

Introducción al PLN

- Ingeniería Lingüística
 - Tareas
 - Recursos
 - Aplicaciones
- Niveles de procesamiento lingüístico

Introducción al PLN

Del PLN a la Ingeniería Lingüística

- La Ingeniería Lingüística es la aplicación del conocimiento de la lengua al desarrollo de sistemas informáticos capaces de reconocer, comprender, interpretar y generar lenguaje humano en todas sus formas.
- Las *tecnologías de la lengua* son un elemento fundamental para el éxito de la llamada *sociedad de la información*
- La comunicación entre personas y máquinas, a medio plazo, podrá ser llevada a cabo en lenguaje natural, en nuestra propia lengua.

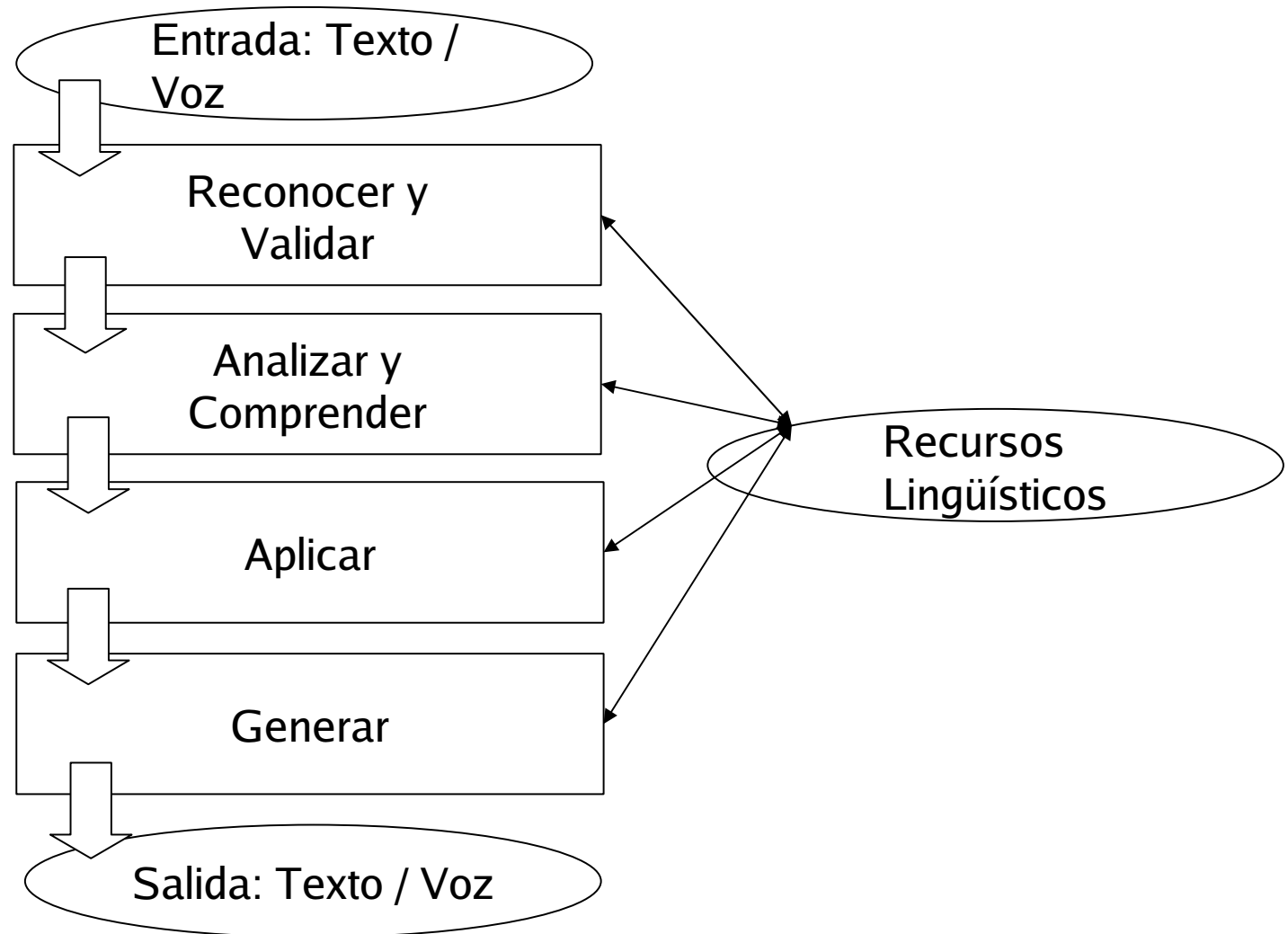
Introducción al PLN

Del PLN a la Ingeniería Lingüística

- La IL comprende:
 - Métodos, Técnicas y Herramientas
 - Recursos (Lingware)
 - Aplicaciones

Introducción al PLN

Componentes de un sistema de PLN (genérico)



Introducción al PLN

Tareas principales

- Identificación del hablante. Identificación.
- Reconocimiento del habla.
- Reconocimiento de caracteres.
- Comprensión de la lengua. Sistemas de diálogo.
- Generación de lenguaje.
- Síntesis del habla.

Introducción al PLN

Identificación del hablante. Identificación.

- Teoría de la señal y comunicaciones.
- Problemas:
- Hablante desconocido
- Ruido (en el entorno o en el medio de transmisión)
- Cambios temporales en el hablante.

Introducción al PLN

Reconocimiento del habla

- Teoría de la señal y comunicaciones
- Adquisición y uso de modelos estadísticos de fonemas y palabras
- Problemas:
- Reconocimiento de cualquier locutor
- Ruido
- Acentos, dialectos, agramaticalidades

Introducción al PLN

Reconocimiento de caracteres

- Reconocimiento de caracteres impresos (OCR)
- Problemas:
 - Tipo de letra desconocido
