DESARROLLO DE UN CONVERSOR DE TEXTO A VOZ EN ESPAÑOL DENTRO DE UNA ARQUITECTURA MULTILINGÜE

José Manuel Conejo
Bert Van Coile

Lernout & Hauspie
Speech Products,

Belgium.

0. RESUMEN.

Esta comunicación versa sobre el desarrollo de sistemas de conversión de texto en voz. Se describe someramente un diseño general para tales sistemas, prestando especial atención a aspectos lingüísticos (conversión del texto ortográfico en una representación de sonidos) y fonéticos (síntesis de segmentos y prosodia). Por último, tratamos de una herramienta para el desarrollo de sistemas de síntesis de texto llamada DEPES. En relación a todas estas cuestiones se presta especial atención a la lengua española.

1. INTRODUCCION.

El interés por las técnicas de síntesis del lenguaje y reconocimiento de voz es ya considerable. La demanda de aplicaciones de estas tecnologías crece día a día. La comunicación verbal con la computadora, el hombre hablando a la máquina, la máquina hablando al hombre, es una realidad. Sin embargo, aún es necesario romper barreras lingüísticas y hacer llegar estas tecnologías a la sociedad.

En Lernout & Hauspie Speech Products (L&H) desarrollamos sistemas multilingües para reconocimiento y síntesis de habla. Este trabajo trata sobre aspectos de síntesis y especialmente sobre el desarrollo de un conversor de texto en voz (CTV) para el español.

2. DE TEXTO A HABLA.

Un sistema de conversión texto-voz debe ser capaz de transformar cualquier texto de entrada en una salida hablada, que se comprenda perfectamente y sea lo más parecida posible al habla natural. En otras palabras,debe imitar la compleja actividad humana de leer en voz alta. La mayoría de los sistemas de CTV que existen pueden dividirse en tres partes principales (véase fig.1): una lingüística, una fonética y el sintetizador propiamente dicho. Prestaremos atención a las dos primeras.

La parte lingüística

Esta sección analiza el texto de entrada y proporciona su transcripción fonética e información de tipo léxico, sintáctico y semántico.

Una tarea importante consiste en normalizar el texto, resolviendo elementos tales como abreviaturas, secuencias de números, indicaciones de tiempo, monedas, códigos postales, números romanos, símbolos especiales, etc. La importancia de este proceso no debe subestimarse. Cuanto mayor sea la libertad del usuario, más poderoso debe ser este procesamiento. Por ejemplo, en español hemos documentado multitud de abreviaturas para la palabra "teléfono/s": Tl., Tlf., Tls., Tlfs., etc.

