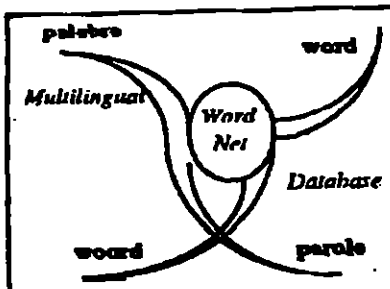


EuroWordNet

Building a Multilingual Wordnet Database with Semantic Relations between Words

LE-2 4003

Project Synopsis

Application Area	Language Resources, Language Engineering
Start Date	March 1996
Duration	36 Months
Total Effort	149 Person Months

Consortium	**Organisation**	**Short Name**	**Role**	**Nat. Code**	**Task**
	• University of Amsterdam	AMS	C	NL	provider
	• Istituto Di Linguistica Computazionale Pisa	ILC	P	IT	provider
	• Fundacion Universidad Empresa	FUE	P	ES	provider
	• Novell Belgium NV	NOV	P	BE	user
	• University of Sheffield	SHE	P	GB	provider

Contact person

Dr Piek Vossen
Project Manager
Computer Centrum Letteren
University of Amsterdam
Spuistraat 134
1012 VB Amsterdam
The Netherlands

Tel: +31 20 525 4624
Fax: +31 20 525 4429
Email: Piek.Vossen@let.uva.nl

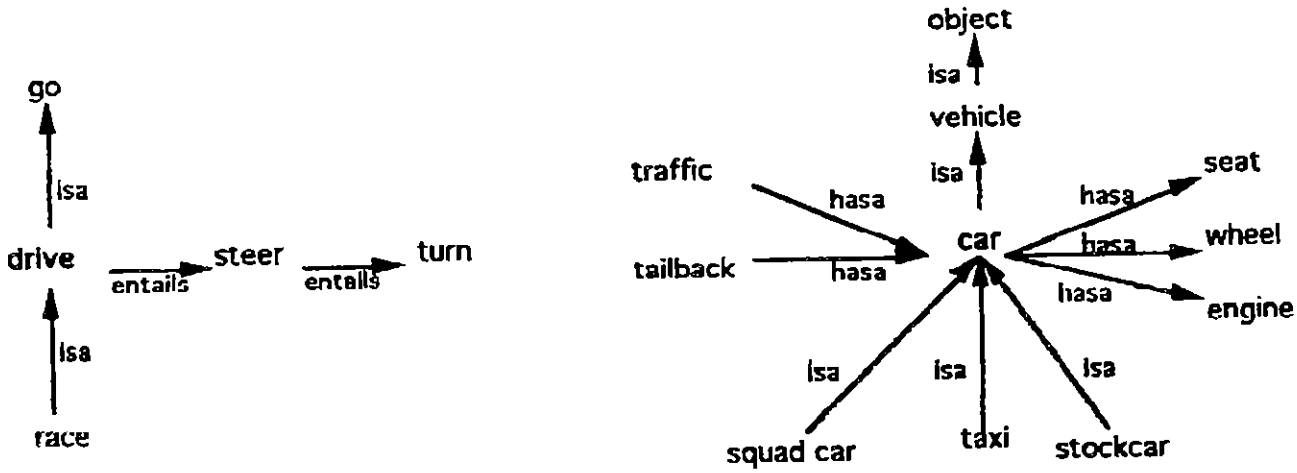
Abstract

The project aims at developing a multilingual database with basic semantic relations between words for several European languages (Dutch, Italian and Spanish). The wordnets will be linked to the American wordnet for English and a shared top-ontology will be derived, while language specific properties are maintained in the individual wordnets. The database can be used for multilingual information retrieval which will be demonstrated by Novell Linguistic Development.

1 Objectives

Currently, information is massively stored in electronic form and can be accessed from anywhere in the world via electronic networks. Although access to this information is constantly being improved by new interfaces and facilities, information retrieval from large electronic resources is still mainly determined by key word matching or fixed indexing and menu systems. Likewise, a user cannot simply use his own words to find information but has to make use of the wordings and rationale of the classification system. As the detail and amount of information increases a non-expert user will have more and more difficulty to use the right terminology to gain access to it. The situation in Europe is even worse since its diversity of languages and cultures constitutes an extra barrier, while the available linguistic tools to support textual search are mostly restricted to English. As a result of this, the information society is becoming restricted to a small group of people that speak English and have good knowledge of the access system and the stored data.