 - Mala calidad del texto

Introducción al PLN

Comprensión de la lengua

- Técnicas de lingüísticas, estadísticas, híbridas
- Problemas:
 - Ámplia cobertura / dominios restringidos
 - Comprensión completa o parcial
 - Modelos semánticos
 - etc.

Introducción al PLN

Sistemas de diálogos (1)

- Alto nivel de comprensión!
- Detección del contenido ilocutivo de las intervenciones del interlocutor humano
- Detección de los actos del habla directos e indirectos

Introducción al PLN

Sistemas de diálogos (2)

- Actos del habla: un acto de habla puede ser solicitar información, ofrecerla, disculparse, expresar indiferencia, expresar agrado o desagrado, amenazar, invitar, rogar, etc.
 - Acto locutivo: es la idea o el concepto de la frase, es decir, aquello que se dice.
 - Acto ilocutivo: es la intención o finalidad concreta del acto de habla.
 - Acto perlocutivo: es el (o los) efecto(s) que el enunciado produce en el receptor en una determinada circunstancia.
- Directos: son aquellos enunciados en los que el aspecto locutivo e ilocutivo coinciden, es decir, se expresa directamente la intención.
- Indirectos: son aquellas frases en las que el aspecto locutivo e ilocutivo no coinciden, por lo tanto la finalidad de la oración es distinta a lo que se expresa directamente.

Introducción al PLN

Generación del lenguaje

- Representación semántica del texto
- Qué decir y cómo decirlo
- Planificación del contenido y de la forma
- Elementos retóricos

Introducción al PLN

Síntesis del habla

- Representación del texto
- Forma: intensidad, entonación, pausas, duración
- Generación a partir de piezas pregrabadas

Introducción al PLN

Recursos

- From Cyc
 - Fred saw the plane flying over Zurich.
 - Fred saw the mountains flying over Zurich.