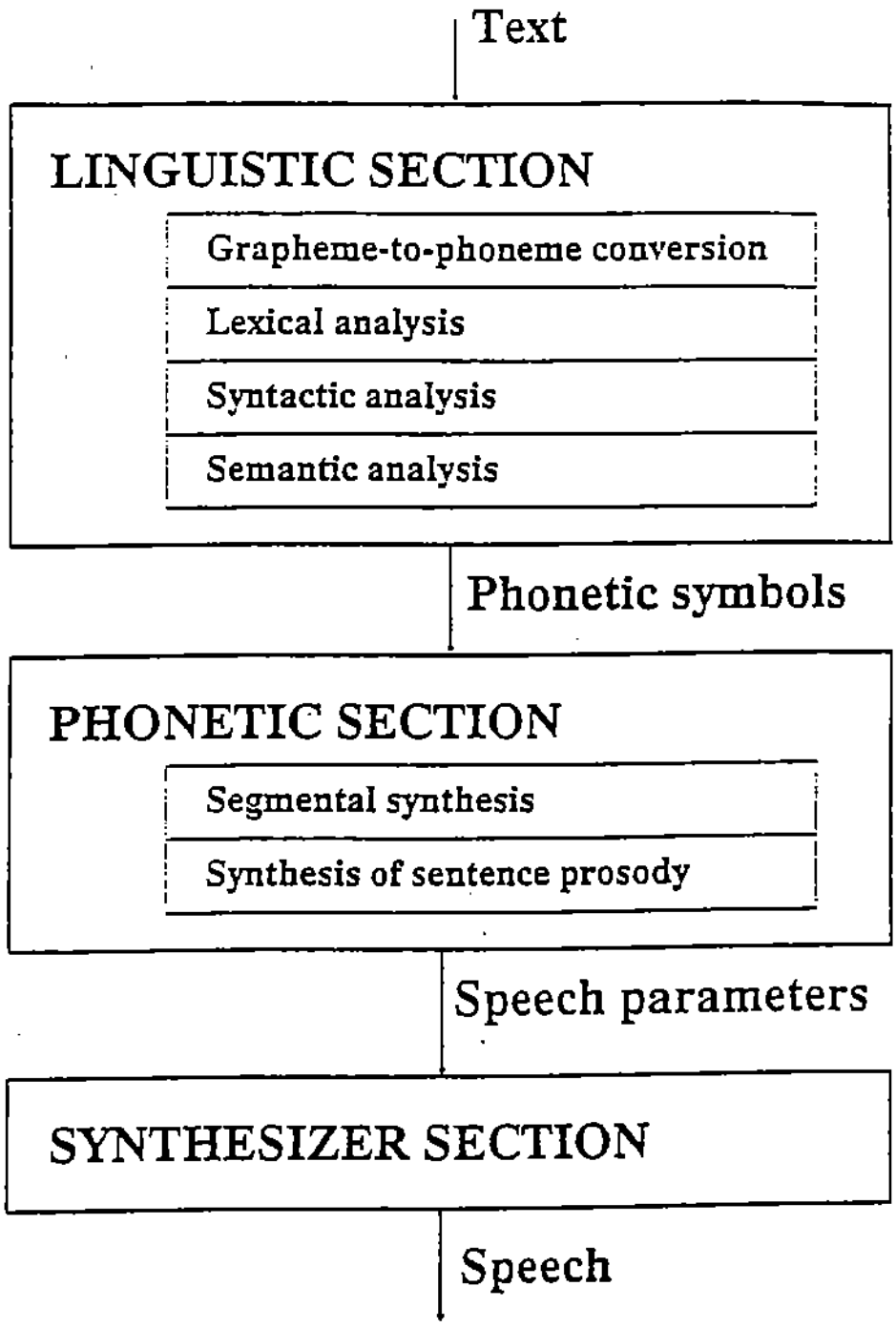
La conversión de secuencias numéricas en español puede entrañar en ocasiones un análisis de tipo léxico, dado que la pronunciación de algunos números varía según el género del sustantivo o abreviatura a los que acompaña (p.ej. 200.000 personas vs. 200.000 coches).

La principal misión de la sección lingüística es la conversión de signos ortográficos en símbolos fonéticos. Aquí se incluyen la conversión de grafemas en fonemas, silabificación y determinación del acento. Es sabido que este procedimiento no resulta demasiado difícil en español (véase más abajo para ejemplos), aunque el grado de complejidad de las reglas dependerá en gran manera del número de fonemas y alófonos incluidos en el alfabeto fonético que utilice el sistema (el nuestro consta en la versión actual de 28 elementos).

Además, como ya quedó indicado, la parte lingüística debe realizar un análisis léxico y sintáctico.

Para llevar a cabo el procesamiento del texto todos nuestros sistemas utilizan una estrategia híbrida que combina diccionarios (de palabras y de morfemas) y gramáticas constituidas por reglas de reescritura dependientes del contexto (véase más abajo). Para cada lengua se crean diccionarios de clases cerradas, de palabras más frecuentes, de excepciones (esto es, casos que no son resueltos correctamente por las reglas) y abreviaturas. Las reglas de reescritura se utilizan para la conversión de grafemas en símbolos fonéticos, para el componente de reglas de la descomposición morfológica y para realizar un análisis sintáctico localizado.

Figure 1.


Text
LINGUISTIC SECTION
Grapheme-to-phoneme conversion
Lexical analysis
Syntactic analysis
Semantic analysis
Phonetic symbols
PHONETIC SECTION
Segmental synthesis
Synthesis of sentence prosody
Speech parameters
SYNTHESIZER SECTION
Speech

La parte fonética

Esta sección produce los parámetros para el sintetizador, teniendo en cuenta los datos facilitados por la parte lingüística. Se encarga de la síntesis de segmentos (la creación de buenas características espectrales para el texto que se desea sintetizar) y de la prosodia (entonación y duración de los sonidos).

Nuestra síntesis se basa en una técnica de concatenación de segmentos. Un inventario de pequeños segmentos hablados tomadas de habla real se utiliza para sintetizar cualquier mensaje. La elección de las unidades elementales para la síntesis es un factor fundamental a la hora de determinar la calidad y la complejidad del sistema de conversión.