To provide non-expert searchers flexible access to the information society it is therefore crucial to develop tools that can expand his general and common words in a specific language to any possible variant or term in any other language. The user should be able to get around the choice of words in a document or the choice of key words by matching meanings rather than words. Such tools depend on the availability of generic resources with basic semantic relations between words, like the Princeton WordNet (Miller et al 1990). The American WordNet database consists of semantic relations between English word meanings (so-called synsets) which can be accessed as a kind of thesaurus in which words with related meanings are grouped together. For example, a noun like "car" is linked to, among others, all words that have a *hyponymy* or *isa* relation or a *meronymy* or *hasa* relation with it, and a verb like "drive" to, among others, all words that have a *hyponymy* or an *entailment* relation with it[1]:

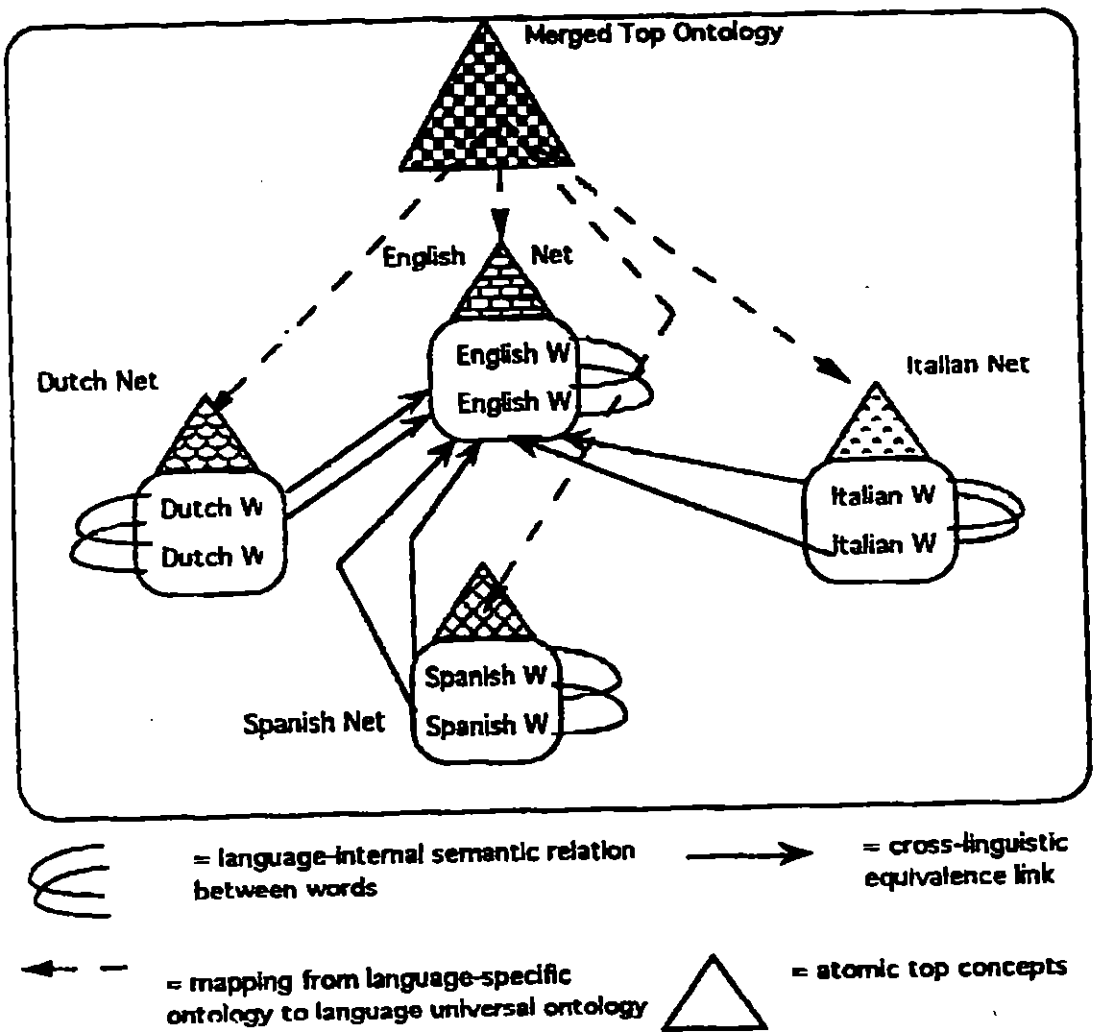


With such a database the query-terms of a user can be expanded to any set of closely related terms in a language, leading to better retrieval of information in terms of recall. For example a query with the terms "drive" and "car" will be expanded to combinations such as "go + car", "race + vehicle", "steer + car", "turn + wheel", "race + engine".

Unfortunately, such resources are not available for other languages than English, let alone a resource in which multiple wordnets are combined and interlinked. This severely holds back developments in language engineering and the information society in Europe. The aim of this project is therefore to develop such a *multilingual* database with wordnets for several European languages (Dutch, Italian and Spanish) which can be used to improve recall of queries via

1 Here a simplified example is given. In practice, different subtypes of "isa" and "hasa" relations will be distinguished as well as various other types of relations.

semantically linked variants in *any* of these languages. These European wordnets will as much as possible be built from available existing resources and databases with semantic information developed in various projects. This will not only be more cost-effective but will also make it possible to combine information from independently created resources, making the ultimate database more consistent and reliable, while keeping the richness and diversity of the vocabularies of the different languages. The wordnets will be stored in a central lexical database system and the word meanings will be linked to meanings in the Princeton WordNet1.5. Furthermore, we will merge the major concepts and words in the individual wordnets to form a common language-independent ontology (an ontology is the set of semantic relations between concepts). This will guarantee compatibility and maximise the control over the data across the different wordnets while language-dependent differences can be maintained in the individual wordnets.