Introducción al PLN

Recursos

- Del PLN a la CLN (Comprensión del Lenguaje Natural)
- Procesamiento semántico a gran escala (conceptos en lugar de palabras)
- Dos problemas complementarios:
 - Cuello de botella de la adquisición
 - porque no sabemos tratar la ambigüedad
 - Ambigüedad
 - porque falta conocimiento

Introducción al PLN

Recursos

- Qué conocimiento es necesario para un sistema de NLP?
- Dónde podemos localizar este conocimiento?
- Qué procedimientos automáticos podemos aplicar?

Introducción al PLN

Qué conocimiento es necesario para un sistema de NLP?

- Tokenización: puntuación, palabras, números, fechas, ...
- Fonológico: fonemas, acentos, ...
- Morfológicos: categoría, género, concordancia, ...
- Sintaxis: categoría, subcategorización, estructura argumental, ...
- Semántica: clase semántica, preferencias de selección, ...
- Pragmática: uso, registro, dominio, ...

Introducción al PLN

Dónde podemos localizar este conocimiento?

- Cerebro humano
- Fuentes léxicas estructuradas:
 - Diccionarios monolingües y bilingües
 - Tesauros
- Fuentes no estructuradas:
 - Corpus monolingües y bilingües
- Combinación de los anteriores

Introducción al PLN

Qué procedimientos automáticos podemos aplicar?

- Aproximación descriptiva
 - Construcción asistida por un Sistema de Información
- Aproximación prescriptiva
 - Adquisición automática a partir de recursos preexistentes
- Aproximación mixta

Introducción al PLN

Ejemplo: MRDs (diccionarios)

jardín_1_1	Terreno donde se cultivan plantas y flores ornamentales.
florero_1_4	Maceta con flores .
ramo_1_3	Conjunto natural o artificial de flores , ramas o hierbas.
pétalo_1_1	Hoja que forma la corola de la flor .
tálamo_1_3	Receptáculo de la flor .
miel_1_1	Substancia viscosa y muy dulce que elaboran las abejas, en una distensión del esófago, con el jugo de las flores y luego depositan en las celdillas de sus panales.
florería_1_1	Floristería; tienda o puesto donde se venden flores .
florista_1_1	Persona que tiene por oficio hacer o vender flores .
camelia_1_1	Arbusto cameliáceo de jardín, originario de Oriente, de hojas perennes y lustrosas, y flores grandes, blancas, rojas o rosadas (Camellia japonica).
camelia_1_2	Flor de este arbusto.
rosa_1_1	Flor del rosal.

Introducción al PLN

Recursos

- Diccionarios (Machine-Readable Dictionaries)
- Lexicones generales / especializados
- Bases de Conocimiento
- Ontologías
- Gramáticas
- Corpus textuales
- Web / Internet como fuente de información

