado.

Después, para tomar una referencia de tiempos, Javier escribió con el teclado convencional las primeras 90 palabras del mismo texto, con la misma calidad, tardando 4 minutos. Una persona con conocimientos de mecanografía, lo escribe en 2 minutos. Es decir, el acceso directo, para los usuarios que pueden utilizarlo, es el método más rápido.

Se repitió el experimento, utilizando acceso directo y predicción de palabras, ralentizándose ligeramente la escritura, ya que el uso del teclado convencional, para una persona que esté habituada, es un proceso automático, y, al introducir una interrupción entre cada una de las letras para leer la lista de palabras predichas se retrasa la escritura.

Por último, Javier escribió uno de los párrafos anteriores sin utilizar predicción de palabras, tardando aproximadamente media hora (lo mismo que la primera vez que lo escribió, cuando el diccionario personal aún no estaba adaptado).

Sus impresiones sobre la primera sesión utilizando *Predice* son:

- Al principio, exige mucha atención, cuando es necesario adaptarse al barrido, leer las palabras predichas y buscar las letras en la matriz, lo que hace necesario realizar paradas periódicamente.
- Al avanzar la sesión, el esfuerzo requerido disminuye, al adaptarse al barrido y haber memorizado las letras más frecuentes.
- Cuando escribió el texto utilizando el barrido, sin uso de la predicción, el proceso de escritura era más relajado, incluso aburrido.
- Considera útil la predicción, para usuarios de pulsador.

Además, se han podido comprobar los beneficios de la predicción, acelerando la escritura si el diccionario personal está adaptado, cuando se utiliza barrido automático controlado por pulsador. Si el usuario puede utilizar acceso directo, ése será el método más rápido.

En sesiones sucesivas, los usuarios consiguen configuraciones que les permiten aumentar la velocidad de escritura, ya que, al acostumbrarse a la predicción, memorizar la posición de las letras y aumentar su destreza con el barrido, cometerán menos errores y podrán aumentar la velocidad de barrido de las opciones.

4.2.2 Evaluación a largo plazo

A continuación se presentan varios informes sobre el uso del programa *Predice*, (uno de los sistemas en los que está integrada la predicción de palabras), por varios usuarios del Centro *Maset de Frater (Centro de Promoción Personal e Integración Social de Paralíticos Cerebrales Adultos y Minusválidos Físicos Gravemente Afectados)*. Dichos informes han sido realizados por David Carreres Lacasa (psicólogo), basándose en los trabajos dirigidos por el logopeda del centro Mario Brancal Boldori, con la ayuda de Alicia Cucala (psicóloga), en sesiones que han tenido lugar dos veces a la semana, durante varios años.

En el Centro *Maset de Frater*, son 4 las personas que en este momento utilizan *Predice*. Como características generales, los cuatro usuarios:

- Son personas adultas, gravemente afectadas por Parálisis Cerebral, con un grado de minusvalía que supera en todos los casos el 85%. Están confinados en silla de ruedas y necesitan ayuda de terceras personas para realizar las actividades de la vida diaria.
- Presentan trastornos severos del lenguaje (disartria y anartria[6]), que imposibilitan un habla funcional e inteligible. Según la clasificación de Kluin[7], 3 de los usuarios (*Usuarios 1, 2 y 3*) se encuentran en el grado 4, con habla ininteligible, y el otro (*Usuario 4*) en el grado 3, con disartria moderada.
- Utilizan el pulsador como único modo de acceso al ordenador, porque su afectación motriz les impide acceder a cualquier otro tipo de soporte: tienen deficiente control cefálico, imposibilidad de controlar un ratón u otro dispositivo similar y movimientos espásticos continuos.

Con respecto a las características particulares de cada uno, el *Usuario 1* accede a la interfaz con un conmutador de presión, acoplado a su mano izquierda, con el que controla el barrido automático. El *Usuario 2* utiliza el mismo sistema, pero el conmutador es de pie (muy similar al pedal de las máquinas de coser), acoplado a su pie derecho. El *Usuario 1 y el 2* pueden leer y escribir, y *Predice*, además de herramienta de rehabilitación, es el editor de textos con el que realizan todas sus tareas escolares. Aunque ambos poseen el Graduado Escolar, su nivel académico estaría en torno al sexto o séptimo curso de la antigua EGB. Cursan todas las asignaturas de la EPA (Educación Permanente de Adultos: matemáticas, valenciano, inglés, sociales, naturales y lengua), ayudándose del *Predice*.

El *Usuario 4* es un caso similar, aunque su capacidad lecto-escritora es mucho mejor y no tiene problemas en la correspondencia fonema/grafema. Tiene 31 años y esta cursando en un proyecto de integración (más social que académica), 5º de Administrativo de la Rama de Hostelería en un Centro Público, correspondiente a 2º de 2º grado de la antigua

[6]Llamamos *disartria* a las dificultades de la expresión del lenguaje por trastornos del tono y del movimiento de los músculos fonadores debidos a lesiones del sistema nervioso. Etimológicamente, significa dificultad o trastorno de la articulación de la palabra, pero generalmente se entiende como un trastorno de la articulación de los fonemas, como consecuencia de lesiones de las zonas del sistema nervioso que gobiernan los músculos de los órganos fonadores, provocando alteraciones en su tono, movimientos anormales involuntarios o interrupciones de la orden cerebral que dan como resultado una alteración del habla, que, en sus casos más extremos, puede desembocar en la imposibilidad de articular correctamente ningún fonema (*anartria*).

[7]Según Kluin, los grados de severidad de la disartria dependiendo de la inteligibilidad conversacional del habla son:

- Grado 1. Normal: sin desorden motor del habla.
- Grado 2. Leve: habla inteligible aunque su sonido no es normal.
- Grado 3. Moderada: la producción de palabras o expresiones cortas es usualmente inteligible, sin embargo, esta inteligibilidad disminuye cuando se incrementa la longitud de las frases y, frecuentemente, es necesario repetirlas.
- Grado 4. Severa: el habla es ininteligible.

Formación Profesional (último curso antes de extinguirse definitivamente la Formación Profesional). Todas las tareas y trabajos los hace en el Centro (es residente) con *Predice* y asiste a clase en calidad de oyente.

Por último, el *Usuario 4*, que tiene 20 años y utiliza *Predice* para construir palabras sencillas. Al no poder escribir de ninguna otra manera, *Predice* es *su lápiz*. Este alumno sufre un leve-moderado retraso mental y, según nuestra experiencia, no completará el proceso lecto-escritor. Sin embargo, *Predice* le da la oportunidad de escribir palabras sencillas, nombres, números, etc., lo que supone un avance importante, ya que compensa de alguna manera su bajo nivel cognitivo con aprendizajes significativos, muy útiles en la vida cotidiana.

Predice se ha revelado como un procesador de textos apto, también, para personas con problemas motrices asociados a retraso mental no demasiado importante, por la sencillez de manejo de las opciones básicas, que no requieren demasiada destreza.

A continuación se presentan dos informes más detallados con la evolución a largo plazo de la calidad de la escritura de dos de estos usuarios: el *Usuario 1 y el 2*. El trabajo realizado con ellos en el centro ha ido encaminado a la mejora de la producción escrita, que estaba claramente deteriorada, como se puede apreciar en las palabras y frases del principio.

La herramienta utilizada para este trabajo rehabilitador y que ha proporcionado un soporte básico para otros aprendizajes, ha sido *Predice*. En el trabajo puramente rehabilitador, *Predice* se utiliza eliminando la predicción de palabras para que no interfiera, y, así, obtener una imagen fiel de la transcripción fonema-grafema de los usuarios. Sin embargo, en todas las demás actividades escolares, incluyendo las redacciones y los mensajes espontáneos, sí se les permite utilizar la predicción, y, además, se entiende que es aconsejable, ya que les aporta una realimentación importante.

Como estos informes son resultado de un estudio a largo plazo, la predicción de palabras utilizada no incluye los últimos métodos introducidos en el sistema (ni el analizador ni el diccionario personal).

4.2.2.1 Informe 1

El *Usuario 1* tiene 33 años. No empezó una escolarización sistematizada hasta los 25 años, si bien sí había recibido apoyo individualizado desde los 12 a los 22 años en su propio domicilio. En su historial escolar se reseña que aprendió a leer a los 14 años y a escribir hacia los 19. Se puede afirmar que posee un nivel cognitivo muy próximo a la "normalidad" (si bien no se puede valorar con exactitud), aunque haya existido una falta de estimulación importante en su infancia.

Cuando acudió al Centro, a los 25 años, su escritura era apenas inteligible, como se puede apreciar en los ejemplos mostrados a continuación. Estos ejemplos se corresponden con las respuestas que escribía como contestación a adivinanzas que se le planteaban, cuya solución, con seguridad, conocía:

Respuesta correcta	Respuesta escrita
Pablo	Palbo
mueble	muebel
diablo	llador
plátano	palno
jabón	jabo
árbol	albro
jirafa	jifara
pájaro	pajo
planeta	paneta
blando	brado

Posteriormente, utilizando ya *Predice*, se comenzó con la escritura de párrafos, pidiéndole que describiera imágenes.

Contenido de la imagen	Descripción del *Usuario 1*
Un vendedor está en una tienda empaquetando un regalo.	Un hombre estar un tiena mirador una caja.
El vendedor está terminando y pone una cinta con un lazo en la caja.	El señor estar en bala la caja con una cita hacero un lazo.
El vendedor entrega la caja para regalo a una señora.	Una señar ba tiena a compra una caja.

Posteriormente, su escritura comenzó a ser medianamente inteligible, como se puede apreciar en el siguiente párrafo, de Junio del 97, donde construye una redacción utilizando el lenguaje espontáneo.

Viaje a Torrevieje

El viernes cuando vamos a hotel cenamos la a mesa con Andres, Frenando y cuatro mujeres no se sus nombres. Despues de cena llame a mi casa, fuemos a discoteca y bailaba con Marimar, Antonia, Majura y Carmen. Fui a durmir a las dos la noche. El sabado baje las nueve de la mañana, tome un zumo de narrajas y un bocata, despues fuimos a al vuelta para ven en pueblo un grupo. Fuimos a un bar. Yo hablavia una chica pero nada de nada, fuimos a al hotel a comer, por tarde fuimos una vuelta en barco para conocer la costa blanca me gusto, a la cena la mesa con bordes Juan, Sergio y Cosme Jose y luego fuimos a ver el partido entre Barcelona y Real Madrid por television yo soy un borde tome el pelo a Juan y Sergio toda noche porque fuimos a dos discotecas.

A continuación se transcriben textos (lenguaje espontáneo) muy recientes. Las faltas ortográficas y gramaticales continúan, sin embargo, la progresión ha sido constante a lo largo de estos años. Redacción sobre la visita a un museo:

La prima visita al museo "espacio del arte"

Un grupo fuimos ha ver una exposición sobre Movimiento Aparente. El guia se explicados cada un Arterias, primamente habían unos fotos de una Ciudad esa tinencia era como hacer el cine, 24 Por Sungo fotográfias . Después vemos ha ver una alfombra como vesta por el sol, más tadre vemos una película era anteartida tendedia vemos finales. Habían unas fotografías con bombillas y porque hacer el arteria quiera empresada que las fotografías no decir la verdad, había una foto un hombre que empresada una gran pena peor no había perdó

su equipo y entonces saque una cusion que las imágenes no decir la verdad, después el guia dijo a Tino que entrada a cueva de sueños no había nada porque los sueños son nada. Beatriz e yo entramos a nuestra niñeces: Mi opinión está exposición es muy rara. El guia es muy importante.

Ahora, un ejercicio en el área de experiencias:

"Los movimientos son dos. Rotacion girar sobre si misma una vuelta tarda 24 horas por ello dia y noche, traslación es otro movimiento la tierra da vueltas sol tarda una vuelta un año El movimiento de rotación es dia la noche. El movimiento de traslación es que pasa las estaciones y los años."

4.2.2.2 Informe 2

El *Usuario 2* nació el 23/04/74. Es un joven de características similares a las del *Usuario 1*, aunque fue escolarizado a los 8 años, por todo lo que presenta un mejor bagaje. Aprendió a leer a los 10 años y a escribir a los 15, según consta en su expediente. El trabajo realizado con el *Usuario 2* es igual que el del *Usuario 1* (correspondencia fonema/grafema).

A continuación se incluyen dos textos espontáneos del *Usuario 2* (escritos con *Predice*) de este mismo curso. Se puede observar que la calidad de sus textos es superior a la de los del *Usuario 1*.

"Mi aparato de música va cuando quiere. El domingo estaba grabando la cinta y cuando fue mi hermana a darle a play del cd para que empezara otra vez el cd dijo el aparato que ya no iba mas, no pude acabar de grabar tu cinta ni pude grabar la mia. El domingo me levante una ora y media antes para grabate tu cinta. Como era para ti no me sabe nada mal. Me llamo mi madre para decime que nos íbamos al pueblo porque yo le dije que me dijera algo."

"El viernes celebro el cumple mi hermana y fueron algunos primos a cenar a casa y como es aditual en todos los cumpleaños entre todos los perpara una caja grande y meten de todo. A mi hermana le metieron en la caja un muñeco con la cabeza abrierta y rellena de sesos y en otro paquete abia un corazón. Ya bes que animales son mis primos. Ya conoci al novio de mi hermana. Una pregunta puedo embitar a mi chica al cine el domingo que bien".

Y, por último, una carta escrita en noviembre de 2000.

18-11-00

hola, soy un chico que usa tus editores.

Antes de utilizar el predice he gastado tres programa mas. El primero era muy sencillo porque solo se podia escribir. El segundo ya tenia la función de imprimir y nada mas.

El tercero había dos versiones, la primera me dejaba retroceder el cursor sin borrar y la segunda era un poco mas completa que la anterior porque llevaba las hociones como la del diccionario pero si lo queria usar tenia que marca un símbolo para que se active y cuando selañaba la palabra que quería se desactivaba automáticamente y cada vez que lo necesitaba tenía pegarle al símbolo. Había otra oncion que era la de sentintador de voz pero lo use muy poco porque no iba bien.

El sistema de grabacion de los textos fue pesimismo. Ni primero y ni el segundo no se podia guardar los trabajos. En el tercero la forma de achibación era por letras, te lo explico. Cada letras del abecedario eran como pequeños ficheros que se podia guardar unas diez líneas en cada letra, cuando estaban todas llenas tenia que borrar los trabajos para grabar los textos nuevos. Era un jaleo porque me tocaba acordar en que letra estaba vacía para guardar la faena y muchas veees me cargaba algún escrito porque metia encima de otro trabajo.

Menos en el tercer programa y en la última versión que la manejaba con un mando con la cabeza y los otros editores los llevaba con dos comuntadores poro me cansaba bastante cuando éstaba trabajando un rato.

Cuando empece a usar el predice prove a pegar al mando con el pie y me va mucho mejor porque no me canso tan pronto.

Desde conoci este programa escrivo mas rápido que con los otros editores.

Lo que me gusta de el es que se puede crear ficheros, tiene el predilector a la pantalla principal, se puede retroceder el cursor sin borrar,ectr.

El último programa que has creado aun no he podido provar todas las funciones que tiene pero si me leido las posibilidades que hay y me a parecido muy interesantes.

Cuando prueve todas fundiciones ya te escribire otra carta y te contare como me a parecido el wuin.

Ya me despido atentamente.

Se puede apreciar, y la opinión de los expertos lo corrobora, que el uso de *Predice* beneficia a los usuarios a largo plazo, siendo una herramienta muy positiva en el apoyo a la lecto-escritura. Para algunos usuarios, como se ha mostrado, es la única posibilidad de acceder a la escritura.

El uso de la predicción de palabras aumenta la calidad de su escritura, aunque, como se ha visto en el informe correspondiente, su uso en el proceso de educación debe ser supervisado por el profesor, para que no interfiera negativamente (por ejemplo, si el usuario sabe que una determinada palabra aparece en la lista de palabras predichas, puede seleccionarla siempre, sin fijarse en su escritura exacta).

Se puede concluir que el uso del sistema, debidamente controlado en las etapas educativas, es positivo a largo plazo, produciendo un aumento en la calidad de los textos generados, y una disminución del esfuerzo, tanto físico como cognitivo, necesario para producirlos.

Los usuarios manifiestan que el uso del sistema les aporta rapidez y les cansa menos que otras aplicaciones similares, y utilizan la predicción de palabras cuando se les permite (por ejemplo, al escribir texto espontáneo, realizar tareas escolares, etc.,), de lo cual se puede deducir que valoran positivamente el apoyo que les proporciona.

4.3 Comparación con otros sistemas

La comparación con otros sistemas es la forma de situar cada herramienta y su rendimiento con respecto a las demás que componen el estado de la cuestión. Cuando se trata de predicción de palabras, esta comparación es particularmente difícil por una serie de razones que se exponen a continuación, que algunas veces pueden llegar a hacerla imposible.

El primer problema es la *falta de estandarización en las condiciones de prueba*, que provoca que los resultados disponibles acerca de cada programa hayan sido obtenidos en diferentes condiciones: textos de prueba y entrenamiento con características muy diferentes, distintos valores de configuración de la predicción (número de palabras predichas, etc.). En estas condiciones, la comparación no es válida, porque, como se explica en secciones anteriores, la variación en alguno de los parámetros puede modificar los resultados considerablemente.

Además, no siempre se dispone de la misma *medida* en todos los sistemas: unos proporcionan el ahorro de pulsaciones, mientras que en otros se conoce la velocidad de aprendizaje, imposibilitando cualquier comparación.

En la mayoría de los casos, la información de la que se dispone es la evaluación de la predicción *integrada en una interfaz determinada*, y las medidas obtenidas están fuertemente influenciadas por las características de dicha interfaz. Por ejemplo, si la aplicación funciona con barrido controlado por pulsador, las medidas de tiempo son diferentes a las que se obtienen en las aplicaciones que se utilizan con el teclado convencional y el número de pulsaciones que se ahorran por cada carácter se contabilizará de diferente manera. Además, entre las características de la interfaz, pueden estar incluidas o no otras técnicas de ahorro de pulsaciones como la adición de espacios tras los signos, etc. que también influyen en los resultados.

Por otra parte, es muy difícil conseguir resultados de predicción de palabras en *castellano*, ya que la mayor parte de los sistemas son para otros idiomas. Las diferencias intrínsecas entre los distintos idiomas pueden hacer que un mismo método de predicción produzca resultados distintos en ellos, incluso que algunos métodos aplicados para un idioma particular no tengan ningún sentido en otros. Algunos ejemplos de dichas diferencias que pueden producir variaciones significativas en los resultados de la predicción son:

- Diferente comportamiento de los verbos en inglés y en castellano: los verbos en inglés tienen entre tres y cinco formas diferentes, mientras que en castellano cada verbo (tanto los regulares como los irregulares) tiene hasta cincuenta y tres. Este elevado número de formas dificulta en gran medida la predicción de los verbos en castellano, produciendo un gran incremento en el tamaño del diccionario, tiempo de procesamiento y complejidad de los algoritmos de predicción, dependiendo del método elegido para gestionarlo.

- Diferente procedimiento de generación de palabras en alemán y castellano. En alemán, uno de los mecanismos habituales de generación de palabras nuevas es la concatenación. Alguna de las consecuencias de este método con respecto a las características de las palabras generadas y su influencia en la predicción de palabras se muestran a continuación:

  - Se generan palabras mucho más largas, por tanto, cuando se predicen, se ahorran muchas más pulsaciones.
  - Las palabras compuestas pueden ser predichas componente a componente, técnica imposible de utilizar para otros idiomas.
  - Aumenta la complejidad de la etapa de entrenamiento, ya que es necesario un proceso de segmentación de las palabras en fragmentos para determinar qué componentes se incluyen en el diccionario. Si no es posible la segmentación, se consideran las palabras compuestas como un todo, incrementándose el número de palabras diferentes del texto de prueba, del diccionario y la probabilidad de que las palabras no estén incluidas en él.