What this implies is best illustrated with an example. Consider, for example, all the words that are related to *body parts*. All the wordnets will share the top-ontology concept BODY but each language has different lexicalizations for body parts. Whereas English words like "head" and "leg" can name the same parts of "animals" and "humans", in Dutch different words are used for animal parts and humans parts ("kop" (head) and "poot" (leg) for all animals except horses and "hoofd" (head) and "been" (leg) for humans and horses respectively). Similarly, in English and Dutch there are different words for "finger" and "toe" whereas Italian and Spanish have a single word to name both types of body parts ("dito" and "dedo" respectively). Each wordnet will thus reflect a unique lexicalization pattern. Equivalence in the lexicalization will be reflected by parallelism in the wordnet structure and simple equivalence relations with the English words, whereas differences in lexicalization will be reflected by divergence of the wordnet structures and

partial equivalence relations with the closest English word (different types of partial-equivalence relations will be distinguished).

Builders of the wordnets are: the University of Amsterdam (Co-ordinator of the project), the University of Sheffield, the University of Pisa, and a research team belonging to three Spanish Universities: University of Barcelona, Technical University of Catalunya (UPC) and U.N.E.D. The resulting data will be stored in a multilingual database system developed by Novell Linguistic Development in Antwerp in which relations can be traversed, selections can be made and data can be exported. Special facilities are made to get a multinlingual view on the wordnets.

2 User Involvement

The multilingual database will be tested and demonstrated by Novell who will also act as a user in the project and who is interested in developing such a multilingual information retrieval tool for the European Community. After full delivery of the resource Novell will load it into their Information Retrieval System (IRS) and test the adequacy in the demonstration phase. The Novell retrieval system does not use key words. Instead the user can enhance the query with a ConceptNet to search for all possible related terms. At this point, an English ConceptNet has been developed using WordNet1.5 as the main data source.

The development and testing of an IRS as such goes beyond the scope of this project. The user-requirements of such a full system include aspects such as flexibility, on-line help, visual representation of the results, allowing for different retrieval techniques and criteria (such as thesaurus or fixed-indexing systems, automatic key-word browsing, information on authors, publishers, institutes, reviews, date, etc.). The aim of this project is not to develop such a system but to provide a generic basic resource that could be included in such a broader IRS. Given the current State of the Art in IRSs the availability of general, generic resources such as a wordnet is typically expected to help non-expert users when retrieval by indexing is problematic because:

1) the indexing system does not cover the desired aspect or facet that a user is looking for,
2) the words chosen by the user are not included in the indexing key-word list,
3) the user speaks another language.