En L&H utilizamos difonemas y trifonemas para la concatenación de segmentos. La principal ventaja del difonema es que se trata de una unidad que preserva dentro de ella la transición y la mayor parte de los efectos debidos a la coarticulación de sonidos, ya que comienza en el estado estable de un sonido y termina en el estado estable del siguiente. La misma técnica de síntesis de segmentos es utilizada para todas las lenguas sobre las que trabajamos.

No obstante, es obvio que el número de difonemas y trifonemas varía de una lengua a otra y que también depende del número de elementos contemplados en el alfabeto fonético. Para el conversor del español este número es aproximadamente de 1.300 segmentos

La simple concatenación de estas unidades básicas ya produce un habla comprensible. Sin embargo, inteligibilidad y especialmente naturalidad deben ser mejoradas. Esta mejora se realiza por medio de modelos adecuados de duración de los sonidos y entonación.

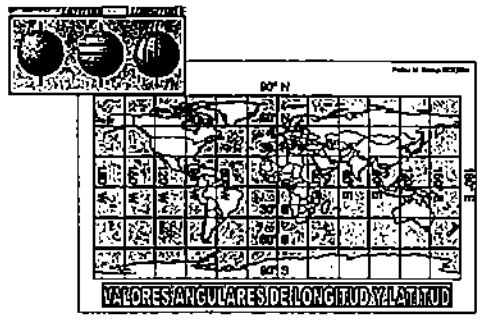
El modelo para la duración se desarrolla después de medir la duración de los sonidos en textos extensos leídos por un hablante nativo. El modelo se construye sobre bases estadísticas y cubre varios factores: duración intrínseca del sonido, efectos de límites sintácticos (p.ej. alargamiento ante pausa), acento (léxico, oracional), categoría de la palabra (funcional vs. no funcional), contexto fónico, etc. El modelo para el español está en fase de desarrollo.

El desarrollo del modelo de entonación se basa en un método de análisis iterativo de perceptibilidad. Se estiliza un conjunto de configuraciones tonales naturales. El proceso de estilización continúa en tanto en cuanto no se perciba una diferencia significativa entre el habla sintetizada con la configuración natural y el habla sintetizada con la configuración estilizada. Un modelo de entonación se deriva de las configuraciones tonales estilizadas. El resultado es un conjunto de reglas que indican como los movimientos tonales elementales (ascensos y descensos estandarizados de tono) pueden combinarse para crear la configuración tonal adecuada para un mensaje completo. Estas reglas toman en consideración el número y la localización de las palabras dominantes y los límites sintácticos más importantes.

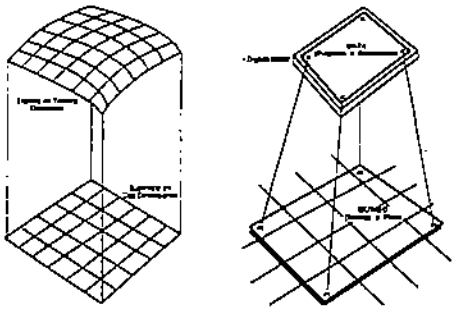
Fundamentos de Geodesia en Cartografía Digital