Introducción al PLN

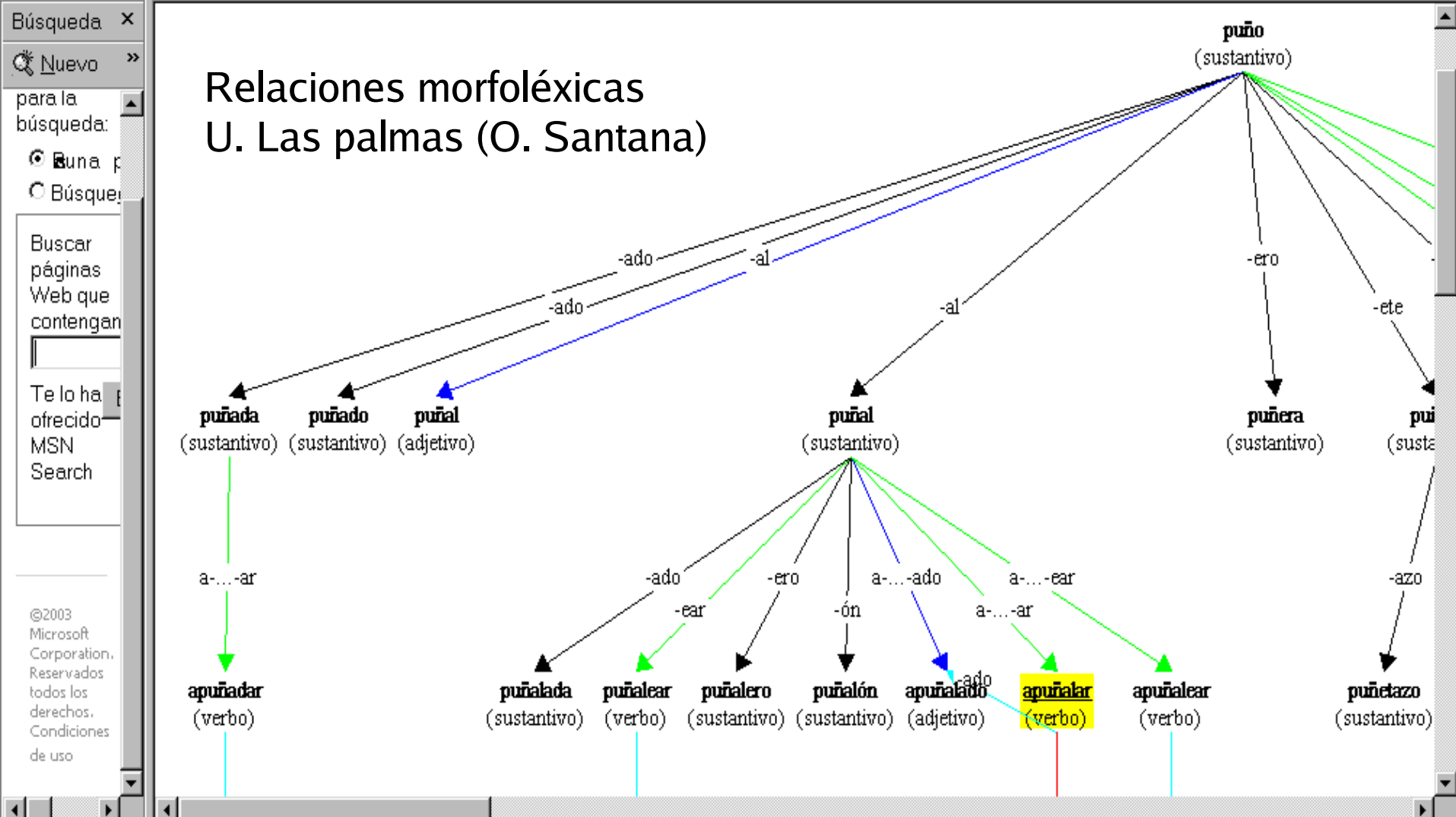
Diccionarios

- MRDs
- Tipos: generales, normativos, de uso, escolares, de aprendizaje, mono/bilingües
- Contenido, tamaño, organización, ...
 - Entrada, acepción, relaciones, consistencia, ...
- Tesauros
- Enciclopedias (Wikipedias)

Introducción al PLN

Lexicones

- Repositorios de información asociada a lemas / palabras
- Vocabularios
- Nombres propios / Gazetteers
- Terminologías
- Locuciones
 - grupo estable de dos o más palabras que funciona como una unidad léxica con significado propio, no derivado de la suma de significados de sus componentes (por ejemplo: locución adjetiva: una verdad como un templo)
- Siglas



Introducción al PLN

Bases de Conocimiento

- WordNet (WN)
- EuroWordNet (EWN)
- ThoughtTreasure, ConceptNet, MindNet, ...
- FrameNet, VerbNet, PropBank, OntoNotes, ...
- Extended WN, MCR, Omega, ...

Introducción al PLN

Ontologías

- Léxicas vs. conceptuales
- Generales vs. dominio
- Contenido, granularidad, relaciones, ...

- CyC
- Top Concept Ontology (EWN)
- Mikrokosmos
- SUMO (Suggested Upper Merged Ontology)

Introducción al PLN

Gramáticas

- Morfológicas
- Sintácticas
- Sintagmáticas vs. unificación
- Probabilísticas
- Cobertura, lengua, categorías, ...