Therefore, the usefulness of the resources will not be tested by end-users in a real-life environment but by the developers of the IRS (in this project Novell) in controlled test situations that reveal the quality and added value for an IRS.

In addition to the direct scope of the project, we will also form a European User-Group of wordnet-builders and users that cover a wider range of languages and applications. The members of the User-Group will have the possibility to give feed-back to early releases of the project results (including sample, databases, documentation, definition of standards and data formats) which will be taken into account in the incremental building of the resources. Furthermore, we hope to create a wider awareness of the project results and to pave the way for the extension of the resources to other languages, larger vocabularies and other types of applications. The User-Group will contribute to a more complete understanding and description of the different user-needs depending on the kind of resources developed in this project (or developed on the basis of the project results).

The User-Group currently comprises:

Publishers	**Application area**
• Van Dale Lexicografie B.V. (NL) (provider of data)	(electronic) dictionaries, language generation tools, learning tools.
• Bibliograf (ES) (provider of data)	(electronic) dictionaries.
• Garzanti (IT)	(electronic) dictionaries.
Software Developers	
• SENA Athens (GR)	information retrieval
• CapVolmac, Utrecht (NL)	authoring tools, Grammar checkers
• INCYTA Barcelona (ES)	machine Translation
• Novell Linguistic Development, Antwerp (BE)	information retrieval, authoring tools, natural language interfaces
• LOGOS (IT)	technical translations, desktop publishing, technical writing.
• EBSCO (ES)	products for automated library systems, publishing of reference databases, retrieval systems for citation and full text databases, document delivery
• BERTIN (FR)	information retrieval in textual databases, concept-based indexing
• DATAMAT (IT)	information retrieval, document processing
Non-profit users	
• RKD, National Institute for Art-Historical Documentation (NL)	information retrieval, electronic libraries
• University of Madrid (ES)	machine translation, corpus linguistics, electronic dictionaries.
• VPRO, Broadcasting Organization (NL)	information retrieval, Internet services
Builders	
• University of Heidelberg (DE)	German wordnet
• University of Tuebingen (DE)	German wordnet
• University of Athens (GR)	Greek wordnet
• University of Goetheborg (ES)	Swedish wordnet
• University of Euskal Herriko	Bask wordnet
• University of Tartu, Estonia	Estonian, Latvian and Lithuanian wordnet
• University of Nantes, (FR)	French wordnet

During the project the user-group will be extended to achieve a maximal coverage in the different interested parties in Europe, where coverage relates to spreading in national interest, organisation type and type of application. Via exhibitions, the distribution of documents, electronic mailings and workshops we want to create an awareness of the electronic-linguistic services that can be developed using the multilingual wordnets as a starting point.

3 Results and exploitation

The most important deliverables will be a user-guide on the tools to develop the resources, the wordnets in each separate language (Dutch, Italian and Spanish) linked to the English WordNet, the shared top-ontology, the database in which all this can be viewed and selections can be exported and a report on the demonstration of the results in information retrieval tasks.

On a longer term the wordnets will become the backbone of any semantic database of the future and will open up a whole range of new applications and services in Europe at a trans-national and trans-cultural level. It will enhance the fundamental understanding of lexicalisation patterns across languages which will be crucial for machine translation and language learning systems. It will give non-native users and non-skilled writers the possibility to navigate or browse through the vocabulary of a language in new ways, giving them an overview of expression which is not feasible in traditional alphabetically organised resources. Finally, it will stimulate the development of sophisticated lexical knowledge bases which are crucial for a whole gamut of future applications, ranging from basic information retrieval to question/answering systems, language understanding and expert systems, summarizers to automatic translation tools and resources.