- Otras características específicas del lenguaje (que no están presentes en otros idiomas), y que pueden ayudar en el proceso de predicción, como, por ejemplo, la concordancia gramatical de género entre el artículo y el nombre en castellano, que no está presente en inglés.

A continuación se muestran los resultados de otros predictores de palabras, especificando, si están disponibles, las condiciones de prueba. Por las razones anteriormente mencionadas, no es posible realizar una comparación rigurosa de los resultados de la predicción de palabra de esta tesis y los de otros modelos, ya que, aunque se conocen resultados de la evaluación de otros predictores, no se especifican todas las condiciones de prueba asociadas a los experimentos con los que obtienen dichos resultados, y, además, la mayoría no se evalúa para castellano.

En [Pala98b] se publican los resultados de un experimento en el que se comparaban dos sistemas, *Predice* y *PredictAbility*, descrito en [Clay96], desarrollado en el departamento Applied Computing de la Universidad de Dundee. La comparación se realizó para cinco idiomas de la comunidad europea, inglés, francés, italiano, alemán y español. Se disponía de dos textos, traducidos a los cinco idiomas, que se utilizaron en el entrenamiento de los diccionarios y en las pruebas. El resultado fue que los resultados de ambos sistemas prácticamente coincidían, y estaban comprendidos entre el 30% y el 37% de pulsaciones ahorradas, dependiendo del idioma.

En [Carl97a] se presenta *Profet*, un sistema de predicción de palabras para sueco, que utiliza un modelo de predicción basado en una combinación de *tripos, bipos y unipos*, y que proporciona un resultado del 46% de pulsaciones ahorradas. En [Boek96] se presenta un predictor para holandés, que utiliza un algoritmo que combina pares y tríos de palabras, con apoyo de pares y tríos de categorías gramaticales, proporcionando unos resultados de hasta un 44% de ahorro en número de pulsaciones.

En [Stew96] se presenta un estudio de Higginbotham en el que se estima que el porcentaje de pulsaciones que ahorra un sistema de predicción en media se encuentra entre 37-45%.

Capítulo 5

Conclusiones

Esta tesis doctoral aborda el estudio de la inclusión de información lingüística en la predicción de palabras con el objetivo de mejorar los sistemas de ayuda a la escritura de personas que pueden tener distintos tipos de discapacidad. Sentado este objetivo, debemos ser conscientes de que las particularidades de esta tarea implican la necesidad de su análisis específico, no contemplado en otros estudios más generales de procesado de lenguaje natural. Otro valor añadido es el tratamiento específico del castellano como lenguaje objetivo, ya que la mayor parte de la literatura disponible sobre esta problemática se refiere a otros idiomas.

Para la inclusión de la información lingüística se propone una arquitectura novedosa que permite desarrollar una metodología original de combinación de las diversas fuentes de información lingüística exploradas, que en el presente trabajo ha demostrado ser óptima. De entre las posibles fuentes de información lingüística, hemos seleccionado un conjunto de ellas, fundamentalmente en los niveles léxicos, morfológicos y sintácticos. En cada uno de los módulos que realizan la inclusión del conocimiento lingüístico en la arquitectura se han realizado también aportaciones específicas, tanto en el diseño y organización de la información, como en la metodología particular de uso de la misma de cara a la predicción de palabras y a la mejor colaboración con otros módulos.

En un primer nivel de abstracción, los métodos de predicción que se han incluido utilizan dos estrategias de modelado de la información lingüística: el modelado estocástico, que considera una pequeña historia anterior en lo escrito, y el formal, que hace uso de una gramática que describe las estructuras válidas del lenguaje.

De nuestro estudio es destacable la evaluación realizada, tanto cuantitativa en laboratorio, como cualitativa en pruebas de campo, tras la implementación real de un sistema de ayuda a la escritura que sigue esta arquitectura. La evaluación objetiva cuantitativa analiza la potencia de cada fuente de información y permite elegir la mejor estrategia de combinación que represente de hecho una ayuda para los usuarios de esta tecnología. Hemos de señalar que en la actualidad no existe ningún estándar a nivel nacional ni internacional que determine los parámetros que se deben evaluar, las pruebas necesarias para hacerlo, ni las condiciones en las que éstas se deben realizar. En esta tesis se aporta un estudio detallado de los diversos factores que influyen en los resultados, sobre los cuales sería necesario realizar un esfuerzo de normalización. Además, hemos realizado una propuesta

sobre las métricas a utilizar considerando aquellas que mejor evalúan los beneficios que cada colectivo particular de usuarios obtiene de la predicción. La evaluación cualitativa por parte de los usuarios del sistema representa una gratificación como fruto del esfuerzo empeñado.

Entrando en un mayor nivel de detalle en la descripción de las aportaciones de esta tesis, se mostrarán en primer lugar las referentes a la arquitectura propuesta, la organización de la información y las fuentes de conocimiento utilizadas, pasando a continuación a presentar las aportaciones referidas al modelado formal y los estudios sobre modelos estocásticos categoriales, para finalizar con un resumen y discusión de los resultados de la evaluación llevada a cabo sobre una implementación de la arquitectura.

Una vez estudiadas y adaptadas las diferentes fuentes de información lingüística, nos hemos aprovechado de sus características para elaborar una arquitectura global que permite la original combinación de la información que hemos encontrado como óptima para sistemas de ayuda a la escritura. En la arquitectura proponemos un módulo de gestión de los flujos de información que permite integrar la potencia desplegada por los distintos módulos lingüísticos (diccionario general, diccionario personalizado de la sesión en curso, diccionarios temáticos, modelos estocásticos con unigramas, bigramas, bipos y tripos, analizador sintáctico que sigue las reglas de la gramática formal, filtro de palabras rechazadas y predictor de sufijos).

Nuestra arquitectura tiene como características destacadas:

- Su modularidad, que permite integrar fácilmente nuevos métodos de predicción y fuentes de conocimiento adicionales, sin necesidad de invertir esfuerzo en el rediseño de la arquitectura.
- Su flexibilidad, por una parte, dada la independencia de la arquitectura de la tarea concreta de aplicación de la predicción de palabras, y por otra, del idioma objetivo, en sus aspectos procedimentales.
- Su potencia, por su capacidad integradora de múltiples flujos de información.

Para llegar a la predicción de un conjunto limitado de palabras con apoyo de fuentes de información lingüística, las palabras se han descrito en los diccionarios consignando información de muy diverso carácter. Cada significante está acompañado por sus posibles lemas, categorías y rasgos, información que ha sido diseñada y organizada minuciosamente para su utilización en la gramática formal. Además, se incluye toda la información probabilística asociada a cada palabra. Consideramos también una contribución relevante de esta tesis la metodología de diseño y organización de esta información, destacando la aportación del conjunto y la descripción de las categorías definidas en la arquitectura, que, como conclusión del trabajo, recomendamos que se aparten en algunos detalles de las clasificaciones tradicionales, para conectar mejor con los comportamientos sintácticos reales observados y que lograremos describir con la gramática formal propuesta. También se aporta el diseño del conjunto de rasgos, junto a originales mecanismos de manejo de los mismos que hemos incluido en el diseño de la gramática formal.

La capacidad de utilizar información procedente de los diccionarios temáticos y del personal (genéricamente menos entrenada que la del diccionario general) ha sido posible gracias al conveniente desacoplamiento en diferentes módulos entre la información asociada a cada palabra y los modelos gramaticales de predicción.

Con respecto al modelado formal, el estudio detallado de los fenómenos lingüísticos tanto teórica como empíricamente, nos ha llevado a diseñar una gramática probabilística independiente de contexto que contempla una original imbricación de mecanismos que la dotan de una importante potencia descriptiva del lenguaje:

- Potente gestión de rasgos en los símbolos terminales, más allá de la unificación, permitiendo modelar no solamente la concordancia entre rasgos, sino también la posibilidad de imponer y prohibir ciertos valores de los mismos.
- Manejo de rasgos en símbolos no terminales, lo que proporciona gran potencia expresiva permitiendo no solamente imponer restricciones a sus valores (concordancia, imposición o prohibición), sino también incluir mecanismos de control como, por ejemplo, la restricción del conjunto de reglas desplegado, dependiendo del valor de los rasgos del símbolo no terminal.
- Gestión de excepciones en la gramática, a través de la posibilidad de imponer y prohibir tanto significantes como lemas, para modelar aquellas expresiones en las que los significantes o lemas tienen un comportamiento concreto, diferente al de cualquiera de las categorías a las que pertenecen.
- Opcionalidad de elementos, lo que permite describir los casos en los que la validez de una regla es independiente de la presencia o ausencia de un determinado símbolo.

Por supuesto, el diseño de los diccionarios ha sido tal que estos contienen toda la información exigida por la gramática. El manejo de modelos estocásticos a nivel de categoría y de una gramática formal también fundamentalmente categorial permiten evitar la dependencia explícita de los significantes. Se ha tenido especial cuidado en la decisión de trasladar una importante proporción de información de una palabra de las categorías a los rasgos, de manera que se racionalice el diseño de las reglas de la gramática.

Como muestra de la potencia conseguida con la división entre categorías y rasgos, junto con los mecanismos de manejo aportados en esta tesis, hemos descrito ejemplos en los que necesitaríamos más de 200 reglas para describir las mismas estructuras con una gramática independiente de contexto tradicional y que en nuestra propuesta quedan reducidos a una sola regla.

La información de categorías y rasgos asociada a cada palabra se ha aprovechado también para el modelado estocástico, permitiéndonos describir modelos categoriales (bipos, tripos) con unas exigencias menores de entrenamiento que aquellos que dependen directamente de los significantes (como bigramas y trigramas). En estos modelos hemos añadido un mecanismo de filtrado por comprobación de rasgos, para compensar parcialmente la pérdida de potencia que supone el desplazamiento de parte de la carga expresiva de las categorías a los rasgos.

Para el problema de la categorización de palabras desconocidas hemos contribuido también con una metodología que integra distintas fuentes de conocimiento (terminacio-

nes distinguidas, prefijos o sufijos, conjugación de verbos regulares y manejo de clíticos), usando, en última instancia, una solución basada en el aprendizaje automático de la distribución más probable de categorías en las palabras desconocidas.

En este trabajo no nos hemos limitado a realizar un planteamiento teórico, sino que hemos implementado un sistema según la arquitectura propuesta, lo que nos ha permitido evaluar cada uno de los métodos, modelos y sus combinaciones, y conseguir la metodología de combinación óptima. En dicha implementación se han tenido en cuenta, además, consideraciones específicas para el diseño de la interfaz de usuario, en tres niveles: un buen diseño de la interfaz genérica del programa, su adaptación a usuarios con discapacidad y, por último, la inclusión de mecanismos adaptativos, algunos específicos de las interfaces con barrido.

Con respecto a la evaluación, y, a partir del estudio de los factores que influyen en ella y cómo lo hacen, hemos establecido una metodología para abordarla. Proponemos como parámetros a evaluar el porcentaje de palabras predichas, importante para las personas con problemas lingüísticos, y el de pulsaciones ahorradas, más relevante para las personas con discapacidad física, por su relación directa con el esfuerzo ahorrado.

Tras el proceso de evaluación, se propone la combinación óptima de la información procedente de los distintos diccionarios y métodos de predicción. En esta combinación se da prioridad a las palabras que aporta el modelo basado en bigramas de los diccionarios temático y personal. A continuación se utilizan los modelos estocásticos categoriales, aplicados primero al diccionario temático correspondiente y, posteriormente, con una adecuada ponderación, a los diccionarios personal y general. Por supuesto, las palabras que incluimos en la lista deben cumplir las condiciones impuestas por las letras iniciales que ya haya introducido el usuario. El valor de los pesos con que se integra la información general con la personalizada se ha optimizado para textos de pequeña longitud, que son los que pueden ser generados en una sola sesión por una persona que necesita de ayudas técnicas para escribir, para ser consecuentes con las condiciones reales de trabajo del sistema.

En cuanto al método de predicción basado en la gramática formal, el conjunto de aportaciones realizadas ha permitido obtener resultados cercanos a los de los métodos basados en modelos estocásticos categoriales, quedando para las líneas de trabajo futuras tanto el completar su capacidad descriptiva, como un análisis más profundo de la inclusión de otras fuentes de conocimiento, por ejemplo de corte semántico, que apoye al análisis gramatical y aporte incluso mejoras significativas. La modularidad y flexibilidad con que se ha dotado a la arquitectura permitirán realizar esta investigación aprovechándose del gran esfuerzo ya invertido aquí.

Capítulo 6

Líneas de trabajo futuras

La predicción de palabras en castellano aplicada a la comunicación alternativa y aumentativa es un campo en el cual están abiertos aún numerosos temas que pueden tomarse como base para establecer líneas futuras de investigación.

Realmente son muchas las que se nos ocurre plantear tras la realización de un trabajo como el abordado en esta tesis doctoral, pero nos limitaremos a destacar aquellos que nos parecen más interesantes con el objetivo fundamental de alcanzar una base de conocimiento y experiencia suficiente que permitan desarrollar en el futuro sistemas de predicción de palabras cada vez más potentes.

6.1 Diccionarios

El diccionario general actual, aunque por su volumen represente una parte importante de las palabras del castellano, puede ser completado, no sólo para contener mayor cantidad de palabras, sino para admitir otro tipo de entradas, como unidades multipalabra (por ejemplo, expresiones, perífrasis, etc.), abreviaturas, siglas, etc., que podrían aumentar la potencia del sistema global.

Igualmente sería muy interesante integrar otras fuentes de conocimiento léxico más potentes, como las aparecidas en los últimos años para el idioma castellano, entre las que destacamos las descritas en [Goñi97], basadas en morfología inflexional y dotadas de un extenso conjunto de herramientas para facilitar la creación y gestión de diccionarios electrónicos, o las mostradas en [Rios99], que constan de más de 600000 formas (92000 lemas) con información de categorías y rasgos.

Además, es necesario diseñar estrategias más potentes de entrenamiento de la información frecuencial que incluye, especialmente la correspondiente al reparto de frecuencias entre los distintos significados de una misma palabra, usando una mayor cantidad de textos etiquetados sin ambigüedad.

En este sentido sería también posible utilizar herramientas de etiquetado automático, específicas para castellano, como las que se describen en [Subi00], que hace uso de autómatas finitos, apoyándose en una estructura léxica muy completa.

Podría, también, refinarse la metodología actual de combinación de diccionarios, extendiendo su capacidad para que permita utilizar todos los diccionarios disponibles simultáneamente. Entre los criterios a considerar en la metodología de combinación podemos plantear el grado de entrenamiento o la adaptación de cada diccionario al texto.

Sería necesario investigar sobre las posibles funciones distancia entre diccionarios y entre los diccionarios y el texto escrito, que permitan decidir, dinámicamente, cuál de ellos sería el más conveniente para la escritura de un texto determinado.

6.2 Métodos de predicción

En el estado actual, es importante aumentar la capacidad descriptiva de la gramática, para modelar un conjunto más amplio de fenómenos sintácticos. Se puede estudiar el refinamiento del conjunto de categorías y rasgos disponible, para soportar la información adicional requerida por las nuevas reglas, por ejemplo, si se decide modelar ciertos comportamientos de los verbos como la transitividad, etc., o incluso para gestionar información semántica. En este sentido podemos citar la utilidad de adaptar propuestas como las descritas en [Cola99], en las que se hace un uso intensivo del modelado formal en sistemas de comprensión de habla en castellano, integrando gramáticas formales sintácticas y semánticas y haciendo uso de interesantes mecanismos de manejo de rasgos semánticos.

La potencia del analizador podría ser ampliada, aumentando su flexibilidad, por ejemplo, añadiendo un esquema basado en *valencias* para controlar los elementos opcionales.

La robustez del sistema puede ser ampliada, aumentando la potencia del módulo de recuperación de errores. El retroceso al modelo basado en *tripos* utilizado en la actualidad puede ser sustituido por un sistema que no interrumpa el análisis en caso de error, sino que lo flexibilice. Apuntaremos como interesantes las ideas sobre modelos de lenguajes estructurados que hacen una combinación de análisis sintáctico parcial usando un analizador tradicional con esquemas estocásticos, en los que los elementos considerados son los resultados parciales de dicho proceso de análisis [Jeli99].

También sería posible plantear un módulo de gestión de errores que sea capaz de distinguir entre diferentes tipos de error (o al menos lanzar todas las posibilidades, en una primera aproximación, hasta que la información posterior permita seleccionar cual es la correcta), asignando unos costes a cada uno de los tipos de error (inserción, borrado o sustitución de una categoría o palabra, o problemas de consistencia de los rasgos, lemas o significantes con las restricciones concretas de cada regla).

Igualmente, se podría estudiar la manera de aumentar la flexibilidad del modelo, adaptando el conjunto de reglas al texto escrito o al usuario, considerando métodos automáticos de entrenamiento de la frecuencia de las reglas (más refinados que los usados en esta tesis, usando técnicas como la descrita en [Sanc99]), de los elementos opcionales, entrenamiento del peso a asociar a los errores, tanto para fases previas a la escritura como para adaptar las reglas al usuario, mientras escribe, en los casos en los que sea razonable. En este sentido, la inferencia automática de gramáticas es un campo de investigación que podría repercutir positivamente en nuestros estudios.

Para aumentar la potencia del analizador y su flexibilidad, una línea futura importante podría consistir en el estudio de la posibilidad de establecer reglas en varios niveles de detalle. Un nivel en el cual se tratan los significantes, de forma que cuando se disponga de información detallada sobre una palabra concreta, se dé prioridad a la regla que modela dicho caso. En un segundo nivel, se plantearán reglas más generales, para categorías gramaticales, que se utilizarán cuando no se disponga de ninguna regla específica sobre la palabra analizada.

La potencia de la gramática propuesta en esta tesis podría ser igualmente aprovechada en otras áreas, como el reconocimiento de habla, que tiene algunas características similares a las de la predicción de palabras (nos referimos, por ejemplo, a la necesidad de tratar construcciones agramaticales por errores en el reconocedor, lo que exige un considerable aumento de la robustez necesaria).

El criterio de clasificación propuesto para la elaboración del repertorio de categorías y rasgos admite aún más refinamiento, como se ha comentado anteriormente. Queda pendiente el estudio de técnicas de agrupamiento automático (*clustering*), como las planteadas en [Nies98] para confeccionar el esquema categorial y evaluar su potencia en comparación con el que parte de conocimiento experto.

Quedaría también por explotar aún más la potencia de modelos *n-pos*, con un valor de n superior a 3. Dicha tarea sería abordable mientras mantuviéramos reducido el número de categorías definidas, pero sería necesario rediseñar y evaluar cuidadosamente el impacto de dicha extensión en algunos de los mecanismos propuestos (aprovechamiento de la concordancia de rasgos, por ejemplo).

Como alternativa a los modelos estocástico y formal propuestos, es viable la utilización de mecanismos basados en el uso de redes neuronales, como estimadores de la palabra o categoría siguiente, en la línea de lo publicado en [Pala98c] donde se mostraban prometedores resultados de esta alternativa en comparación con el uso de *bipos* y *tripos*.