Prof.: Alfredo Ramón Morte
Universidad de Alicante

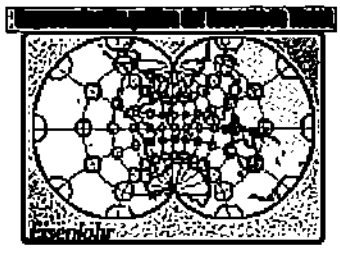
Latitud y Longitud



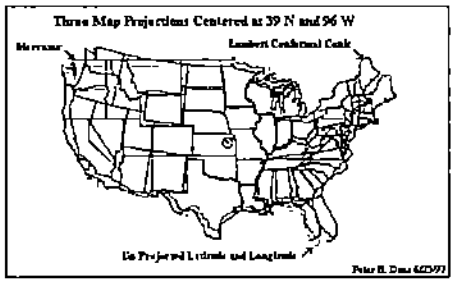
El "viaje" de la Tierra al plano



Indicatriz de Tissot



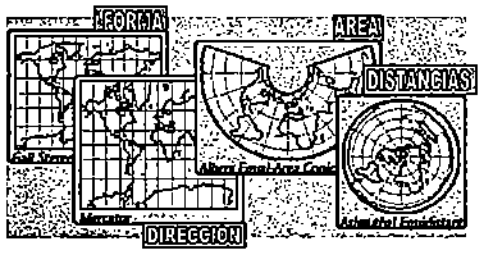
Solape de varias proyecciones



I.- Por sus cualidades

- Conservan las FORMAS:
  - Isógonas o equiangulares
- Conservan las Superficies o ÁREAS:
  - Equiáreas
- Subgrupos del resto: Afilácticas
  - Equidistantes: conservan DISTANCIAS
  - Navegación: Conservan el RUMBO

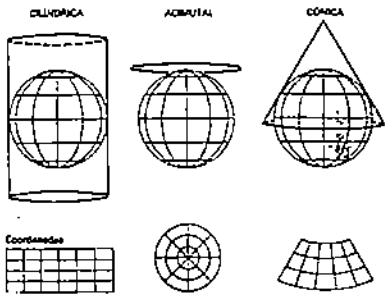
Distorsiones de la Proyección



II.- Por sus fundamentos

- PLANAS, AZIMUTALES O CENITALES
- CÓNICAS
- CILÍNDRICAS O SEUDOCILÍNDRICAS

Sistemas Cartesianos Decimales

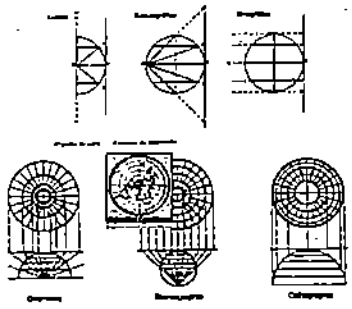


Proyecciones PLANAS

- Central o Gnómica
- Ortográfica
- Estereográfica

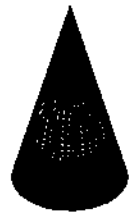
Planar Projection Surface

Factura de las Proy. Planas



Proyecciones Cónicas

- Extensión de la superficie de contacto:

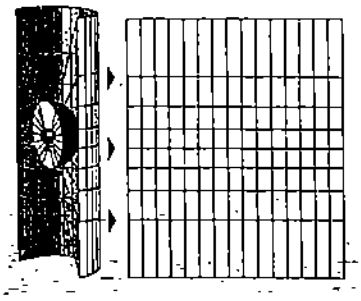


Conical Projection Surface

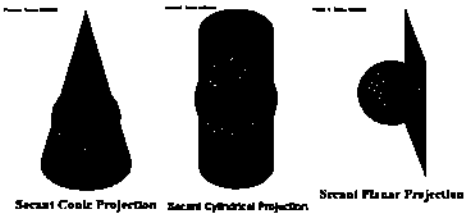
Proyecciones Cilíndricas

Cylindrical Projection Surface

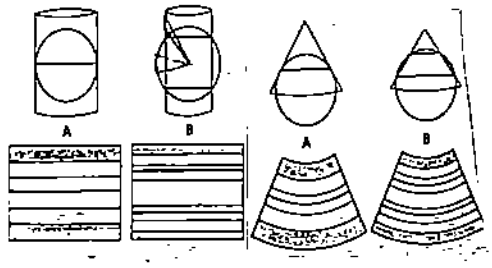
Factura de las Proy. Cilíndricas



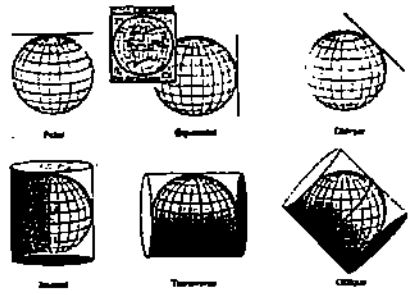
Secantes



Tangentes vs. secantes

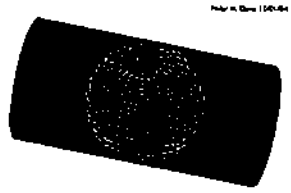


Normales, Transversas y Oblicuas



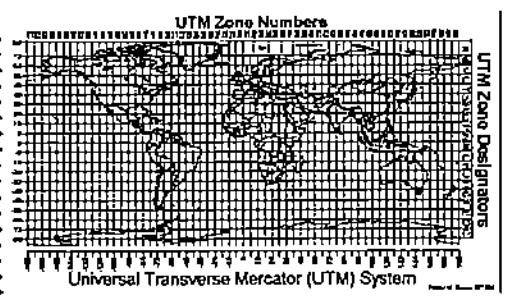
PROYECCIÓN UTM

- Universal Transversa de Mercator (cilíndrica):