The results of the project will be publicly available where licensing contracts for background and foreground material will be drawn up at an early stage of the project. Non-commercial use will be free, commercial use will be charged for the background costs. The results will be stored on a CD and will be announced and distributed via commercial channels and via the academic networks. Information on the project can be obtained from a WWW home-page (see the contact person above), such as:

- general information on the project, such as progress, partnership, goals and aims, public documents, sponsorship
- forms to become registered as a member of the User Group.
- licensing forms for obtaining the project results
- public data samples, tools and databases

4 Work Parts and Time Schedule

The project can be characterized both as an LE-resource project and as a longer-term 'leading-edge' application project. The main focus will therefore be on Stage II (Development and Verification) of a project life-cycle, with minimal work parts for Stage I (Preparatory Activities) and Stage III (Demonstration). The main body of work will involve the building of the wordnets and the verification and demonstration in an information retrieval setting. The work packages (WPs) are organized around the 5 stages of a life-cycle:

Phase 1& 2:	User requirements and functional specification	WP1
Phase 3:	Building of the demonstrator	WP2, WP3, WP4, WP5, WP6
Phase 4:	Validation	WP7
Phase 5:	Exploitation	WP8

Separate work packages are devoted to managament (WP0), awareness and dissemination (WP9) and concertation (WP10).

Los procesos fonológicos y su manifestación fonética en diferentes situaciones comunicativas: la alternancia vocal/ semiconsonante/ consonante

Lourdes Aguilar
Departament de Filologia Espanyola, Universitat Autònoma de Barcelona, Bellaterra, Barcelona.

Tesis doctoral dirigida por la dra. Dolors Poch y presentada el día 25 de mayo de 1994 ante el tribunal constituido por: Dr. Joan Argente Giralt, Dr. Bernard Harmegnies, Dr. Josep Martí i Roca, Dra. Emilia Enríquez Carrasco, Dra. Lourdes Oñederra Olaizola.

Los problemas en torno a la interpretación de elementos como las semiconsonantes aparecen de forma recurrente en el desarrollo de la teoría fonológica pero, en general, las consideraciones de adscripción fonémica o de explicación de la alternancia vocal-semiconsonante-consonante no dan cuenta de los comportamientos fonéticos. Por otra parte, en el dominio de la fonética se utilizan los términos de "semiconsonante" y de "semivocal" a pesar de hacer referencia a la silabicidad, propiedad de carácter fonológico.

En el trabajo que presentamos se acude al método experimental con el fin de obtener datos válidos para la descripción fonética de semiconsonantes y semivocales frente a vocales y consonantes. El análisis desde el dominio de la fonética acústica se concibe como una fuente de información que permite dilucidar las propiedades básicas de los segmentos así como observar su manifestación en diversas situaciones de habla. El objetivo último es integrar la información obtenida en el experimento con otras informaciones procedentes del ámbito fonológico, tales como las propiedades funcionales y distribucionales de los segmentos.

El trabajo se estructura en dos grandes áreas, referidas a la caracterización teórica de los elementos objeto de estudio, desde ambos puntos de vista fonético y fonológico, y a la aplicación de un método experimental para la obtención de datos acústicos.

En lo que se refiere a los segmentos observados, al abordar los fenómenos de la alternancia vocal-semiconsonante-consonante y el contraste semivocal-semiconsonante, se implica a un número elevado de unidades: elemento palatal integrante de un diptongo en posición post- o prevocálica, consonante africada palatal sonora, consonante fricativa palatal sonora, consonante aproximante palatal sonora, consonante oclusiva palatal sonora, elemento velar integrante de un diptongo en posición post- o prevocálica, consonante fricativa velar sonora, consonante fricativa velar labializada sonora,

consonante aproximante velar sonora y consonante aproximante labiovelar sonora.

El corpus de análisis contiene muestras de habla procedentes de dos actividades comunicativas: la conversación y la lectura. En cuanto a la conversación, el objetivo es conseguir un conjunto de diálogos amplio pero de diseño restringido, de manera que posibilite el estudio del habla natural, con un importante grado de espontaneidad, pero a la vez permita manipular las variables de interés: la grabación de dos informantes mientras comparten una tarea específica parece satisfacer este requisito. En nuestro caso, se pide a los hablantes que colaboren verbalmente para reproducir en el mapa de uno de los participantes una ruta impresa en el del otro. Un informante tiene el dibujo del mapa de una zona imaginaria con una ruta marcada; frente a él, el otro informante dispone de una copia del mismo mapa pero sin ninguna ruta. Se le pide al primer sujeto que describa en detalle la ruta indicada en el mapa de tal modo que el interlocutor pueda reproducirla en el suyo. Esta estrategia permite elegir los nombres de los topónimos de forma que contengan las combinaciones vocálicas deseadas.