6.3 Métodos de evaluación

En primer lugar, sería necesario establecer un estándar que estableciera las condiciones en las que se debe evaluar cada sistema de predicción, de forma que se puedan realizar comparaciones objetivas entre sistemas, que pueden incluso predecir en diferentes idiomas.

Para ello se debería estudiar cual es el método de evaluación que mejor caracteriza el sistema, considerando métricas nuevas que representen de forma más fiable el beneficio que proporciona la predicción, según nuevos criterios dependientes de cada usuario en concreto. Por ejemplo, si una persona utiliza la predicción como ayuda porque comete frecuentes faltas de ortografía y solamente la utiliza ante palabras de ortografía dudosa, esas serían las únicas a contabilizar en la nueva métrica. Sería también necesario valorar la carga cognitiva que el sistema exige del usuario, dependiendo de su configuración: con más o menos palabras predichas o con distintos diseños de la interfaz.

Se deben establecer criterios que permitan valorar de forma objetiva el aumento de la

calidad de los textos generados: aumento de la inteligibilidad, disminución en el número de faltas de ortografía, uso de construcciones gramaticales más complejas, etc.

En el proceso de evaluación, ayudaría disponer de un modelo de usuario, basado en datos obtenidos de sesiones de escritura utilizando el sistema, que permita realizar medidas de tiempo y estudiar la influencia de ciertos factores, como la longitud de la lista de palabras predichas, la influencia de los errores, etc. En dicho modelo debería considerarse, entre otras cosas, el tiempo necesario para leer la lista de palabras predichas y decidir si alguna es la correcta, la posibilidad de que una pulsación pueda ser errónea y la necesidad de corregir sus consecuencias (incluyendo errores en el intento de corrección).

6.4 Aplicación de la arquitectura a otros idiomas y nuevos entornos de aplicación

Una interesante línea de investigación futura es la que se refiere a la aplicación de la arquitectura y la metodología propuesta a idiomas distintos del castellano.

La filosofía de diseño de la arquitectura en cuando a modularidad y flexibilidad hacen razonablemente sencilla la extensión a otros idiomas de algunos módulos de la misma, pero es necesario hacer un esfuerzo muy significativo para introducir en el modelo el conocimiento específico de los nuevos idiomas, sobre todo en lo que se refiere al modelado formal.

Por el mismo motivo, es posible plantear la extensión a otros tipos de gramáticas, como, por ejemplo, la de los mensajes que producen las personas que escriben con símbolos o palabras globales, o la de las personas sordas al escribir. Sería muy interesante ir más allá y diseñar mecanismos de traducción y corrección de la salida de dichos sistemas, lo que permitiría mejorar aún más la calidad de los textos y mensajes generados por estos colectivos de usuarios.

Apéndice A

Descripción de los textos utilizados para entrenamiento y evaluación

Es este apéndice se describen las características principales de los textos utilizados para entrenamiento y evaluación, describiendo previamente algunos de los criterios usados para seleccionarlos.

A.1 Criterios de selección de textos

En la actualidad, la disponibilidad de textos de todo tipo en formato electrónico es muy alta, con lo que es posible elegir entre una gran cantidad de alternativas a la hora de analizar sistemas de procesamiento de lenguaje natural basados en texto[1].

En este trabajo de tesis se hizo una selección que trataba de cubrir los siguientes aspectos:

- textos pertenecientes a dominios razonablemente disjuntos, entre sí y con respecto a los de los textos de entrenamiento de las fuentes de información de los sistemas analizados. Se seleccionaron textos de dominios periodístico, técnico (parte del contenido de este documento) y literario (novelas clásicas y cuentos modernos).
- textos de suficiente longitud como para asegurar un adecuado entrenamiento de los elementos entrenables y garantizar, siempre que sea posible, la validez estadística de los resultados obtenidos.
- textos razonablemente cortos, para evaluar el sistema en condiciones más cercanas a las que se encontrará en la realidad: los usuarios con algún tipo de discapacidad son capaces de escribir poca cantidad de texto en cada sesión.

En los siguientes apartados se ofrecerá información genérica del contenido de cada texto, así como información detallada cuantitativa de cada uno de ellos.

[1] Aunque es de destacar que dichos textos están disponibles, en su mayoría, sin ningún tipo de información adicional: carecen de características normalizadas, de información lingüística asociada, etc.

A.2 Características generales

En la tabla A.1 de la página 180 se incluye una descripción general de los textos utilizados en los distintos experimentos realizados. Los campos descritos son los siguientes:

- Identificador: El nombre utilizado en las referencias que se hacen al texto dado.
- Tipo: E si se trata de un texto usado en entrenamiento del sistema o T si se trata de un texto usado en experimentos de evaluación.
- Descripción: Descripción general del contenido y procedencia del texto dado.

Identificador	Tipo	Descripción
MUNDO3-6	E	Textos procedentes de la edición electrónica del periódico *El mundo*, correspondientes a los meses Marzo a Junio de 1994. No etiquetado. Alrededor de 8 millones de palabras.
MUNDO1M	E	Texto de un millón de palabras procedentes de la edición electrónica del periódico *El mundo*, categorizado con ambigüedad. Alrededor de 1 millón de palabras.
FEB100K	T	Texto procedente del mes de Febrero de 1994 de la edición electrónica del periódico *El mundo*, compuesto por alrededor de 100.000 palabras. No etiquetado
FEB1K	T	Texto de alrededor de 1.000 palabras, procedente del mes de Febrero de 1994 de la edición electrónica del periódico *El Mundo*.
PAIS	E	Texto categorizado de forma semiautomática, procedentes del periódico *El País*, compuesto por alrededor de 50.000 palabras.
TESIS	T	Texto perteneciente a los primeros capítulos de esta tesis[2], compuesto por alrededor de 35.000 palabras sin etiquetar.
TESIS2	T	Texto perteneciente a los primeros apartados de esta tesis, compuesto por alrededor de 2.000 palabras sin etiquetar preprocesada para que su contenido se ajuste de forma más adecuada a la gramática disponible en el sistema. Este texto se incluye en el apéndice B (página 187)
CUENTO1	T	Cuento corto. Sin etiquetar. Alrededor de 1.000 palabras. Los cuentos han sido seleccionados de entre los disponibles en línea en http://www.eltunel.com.ar/.
CUENTO2	T	Cuento corto. Sin etiquetar. Alrededor de 3.000 palabras
CUENTO2	T	Cuento corto. Sin etiquetar. Alrededor de 1.500 palabras.
QUIJOTE	T	Texto completo del Quijote. Sin etiquetar. Alrededor de 400.000 palabras. Puede encontrarse en varios sitios en Internet, entre ellos el "Proyecto Cervantes 2001" (http://www.csdl.tamu.edu/cervantes/spanish/spanindex.html).
MOLINOS	T	Texto del capítulo de los Molinos de El Quijote. Sin etiquetar. Alrededor de 3.000 palabras.
QUIJOTE15-13K	E	Texto de alrededor de 13000 palabras, procedente del capitulo XV del Quijote

Tabla A.1: Tabla de características generales de los textos usados en castellano.

A.3 Características cuantitativas de los textos de entrenamiento

En esta sección se ofrece información cuantitativa sobre los textos usados en entrenamiento. En concreto:

- Parámetros relacionados con el número de elementos encontrados (palabras, signos, números, palabras distintas) en la tabla A.3 de la página 181).
- Parámetros relacionados con el número de palabras que no están incluidas en el diccionario general, en la tabla A.3 de la página 181).

Identificador	NET	NPAL	NSIG	NNUM	NPALD
MUNDO1M	899.111	770.500	115.713	12.898	53.143
MUNDO3-6	9.413.779	7.933.207	1.310.6794	169.893	170.379
PAIS	54.176	47.327	6.253	596	9.970
QUIJOTE[3]	441.940	374.225	67.702	13	23.435
QUIJOTE15-13K	13.935	11.701	2.234	0	2.756

Tabla A.2: Tabla de características cuantitativas de los textos de entrenamiento. Elementos encontrados.

Donde los campos descritos son:

- Identificador: Identificador del texto
- NET: Número de elementos totales (palabras+signos+números)
- NPAL: Número de palabras totales
- NSIG: Número de signos totales
- NNUM: Número de cadenas numéricas
- NPALD: Número de palabras distintas (vocabulario del texto)

Identificador	OOVT	OOVTD	OOVPT	OOVPTD
MUNDO1M	67.124 (8% NPAL)	21.468 (41% NPALD)	35.035 (52% OOVT)	9.135 (42% OOVTD)
MUNDO3-6	749.198 (9% NPAL)	93.957 (56% NPALD)	432.343 (57% OOVT)	46.813 (49% OOVTD)
PAIS	2.423 (5% NPAL)	1.541 (15% NPALD)	1.442 1.442 (59% OOVT)	764 (49% OOVTD)
QUIJOTE	29040 (7% NPAL)	9.187 (39% NPALD)	10.387 (35% OOVT)	1.581 (17% OOVTD)
QUIJOTE15-13K	1.045 (8% NPAL)	586 (21% NPALD)	380 (36% OOVT)	107 (18% OOVTD)

Tabla A.3: Tabla de características cuantitativas de los textos de entrenamiento. Palabras fuera de vocabulario

Los campos son los siguientes:

- Identificador: Identificador del texto
- OOVT: Número total de palabras fuera de vocabulario
- OOVTD: Número de palabras fuera de vocabulario distintas
- OOVPT: Número de nombres propios fuera de vocabulario
- OOVPTD: Número de nombres propios distintos fuera de vocabulario

A.4 Características cuantitativas de los textos de evaluación

En esta sección se incluye información cuantitativa sobre los textos usados en las pruebas de esta tesis (los mismos que se detallan en la sección anterior).

Identificador	NET	NPAL	NSIG	NNUM	NPALD
FEB100K	118.386	103.583	13.341	1.462	14.859
FEB1K	1.461	1.286	163	12	648
TESIS	42.740	37.496	5.166	78	3.736
TESIS2	2.240	2.058	180	2	674
CUENTO1	1.204	986	218	0	498
CUENTO2	3.036	2.672	364	0	829
CUENTO3	1.610	1.416	190	4	605
MOLINOS	3.541	2.980	561	0	999

Tabla A.4: Tabla de características cuantitativas de los textos de evaluación. Elementos encontrados.

Donde los campos son:

- Identificador: Identificador del texto
- NET: Número de elementos totales (palabras+signos+números)
- NPAL: Número de palabras totales
- NSIG: Número de signos totales
- NNUM: Número de cadenas numéricas
- NPALD: Número de palabras distintas (vocabulario del texto)

Identificador	OOVT	OOVTD	OOVPT	OOVPTD
FEB100K	7.055 (7% NPAL)	3.804 (26% NPALD)	3.816 (53% OOVT)	1.647 (43% OOVTD)
FEB1K	80 (6% NPAL)	72 (11% NPALD)	29 (36% OOVT)	24 (33% OOVTD)
TESIS	3.658 (9% NPAL)	807 (21% NPALD)	285 (8% OOVT)	190 (23% OOVTD)
TESIS2	107 (5% NPAL)	76 (11% NPALD)	6 (5% OOVT)	6 (7% OOVTD)
CUENTO1	81 (8% NPAL)	69 (14% NPALD)	16 (19% OOVT)	9 (13% OOVTD)
CUENTO2	199 (7% NPAL)	118 (14% NPALD)	80 (40% OOVT)	25 (21% OOVTD)
CUENTO3	120 (8% NPAL)	91 (15% NPALD)	39 (32% OOVT)	21 (23% OOVTD)
MOLINOS	232 (7% NPAL)	134 (13% NPALD)	89 (38% OOVT)	21 (15% OOVTD)

Tabla A.5: Tabla de características cuantitativas de los textos de evaluación. Palabras fuera de vocabulario

Donde los campos son:

- Identificador: Identificador del texto
- OOVT: Número total de palabras fuera de vocabulario
- OOVTD: Número de palabras fuera de vocabulario distintas
- OOVPT: Número de nombres propios fuera de vocabulario
- OOVPTD: Número de nombres propios distintos fuera de vocabulario

A.5 Distribución frecuencial de las pulsaciones

En este apartado se ofrecen los histogramas correspondientes a la distribución frecuencial del número de palabras encontrados en función del número de pulsaciones necesarias para escribirlas.

A modo de ilustración de cada dominio tratado, se incluyen únicamente las correspondientes a los textos TESIS (figura A.1 de la página 184), QUIJOTE15-13K (figura A.2 de la página 184) y FEB100K (figura A.3 de la página 185).

De ellas puede observarse una razonable similitud, con lo que se asume que las diferencias existentes no deben influir significativamente en los resultados obtenidos en cada caso.

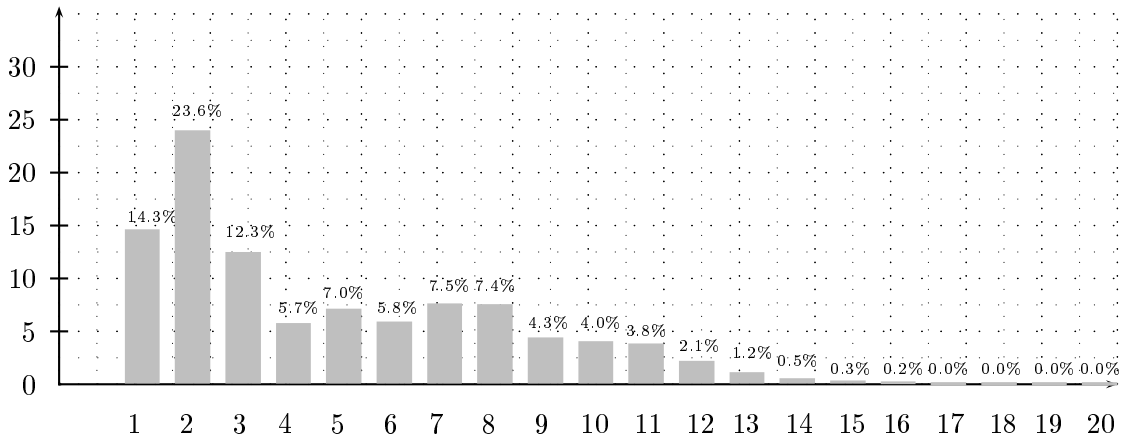


Figura A.1: Distribución de palabras en función de pulsaciones necesarias para el texto TESIS

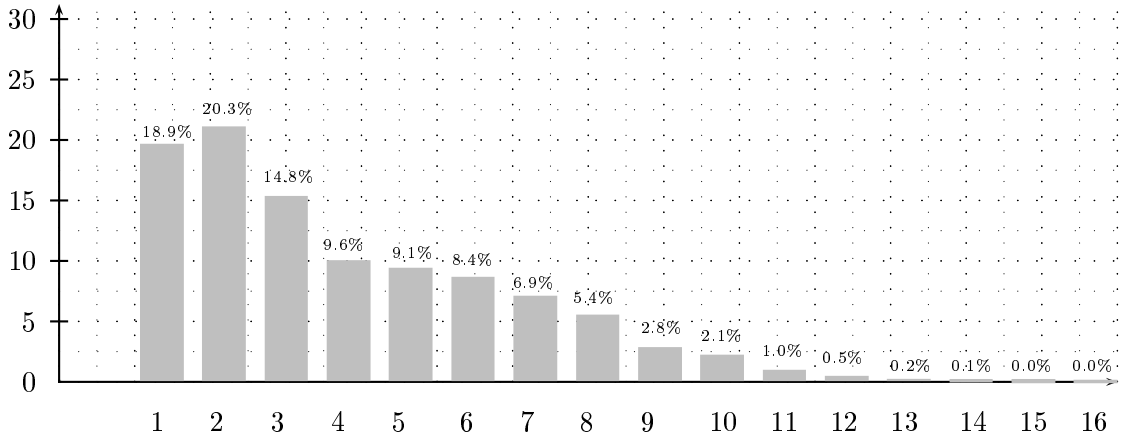


Figura A.2: Distribución de palabras en función de pulsaciones necesarias para el texto QUIJOTE15-13K

A.6 Distribución estadística categorial de los textos de PAIS

En este apartado se ofrece información acerca de la distribución estadística de los textos que forman la base de datos PAIS, descrita en el apartado anterior.

El histograma de dicha distribución está incluido en la figura A.6 de la página 185.

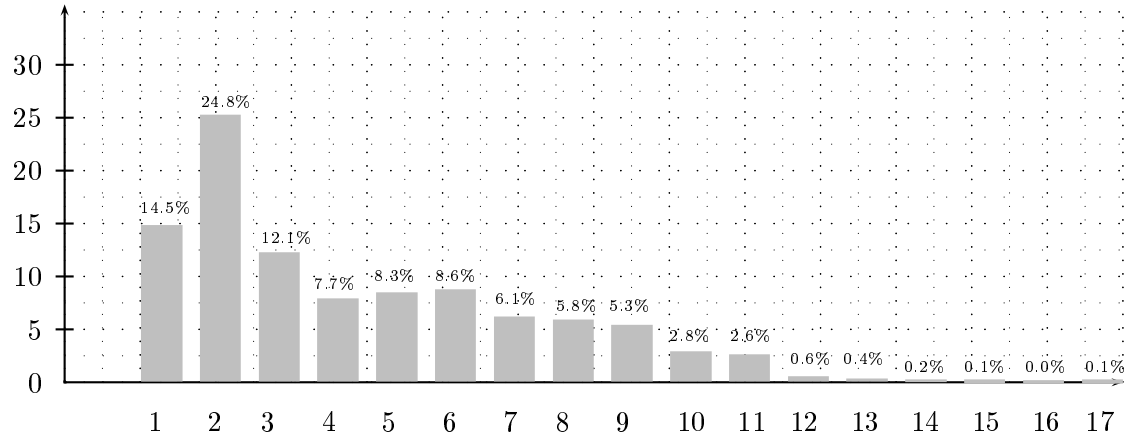


Figura A.3: Distribución de palabras en función de pulsaciones necesarias para el texto FEB100K

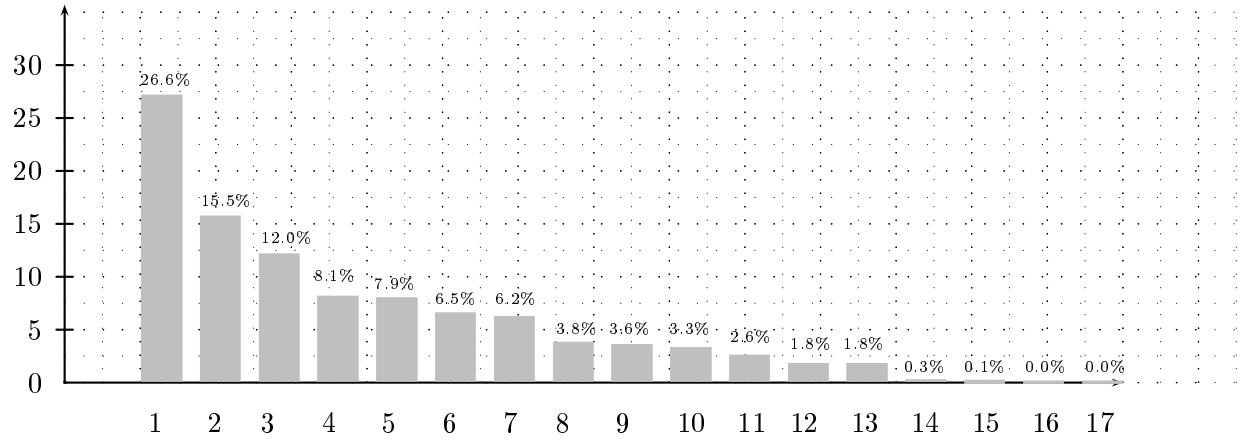


Tabla A.6: Distribución frecuencial de las categorías del texto PAIS.