Introducción al PLN

Corpus

- Textuales vs. orales
- Monolingües vs. multilingües
- Gran tamaño (1Mw – 1Gw - 1Tw)
- Estructurados vs. no estructurados
- Anotados vs. no anotados
- Útiles para adquisición de:
 - Colocaciones, estructura argumental, contextos, inducción gramatical, relaciones léxicas, preferencias de selección, ...

Introducción al PLN

Corpus anotados

- Categoría gramatical (Part-of-speech tagging)
- Lematizados
- Desambiguados, con sentidos de un diccionario
- Parentizados
- Analizados sintácticamente
- Alineados

Introducción al PLN

Corpus ejemplos

- Brown corpus
 - Wall Street Journal
 - British National Corpus
 - Penn Treebank
 - Susanne
 - SemCor
 - Parlamento europeo
-
- ELRA/ELDA
 - LDC (Linguistic Data Consortium)

Introducción al PLN

Corpus ejemplos

- CREA, recopilado por RAE 200Mw. Etiquetado y lematizado
- CRATER, castellano, inglés, frances 5.5Mw. Etiquetado y alineado
- LEXESP, castellano 5Mw. Etiquetado y lematizado
- 3LB, castellano, catalán, euskara. Sintáctico y semántico.

- Instituto Cervantes. Observatorio Español de Industrias de la Lengua.
- <http://www.cervantes.es/oeil/Oeil0.htm>

Introducción al PLN

Web corpus

- Fuente heterogénea de información multilíngüe
- Heterogénea: contenido, lengua (70% inglés), formatos
- Metabuscadores Lingüísticos
 - Webcorp <http://www.webcorp.org.uk/>
 - Linguistic's Search Engine
<http://lse.umiacs.umd.edu:8080/>

Introducción al PLN

Aplicaciones

En la actualidad ya hay múltiples servicios lingüísticos que facilitan el trabajo al usuario humano:

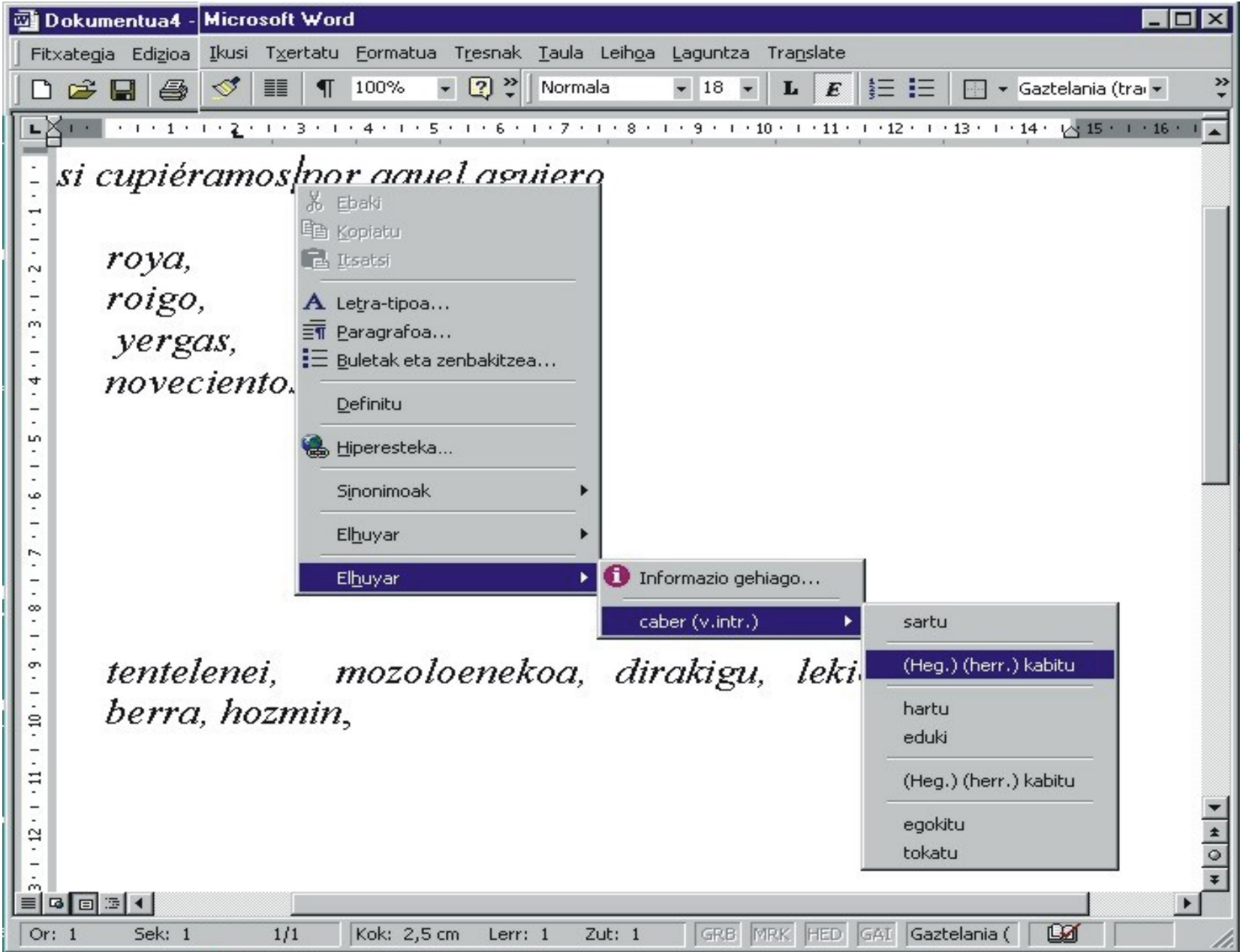
- Ayuda a la edición y comprensión de textos
- Traducción automática
- Tratamiento de grandes volúmenes de texto
- Tratamiento de voz
- Enseñanza de segundas lenguas
- ...