Con el fin de comparar las realizaciones obtenidas en una situación de diálogo y las procedentes de la lectura, las palabras del corpus se incluyen en frases de referencia.

Las secuencias del corpus se analizan en el dominio de la frecuencia por medio de procedimientos que dan cuenta del carácter dinámico de los diptongos. Las pruebas estadísticas aplicadas sobre los datos muestran que el grado de curvatura de la trayectoria formántica es un índice de discriminación de los diptongos frente a los hiatos. Se obtienen además ecuaciones fácilmente incorporables en sistemas automáticos.

Las datos derivados de la descripción fonética junto con los procedentes del ámbito fonológico permiten cuestionar el inventario fonético del español y evaluar su repercusión en el inventario fonológico. La reflexión en torno a ambos inventarios es de gran utilidad en el ámbito de la tecnología del habla. La constitución de diccionarios de unidades de síntesis así como la transcripción automática grafía-sonido requiere la elección de los fonemas y alófonos adecuados; en cuanto a los sistemas de reconocimiento, una descripción detallada de las variantes fonéticas facilita la tarea de identificación. Por otro lado, la comparación de las secuencias en dos situaciones de habla permite establecer niveles de transcripción en los algoritmos de conversión grafía-sonido o variaciones estilísticas en los sistemas de conversión de texto a habla.

TRATAMIENTO AUTOMÁTICO DE LA MORFOLOGÍA VASCA

Diseño y construcción de un procesador morfológico robusto para el euskara y del corrector ortográfico realizado con esa base.

Iñaki Alegria Loinaz

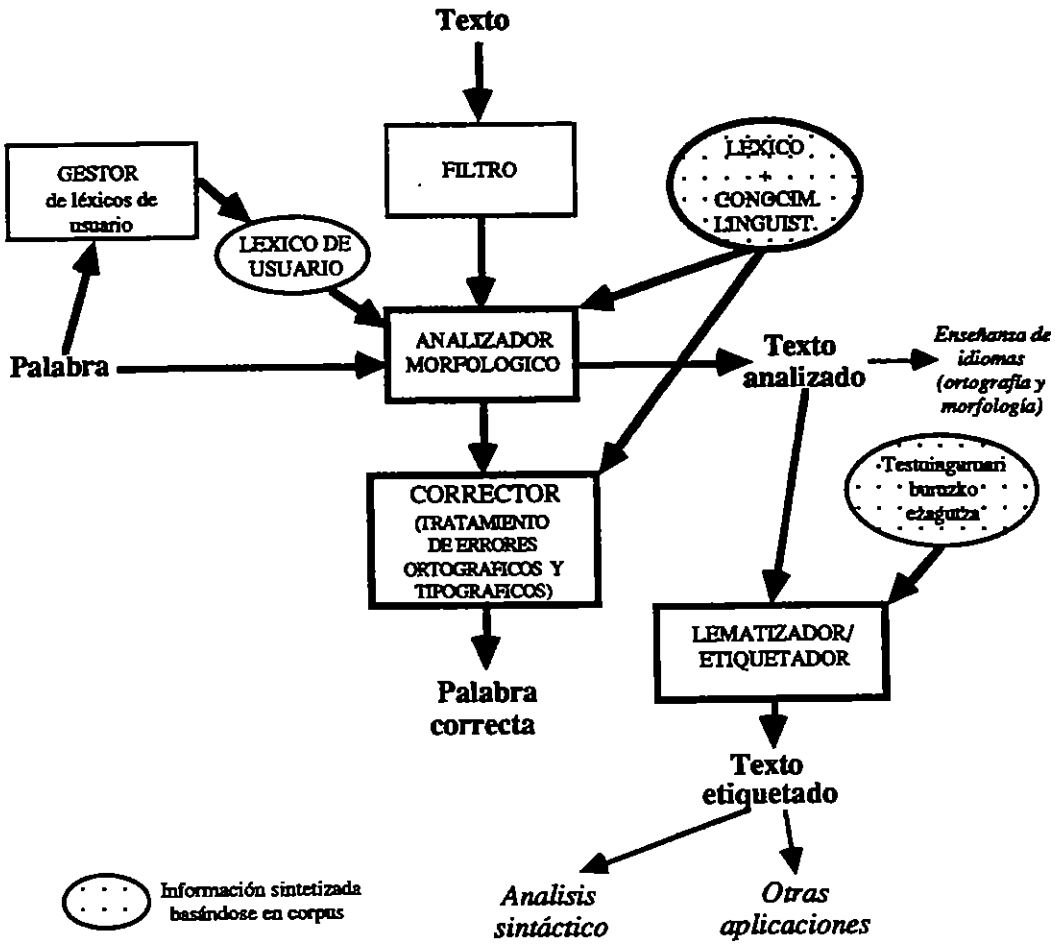
Directores: Xabier Artola Zubillaga
Kepa Sarasola Gabiola