Leyenda:

- 1: Nombre (n)
- 2: Determinante (det)
- 3: Verbo (v)
- 4: Signos de puntuación (signo)
- 5: Preposición *a* y *de* (prepade)
- 6: Adjetivo (adj)
- 7: Preposición (prep)
- 8: Conjunción (conj)
- 9: Punto (punto)

- 10: Adverbio (adv)
- 11: Pronombre (pron)
- 12: Artículo contracto (arco)
- 13: Pronombres clíticos (pronclit)
- 14: Numeral ordinal (numo)
- 15: Cajón de sastre (cjs)
- 16: Pronombre que sigue a preposición (propre)
- 17: Pronombres *Contigo*, *conmigo* y *consigo* (sprep)
- 18: Interjección (no aparece ninguna en los textos de entrenamiento).

Apéndice B

Comparación entre los textos TESIS y TESIS2

Dadas las características del texto TESIS2, que ha sido generado a partir de TESIS[1], de forma que su contenido se ajuste de forma más adecuada a la gramática disponible en el sistema, se ha decidido incluir una comparación de ambos en este apartado, para que el lector pueda comprobar que no se ha introducido demasiada artificialidad en el mismo, habiéndose realizado pocos cambios con respecto al original.

Como se comentó en el apartado 4.1.3.4.2 (página 145) del capítulo de evaluación, los cambios han consistido en:

- Eliminación de referencias bibliográficas insertadas en las frases, ya que son cadenas no incluidas en el diccionario, que, por lo tanto, interrumpen en análisis.
- Eliminación de la palabra "etc." insertada dentro de una frase, ya que el algoritmo de segmentación de oraciones considera el punto como fin de frase, y el analizador vuelve a situarse en el comienzo de la análisis.
- Eliminación o reescritura de algunos incisos escritos entre paréntesis.
- Eliminación de números en las enumeraciones.
- Las frases excesivamente largas, con concatenaciones complejas, han sido segmentadas en oraciones de menor longitud. La longitud media de las frases de TESIS2 es de 18.56 palabras.

En la tabla B.1 se incluye la comparación, párrafo a párrafo entre ambos textos.

[1]Como se indicó anteriormente, el texto no es exactamente igual al que aparece en este documento, habiendo sido extraído de una versión preliminar del mismo

Tabla B.1: Comparación entre los textos TESIS y TESIS2

Texto original (TESIS)	Versión modificada (TESIS2)
La predicción de palabras consiste en intentar adivinar la palabra que el usuario está escribiendo, basándose en las letras que están ya escritas y en información adicional proveniente del texto que está siendo introducido.	La predicción de palabras consiste en intentar adivinar la palabra que el usuario está escribiendo. Se basa en las letras que ya están escritas y en información adicional proveniente del texto que está siendo introducido.
Las palabras predichas se muestran al usuario para que seleccione entre ellas la que desea escribir (si está en la lista), permitiéndole así introducir dicha palabra de forma rápida y con muy poco esfuerzo.	Las palabras que se predicen se muestran al usuario para que seleccione entre ellas la que desea escribir. Así, se le permite introducir dicha palabra de forma rápida con muy poco esfuerzo.
Con el sistema de predicción de palabras se pretende ayudar a usuarios con problemas fundamentalmente físicos (por ejemplo discapacidad motriz, esclerosis lateral amiotrófica, parálisis cerebral, parapléjicos o accidentados de coche) pero también lingüísticos (como personas con dislexia) a la escritura de textos de toda índole.	Con el sistema de predicción de palabras se pretende ayudar a usuarios con problemas a la escritura de textos de toda índole.
Por ello la predicción se incluye en programas de ayuda a usuarios con problemas, como comunicadores y editores de texto que pueden ser utilizados con pulsador.	La predicción se incluye en programas de ayuda a usuarios con problemas que pueden ser utilizados con pulsador.
En esta tesis se describen los diferentes métodos de predicción de palabras en castellano que se están utilizando en una aplicación para discapacitados motrices, empleando un modelo híbrido que conjuga métodos estadísticos con reglas declarativas, sistemas estáticos y dinámicos.	En esta tesis se describen los diferentes métodos de predicción de palabras en español que se están utilizando en una aplicación para discapacitados motrices. Se emplea un modelo híbrido que conjuga métodos estadísticos con reglas declarativas, sistemas estáticos y dinámicos.
Con este sistema se consigue incrementar la velocidad de escritura desde 2-3 palabras por minuto a 8-10.	Con este sistema se consigue incrementar la velocidad de escritura desde dos palabras por minuto a ocho.
Introducción.	-
Los ordenadores están cobrando cada vez mayor importancia en nuestra sociedad. Está aumentando vertiginosamente el número de ordenadores en los hogares y su inserción en el mundo laboral es indudable.	Los ordenadores están cobrando mayor importancia en nuestra sociedad. Está aumentando vertiginosamente el número de ordenadores en los hogares y su inserción en el mundo laboral es indudable.
La integración de Internet en la vida cotidiana está cambiando paulatinamente la forma de acceso a la información, comunicación, compras, reserva de billetes, etc. Lo que en principio es solamente una herramienta, está modificando sustancialmente la forma de vida de muchas personas.	La integración de Internet en la vida cotidiana está cambiando paulatinamente la forma de acceso a la información, comunicación, compras y reserva de billetes. Lo que es solamente una herramienta está modificando sustancialmente la forma de vida de muchas personas.
Hay un sector de la sociedad para el que los ordenadores son especialmente importantes, por la trascendencia de las tareas que les permite realizar: las personas con discapacidad.	Hay un sector de la sociedad para el que los ordenadores son especialmente importantes por la trascendencia de las tareas que les permite realizar: las personas con discapacidad.
Dependiendo del tipo de discapacidad y de su grado, estas personas pueden tener dificultades e incluso estar totalmente imposibilitadas para moverse, comunicarse o controlar su entorno por sí mismas.	Dependiendo del tipo de discapacidad y de su grado, estas personas pueden tener dificultades para moverse, comunicarse o controlar su entorno por sí mismas.

— Continúa en la página siguiente —

Tabla B.1: Comparación textos TESIS vs. TESIS2 (cont.)

Texto original (TESIS)	Versión modificada (TESIS2)
Los problemas en actividades tan básicas como las mencionadas dificultan en mayor o menor medida su vida diaria.	Los problemas en actividades tan básicas como las mencionadas dificultan en mayor o menor medida su vida diaria.
Se puede pensar, por ejemplo, en las dificultades con las que se encuentran las personas con parálisis cerebral, esclerosis lateral amiotrófica, las víctimas de accidentes, etc. Es ahora, con la flexibilidad de los ordenadores, cuando se pueden construir sistemas capaces de ayudar a estas personas de forma eficiente, con soluciones a la vez sencillas de utilizar y suficientemente potentes.	Se puede pensar en las dificultades con las que se encuentran las personas con parálisis cerebral. con la flexibilidad de los ordenadores se pueden construir sistemas capaces de ayudar a estas personas de forma eficiente.
La versatilidad de los ordenadores permite adaptarlos para que personas con muy diversas necesidades los puedan utilizar, aumentando su calidad de vida y su grado de autonomía personal.	La versatilidad de los ordenadores permite adaptarlos para que personas con muy diversas necesidades los puedan utilizar, aumentando su calidad de vida y su grado de autonomía personal.
Piénsese, por ejemplo, en las escasas posibilidades que hubiera tenido hace unos años Stephen Hawking, aquejado de esclerosis lateral amiotrófica, que le provoca graves problemas físicos y de comunicación.	Piénsese en las escasas posibilidades que hubiera tenido hace unos años Stephen Hawking, que le provoca graves problemas físicos y de comunicación.
En la actualidad, utilizando su ordenador personal con un solo dedo, es capaz de comunicarse, impartir conferencias e, incluso, escribir libros por sí mismo.	En la actualidad, utilizando su ordenador personal con un solo dedo, es capaz de comunicarse, imparte conferencias y escribe libros por sí mismo.
La característica que hace al ordenador especialmente adecuado para servir de base a las ayudas técnicas, es su posibilidad para adaptarse a las necesidades de los usuarios.	La característica que hace al ordenador especialmente adecuado para servir de base a las ayudas técnicas, es su posibilidad para adaptarse a las necesidades de los usuarios.
Habitualmente el ordenador se utiliza manejando el teclado y el ratón, visualizando datos en la pantalla y escuchando mensajes sonoros. Sin embargo, para las personas con discapacidad, que tienen problemas con algunos o todos estos dispositivos, es posible modificar dicha interacción, para evitar las partes del sistema que el usuario no pueda manejar.	Habitualmente el ordenador se utiliza manejando el teclado y el ratón, visualizando datos en la pantalla y escuchando mensajes sonoros. Sin embargo, para las personas con discapacidad que tienen problemas con algunos o todos estos dispositivos, es posible modificar dicha interacción, para evitar las partes del sistema que el usuario no pueda manejar.
Las adaptaciones que se deben realizar dependen de la discapacidad del usuario: a las personas sordas se les muestran por escrito todos los mensajes audibles; para las personas ciegas se transforma la información visual en auditiva (mediante síntesis de voz) o táctil (utilizando líneas braille), y para las personas con discapacidades físicas (con dificultades para utilizar el ratón o el teclado), se buscan formas de entrada alternativas.	Las adaptaciones que se deben realizar dependen de la discapacidad del usuario. a las personas sordas se les muestran por escrito todos los mensajes audibles. Para las personas ciegas se transforma la información visual en auditiva o táctil y para las personas con discapacidades físicas se buscan formas de entrada alternativas.

— Continúa en la página siguiente —

Tabla B.1: Comparación textos TESIS vs. TESIS2 (cont.)

Texto original (TESIS)	Versión modificada (TESIS2)
Las dificultades de las personas con discapacidad física a la hora de utilizar un ordenador son muy variadas, tanto en su tipología como en su grado. Dependiendo de las particularidades de su discapacidad, la persona puede tener problemas en el uso del teclado, del ratón o en ambos, y es necesario buscar una solución alternativa para sustituirlos, manteniendo su funcionalidad.	Las dificultades de las personas con discapacidad física a la hora de utilizar un ordenador son muy variadas. Dependiendo de las particularidades de su discapacidad, la persona puede tener problemas en el uso del teclado, del ratón o en ambos, y es necesario buscar una solución alternativa para sustituirlos, manteniendo su funcionalidad.
En [Cand97] se presentan las ayudas técnicas más comunes que pueden utilizar las personas con diversas discapacidades para el acceso al ordenador.	Se presentan las ayudas técnicas más comunes que pueden utilizar las personas con diversas discapacidades para el acceso al ordenador.
En los casos extremos, el usuario puede que sólo sea capaz de controlar con precisión una parte de su cuerpo, como puede ser un brazo, un dedo, una pierna, la lengua, un párpado o, incluso, su respiración.	En los casos extremos, el usuario puede que sólo sea capaz de controlar con precisión una parte de su cuerpo, como puede ser un brazo, un dedo, una pierna, la lengua, un párpado o, incluso, su respiración.
El problema a resolver es la transformación de ese movimiento controlado en una orden comprensible por el ordenador. La solución habitual para estos casos extremos es instalar en el ordenador un pulsador (por ejemplo, conectándolo a alguno de sus puertos) que el usuario pueda controlar con su movimiento útil (cabeza, brazo, pierna, soplidos, etc.), y que envía al ordenador información cuando es pulsado.	El problema a resolver es la transformación de ese movimiento controlado en una orden comprensible por el ordenador. La solución habitual para estos casos extremos es instalar en el ordenador un pulsador que el usuario pueda controlar con su movimiento útil.
Para transformar esa información en una orden, se instalan programas específicos: por ejemplo, un editor de texto utilizable con pulsador tiene, además de los menús habituales de tratamiento de texto, ficheros, etc., una matriz con las letras que se barre automáticamente: un cursor se desplaza, resaltando secuencialmente las letras, de forma que el usuario solamente necesita presionar el pulsador cuando la letra que desea escribir es resaltada.	Para transformar esa información en una orden se instalan programas específicos. Un editor de texto utilizable con pulsador tiene una matriz con las letras que se barre automáticamente. Un cursor se desplaza de forma que el usuario solamente necesita presionar el pulsador cuando la letra que desea escribir es resaltada.
Así, el usuario puede escribir sin necesidad de utilizar el teclado convencional. De la misma forma (barrido automático más pulsador) se accede a todos los demás menús del programa y de cualquier otra aplicación de este tipo.	Así, el usuario puede escribir sin necesidad de utilizar el teclado convencional. De la misma forma se accede a todos los demás menús del programa y de cualquier otra aplicación de este tipo.
El problema obvio que tiene este modo de acceso es la velocidad: el barrido automático tiene que ser lo suficientemente lento como para poder ser controlado por la persona, que incluso puede tener temblores.	El problema obvio que tiene este modo de acceso es la velocidad. El barrido automático tiene que ser lo suficientemente lento como para poder ser controlado por la persona que incluso puede tener temblores.
El tiempo que debe permanecer resaltada cada opción para asegurar que el usuario sea capaz de pulsar y elegirla, suele ser superior al medio segundo, es decir, para escribir cada letra, es necesario esperar al menos medio segundo por cada una de las opciones que la precedan en el menú.	El tiempo que debe permanecer resaltada cada opción para asegurar que el usuario sea capaz de pulsar y elegirla, suele ser superior al medio segundo, es decir, para escribir cada letra, es necesario esperar al menos medio segundo por cada una de las opciones que la precedan en el menú.

— Continúa en la página siguiente —

Tabla B.1: Comparación textos TESIS vs. TESIS2 (cont.)

Texto original (TESIS)	Versión modificada (TESIS2)
Aunque hay diversos sistemas para intentar acelerar este acceso (por ejemplo, barrido por filas y columnas), una velocidad de escritura de cuatro o cinco palabras por minuto suele ser habitual en estos casos, según hemos encontrado en nuestros estudios [Pala96].	Aunque hay diversos sistemas para intentar acelerar este acceso (por ejemplo, barrido por filas y columnas) , una velocidad de escritura de cuatro palabras por minuto suele ser habitual en estos casos, según hemos encontrado en nuestros estudios.
Por supuesto, no se puede comparar esta velocidad con las 200-400 pulsaciones por minuto (40-80 palabras por minuto) que puede alcanzar una persona experimentada que trabaja con ordenador, por lo cual, es necesario invertir todo el esfuerzo posible en intentar acelerarlo, especialmente porque hay un colectivo de personas para quienes el pulsador es su único modo de acceso al ordenador.	Por supuesto, no se puede comparar esta velocidad con las 200 pulsaciones por minuto que puede alcanzar una persona experimentada que trabaja con ordenador. Es necesario invertir todo el esfuerzo posible en intentar acelerarlo. Hay un colectivo de personas para quienes el pulsador es su único modo de acceso al ordenador.
Los objetivos de esta tesis son estudiar y desarrollar mecanismos para aumentar la velocidad de escritura con el ordenador, y reducir el esfuerzo necesario en dicho proceso para aquellas personas que sólo pueden utilizar un pulsador como medio de acceso.	Los objetivos de esta tesis son estudiar y desarrollar mecanismos para aumentar la velocidad de escritura con el ordenador, y reducir el esfuerzo necesario en dicho proceso para aquellas personas que sólo pueden utilizar un pulsador como medio de acceso.
Los dos puntos en los que se incidirá serán:	Los dos puntos en los que se incidirá serán los siguientes.
1. Reducir el número de pulsaciones necesarias para escribir un texto. Con ello se consigue disminuir el esfuerzo necesario para introducirlo.	Reducir el número de pulsaciones necesarias para escribir un texto. Con ello se consigue disminuir el esfuerzo necesario para introducirlo.
2. Disminuir el número medio de opciones que se deben recorrer en un sistema de barrido hasta llegar a la opción deseada, reduciendo así el tiempo necesario para introducir los textos.	Disminuir el número medio de opciones que se deben recorrer en un sistema hasta llegar a la opción deseada, reduciendo así el tiempo necesario para introducir los textos.
Una de las estrategias utilizadas para conseguir el primer objetivo se basa en la predicción de palabras: el propio programa usado para escribir procesa el texto e intenta predecir cuál es la palabra que el usuario está intentando introducir en ese momento.	Una de las estrategias utilizadas para conseguir el primer objetivo se basa en la predicción de palabras. El propio programa usado para escribir procesa el texto e intenta predecir cuál es la palabra que el usuario está intentando introducir en ese momento.
Las candidatas predichas son mostradas y barridas junto con las letras.	Las candidatas que se predicen son mostradas y barridas junto con las letras.
Si alguna de las ellas es la adecuada, puede ser seleccionada e insertada con el mismo método usado para escribir una sola letra, ahorrando tiempo y esfuerzo.	Si alguna de las ellas es la adecuada, puede ser seleccionada e insertada con el mismo método usado para escribir una sola letra, ahorrando tiempo y esfuerzo.
El ahorro depende de la calidad de los algoritmos de predicción y de la disposición de la interfaz de usuario.	El ahorro depende de la calidad de los algoritmos de predicción y de la disposición de la interfaz de usuario.

— Continúa en la página siguiente —

Tabla B.1: Comparación textos TESIS vs. TESIS2 (cont.)

Texto original (TESIS)	Versión modificada (TESIS2)
Una predicción perfecta es aquella que siempre muestra al usuario la palabra que desea escribir en la lista de candidatas, en cuyo caso sólo es necesario seleccionar las palabras a medida que van apareciendo en el menú. Este caso ideal está muy lejos de la situación actual, aunque ahora pueden conseguirse velocidades de 8-10 palabras por minuto, lo que supone aproximadamente duplicar la velocidad con respecto a cuando no se usa predicción de palabras.	Una predicción perfecta es aquella que siempre muestra al usuario la palabra que desea escribir en la lista de candidatas, en cuyo caso sólo es necesario seleccionar las palabras a medida que van apareciendo en el menú. Este caso ideal está muy lejos de la situación actual, aunque ahora pueden conseguirse velocidades de 8 palabras por minuto, lo que supone aproximadamente duplicar la velocidad con respecto a cuando no se usa predicción de palabras.
Además de lo visto, la predicción de palabras no es útil solamente para las personas con discapacidad física.	Además de lo visto, la predicción de palabras no es útil solamente para las personas con discapacidad física.
Puede ser utilizada también como ayuda para personas que tengan problemas para escribir correctamente: como apoyo para el aprendizaje de la escritura o para reducir el número de faltas de ortografía, por algún tipo de discapacidad lingüística, dislexia, etc. También puede ser utilizada por estudiantes de una segunda lengua, aumentando la calidad del texto generado, la riqueza de vocabulario e, incluso, la cantidad, por la confianza que produce en el usuario el apoyo gramatical y ortográfico, como se explica en [Carl97] y [Magn98].	Puede ser utilizada también como ayuda para personas que tengan problemas para escribir correctamente: como apoyo para el aprendizaje de la escritura o para reducir el número de faltas de ortografía, por algún tipo de discapacidad lingüística, dislexia, etc. También puede ser utilizada por estudiantes de una segunda lengua, aumentando la calidad del texto generado, la riqueza de vocabulario e, incluso, la cantidad, por la confianza que produce en el usuario el apoyo gramatical y ortográfico, como se explica en ciertos estudios.
Si se añade conocimiento lingüístico a la predicción de palabras, puede ayudar a personas con problemas para formar frases gramaticalmente correctas.	Si se añade conocimiento lingüístico a la predicción de palabras, puede ayudar a personas con problemas para formar frases gramaticalmente correctas.
La predicción de palabras está enmarcada en los estudios sobre Comunicación Alternativa y Aumentativa, encargada de todas las disciplinas destinadas a permitir la comunicación de personas con problemas: desde los sistemas más sencillos, utilizando gestos o lápiz y papel para comunicarse más fácilmente, hasta adaptaciones en la interfaz de los programas o el uso de sofisticadas técnicas de procesamiento de lenguaje natural para intentar predecir lo que se está intentando escribir.	La predicción de palabras está enmarcada en los estudios sobre Comunicación Alternativa y Aumentativa, encargada de todas las disciplinas destinadas a permitir la comunicación de personas con problemas: desde los sistemas más sencillos, utilizando gestos o lápiz y papel para comunicarse más fácilmente, hasta adaptaciones en la interfaz de los programas o el uso de sofisticadas técnicas de procesamiento de lenguaje natural para intentar predecir lo que se está intentando escribir.
En [Baum90], [Chur92], [Dema89], [Dema92], [Dema94], [Lang99b] y [McCo95] se muestran investigaciones sobre la aplicación de diversas disciplinas a la comunicación alternativa y aumentativa.	En algunos artículos se muestran investigaciones sobre la aplicación de diversas disciplinas a la comunicación alternativa y aumentativa.
En la actualidad, como se verá posteriormente, están disponibles varios sistemas de ayuda a personas con discapacidad que usan predicción de palabras.	En la actualidad, como se verá posteriormente, están disponibles varios sistemas de ayuda a personas con discapacidad que usan predicción de palabras.
La mayoría están diseñados para el idioma inglés, aunque algunos utilizan técnicas automáticas de entrenamiento para predecir en cualquier idioma.	La mayoría están diseñados para el idioma inglés, aunque algunos utilizan técnicas automáticas de entrenamiento para predecir en cualquier idioma.