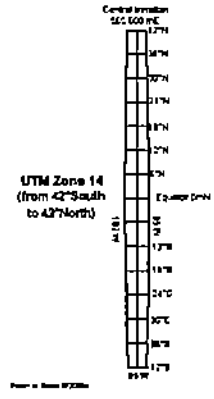


Transverse Cylindrical Projection Surface

Posiciones UTM (husos)



Detalle de un Huso o Zona UTM



3. EL SISTEMA DEPES.

El sistema que acabamos de describir está siendo aplicado a varias lenguas. Esta es una tarea ambiciosa que requiere herramientas de desarrollo adecuadas y flexibles, tales como editores de señales acústicas, programas de análisis del habla, programas para la segmentación automática de sonidos y otros. A continuación describimos una de ellas.

DEPES (Development Environment for Pronunciation Expert Systems) tiene como características principales flexibilidad y facilidad de uso. DEPES ofrece un poderoso lenguaje de representación que combina eficientemente elementos tomados de lenguajes de programación de alto nivel, como Pascal, con notaciones puramente lingüísticas de la Fonología Generativa.

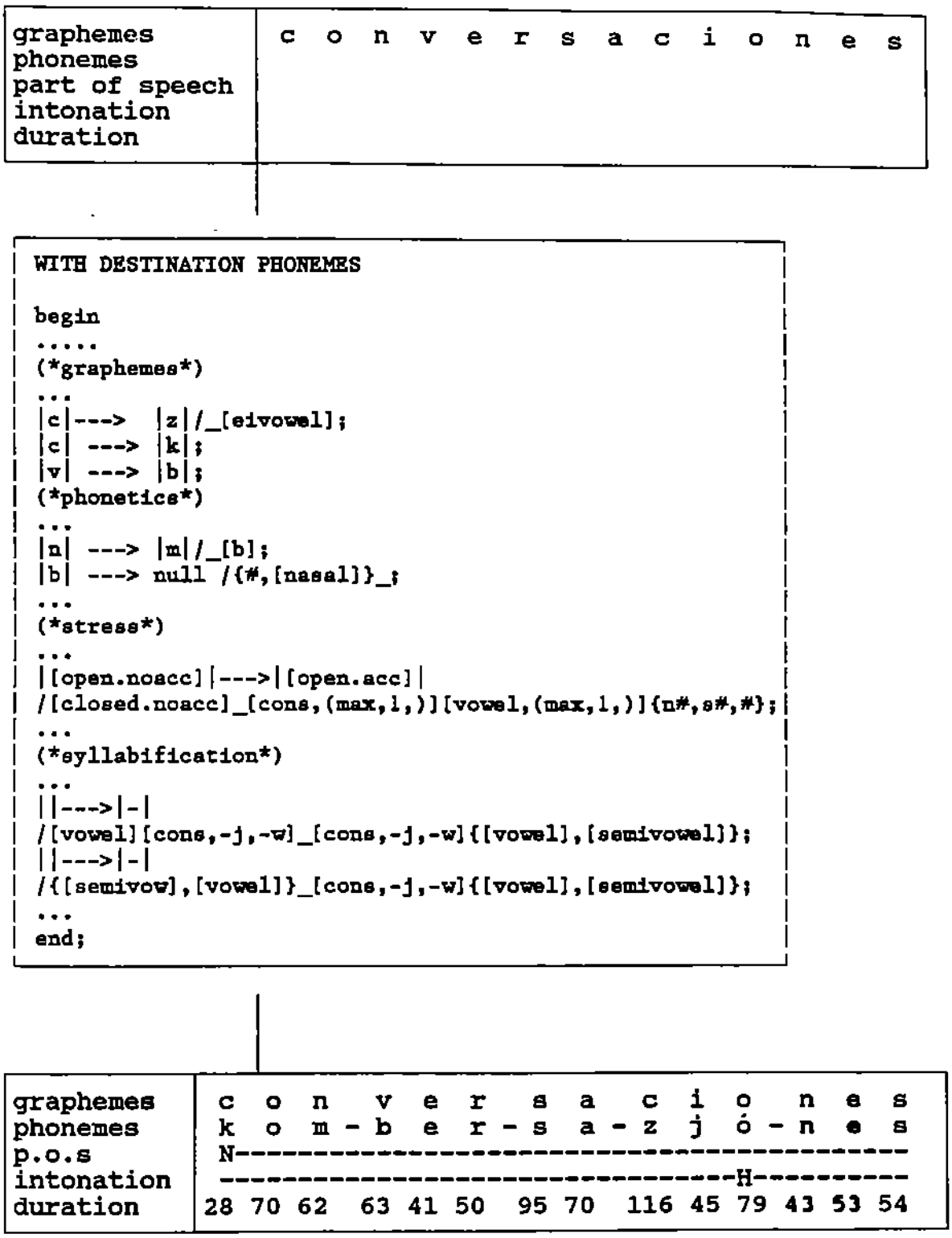
En el marco de desarrollo de DEPES el lingüista puede crear (incluso sin ser especialista en computadoras) gramáticas para cualquier nivel de análisis (fonético, morfológico, léxico, sintáctico). Estas gramáticas pueden ser desarrolladas, llevadas a la práctica y puestas a prueba en un corto plazo de tiempo. También puede ocuparse de problemas de normalización del texto como los mencionados más arriba. El rendimiento de las reglas escritas en el formalismo de DEPES es evaluado sobre amplias bases de datos y textos introducidos al sistema por medio de procedimientos estadísticos especialmente diseñados al efecto. La base de datos para la lengua española fue desarrollada por nosotros mismos.