Donostia, Junio de 1995.

Euskal morfologiaren tratamendu automatiko tresnak (Herramientas para el tratamiento automático de la morfología del euskara) quiere ser un primer paso en el procesamiento automático del euskara.

Este trabajo debe ubicarse dentro de un amplio proyecto a largo plazo para impulsar y desarrollar el procesamiento automático del euskara, y para ello estamos trabajando en equipo un grupo de lingüistas e informáticos y recoge el diseño y ejecución de dos herramientas básicas: un procesador morfológico y un corrector ortográfico, siendo ambos de escala real y reusables, basados en normas y conocimiento morfológico pero verificados con corpus reales. Para el trabajo en la morfología ha sido fundamental la preparación de la base de datos léxica (EDBL) y el manejo de un conjunto de textos que han servido como fuente de conocimiento y como banco de pruebas para la evaluación de resultados.

El esquema general del trabajo realizado se refleja en la figura siguiente:



La herramienta básica: el procesador morfológico automático.

El procesador morfológico se ha desarrollado en dos etapas, por un lado la construcción de un procesador morfológico estándar y posteriormente su complementación para conseguir un procesador robusto que reconoce un mayor número de palabras —mayor *coverage* o cobertura.

Las técnicas utilizadas en las dos fases se basan en la morfología de dos niveles evitando las soluciones particulares y consiguiendo un sistema totalmente homogéneo. Sobre la base del formalismo de dos niveles se han realizado tres extensiones: por un lado se le ha añadido la posibilidad de léxico(s) particular(es), por otro lado, se le ha dado una "nueva" aplicación en el tratamiento de variantes lingüísticas no estándares y por último se ha robustecido el sistema con el "análisis sin léxico" que anteriormente se había utilizado en fonología pero no en morfología.

Se han propuesto diversas mejoras del modelo de morfología de dos niveles durante el desarrollo de este proyecto.

El producto comercial: el corrector ortográfico Xuxen.

Desde un principio la realización de un corrector ortográfico para el euskara era uno de nuestros objetivos centrales ya que coincidía plenamente con los objetivos de aplicación y escala real que nos habíamos marcado y es una herramienta de gran importancia para el euskara debido al proceso de normalización que se está desarrollando.

Teniendo en cuenta esos objetivos se ha dado prioridad al tratamiento de errores que surgen del incipiente proceso de normalización —los llamaremos variaciones lingüísticas y errores de competencia— dejando el tratamiento del resto de errores —los llamados errores tipográficos— en un segundo plano.

El proceso de corrección se compone de dos módulos complementarios: el de los errores de competencia y el de los errores tipográficos. Mientras que para el tratamiento de los primeros se ha llevado a cabo un tratamiento innovador basado en la morfología de dos niveles el tratamiento de los errores tipográficos se hace de forma bastante convencional centrándose el trabajo en cuestiones de eficiencia.

ANHITZ: SISTEMA DICCIONARIAL MULTILINGÜE DE AYUDA EN LA TRADUCCION

Tesis doctoral. Facultad de Informática de Donostia, Universidad del País Vasco. Julio, 1995.

Autor: Xabier Arregi (e-mail: jiparipx@si.ehu.es)

Resumen.