— Continúa en la página siguiente —

Tabla B.1: Comparación textos TESIS vs. TESIS2 (cont.)

Texto original (TESIS)	Versión modificada (TESIS2)
La mayor parte de los sistemas de predicción disponibles en nuestro idioma han sido desarrollados en el Laboratorio de Tecnologías de Rehabilitación (LTR), en el cual se está realizando esta tesis doctoral.	La mayor parte de los sistemas de predicción disponibles en nuestro idioma han sido desarrollados en el Laboratorio de Tecnologías de Rehabilitación donde se está realizando esta tesis doctoral.
El segundo objetivo, disminuir el número medio de opciones que se deben recorrer en un sistema de barrido para escribir un texto, se puede resolver con un buen diseño de la interfaz de usuario, teniendo en cuenta la mayor cantidad posible de información.	El segundo objetivo es disminuir el número medio de opciones que se deben recorrer en un sistema de barrido para escribir un texto, se puede resolver con un buen diseño de la interfaz de usuario, teniendo en cuenta la mayor cantidad posible de información.
Además, es posible personalizar el programa, considerando el uso que hace cada usuario de cada opción.	Además, es posible personalizar el programa, considerando el uso que hace cada usuario de cada opción.
Las técnicas de diseño de interfaces adaptativas pueden modificarse y aplicarse a los programas orientados a usuarios con discapacidad física, mejorando la interacción entre hombre y máquina.	Las técnicas de diseño de interfaces adaptativas pueden modificarse y aplicarse a los programas orientados a usuarios con discapacidad física, mejorando la interacción entre hombre y máquina.
Con estos dos objetivos, se reducirá tanto el tiempo como el esfuerzo necesarios para que el usuario realice cualquier tarea que requiera acceso al ordenador: escribir un texto, comunicarse, controlar su entorno, etc.	Con estos dos objetivos, se reducirá tanto el tiempo como el esfuerzo necesarios para que el usuario realice cualquier tarea que requiera acceso al ordenador.
Predicción de palabras.	Predicción de palabras.
La predicción de palabras es uno de los métodos más utilizados para apoyar la escritura de personas con diversos problemas.	La predicción de palabras es uno de los métodos más utilizados para apoyar la escritura de personas con diversos problemas.
Sirve como ayuda para acelerar la escritura de las personas con discapacidades físicas y aquellas con problemas lingüísticos la utilizan para aumentar la calidad de sus textos, disminuyendo las faltas de ortografía y los errores gramaticales como se indica en [PAL00].	Ayuda a acelerar la escritura a las personas con discapacidades físicas y aquellas con problemas lingüísticos la utilizan para aumentar la calidad de sus textos, disminuyendo las faltas de ortografía y los errores gramaticales, como se indica en un artículo.
Los beneficios que se produzcan en ambos casos dependen de la calidad de la predicción de palabras, por lo cual es necesario utilizar los métodos más sofisticados a nuestro alcance, considerando la mayor cantidad posible de información a la hora de hacer las propuestas.	Los beneficios que se produzcan en ambos casos dependen de la calidad de la predicción de palabras, por lo que es necesario utilizar los métodos más sofisticados a nuestro alcance, considerando la mayor cantidad posible de información a la hora de hacer las propuestas.
Además del ámbito de la discapacidad, la predicción de palabras puede utilizarse en aplicaciones reconocimiento de voz (para restringir la lista de candidatas). En la actualidad hay versiones sencillas de predictores incluidas en algunos sistemas, como el Microsoft Word y algún modelo de teléfono móvil, para ayudar a acelerar la escritura.	La predicción de palabras puede utilizarse en aplicaciones reconocimiento de voz para restringir la lista de candidatas. En la actualidad hay versiones sencillas de predictores incluidas en algunos sistemas para ayudar a acelerar la escritura.
Encuadre científico-tecnológico.	Encuadre científico-tecnológico.
En este apartado se presenta el estado de la cuestión de diversos temas relacionados con la predicción de palabras.	En este apartado se presenta el estado de la cuestión de diversos temas relacionados con la predicción de palabras.

— Continúa en la página siguiente —

Tabla B.1: Comparación textos TESIS vs. TESIS2 (cont.)

Texto original (TESIS)	Versión modificada (TESIS2)
En la primera parte se explican una serie de modelos de lenguaje que pueden ser empleados en la predicción de palabras.	En la primera parte se explican una serie de modelos de lenguaje que pueden ser empleados en la predicción de palabras.
Se comienza con los métodos más sencillos, basados en estadísticas sobre secuencias de palabras, y se aumenta la potencia de los modelos, introduciendo información gramatical con diversos grados de complejidad.	Se comienza con los métodos más sencillos, basados en estadísticas sobre secuencias de palabras, y se aumenta la potencia de los modelos, introduciendo información gramatical con diversos grados de complejidad.
Posteriormente se muestran las consideraciones a tener en cuenta en la etapa de evaluación, y, por último, se describen un conjunto de sistemas comerciales de ayuda a personas con discapacidad lingüística o física, que integran la predicción de palabras como apoyo a la escritura.	Posteriormente se muestran las consideraciones a tener en cuenta en la etapa de evaluación, y se describen un conjunto de sistemas comerciales de ayuda a personas con discapacidad lingüística o física que integran la predicción de palabras como apoyo a la escritura.
Modelos de lenguaje.	Modelos de lenguaje.
La predicción de palabras se basa en un adecuado modelado del idioma, para lo que se pueden utilizar diversas técnicas a las que se dedican los próximos apartados.	La predicción de palabras se basa en un adecuado modelado del idioma, para lo que se pueden utilizar diversas técnicas a las que se dedican los próximos apartados.
Los modelos de lenguaje son formalismos que permiten expresar las relaciones entre las palabras.	Los modelos de lenguaje son formalismos que permiten expresar las relaciones entre las palabras.
A continuación, se presentan varios modelos que se pueden utilizar para predecir, basados en diferentes fuentes de información, desde temporal hasta probabilística y gramatical (con distintos grados de complejidad).	A continuación, se presentan varios modelos que se pueden utilizar para predecir, basados en diferentes fuentes de información desde temporal hasta probabilística y gramatical con distintos grados de complejidad.
Se puede encontrar más información en [Alle94], [Bunt96] y [Krul91], entre otros. En [Bert95], [Boek96], [Gara94a], [Gara94b], [Hunn85], [Hunn86], [Hunn87] y [Vand91] se explican otros modelos utilizados para predecir	Se puede encontrar más información en estos artículos. En estos artículos se explican otros modelos utilizados para predecir.
En algunos casos, como en [Hunn89], se ha intentado incluir también información semántica, pero para un vocabulario muy reducido, ya que es un método de predicción que precisa gran cantidad de información.	En algunos casos se ha intentado incluir información semántica, pero para un vocabulario muy reducido, ya que es un método de predicción que precisa gran cantidad de información.
Modelos de lenguaje estadísticos.	Modelos de lenguaje estadísticos.
Son los más sencillos conceptualmente y consideran como única información la frecuencia absoluta y/o relativa de las palabras y las secuencias de palabras.	Son los más sencillos conceptualmente y consideran como única información la frecuencia absoluta de las palabras y las secuencias de palabras.
En su versión más simple, modelo unigram , sólo se considera la frecuencia absoluta de las palabras.	En su versión más simple sólo se considera la frecuencia absoluta de las palabras.
Cuando se utiliza este modelo, en cada momento se predicen las palabras más frecuentes que comiencen por las letras escritas de la palabra en curso, sin aplicar ningún filtrado adicional.	Cuando se utiliza este modelo, en cada momento se predicen las palabras más frecuentes que comiencen por las letras escritas de la palabra en curso, sin aplicar ningún filtrado adicional.

Apéndice C

Conjunto de reglas de reescritura del sistema

A continuación se lista el conjunto de reglas que está siento utilizado en el módulo de predicción gramatical basado en la gramática independiente de contexto.

Los símbolos no terminales corresponden a las categorías gramaticales explicadas en la sección 3.2.1.4 (39):

Nemónico	Categoría correspondiente
v	verbo
adj	adjetivo
n	nombre
pron	pronombre
pronclit	clítico
det	determinante
numo	numeral ordinal
propre	pronombre que sigue a preposición
sprep	pronombres con preposición
prep	preposición
arco	artículo contracto
prepade	preposición a y de
adv	adverbio
conj	conjunción
interj	interjección
signo	signo
cjs	cajón de sastre
punto	punto

Tabla C.1: Tabla de correspondencia entre los nemónicos utilizados y las categorías.

Los nemónicos correspondientes con los nombres de los rasgos y sus valores se muestran en la siguiente tabla:

Nemónico	Rasgo correspondiente
gen	género
m	masculino
f	femenino
a	género de los nombres que comienzan por a tónica
num	número
s	singular
p	plural
per	persona
1	primera
2	segunda
3	tercera
tipo-v	tipo de verbo
cop	copulativo
mas_par	admite verbo siguiente en participio
mas_ger	admite verbo siguiente en gerundio
mas_inf	admite verbo siguiente en infinitivo
mas_a_inf	admite a continuación *a* más verbo en infinitivo
mas_que_inf	admite a continuación *que* más verbo en infinitivo
mas_de_inf	admite a continuación *de* más verbo en infinitivo
q_subj	admite a continuación *que* más verbo en subjuntivo
q_indi	admite a continuación *que* más verbo en indicativo
tipo-tv	tipo de terminación verbal
fl	flexivo
ger	gerundio
par	participio
inf	infinitivo
tv-fl-md	modo de verbo flexivo
ind	indicativo
sub	subjuntivo
tipo-a	tipo de adjetivo
a1	adjetivo antepuesto
a3	adjetivo pospuesto
tipo-n	tipo de nombre
pr	propio
tipo-pron	tipo de pronombre
pers	personal
npers	no personal

Nemónico	Rasgo correspondiente
st-cli	tipo de clítico
cl1	clítico que sólo puede aparecer en primera posición
cl2	clítico que sólo puede aparecer en segunda posición
t-prondet	tipo de pronombre o determinante
def	definido
ndef	no definido
rel	relativo
ard	determinado
card	cardinal
interg	interrogativo
t-npers-prep	tipo de pronombre
sin-prep	el pronombre no puede seguir a una preposición
tadv	tipo de adverbio
com	comparativo
gr	de grado
tadjsup	tipo de adjetivo (superlativo o no)
superl	superlativo

Tabla C.2: Tabla de correspondencia entre los nemónicos utilizados y los rasgos.