Introducción al PLN

Aplicaciones

- Ayuda a la edición y comprensión de textos
 - Correctores ortográficos y de estilo
 - Sistemas de consulta de diccionarios
 - On-line con editor de textos. Elhuyar(Cast-Eusk), UZEI sinón.
 - Muchísimos diccionarios: www.yourdictionary.com
- Traducción automática
- Tratamiento de grandes volúmenes de texto
- Tratamiento de voz
- Enseñanza de segundas lenguas
- ...

Introducción al PLN



Introducción al PLN

Aplicaciones

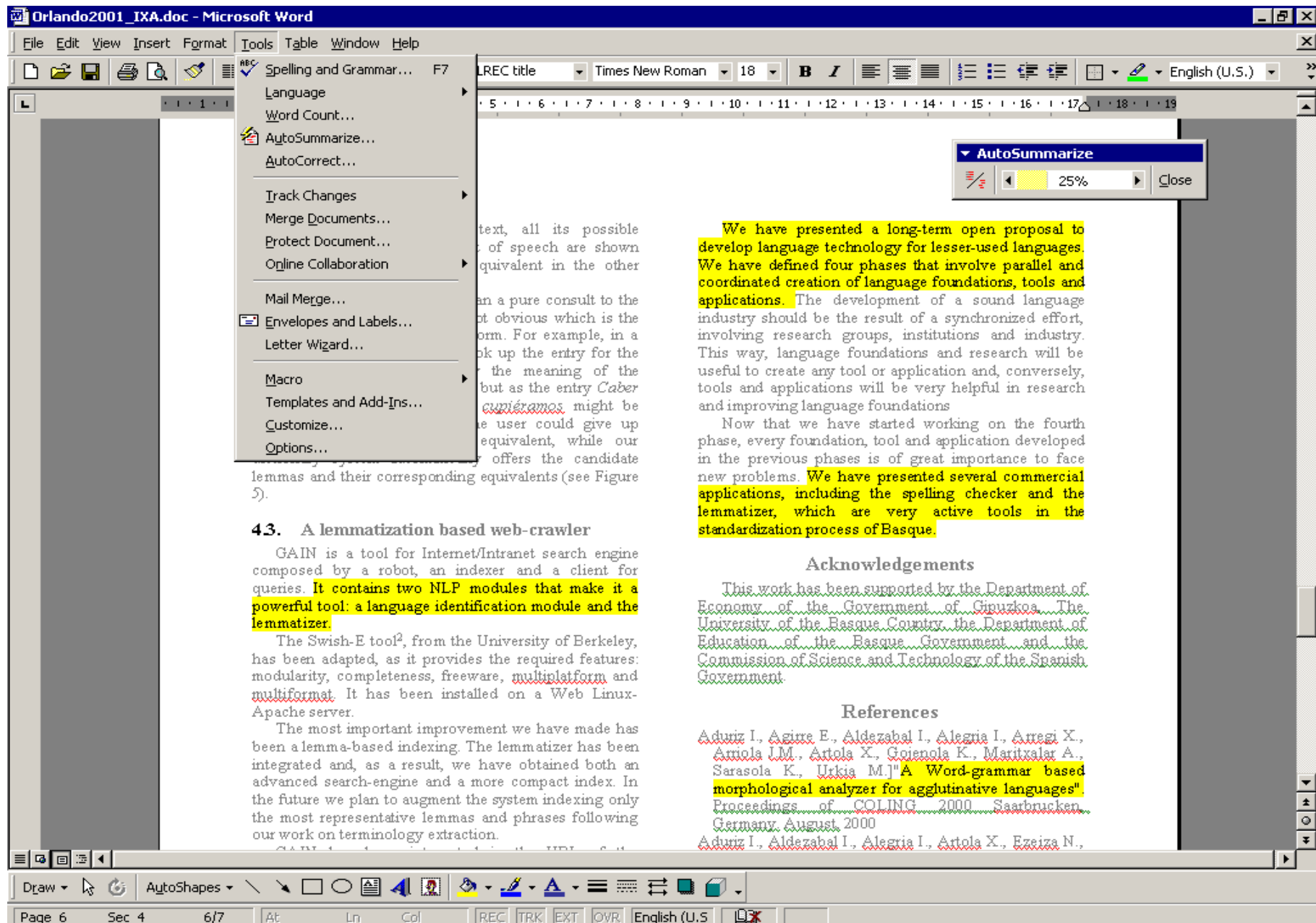
- Ayuda a la edición y comprensión de textos
- Traducción automática
 - Generación de borradores para traducción
 - Comprensión superficial de documentos (Web)
 - Memorias de traducción
- Tratamiento de grandes volúmenes de texto
- Tratamiento de voz
- Enseñanza de segundas lenguas
- ...

Introducción al PLN

Aplicaciones

- Ayuda a la edición y comprensión de textos
- Traducción automática
- Tratamiento de grandes volúmenes de texto
 - Recuperación de información (~ Google)
 - Extracción de información
 - Resúmenes automáticos
 - Sistemas de pregunta-respuesta (Question-Answering)
- Tratamiento de voz
- Enseñanza de segundas lenguas
- ...

Introducción al PLN



Introducción al PLN

Aplicaciones

- Ayuda a la edición y comprensión de textos
- Traducción automática
- Tratamiento de grandes volúmenes de texto
- Tratamiento de voz
 - Síntesis de voz
 - Reconocimiento del habla (<http://www.nuance.com>)
- Enseñanza de segundas lenguas
- ...

Introducción al PLN

Aplicaciones

- ...
- **Interactive online CL Demos**
- <http://www.ifi.unizh.ch/CL/InteractiveCLtools>
- <http://www.lt-world.org/>
- <http://registry.dfki.de/>

Introducción al PLN

Aplicaciones

- ACL (Association for Computational Linguistics)
- SEPLN
- IXA taldea (Donostia)
- TALP Research group (Barcelona)
- GPLSI (Alicante)
- UNED NLP group (Madrid)
- ... y ~30 grupos más.