En la tesis se presenta el sistema diccionarial multilingüe ANHITZ, cuyo objetivo es ayudar de una manera activa a los traductores humanos en la traducción léxica. En la memoria se describen los aspectos metodológicos, de diseño, la representación del conocimiento diccionarial multilingüe y las características funcionales del sistema.

El diseño del sistema integra aspectos de la traducción asistida por ordenador, de la lexicografía computacional y de los sistemas de ayuda basados en el conocimiento. Esta integración ha resultado enriquecedora y permite abordar la asistencia a los traductores humanos en el uso de los diccionarios yendo más allá de los diccionarios convencionales. No se trata de un simple cambio de soporte, sino de que los diccionarios se conviertan en herramientas especializadas y activas.

Para la integración de esas tres disciplinas se han seguido las ideas y directrices de la metodología KADS. Esta metodología representa, en cierta medida, un intento de llevar a la Ingeniería del Conocimiento fundamentos que ya se han usado en la Ingeniería del Software. Se defiende el uso de modelos formales y la distinción entre la especificación e implementación del conocimiento. KADS propone, además, la clara separación entre los distintos tipos de conocimiento en la adquisición y uso del mismo. Asumiendo como nuestro ese principio, en el diseño de ANHITZ se han distinguido cuatro niveles de conocimiento: de dominio, de inferencia, de tarea y estratégico.

En la construcción de ANHITZ hemos recurrido a una orientación basada en tareas. En primer lugar se ha modelizado el uso de los diccionarios en la traducción léxica. La metodología usada en la modelización parte de un estudio empírico basado en la observación directa y en protocolos orales. Con los datos recogidos se ha procedido a las fases de conceptualización, especificación y operacionalización. Como resultado de este proceso se ha expresado en un lenguaje de modelización ("CommonKADS Conceptual Modelling Language") el conocimiento adquirido sobre el uso de los diccionarios en la traducción léxica, que es la tarea principal del modelo.

En el nivel de dominio del sistema se ha almacenado el conocimiento diccionarial y se ha propuesto un esquema de representación del conocimiento léxico-semántico multilingüe. Se ha trabajado especialmente sobre los aspectos relacionados con el carácter multilingüe de la información. Se ha optado por una aproximación de transferencia, pero las relaciones bilingües se han surtido de información complementaria en el diccionario de transferencia. Según nuestro criterio, en el esquema de representación que se propone son relevantes el tratamiento dado a la información bilingüe y al problema de correspondencia entre acepciones de distintas lenguas, así como la propuesta para incorporar información específica orientada a la traducción. Consideramos que esta propuesta es consistente y generalizable a cualquier lengua. En cuanto a las dimensiones de la Base de Conocimiento Diccionarial hemos de decir que se ha recogido información de dos lenguas —euskara y francés—, tres diccionarios —dos monolingües y uno bilingüe—, se ha estructurado en nueve submódulos y se han almacenado más de mil conceptos y relaciones en el actual prototipo. El hecho de seleccionar acepciones de un dominio concreto —relacionadas con la administración municipal— le ha conferido cierta cohesión al conocimiento diccionarial y ha permitido probar las distintas posibilidades de uso.

El conocimiento diccionarial también se ha dotado de "comportamiento", mediante mecanismos para el razonamiento léxico. Tanto los mecanismos de enriquecimiento, como la reglas léxicas para la deducción dinámica, como también las funciones básicas son parte del conocimiento de nivel de inferencia. Las funciones básicas pueden considerarse como conocimiento sobre el comportamiento que se asigna a las unidades diccionariales y mediante estas funciones se establece la relación entre el nivel de inferencia y el nivel de tarea.

Las descripciones y la estructuración jerárquica de las tareas para consulta diccionarial se utilizan en el nivel estratégico, soportadas por la arquitectura que para ello se ha diseñado. La arquitectura del sistema se adecúa al manejo de los conocimentos de distintos niveles. En último término, el sistema puede simular los procesos cognitivos del traductor en la consulta diccionarial.

Dicha arquitectura se ha adaptado, en lo referente al funcionamiento, al esquema de los sistemas de ayuda activos. Se ha propuesto un estilo de utilización que pretende reunir las características idóneas de los sistemas de ayuda. El comportamiento del sistema pretende ser activo, amoldable e inteligente.