A continuación se presenta el conjunto de reglas del sistema.

```
S:
    SPrep ? 12
    signo            [sgte{,}] ? 2
    ElemPpioFrase ? 50
    signo            [sgte{,}] ? 2
    SAdvPpio ? 69
    SPrep ? 230
    signo            [sgte{,}] ? 2
    SSencilla
    & 1000 ;

Frase  [st-prondet{X}tipo-tv{M}tv-fl-md{Z}]:
    SNCoord          [st-prondet{X}] ? 300
    SNCoord ? 24
    SAdvCentral ? 38
    SFormaVerbalCl[tipo-tv{M}tv-fl-md{Z}]
    SAdvCentral ? 7
    SPrep ? 380
    SPrepSig ? 3
    SNCoord ? 720
    SNCoord ? 50
    SAdvCentral ? 2
```

```
        & 1000 ;

SRel:
        pron            [sgte{que}]
        Frase
        & 1000 ;

SSencilla   [st-prondet{X}]:
        Frase           [tipo-tv{fl}st-prondet{X}]
        signo           [sgte{,}] ? 2
        SSencilla ? 230
        & 804

    |   SN              [st-prondet{X}gen{M}num{Y}per{Z}] ? 500
        SFormaVerbalCop   [tipo-tv{fl}per{Z}num{Y}]
        SAdvCentral ? 5
        SAdj            [gen{M}num{Y}]
        SPrep ? 15
        signo           [sgte{,}] ? 2
        SSencilla ? 230
        & 14

    |   SN              [st-prondet{X}gen{M}num{Y}per{Z}] ? 500
        SFormaVerbalCop   [tipo-tv{fl}per{Z}num{Y}]
        SAdvCentral ? 5
        SNCoord         [gen{M}num{Y}]
        SPrep
        signo           [sgte{,}] ? 2
        SSencilla ? 230
        & 90

    |   Conjsub
        Frase           [tipo-tv{fl}tv-fl-md{sub}st-prondet{X}]
        signo           [sgte{,}] ? 2
        SSencilla ? 230
        & 1

    |   Conjind
        Frase           [tipo-tv{fl}tv-fl-md{ind}st-prondet{X}]
        signo           [sgte{,}] ? 2
        SSencilla ? 230
        & 14

    |   Enlace
        Frase
        signo           [sgte{,}] ? 2
        SSencilla ? 230
        & 104 ;

Enlace:
        conj
        & 990
```

```
|   cjs           [sgte{sino}]
    pron          [sgte{que}]
    & 2

|   prep          [sgte{sin}]
    cjs           [sgte{embargo}]
    signo         [sgte{,}] ? 50
    & 2

|   adv           [sgte{no}]
    cjs           [sgte{obstante}]
    signo         [sgte{,}] ? 50
    & 2

|   prep          [sgte{a}]
    cjs           [sgte{priori}]
    signo         [sgte{,}] ? 50
    & 2

|   prep          [sgte{con}]
    cjs           [sgte{respecto}]
    prep          [sgte{a}]
    SN            [t-npers-prep{~sin-prep}]
    signo         [sgte{,}] ? 50
    & 2 ;

SAdvPpio:
    SAdv
    signo         [sgte{,}] ? 10
    & 1000 ;

SAdvCentral:
    signo         [sgte{,}] ? 10
    SAdv
    signo         [sgte{,}] ? 10
    & 1000 ;

ElemPpioFrase:
    SSencilla     [st-prondet{rel}]
    & 529

|   SSencilla     [st-prondet{interg}]
    & 470

|   SFormaVerbalCl[tipo-tv{ger}]
    SAdvCentral ? 7
    SPrep ? 380
    SPrepSig ? 3
    SNCoord ? 720
    SNCoord ? 50
    & 10

|   SFormaVerbalCop   [tipo-tv{ger}]
```

```
        SAdj
        SPrep ? 250
        & 2

   |    SFormaVerbalCop   [tipo-tv{ger}]
        SNCoord
        SPrep ? 250
        & 2 ;

SAdv:
        adv             [tadv{gr}] ? 10
        adv             [lema{ ~muy ~tan ~no ~sí}]
        & 1000 ;

adjpart   [gen{X}num{Y}tadjsup{Z}tipo-a{M}]:
        adj             [gen{X}num{Y}tadjsup{Z}tipo-a{M}]
        & 990

   |    v               [tipo-tv{par}gen{X}num{Y}]
        & 10 ;

SAdj1   [gen{X}num{Y}tadjsup{Z}]:
        adjpart         [gen{X}num{Y}tadjsup{Z}tipo-a{~a1}]
        SPrep ? 4
        & 980

   |    adjpart         [gen{X}num{Y}tadjsup{Z}tipo-a{~a1}]
        conj            [sgte{y}]
        adjpart         [gen{X}num{Y}tadjsup{Z}tipo-a{~a1}]
        & 10

   |    adjpart         [gen{X}num{Y}tadjsup{Z}tipo-a{~a1}]
        conj            [sgte{o}]
        adjpart         [gen{X}num{Y}tadjsup{Z}tipo-a{~a1}]
        & 10 ;

SintTrasComo:
        SN
        & 333

   |    SAdj1
        & 333

   |    SPrep
        & 333 ;

PPioSAdj:
        pron            [sgte{mucho}]
        & 500

   |    det             [sgte{un}] ? 400
        pron            [sgte{poco}]
        & 500 ;
```

```
TrasPpio:
        adv             [sgte{más}]
        & 500

  |     adv             [sgte{menos}]
        & 500 ;

FinSAdj:
        pron            [sgte{que}]
        SintTrasComo
        & 500

  |     prepade         [sgte{de}]
        pron            [sgte{lo}]
        pron            [sgte{que}]
        Frase
        & 500 ;

SAdj   [gen{X}num{Y}]:
        adv             [tadv{gr}] ? 7
        SAdj1           [gen{X}num{Y}tadjsup{~superl}]
        & 519

  |     SAdj1           [tadjsup{superl}gen{X}num{Y}]
        & 479

  |     adv             [tadv{com}]
        SAdj1           [gen{X}num{Y}]
        conj            [sgte{como}]
        SintTrasComo
        & 1

  |     PPioSAdj
        TrasPpio
        SAdj1
        FinSAdj
        & 1

  |     det             [lema{el}gen{X}num{Y}]
        TrasPpio
        SAdj1
        prepade         [sgte{de}]
        SN              [t-prondet{def}]
        & 1 ;

SPrepde:
        prepade         [sgte{de}]
        SN              [t-npers-prep{~sin-prep}]
        & 1000 ;

SoloMismo   [gen{X}num{Y}]:
        adj             [lema{mismo}gen{X}num{Y}]
```

```
        & 500

    |   adj             [lema{solo}gen{X}num{Y}]
        & 500 ;

SPrepSig:
        conj            [sgte{y}]
        SPrep
        & 1000 ;

SPrep:
        prepade
        SN              [t-npers-prep{~sin-prep}]
        & 418

    |   prep
        SN              [t-npers-prep{~sin-prep}]
        & 442

    |   prepade         [sgte{de}]
        prep            [sgte{entre}]
        SN              [num{p}t-npers-prep{~sin-prep}]
        & 1

    |   prep            [sgte{entre}]
        SN
        conj            [sgte{y}]
        SN
        & 1

    |   prep            [sgte{para}]
        prep            [sgte{con}]
        SN              [t-npers-prep{~sin-prep}]
        & 1

    |   arco
        numo            [gen{m}num{s}]
        SPrepde ? 30
        & 3

    |   prep            [sgte{ ~según ~bajo ~cabe ~so ~hacia}]
        SFormaVerbal    [tipo-tv{inf}]
        & 55

    |   prepade
        adv             [sgte{no}] ? 1
        SFormaVerbal    [tipo-tv{inf}]
        & 80

    |   arco
        numo            [gen{m}num{s}] ? 1
        adj             [gen{m}num{s}tipo-a{~a3}] ? 17
        n               [gen{m}num{s}tipo-n{~pr}]
```

```
        adj             [gen{m}num{s}tipo-a{~a1}] ? 50
        adj             [gen{m}num{s}tipo-a{~a1}] ? 3
        SPrep ? 560
        SRel ? 136
        & 85

  |     arco
        adv             [sgte{no}] ? 1
        SFormaVerbal    [tipo-tv{inf}]
        & 3

  |     prepade
        propre
        SoloMismo       [num{s}] ? 50
        & 10

  |     prep            [sgte{ ~con}]
        propre
        SoloMismo       [num{s}] ? 50
        & 1

  |     sprep
        adj             [lema{mismo}num{s}] ? 50
        & 1 ;

SNSiguiente:
        signo           [sgte{,}]
        SN
        & 250

  |     conj            [sgte{y}]
        SN
        & 250

  |     conj            [sgte{e}]
        SN
        & 250

  |     conj            [sgte{o}]
        SN
        & 250 ;

SNCoord   [st-prondet{X}gen{M}num{Y}]:
        SN              [st-prondet{X}gen{M}num{Y}]
        SNSiguiente ? 25
        SNSiguiente ? 5
        SNSiguiente ? 5
        SNSiguiente ? 5
        & 1000 ;

Antenpr   [gen{X}]:
        cjs             [sgte{don}gen{m}gen{X}]
        & 250
```

```
|   cjs           [sgte{san}gen{m}gen{X}]
    & 250

|   cjs           [sgte{doña}gen{f}gen{X}]
    & 250

|   cjs           [sgte{santa}gen{f}gen{X}]
    & 250 ;

SN  [num{s}per{3}gen{X}]:
    Antenpr       [gen{X}] ? 1
    n             [tipo-n{pr}gen{X}]
    n             [tipo-n{pr}] ? 128
    n             [tipo-n{pr}] ? 127
    n             [tipo-n{pr}] ? 115
    & 15 ;

SN  [gen{m}num{s}per{3}t-prondet{def}st-prondet{ard}]:
    pron          [sgte{lo}]
    SAdj1         [gen{m}num{s}]
    & 5 ;

|   pron          [lema{el}]
    conj          [sgte{que}]
    Frase
    & 150 ;

|   det           [sgte{el}] ? 8
    v             [tipo-tv{inf}]
    SPrep ? 31
    adv ? 10
    & 146 ;

SN  [gen{f}num{s}per{3}st-prondet{M}t-prondet{Z}]:
    det           [gen{m}num{s}st-prondet{M}t-prondet{Z}] ? 165
    adv           [tadv{gr}] ? 10
    adj           [gen{m}num{s}tipo-a{~a3}] ? 20
    n             [gen{a}num{s}]
    SAdj1         [gen{f}num{s}tipo-a{~a1}] ? 21
    SRel ? 21
    & 6 ;

SN  [num{p}per{3}gen{X}t-prondet{Z}st-prondet{M}]:
    det           [t-prondet{def}num{p}st-prondet{M}t-prondet{Z}
                   gen{X}lema{ ~ambos}st-prondet{~card}] ? 5
    det           [st-prondet{card}num{p}gen{X}]
    numo          [num{p}gen{X}num{Y}] ? 1
    adj           [num{p}gen{X}tipo-a{~a3}] ? 14
    n             [num{p}gen{X}]
    SAdj1         [num{p}gen{X}tipo-a{~a1}] ? 23
    SPrep ? 21
    SRel ? 21
```

```
        & 55 ;

SN      [per{3}t-prondet{def}st-prondet{ard}gen{X}num{Y}]:
        det             [lema{todo}gen{X}num{Y}]
        SN              [per{3}t-prondet{def}st-prondet{ard}gen{X}num{Y}]
        & 1

  |     det             [lema{todo}num{p}gen{X}num{Y}]
        pron            [tipo-pron{pers}num{p}gen{X}per{M}t-npers-prep{Z}]
        & 1

  |     pron            [lema{el}gen{X}num{Y}]
        numo            [gen{X}num{Y}]
        SPrepde ? 500
        & 1

  |     pron            [lema{el}num{p}gen{X}num{Y}]
        det             [t-prondet{ndef}st-prondet{card}num{p}gen{X}num{Y}] ? 50
        numo            [num{p}gen{X}num{Y}]
        & 1 ;

SN      [per{3}gen{X}num{Y}st-prondet{Z}]:
        pron            [gen{X}num{Y}st-prondet{Z}tipo-pron{~pers}st-prondet{~ard}]
        & 47

  |     pron            [gen{X}num{Y}st-prondet{Z}lema{ ~qué ~cómo ~cuando}
                         tipo-pron{~pers}]
        SPrepde
        & 5 ;

SN      [gen{X}num{Y}per{M}t-npers-prep{Z}]:
        pron            [tipo-pron{pers}gen{X}num{Y}per{M}t-npers-prep{Z}]
        det             [t-prondet{ndef}st-prondet{card}gen{X}num{Y}] ? 50
        SoloMismo       [gen{X}num{Y}] ? 1
        & 50

  |     det             [gen{X}num{Y}st-prondet{Z}lema{ ~cuanto}] ? 527
        adv             [tadv{gr}] ? 10
        adj             [gen{X}num{Y}tipo-a{~a3}] ? 54
        n               [gen{X}num{Y}]
        adj             [gen{X}num{Y}tipo-a{~a1}] ? 163
        adj             [gen{X}num{Y}tipo-a{~a1}] ? 23
        SPrep ? 560
        signo           [sgte{,}] ? 2
        SRel ? 76
        & 700 ;

SNegaAfirm:
        adv             [sgte{no}]
        & 490

  |     adv             [sgte{tampoco}]
        & 4
```

```
|   adv            [sgte{jamás}]
    & 4

|   adv            [sgte{nunca}]
    & 4

|   adv            [sgte{sí}]
    & 4

|   adv            [sgte{ya}]
    & 4

|   adv            [sgte{también}]
    & 490 ;

SFormaVerbalCl   [tv-fl-md{X}num{Y}tipo-tv{M}]:
    SNegaAfirm ? 5
    pronclit       [st-cli{~cl2}] ? 159
    pronclit       [st-cli{~cl1}] ? 8
    SFormaVerbal   [tv-fl-md{X}num{Y}tipo-tv{M}]
    & 1000 ;

SFormaVerbalCop   [per{Z}num{Y}tipo-tv{M}]:
    v              [tipo-v{cop}num{Y}per{Z}tipo-tv{M}]
    & 995

|   v              [lema{haber}num{Y}per{Z}tipo-tv{M}]
    SFormaVerbal   [tipo-tv{par}gen{m}num{s}tipo-v{cop}]
    & 4 ;

SFormaVerbal   [tv-fl-md{X}num{Y}tipo-tv{M}]:
    v              [tipo-tv{M}tv-fl-md{X}num{Y}lema{ ~haber ~ser}]
    & 600

|   v              [lema{haber}tipo-tv{fl}num{s}per{3}tipo-tv{M}
                    tv-fl-md{X}num{Y}sgte{ ~ha}]
    & 9

|   v              [lema{haber}tipo-tv{M}tv-fl-md{X}num{Y}]
    SFormaVerbal   [tipo-tv{par}gen{m}num{s}]
    & 4

|   v              [lema{ser}tipo-tv{M}tv-fl-md{X}num{Y}]
    SFormaVerbal   [tipo-tv{par}num{Y}]
    & 7

|   v              [tipo-v{mas_de_inf}tipo-tv{M}tv-fl-md{X}num{Y}]
    prep           [sgte{de}]
    SFormaVerbal   [tipo-tv{inf}]
    & 1

|   v              [tipo-v{mas_a_inf}tipo-tv{M}tv-fl-md{X}num{Y}]
```

```
      prep            [sgte{a}]
      SFormaVerbal    [tipo-tv{inf}]
      & 50

|     v               [lema{consistir}tipo-tv{M}tv-fl-md{X}num{Y}]
      prep            [sgte{en}]
      SFormaVerbal    [tipo-tv{inf}]
      & 1

|     v               [tipo-v{q_subj}tipo-tv{M}tv-fl-md{X}num{Y}]
      conj            [sgte{que}]
      adv             [sgte{no}] ? 1
      SFormaVerbal    [tipo-tv{fl}tv-fl-md{sub}]
      & 1

|     v               [tipo-v{q_indi}tipo-tv{M}tv-fl-md{X}num{Y}]
      conj            [sgte{que}]
      adv             [sgte{no}] ? 1
      SFormaVerbal    [tipo-tv{fl}]
      & 1

|     v               [tipo-v{mas_que_inf}tipo-tv{M}tv-fl-md{X}num{Y}]
      conj            [sgte{que}]
      adv             [sgte{no}] ? 1
      SFormaVerbal    [tipo-tv{inf}]
      & 1

|     v               [tipo-v{mas_inf}tipo-tv{M}tv-fl-md{X}num{Y}]
      adv             [sgte{no}] ? 1
      SFormaVerbal    [tipo-tv{inf}]
      & 153

|     v               [tipo-v{mas_par}tipo-tv{M}tv-fl-md{X}num{Y}]
      SFormaVerbal    [tipo-tv{par}gen{m}num{s}]
      & 51

|     v               [tipo-v{mas_ger}tipo-tv{M}tv-fl-md{X}num{Y}]
      SFormaVerbal    [tipo-tv{ger}]
      & 151

|     v               [lema{haber}tipo-tv{M}tv-fl-md{X}num{Y}tv-fl-tp{~impf}]
      prep            [sgte{de}]
      SFormaVerbal    [tipo-tv{inf}]
      & 1 ;

Conjsub:
      prep            [sgte{a}]
      n               [sgte{fin}]
      prep            [sgte{de}]
      conj            [sgte{que}]
      & 10

|     prep            [sgte{con}]
```

```
    pron        [sgte{el}]
    n           [sgte{fin}]
    prep        [sgte{de}]
    conj        [sgte{que}]
    & 10

|   prep        [sgte{con}]
    pron        [sgte{el}] ? 500
    n           [sgte{objeto}]
    prep        [sgte{de}]
    conj        [sgte{que}]
    & 10

|   adv         [sgte{no}]
    v           [sgte{vaya}]
    prep        [sgte{a}]
    v           [sgte{ser}]
    conj        [sgte{que}]
    & 10

|   adv         [sgte{no}]
    v           [sgte{fuera}]
    prep        [sgte{a}]
    v           [sgte{ser}]
    conj        [sgte{que}]
    & 10

|   adv         [sgte{no}]
    v           [sgte{sea}]
    conj        [sgte{que}]
    & 10

|   adv         [sgte{no}]
    v           [sgte{fuera}]
    conj        [sgte{que}]
    & 10

|   prep        [sgte{con}]
    adv         [sgte{tal}]
    prep        [sgte{de}] ? 600
    conj        [sgte{que}]
    & 10

|   prep        [sgte{a}]
    n           [sgte{cambio}]
    prep        [sgte{de}]
    conj        [sgte{que}]
    & 10

|   prep        [sgte{a}]
    n           [sgte{condición}]
    prep        [sgte{de}]
    conj        [sgte{que}]
```

& 10

| prep [sgte{a}]
adv [sgte{menos}]
conj [sgte{que}]
& 10

| prep [sgte{a}]
adv [sgte{no}]
v [sgte{ser}]
conj [sgte{que}]
& 10

| prep [sgte{de}]
adv [sgte{no}]
v [sgte{ser}]
conj [sgte{que}]
& 5

| prep [sgte{a}]
adv [sgte{poco}]
conj [sgte{que}]
& 10

| prep [sgte{con}]
pron [sgte{el}] ? 50
n [sgte{propósito}]
prep [sgte{de}]
conj [sgte{que}]
& 30

| prep [sgte{de}]
adv [sgte{ahí}]
conj [sgte{que}]
& 30

| prep [sgte{de}]
adv [sgte{aquí}]
conj [sgte{que}]
& 30

| prep [sgte{en}]
pron [sgte{el}] ? 60
n [sgte{caso}]
prep [sgte{de}]
conj [sgte{que}]
& 30

| prep [sgte{en}]
pron [sgte{el}]
n [sgte{supuesto}]
n [sgte{caso}] ? 50
prep [sgte{de}]

```
    conj        [sgte{que}]
    & 30

|   prep        [sgte{en}]
    pron        [sgte{la}tipo-pron{npers}]
    n           [sgte{medida}]
    prep        [sgte{en}]
    conj        [sgte{que}]
    & 30

|   prep        [sgte{en}]
    n           [sgte{lugar}]
    prep        [sgte{de}]
    conj        [sgte{que}]
    & 30

|   prep        [sgte{en}]
    n           [sgte{vez}]
    prep        [sgte{de}]
    conj        [sgte{que}]
    & 30

|   adv         [sgte{excepto}]
    conj        [sgte{que}]
    & 30

|   adv         [sgte{salvo}]
    conj        [sgte{que}]
    & 30

|   adv         [sgte{ni}]
    conj        [sgte{que}]
    & 30

|   prep        [sgte{por}]
    adv         [sgte{mucho}]
    conj        [sgte{que}]
    & 30

|   prep        [sgte{por}]
    adv         [sgte{poco}]
    conj        [sgte{que}]
    & 30

|   prep        [sgte{por}]
    pron        [lema{mucho}num{X}gen{Y}]
    n           [num{X}gen{Y}]
    conj        [sgte{que}]
    & 30

|   prep        [sgte{por}]
    adv         [sgte{muy}] ? 10
    pron        [lema{poco}num{X}gen{Y}]
```

```
      n          [num{X}gen{Y}]
      conj       [sgte{que}]
      & 30

  |   prep       [sgte{por}]
      adv        [sgte{muy}] ? 10
      adv        [sgte{poco}] ? 100
      adj        [num{X}gen{Y}tipo-a{~a1}]
      conj       [sgte{que}]
      & 30

  |   adv        [sgte{siempre}]
      conj       [sgte{que}]
      & 30

  |   adv        [sgte{siempre}]
      conj       [sgte{y}]
      adv        [sgte{cuando}]
      & 30

  |   prep       [sgte{sin}]
      conj       [sgte{que}]
      & 30

  |   prep       [sgte{para}]
      conj       [sgte{que}]
      & 295 ;

Conjind:
      conj       [sgte{pues}]
      & 55

  |   v          [sgte{puesto}]
      conj       [sgte{que}]
      & 55

  |   v          [sgte{dado}]
      conj       [sgte{que}]
      & 55

  |   adv        [sgte{ya}]
      conj       [sgte{que}]
      & 105

  |   adv        [sgte{así}]
      conj       [sgte{pues}]
      & 55

  |   adv        [sgte{así}]
      conj       [sgte{que}]
      & 55

  |   adv        [sgte{aun}] ? 200
```

```
    prep        [sgte{a}]
    cjs         [sgte{sabiendas}]
    prep        [sgte{de}]
    conj        [sgte{que}]
    & 5

|   prep        [sgte{de}]
    n           [sgte{forma}]
    conj        [sgte{que}]
    & 105

|   prep        [sgte{de}]
    n           [sgte{manera}]
    conj        [sgte{que}]
    & 55

|   prep        [sgte{de}]
    n           [sgte{modo}]
    conj        [sgte{que}]
    & 55

|   prep        [sgte{por}]
    adj         [sgte{consiguiente}]
    & 5

|   prep        [sgte{por}]
    pron        [sgte{eso}]
    conj        [sgte{que}]
    & 5

|   prep        [sgte{por}]
    pron        [sgte{esto}]
    conj        [sgte{que}]
    & 5

|   prep        [sgte{por}]
    pron        [sgte{ello}]
    conj        [sgte{que}]
    & 5

|   prep        [sgte{por}]
    pron        [sgte{lo}]
    adv         [sgte{tanto}]
    & 105

|   conj        [sgte{porque}]
    & 155

|   conj        [sgte{si}]
    adv         [sgte{bien}]
    & 55

|   conj        [sgte{y}]
```

pron [sgte{eso}]
conj [sgte{que}]
& 55 ;

Apéndice D

Cartas escritas por usuarios de *Predice*

A continuación se muestran algunas de las cartas enviadas por usuarios del editor predictivo *Predice*.

Zaragoza 22 de Octubre de 2000

Sira Elena Palazuelos Cagigas
Departamento de Ingeniería Electrónica
Escuela Técnica Superior de Ingenieros
de Telecomunicación
Ciudad Universitaria s/n
28040 MADRID

Srt. Sira Elena Palazuelos :

Soy Montse Pellegero Ponsa, afectada de Esclerosis Lateral Amiotrofica (desde hace trece años) y usuaria del Programa EDITOR PREDICTIVO de ustedes, desde hace más de cuatro años, con el cual estoy encantada, porque hace diez años que perdí el habla, y todo el día estoy trabajando con el. Lo uso, para comunicarme, y escribir, tengo infinidad de frases, palabras y encabezamientos largos de escrituras, grabadas en los macros, y expresiones, las cuáles me dan mucha rapidez, para pedir las cosas, y escribir.

Me vuelvo a dirigir a usted, para que me informen si ustedes han hecho algún otro programa más moderno que el tengo yo. Ya que a mí me gustaría que tener voz, porque con el programa que me mando usted hace dos años, muy similar al EDITOR PREDICTIVO, tampoco conseguimos que hablará.

Reciba un cordial saludo, y perdone la molestia.

MONTSE PELLEGERO

Montse Pellegero Ponsa
C/J. Luis Pomaron 11, Residencia D. F. A.
(entrada C/Andres Gurpide, frente al nº. 7)
50008 ZARAGOZA

Zaragoza 25 de Enero 1999

Sira Elena Palazuelos Cagigas
Departamento de Ingeniería Electrónica
Escuela Técnica Superior de Ingenieros
de Telecomunicación
Ciudad Universitaria s/n
28040 MADRID

Srt. Sira Elena Palazuelos :

Soy Montse Pellegero Ponsa, afectada de Esclerosis Lateral Amiotrofica (desde hace once años) y usuaria del Programa EDITOR PREDICTIVO de ustedes, desde hace dos años y medio, con el cual estoy muy familiarizada, porque hace ocho años que perdí el habla, y todo el día estoy trabajando con el. Lo uso para comunicarme, y escribir, tengo infinidad de frases, palabras y encabezamientos largos de escrituras, grabadas en los macros, y expresiones, las cuáles me dan mucha rapidez, para pedir las cosas, y escribir.

El motivo de dirigirme a usted, es porque deseo que me mande el presupuesto de la tarjeta VISHA de voz, ya que es lo único que me falta, para completar el programa.

Reciba un cordial saludo.

MONTSE PELLEGERO PONSA

Montse Pellegero Ponsa
C/J. Luis Pomaron 7 (Residencia D. F. A.)
50001 ZARAGOZA

Zaragoza 19 de Diciembre de 2000

Sira Palazuelos
Dpto. Ingeniería Electrónica
Escuela Técnica Superior de Ingenieros de Telecomunicación
Ciudad Universitaria S/N
28040 MADRID

Estimada Sira :

Te estoy muy agradecida por lo amable que has sido con mi hijo y conmigo, y por la ayuda que me das.

Te hago una valoración de mi experiencia sobre el Programa EDITOR PREDICTIVO que tengo, porque aún no he tenido tiempo de valorar a fondo la nueva versión de Windows que me mandaste tú. Es un poco distinto para mí, ya que tengo 52 años y por mi minusvália, yo no he podido manejar el teclado de un programa normal de ordenador. Pero con las mejoras que habéis hecho, pronto me familiarizare con el, porque es fenomenal que se vea toda la página que escribes entera, que tengas ayudas de información, calculadora, matemáticas, y que con el pulsador se puedan grabar los ficheros en disquetes, y otras mejoras más que me falta descubrir. Es decir, habéis hecho un trabajo fabuloso con vuestro esfuerzo.

Te adjunto mi valoración sobre el EDITOR PREDICTIVO que tengo.

Recibe un fuerte abrazo, y te deseo mucha suerte con la tesis doctoral.

MONTSE PELLEGERO

Montse Pellegero Ponsa
C/J. Luis Pomaron 11, Residencia D. F. A.
(entrada C/Andres Gurpide, frente al nº. 7)
50008 ZARAGOZA

Zaragoza 19 de Diciembre de 2000

Estimada Sira :

Hace más de 13 años que padezco Esclerosis Lateral Amiotrofica (ELA enfermedad que padece Stephen Hawking). Cuando enferme yo era una mujer muy activa de 39 años (ahora tengo 52) estaba casada, tenia dos hijos de 8 y 13 años, y trabajaba de peluquera de señoras. Pero cuando perdi la movilidad de las manos, piernas, y el habla (hace diez años) me hundí, porque yo era muy habladora, y entonces me comunicaba con un abecedario, que es un sistema muy lento de comunicación, y no me podia expresar cómo quería (ahora solo lo uso para hablar cosas sueltas, cuando no tengo el ordenador). Más tarde utilice un Comunicador (CANON) con el cual note bastante autonomía de comunicación porque mi hermano que es ingeniero (ésta en Barcelona) me hizo un pulsador muy sensible (lo manejo con un poco de movilidad que tengo en tres dedos) y también me acoplo una pequeña pantalla que tiene cuatro documentos con memoria, y con la impresora puedo pasar los escritos a folios, porque con el Comunicador CANON solo se puede escribir en una estrecha tira de papel. Ahora solo lo uso para comunicarme cuando salgo de casa, porque desde el año 1997 con el EDITOR PREDICTIVO tengo una independencia total de comunicación, y de escritura, porque es un programa especial, con el cual yo puedo manejar el ordenador con un pulsador (uso el mismo pulsador que me hizo mi hermano, para el CANON). Paso más de diez horas al día con el EDITOR PREDICTIVO, y le saco un redimiento tremendo, tengo más de 470 expresiones grabadas en nombres frases cortas, nombres y palabras, y 58 macros, con direcciones, teléfonos encabezamientos y terminaciones, de cartas, articulos, etc. y unas frases larguísimas (para facilitarles ayuda a las personas que me atienden) con lo cuál gano muchísima rapidez de comunicación, para hablar con la gente, y escribir. También escribo articulos en algún periódico de vez en cuando, y en las revistas ADELA de Madrid (asociación de ELA) y en ZANGALLETA (asociación de Disminuidos Físicos de Aragon).

Fallos que le encuentro al EDITOR PREDICTIVO que tengo :

1º. Fallo --- Con el pulsador yo puedo abrir los ficheros que quiero sin embargo no puedo borrar los titulos de ellos. Es decir, para borrar el contenido de un fichero, tengo que borrarlo yo con el pulsador, y seguidamente tengo que escribir una palabra cualquiera, y grabarla, entonces ya tengo el fichero vacio, y el día que quiero borrar el titulo, mi familia me lo borra con el teclado.

2º. Fallo --- El signo más de sumar se escribe normal, pero no se graba ni en el fichero, ni en el folio.

Mejorías que me gustaría que le hicieran al EDITOR PREDICTIVO :

1º. --- Que con el pulsador se pudiera mandar correo electrónico a ordenadores, y a teléfonos móviles. Esto sería fenomenal, para que las personas que no tenemos habla, ni movilidad, nos pudieramos comunicar con el exterior sin que nadie nos tenga que poner el papel de la impresora, y sin que nos tengan que escribir los sobres de las cartas.

2º. --- Que tuviera corregidor de faltas de ortografía.

3º. --- Que se pudiera poner el margen que quisieras, al imprimir en el folio.

4º. Que se pudiera jugar con algún juego, para distraernos en los ratos libres, porque las otras dos mejoras que yo ambiezaba, eran la calculadora, y que con el pulsador se pudieran grabar los ficheros en disquetes, que ya lo habéis puesto en la nueva versión de Windows.

Recibe un fuerte abrazo, y gracias por tú ayuda.

MONTSE PELLEGERO

D. SANTIAGO AGUILERA

ESCUELA SUPERIOR INGENIEROS TELECOMUNICACIONES
Avda. Ciudad Universitaria, s/n
Edificio B, despacho 111

28040 - MADRID

Madrid, 29 de Mayo de 1.996

Muy señor mío:

Me puse en contacto con Vd. a raiz de vez en AULA-96, un editor de textos "PREDICE" en el Stand del Ministerio, que sería de gran utilidad para nuestra hija.

Su respuesta fué inmediata, lo cual le agradecemos sinceramente, ya que nos facilitó los diskettes del programa, que fué a recoger a su despacho, mi marido Pedro Antonio Peña Susaeta.

La niña ha mejorado de manera espectacular en su escritura, sobre todo en rapidez y agilidad, ya que el manejo de los diversos menús es muy asequible para ella, lo cual la ha ido estimulando cada vez más y se encuentra muy independiente para realizar sus tareas.

No se puede hacer idea de la ayuda que nos ha prestado con su programa, ya que es muy versátil y de fácil ejecución para estos niños que están tan disminuídos en sus facultades. Esto es lo que necesitábamos, por ello le doy muy sinceramente las gracias y le felicito por haber tenido en cuenta a este sector de la sociedad que está tan necesitado de ideas y ayudas como la que Vd. ha elaborado, y le animo a que siga investigando es este campo, que sé que es una tarea ardua, pero cientos de padres le estaremos infinitamente agradecidos, ya que les da a nuestros hijos unas posibilidades que nunca soñaron conseguir.

Es un editor muy completo y no se me ocurre, por el momento ninguna objección que poner, ni posibilidades de ampliación, ya que está perfecto. Unicamente, como una pequeña

sugerencia, le comentaré que estamos en contacto con compañeros suyos de otra Universidad española, que están trabajando en un emulador de ratón y de esa forma, poder acceder a Windows, por lo que lo único que se me ocurre, es que se pudiera acceder al "PREDICE" a través de un Icono, sin tener que escribir desde c:/PREDICE, para entrar, ya que entonces la autonomía sería total, pero reconozco que esta sugerencia es totalmente secundaria, ya que no cuesta ningún trabajo ponerlo y entrar en el Editor y la niña ponerse a trabajar.

Muy agradecidos por su ayuda y amabilidad, le enviamos nuestro más cordial saludo.

PALMIRA CORTES PEREZ

Nueva Dirección:

c/Mirlo,7
28760 - TRES CANTOS - MADRID

Tef. 804.42.35

AUXILIA

AUXILIA
POR LA INTEGRACIÓN DEL DISMINUIDO FÍSICO EN LA SOCIEDAD
PER LA INTEGRACIÓ DEL DISMINUÏT FISIC EN LA SOCIETAT

C/ Progreso, 126 - Baix - Telèfon 372 50 82 46011 VALÈNCIA

Juan Lestani
ETSI Telecomunicación
Dpto. Ingeniería Electrónica
Ciudad Universitaria s/n
28040 MADRID

Valencia, 22 de mayo de 1996

Estimado Juan:

Hace ya dos años que recibimos vuestro programa "Predice" de escritura predictiva que tiene la posibilidad de ser controlado a través de un solo pulsador. Desde entonces algunos beneficiarios de nuestra asociación han podido aprovecharse de las posibilidades del programa.

El motivo de la presente carta es para agradeceros la donación del programa en nombre de la asociación Auxilia de Valencia, y comunicaros el aprovechamiento que de éste venimos haciendo.

Inicialmente Joaquín Ferre empezó a utilizarlo a través de un pulsador de barbilla que le diseñaron y construyeron en el Centro Estatal de Autonomía Personal y Ayudas Técnicas (CEAPAT). Aunque los avances son lentos, el programa es para Joaquín uno de los pocos instrumentos que le sirven para poder comunicarse.

Vistos los resultados con Joaquín, este curso otro de nuestros beneficiarios, Julián Triguero, también ha empezado a utilizarlo con un pulsador de barbilla construido esta vez por uno de los voluntarios de la asociación. Julián emplea el programa para escribir documentos.

Atentamente,

Salvador Añó
responsable Taller de Informática

Marisa Vidal
Coordinador Auxilia Valencia

Bibliografía

[**Acce00**] Access Writer. http://snow.utoronto.ca/cgi/tad/showdev.cgi. Febrero de 2000.

[**AENO98a**] AENOR. Norma UNE 139801:1998 EX. "Informática para la salud. Aplicaciones informáticas para personas con discapacidad. Requisitos de accesibilidad de las plataformas informáticas. Soporte físico". 1998.

[**AENO98b**] AENOR. Norma UNE 139802:1998 EX. "Informática para la salud. Aplicaciones informáticas para personas con discapacidad. Requisitos de accesibilidad de las plataformas informáticas. Soporte lógico". 1998.

[**Agui99**] Aguilar, A. y Saumenll, C. "Adaptaciones del ordenador para facilitar la inclusión escolar de un alumno con parálisis cerebral de 12 años de edad". Comunicación y Pedagogía, 162. Páginas 55-58. 1999.

[**Aïtm95**] Aït-Mokhtar, S., Rodrigo-Mateos, J. L. "Segmentación de textos y análisis morfológico de textos en español con el sistema SMORPH". SEPLN nº 17. Páginas: 29-41. 1995.

[**Alfo90**] Alfonseca, M., Sancho, J., Martínez Orga, M. "Teoría de Lenguajes, Gramáticas y Autómatas". Universidad y Cultura. 1990.

[**Alle94**] Allen, J. "Natural language Understanding". Benjamin/Cummings Publishing Company, Inc. 2ª Ed. 1994.

[**Auro99**] Aurora. http://www2.edc.org/NCIP/library/wp/Aurora.html. Junio de 1999.

[**Auro01**] Aurora. http://www.djtech.com/Aurora/info/compare.html. Enero de 2001.

[**Basi98**] Basil, C. Soro-Carnats, E. y Rosell, E. "Sistemas de signos y ayudas técnicas para la comunicación aumentativa y la escritura: Principios teóricos y aplicaciones". 1ª edición. ISBN 844580716-1. Grupo Masson. 1998.

[**Bate78**] Bates, M. "The Theory and Practice of Augmented Transition Network Grammars". Natural Language Communication with Computers. Nueva York. Springer. Páginas: 191-259. 1978.

[**Baum90**] Baumgart, D., Johnson, J., and Helmesteter, E. "Augmentative and Alternative Communication Systems for persons with Moderate and Severe Dissabilities". Baltimore: Brookes. 1990.

[Berg00] Bergman, E., Johnson, E. "Towards Accessible Human-Computer Interaction". http://www.sun.com/tech/access/updt.HCI.advance.html. 2000.

[Bert95] Bertenstam, J., Hunnicutt, S. "Including Grammatical Information in word prediction" (Traducción de "Användning av grammatik y ordpredicering"). Research Conference on Människor-handikapp-livsvillkor (Man-handicap-conditions of life). Habiliteringsförvaltningen. Örebro. Suecia. 1995.

[Boek96] Boekestein, M. "MA Thesis: Word Prediction". Dept. of Language and Speech. Katholieke Universiteit Nijmegen. Agosto de 1996.

[Bunt96] Bunt, H., Tomita, M. "Recent Advances in Parsing Tecnology". Ed.: Kluwer Academic Publishers. 1996.

[Cald87] Calder, J. "Typed Unification for natural language processing" en Klein E. & J. van Benthem (eds) "Categories, Polymorphism and Unification". 1987.

[Cald88] Calder J., Klein, E. & Zeevat H. "Unification Categorial Grammar: A Concise, Extendable Grammar For Natural Language Processing". Proceedings of the 12th International Conference of Computational Linguistics and the 24th Annual Meeting of the Association for Computational Linguistics. Budapest. 1988.

[Cand97] Candelas Arnao, A., Lobato Galindo, M. "Guía de acceso al ordenador para personas con discapacidad". Editado por el Ministerio de Trabajo y Asuntos Sociales. ISBN: 84-88986-71-8. 1997.

[Carl97a] Carlberger, J. "Design and Implementation of a Probabilistic Word Prediction Program". Master Thesis. http://www.speech.kth.se/~johanc/thesis/thesis.html. Estocolmo (Suecia). 1997.

[Casa89] Casado, Enríquez, E.V. "The Spanish Category System". Informe final Esprit-860. Vol I. Sección 1.2.1.1. UN-CAT1588. 1989.

[Cher94] Cherry, A., Hawley, M., Freeman, M., Cudd, P. "Human-Computer Interfacing for the Severely Physically disabled´´. Eds: Zagler, W.L., Busby, G. Wagner, R.R. "Computers for Handicapped Persons". Proceedings of the 4th International Conference ICCHP'94. Vienna. Austria. Septiembre de 1994.

[Chur92] Church, G., Glennen, S. "The Handbook of Assistive Technology". Chapman and Hall. Londres. 1992.

[Clar97] Clarkson, P. y Rosenfeld, R. "Statistical language modeling using the CMU-Cambridge toolkit". Proceedings de EUROSPEECH'97. Páginas: 147-148. 1997

[Cola99] Colás, J. "Estrategias de incorporación de conocimiento sintáctico y semántico en sistemas de comprensión de habla continua en castellano". Tesis Doctoral. ETSI Telecomunicación Universidad Politécnica de Madrid. 1999.

[Coll96] Collins, M. "A New Statistical Parser Based on Bigram Lexical Dependencies". Proceedings de la 34ª Reunión anual de la Association for Computational Linguistics. 1996.

[**Cope98**] Copestake, A., Flickinger, D. "Evaluation of NLP technology for AAC using logged data". ISAAC Research Symposium: Natural Language Processing and AAC. Dublin. 1998.

[**CoWr00**] CoWriter. Junio de 2000. http://www.ndcd.org/ndcpd/people/staff/fifield/littech/tools/software/cowriter.htm.

[**CoWr97**] CoWriter. http://www2.edc.org/NCIP/library/wp/Cowriter.htm. Septiembre de 1997.

[**Dema89**] Demasco, P., McCoy, K., Gong, Y., Pennington, C., Rowe, C. "Towards more intelligent AAC interfaces: The use of natural language processing". RESNA Press. 1989. http://www.asel.udel.edu/nli/pubs/1989/DemaMcCo89.txt.

[**Dema92**] Demasco, P., McCoy, K. "Generating text from compressed input: An intelligent interface for people with severe motor impairments". Comms. of the ACM. Volumen 35. Nº 5. http://www.asel.udel.edu/nli/pubs/1992/DemaMcCo92.txt. Mayo de 1992.

[**Dema94**] Demasco, P. "Human Factors Considerations in the Design of Language Interfaces in AAC". Assistive Technology. Volumen 6.1 Páginas: 10-25. 1994.

[**Donn95**] Donnelly, C., Stallman, R. "Bison. The YACC compatible Parser Generator. Bison Version 1.25". http://www.gnu.org/manual/bison/html_mono/bison.html. Noviembre de 1995.

[**Donn99**] Donnelly, C., Stallman, R. "Bison. The YACC compatible Parser Generator. Bison Version 1.27". http://www.delorie.com/gnu/docs/bison/bison_toc.html. 12 de Febrero de 1999.

[**DRAE00**] Diccionario de la Real Academia de la Lengua. 2000. http://www.rae.es/nivel1/buscon/AUTORIDAD2.HTM

[**Eagl98**] EAGLES Group. "Evaluation of Natural Language Processing Systems". Final Report. Septiembre de 1995.

[**Earl70**] Earley, J. "An Efficient Context-Free Parsing Algorithm". Comm. of the ACM. Volumen 13. Nº 2. Páginas: 94-102. Febrero de 1970.

[**Esco99**] Escoin, J. "PARLA y TaP: Programa de aceleración del rendimiento en lenguaje asistido y síntesis de voz en castellano". Comunicación y Pedagogía, 162. Páginas 67-73. 1999

[**Ferr96**] Ferreiros, J. "Aportación a los métodos de entrenamiento de Modelos de Markov para reconocimiento de habla continua". Tesis Doctoral. ETSI Telecomunicación Universidad Politécnica de Madrid. 1996

[**Gara94a**] Garay-Vitoria, N., González-Abascal, J. "Application of Artificial Intelligence Methods in a Word-Prediction Aid". Eds: Zagler, W.L., Busby, G. Wagner, R.R. "Computers for Handicapped Persons". Proceedings of the 4th International Conference ICCHP'94. Vienna. Austria. Septiembre de 1994.

[Gara94b] Garay-Vitoria, N., González-Abascal, J. "Using statistical and syntatic information in word prediction for input speed enhancement". Basque international workshop on information technology. BIWIT. Biarritz. Francia. 1994.

[Garc00] García Hernández, A. "Implementación de mecanismos adaptativos en la interfaz de usuario de un editor de texto para discapacitados motrices". Proyecto Fin de Carrera. Tutora: Sira E. Palazuelos Cagigas. E.T.S.I. de Telecomunicación. Dpto. de Ingeniería Electrónica. Septiembre de 2000.

[Gibb98] Gibbon, D., Moore, R. & Winsky, R. Eds. "Spoken Language Systems Assessment" (Part of the "Handbook of Standards and Resources for Spoken Language Systems". Volume III). Mouton de Gruyter. Berlin. 1998.

[Gill89] Gillick, L., Coz., S. "Some Statistical Issues in the Comparison of Speech Recognition Algorithms". IEEE International Conference on Acoustics. Speech and Signal Processing. Vol. S1, pp. 532-535. 1989.

[Gome94] Gómez, J. M., Goñi, J. M., González., J. C. "Un analizador sintáctico para gramáticas asociativas por la izquierda". Actas del X Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural. SEPLN 94. Córdoba. Julio de 1994.

[Goñi97] Goñi, J.M., González, J.C. y Montero, A. "ARIES: A lexical platform for engineering Spanish processing tools". Natural Language Engineering 3(4). Páginas: 317-345. Cambridge University Press. 1997.

[Gram96] "Gramática del español correcto". Santillana. ISBN: 84-294-5079-3. 1996.

[Haus87] Hausser, R. "Left-associative Grammar. Theory and implementation". CMU-CMT-87-104. 3 de Junio de 1987.

[Hunn85] Hunnicutt, S. "Lexical Prediction for a Text-to-Speech System". STL-QPSR 2-3/1985. KTH. Estocolmo. Páginas: 47-55. 1985.

[Hunn86] Hunnicutt, S., Neovius, L. "A lexical prediction system". Proceedings of the ICASSP 86. TOKYO. 1986.

[Hunn87] Hunnicutt, S. "Input and Output Alternatives in Word Prediction". STL-QPSR 2-3/1987. KTH. Estocolmo. Páginas: 15-29. 1987.

[Hunn89] Hunnicutt, S. "Using syntactic and Semantic Information in a Word Prediction Aid". Proceedings of the Eurospeech 89. Volumen 2. Paris. Páginas: 191-193. 1989.

[Hunt90] Hunt, M. J. "Figures of Merit for Assesing Connected-Word Recognisers". Speech Communication. Volumen 9. Páginas: 329-336. 1990.

[IBM00] IBM. Software Accessibility. http://www-3.ibm.com/able/accesssoftware.html. 2000.

[Jeli91] Jelinek, F. "Self-organized language modeling for speech recognition". Páginas: 450-506. 1990. En Waibel. A., Lee, K. (eds): "Reading in Speech Recognition". Morgan Kaufman Publishers. 1991.

[**Jeli99**] Jelinek, F. y Chelba, C. "Putting Language Into Language Modeling". Keynote speech 1 en EUROSPEECH'99. Página KN-1. 1999

[**Katz87**] Katz, S. "Estimation of probabilities from sparse data for the language model component of a speech recognizer". IEEE Transactions on Acoustics, Speech and Signal Processing, Marzo 1987, nº 35, pp. 400-401.

[**Kong95**] Kong Joo Lee et all. "A Robust Parser Based on Syntactic Information". Proceedings of the 7th Conference of the EACL. Dublin. Ireland. Páginas: 223-228. 1995.

[**Krul91**] Krulee, G. K. "Computer processing of natural language". Prentice Hall Inc. 1991.

[**Kugl89**] Kugler, "Unification of the word classes of the ESPRIT Project 860". Informe final Esprit-860. Vol I. Sección 1.2.1.2. BU-WKL-0376. 1989.

[**Lang97**] Langley, P. "Machine Learning for Adaptive User Interfaces". Proceedings of the 21st German Annual Conference on Artificial Intelligence. Páginas: 53-62. Freiburg. Alemania. Springer. http://www.isle.org/~langley/adapt.html. 1997.

[**Lang98**] Langley, P., Fehling, M. "The Experimental Study of Adaptive User Interface". Technical Report 98-3. Institute for the Study and Expertise. Palo Alto. CA. http://www.isle.org/~langley/adapt.html. 1998.

[**Lang99a**] Langley, P. "User Modeling in Adaptive Interfaces". Proceedings of the 7th International Conference on User Modeling. Banff. Alberta. Springer. Páginas 357-370. http://www.isle.org/~langley/adapt.html. 1999.

[**Lang99b**] Langer, S., Hunnicutt, S., and Hickey, M. "Languaje processing techniques and resources for communication aids". En "Augmentative and Alternative Communication: New Directions in Research and Practice". Eds: Filip T. Loncke, John Clibbens, Helen H. Arvidson, Lyle L. Lloyd. ISBN: 1 86156 143 1. Páginas: 77-83. Editorial: Whurr Publishers. Londres. 1999.

[**Lawr98**] Lawrence, P. R. "EZ Keys For Windows Uniquely Satisfies AAC and Other Assistive Technology Needs". CSUN 98. http://www.dinf.ch/csun_98/csun98_141.htm. 1998.

[**Lehm96**] Lehmann, S. Oepen, S., "TSNLP Test Suites for Natural Language Processing" in COLING-96. Proceedings of the 16th International Conference on Computational Linguistics. Copenhague. Dinamarca. 5-9 de Agosto de 1996.

[**Lesh98**] Lesher, G. W., Moulton, B. J., Higgimbotham, J., "Techniques for Augmenting Scanning Communication". AAC Augmentative and Alternative Communication. Volume 14. Páginas: 81-101. Junio de 1998.

[**Lowe80**] Lowerre, B. y Reddy R. "The HARPY Speech Understanding System". En "Trends in Speech Recognition", W. Lea, editor. Páginas 340-360. Prentice Hall. 1980.

[Mage91] Magerman, D. y Marcus, M. "Pearl: A Probabilistic Chart Parser". Proceedings de la European ACL Conference. 1991.

[Magn97a] Magnuson, T. "Word Prediction as Linguistic Support for Individuals with Reading and Writing Difficulties". En "The European Context for Assistive Technology". Ed: Placencia Porrero, I., Puig de la Bellacasa, R. IOS Press, Ohmsha. Proceedings of the 2nd TIDE Congress. París. 26-28 de Abril de 1995.

[Magn97b] Magnuson, T. "Profet II, a New Generation of Word Prediction: An Evaluation Study" Proceedings of the AAATE Conference. ISBN: 90 5199 361 7. Páginas: 153-157. Ed. IOS Press(George Anogianakis Et All). Tesalónica. Grecia. 29 Septiembre - 2 Octubre de 1997.

[Magn98] Magnuson, T. "Linguistic evaluation of Profet II: a pilot project". Proceedings ISAAC Dublin. Páginas: 479-480. Ed: Ashfield Publications. 1998.

[Mart96] Martin, S. "Effective Visual Communication for Graphical User Interfaces". www.cs.wpi.edu/~matt/courses/cs563/talks/smartin/int_design.html. Junio de 1996.

[Mart99] Martin, S., Hamacher, C., Liermann. J., Wessel, F., Ney. H. "Assessment of Smoothing Methods and Complex Stochastic Language Modeling". Proceedings on CD-ROM of the Eurospeech 99. Volumen 5. Páginas: 1939-1942. Budapest (Hungría). 1999.

[Meri98] Merino Torres, F. "Implementación de una herramienta para procesamiento de lenguaje natural en entorno Windows". Proyecto Fin de Carrera. Tutora: Sira E. Palazuelos Cagigas. Departamento de Ingeniería de Circuitos y Sistemas de la E.U.I.T. de Telecomunicación. Octubre 1998.

[McCo95] McCoy, K. et al. "Some applications of Natural Language Processing to the Field of Augmentative and Alternative Communication". Technical Report. Computer and Information Sciences Department And Applied Science and Engineering Laboratories. University of Delaware/A.I. duPont Institute. Newark. DE 19617. http://www.asel.udel.edu/nli/pubs/1995/McCoDema95.txt

[Micr99] "Microsoft Windows Guidelines for Accessible Software Design". Diciembre de 1999. http://www.microsoft.com/enable/dev/guidelines/software.htm.

[Mill56] Miller, G. A. "The magic number seven, plus or minus two: some constraints on our capacity for processing information." Psychological Review. Volumen 63. Páginas: 81-97. 1956.

[Morr92] Morris, C., Newell, A. F., Booth, L., Rickets, I. W., Arnott, J. L. "Syntax PAL: A System to improve the Written Syntax of Language-Impaired Users". Assistive Technology. Volumen 4. Páginas: 51-59. 1992.

[Morr98] Morrison, A., Martin, A. "Evaluating the Effectiveness of Word Prediction". Proceedings of the ISAAC 98. Dublin. Irlanda. Páginas: 230-231. 24-27 de Agosto de 1998.

[**Nada84**] Nadas, A. "Estimation of probabilities in the language model of the IBM speech recognition system". IEEE Transactions on Acoustics, Speech and Signal Processing. Agosto 1984, nº 32, pp. 859-861.

[**Newe92**] Newell, A. F., Arnott, J. L., Booth, L., Beattie, W., Brophy, B., Rickets, I. W. "Effect of the PAL Word Prediction System on the Quality and Quantity of Text Generation". Augmentative and Alternative Communication. Volumen 8. Nº 4. (Decker Periodicals Inc. Ontario. Canada (ISSN 0743-4618)). Páginas: 304-311. Diciembre de 1992.

[**Ney91**] Ney, H., Essen, U. "On Smoothing Techniques for Bigram-Based Natural Language Modelling". Proceedings of ICASSP 91. Toronto. Páginas: 825-828. Mayo de 1991.

[**Ney92**] Ney, H., Mergel, D., Noll, A., and Paesler, A. "Data driven search organization for continuous speech recognition". IEEE Transactions on Signal Processing, volumen 40(2). Páginas 272-281. 1992.

[**Ney94**] Ney, H., Essen, U. y Kneser, R. "On Structuring Probabilistic Dependencies in Stochastic Language Modeling". Computer Speech and Language, volumen 8(1). Páginas: 1-28. 1994

[**Nies98**] Niesler T.R., Whittaker E.W.D. y Woodland P.C. "Comparison of Part-Of-Speech and Automatically Derived Category-Based Language Models for Speech Recognitio". Proceeding de ICASSP'98. 1998

[**Ortm96**] Ortmanns, S., Ney, H. y Eiden A. "Language-Model LookAhead for Large Vocabulary Speech Recognition". Proceedings ICSLP'96. 1996.

[**PAL00**] PAL. http://alpha.mic.dundee.ac.uk/acsd/research/pred.html. Febrero 2000.

[**Pere91**] Pereira, F. y Schabes, Y. "Inside-Outside Reestimation from Partially Bracketed Corpora". Proceedings de la 30ª Reunión anual de la Association for Computational Linguistics. Páginas 128-135. 1991.

[**Pred99**] PredictAb1ility. http://www.inclusive.co.uk/catalog/predict.htm. Septiembre de 1999.

[**Rabi89**] Rabiner, L. R. "A tutorial on hidden Markov models and selected applications in speech recognition". Proceedings of the IEEE, 77(2):257-286.

[**RAE00**] Real Academia de la Lengua. 2000. http://www.rae.es/NIVEL1/CONSULTAS/DEMOSTRATIVOS.HTM

[**Raud91**] Raudys, S., Jain., A., "Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners". IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 13, nº 3. Marzo 1991.

[**Ricc96**] Riccardi, G., Pieraccini, R., Bocchieri, E. "Stochastic automata for language modeling". Computer Speech and Language. 10. Páginas: 265-293. 1996.

[Rios99] Ríos, A. "La transcripción fonética automática del Diccionario Electrónico de Formas Simples Flexivas del español: estudio fonológico del léxico". Estudios de Lingüística Española 5. 1999.

[Rodr98] Rodríguez Paíno, L. "Diseño de un editor de textos predictivo para personas discapacitadas bajo entorno gráfico". Proyecto Fin de Carrera. Tutora: Sira E. Palazuelos Cagigas. E.T.S.I. de Telecomunicación. Dpto. de Ingeniería Electrónica. Julio 1998.

[Sanc99] Sánchez, J., Benedí, J., "Learning of stochastic context-free grammars by means of estimation algorithms". Proceedings of the Eurospeech 99. Volumen 4. Budapest (Hungría). Páginas: 1799-1802. 1999.

[Sani99] San Inocente Benito, F. "Actualización e Incorporación de Mejoras a un Editor Predictivo Orientado a Personas con Discapacidades para Entorno Windows". Proyecto Fin de Carrera. Tutora: Sira E. Palazuelos Cagigas. E.T.S.I. de Telecomunicación. Dpto. de Ingeniería Electrónica. Junio de 1999.

[SAW00] SAW. http://www.ace-centre.org.uk/html/sawalt1.html. Noviembre de 2000.

[Simp99] Simpson, R. C., Horstmann Koester, H. "Adaptive One-Switch Row-Column Scanning". IEEE Transactions on Rehabilitation Engineering. Volume 7. Número 4. Páginas: 464-473. Diciembre de 1999.

[Smit96] Smith, S. L., Mosier, J. N. "Guidelines for designing user interface software". http://www.syd.dit.csiro.au/hci/guidelines/sam/guidelines.html. The MITRE Corporation. Bedford. Massachusetts. E.E.U.U. Agosto de 1996.

[Soot00] SoothSayer Word Prediction. http://www.ahf-net.com/sooth.htm. Junio de 2000.

[Stei94] Steinbiss, V., Tran, B.H. y Ney, H. "Improvements in Beam Search". Proceedings de ICSLP'94. Páginas 2143-2146. 1994.

[Stol95] Stolcke, A. "An Efficient Probabilistic Context-Free Parsing Algorithm that Computes Prefix Probabilities". Computational Linguistics. 21 (2). Páginas: 165-201. 1995.

[Suar98] Suárez, M.D., Aguilar, A., Rosell, C. y Basil, C. "Ayudas de alta tecnología para el acceso a la comunicación y la escritura". En Basil, C. Soro-Carnats, E. y Rosell, E. "Sistemas de signos y ayudas técnicas para la comunicación aumentativa y la escritura: Principios teóricos y aplicaciones". 1ª edición. ISBN 844580716-1. Grupo Masson. 1998.

[Subi00] Subirats Rüggeberg, C. y Ortega Gil, M. "Tratamiento automático de la información textual en español mediante bases de información lingüística y transductores". Estudios de Lingüística Española 10. 2000

[Suka92] Sukaviriya, P. N., Foley, J. D. "Built-in User Modelling Support, Adaptive interfaces, and adaptive help in UIDE". Graphics, Visualization and usability center. Georgia Institute of Atlanta. Octubre de 1992.

[**Stew96**] Stewart, H. "Just How Useful is Word Prediction?". ABILITY NEWSLETTER. http://www.abilitycorp.com.au/news_views/newsletter_nov_1996.htm. Noviembre de 1996.

[**Swif87a**] Swiffin, A.L., Arnott, J.L, Pickering, Newell, A.F. "Adaptive and predictive Techniques in a communication Prosthesis". Augmentative and Alternative Communication. Volumen 3. Nº 4. Páginas: 181-191. Diciembre de 1987.

[**Swif87b**] Swiffin, A.L., Arnott, J.L, Newell, A.F. "The use of syntax in a predictive communication aid for the physically handicapped". Proceedings of the 10th Annual Conference on Rehabilitation Technology. San Jose. CA: RESNA. Páginas: 124-126. 1987.

[**Thom94**] Thompson, H. "TEMAA: A testbed study of evaluation methodologies: Authoring aids, Proceedings of the ELSNET Language Engineering Convention". Páginas: 147-148. Paris. 1994.

[**Vall92**] Vallés Botella, M. "Editor Comunicador Predictivo para Personas con Graves Limitaciones Motrices". Proyecto Fin de Carrera. Tutor: Francisco Giménez de los Galanes Cejudo. E.T.S.I. de Telecomunicación. Dpto. de Ingeniería Electrónica. 1992.

[**Vand91**] Van Dyke, J. "Word Prediction for Disabled Users: Applying Natural Language Processing to Enhance communication" BA Thesis. University of Delaware. 1991.

[**Vila98**] Vilaseca, D. "Daniel: el ordenador, una necesidad para el discapacitado, no un capricho". En Basil, C. Soro-Carnats, E. y Rosell, E. "Sistemas de signos y ayudas técnicas para la comunicación aumentativa y la escritura: Principios teóricos y aplicaciones". 1ª edición. ISBN 844580716-1. Grupo Masson. 1998.

[**VOX00**] Diccionario General de la Lengua Española VOX. 2000. http://www.vox.es/consultar.html.

[**Weis93**] Weiss, N.A. y Hasset, M.J. "Introductory Statistics" 3º edición, 1993.

[**Witt91**] Witten, I.H. y Bell, T.C. "The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression". IEEE Transactions on Information Theory, volumen 38(4). 1991.

[**Wood96**] Woods, W.A. "Augmented Transition Networks for Natural Language Analisys". Harvard Computation Laboratory Report No. CS-1. Cambridge. MA: Harvard University. 1996.

[**Writ00**] Write Away 2000. http://www.is-inc.com/product2.htm. Junio de 2000.

Bibliografía del autor

[**Carl97b**] Carlberger, A., Carlberger, J., Magnuson, T., Palazuelos-Cagigas, S.E., Hunnicutt, M.S. & Aguilera, S. "Profet, a New Generation of Word Prediction: An Evaluation Study". Proceedings of the Workshop on NLP for Communication Aids, ACL/EACL'97. Páginas: 23-28. Ed. "Association for Computational Linguistics". Madrid. 1997.

[**Clay98**] Claypool T., Ricketts I., Gregor P., Booth L., Palazuelos S. "Learning Rates of a Tri-Gram Based Gaelic Word Predictor". Proceedings ISAAC Dublin. 1998. Páginas: 178 - 179. Ed: Ashfield Publications. 1998.

[**Pala94**] Palazuelos, S.E. "Incorporación de mejoras ergonómicas y mecanismos predictivos a un editor orientado a personas discapacitadas". Proyecto Fin de Carrera. Tutor: Juan Manuel Montero Martínez. E.T.S.I. de Telecomunicación. Dpto. de Ingeniería Electrónica. Noviembre de 1994.

[**Pala95**] Palazuelos, S.E., Montero, J.M., Gómez S., Aguilera S. "On the Development of a Word Processor With Word Prediction for Severely Physically Handicapped, Non-Vocal Users: PREDICE". Proceedings of ECART 3. Páginas: 119-121. Lisboa. Octubre de 1995.

[**Pala96**] Palazuelos-Cagigas, S. E., Aguilera Navarro, S. "Report on Word Prediction for Spanish". Informe WP7T3D.2IR del Proyecto Europeo "VAESS: Voices, Attitudes and Emotions in Speech Synthesis". TIDE N° 1174. Agosto de 1996.

[**Pala97**] Palazuelos-Cagigas, S. E., Godino-Llorente, J.I., Aguilera Navarro, S. "Comparison Between Adaptive and Non-Adaptive Word Prediction Methods in a Word Processor for Motorically Handicapped Non Vocal User" Proceedings of the AAATE Conference. ISBN: 90 5199 361 7. Páginas: 158-162. Ed. IOS Press(George Anogianakis Et All). Tesalónica. Grecia. 29 Septiembre - 2 Octubre de 1997.

[**Pala98a**] Palazuelos, S., Aguilera S., Rodrigo J., Godino J. "Grammatical and statistical word prediction system for Spanish integrated in an aid for people with disabilities". ISBN: 1-876346-17-5. ICSLP'98. Sydney. 30 Noviembre - 4 Diciembre de 1998.

[**Pala98b**] Palazuelos, S., Aguilera, S., Claypool, T., Ricketts, I., Gregor, P. "Comparison of Two Word Prediction Systems Using Five European Languages". Proceedings ISAAC. Dublin. Páginas: 192 - 193. ISBN: 1 897606 04 4. Ed: Ashfield Publications. 1998.

[Pala98c] Palazuelos, S., Aguilera, S., Ricketts, I., Gregor, P. y Claypool, T. "Artificial Neural Networks applied to Improving Linguistic Word Prediction". Proceedings ISAAC. Dublin. Páginas: 194-195. ISBN: 1 897606 04 4. Ed: Ashfield Publications. 1998.

[Pala99a] Palazuelos, S. E., Rodrigo, J. L., Godino, J. I., Aliaga, F., Martín, J. L., Aguilera, S. "Predicción de palabras en castellano". SEPLN Procesamiento del lenguaje natural. Revista nº 25. ISSN: 1135-5948. Páginas: 151 - 158. Septiembre de 1999.

[Pala99b] Palazuelos Cagigas S. E., Aguilera Navarro, S., Rodrigo Mateos, J. L., Godino Llorente,J. I., Martín Sánchez, J. L. "Considerations on the Automatic Evaluation of Word Prediction Systems". En "Augmentative and Alternative Communication: New Directions in Research and Practice". Eds: Filip T. Loncke, John Clibbens, Helen H. Arvidson, Lyle L. Lloyd. ISBN: 1 86156 143 1. Páginas: 92-104. Editorial: Whurr Publishers. Londres. 1999.

[Pala00] Palazuelos Cagigas S. E., Martín Sánchez, J. L., Godino Llorente, J.I., Rodrigo Mateos, J. L., Arenas García, J., Aguilera Navarro, S. "Estrategias de comunicación utilizando PredWin como comunicador". En: Libro de Actas del Congreso Iberoamericano Iberdiscap 2000. 3º de Comunicación Alternativa y Aumentativa, 1º de Tecnologías de Apoyo para la Discapacidad. Páginas: 277-280. ISBN: 84-699-3253-5. Madrid. 18-20 de Octubre de 2000.