DEPES también facilita varias utilidades como compilador de reglas, linker, debugger y servicio de diccionarios. Recientemente se ha añadido al sistema un programa para el aprendizaje inductivo automático de reglas de pronunciación.

El lenguaje de reglas en DEPES: algunos rasgos

Todas las reglas lingüísticas operan sobre una estructura central dispuesta en niveles. El usuario puede definir tantos niveles como sean necesarios. La figura 2 muestra una estructura de datos en cinco niveles: grafemas, fonemas, parte de la oración, entonación y duración.

Figure 2.


graphemes
phonemes
part of speech
intonation
duration
c o n v e r s a c i o n e s
WITH DESTINATION PHONEMES
begin
.....
(*graphemes*)
...
|c|---> |z|/_[eivowel];
|c| ---> |k|;
|v| ---> |b|;
(*phonetics*)
...
|n| ---> |m|/_[b];
|b| ---> null /{#,[nasal]}_;
...
(*stress*)
...
|[open.noacc]|--->|[open.acc]|
/[closed.noacc]_[cons,(max,1,)][vowel,(max,1,)]{n#,s#,#};
...
(*syllabification*)
...
||--->|-|
/[vowel][cons,-j,-w]_[cons,-j,-w]{[vowel],[semivowel]};
||--->|-|
/{[semivow],[vowel]}_[cons,-j,-w]{[vowel],[semivowel]};
...
end;
graphemes
phonemes
p.o.s
intonation
duration
c o n v e r s a c i o n e s
k o m - b e r - s a - z j ó - n e s
N----------
----------H----------
28 70 62 63 41 50 95 70 116 45 79 43 53 54

Durante el proceso de conversión de texto en voz, las gramáticas lingüísticas y los diccionarios se utilizan para modificar la estructura de datos. Así por ejemplo, las reglas mostradas en la figura realizan operan en el nivel de los "grafemas" y modifican el segundo nivel de los "fonemas". El resultado del sistema proporcionará una transcripción fonética incluyendo la conversión de grafemas en símbolos fonéticos, la asignación del acento y la silabificación.

4.CONCLUSION.

El diseño y desarrollo de un sistema de conversión de texto en voz requiere conocimientos de diversas áreas y ciencias (lingüística, ingeniería, informática, inteligencia artificial). Por consiguiente se requiere una aproximación multidisciplinar. Nosotros utilizamos un diseño general sobre varias e incluso muy diferentes lenguas. Gracias a herramientas flexibles y sofisticadas es posible desarrollar e integrar conocimiento lingüístico específico de una manera rápida y eficiente. Esta estrategia ha probado ser adecuada para el desarrollo de un conversor de texto en voz para la lengua española.

5. BIBLIOGRAFIA.

[1] VAN COILE, B., "The DEPES development system for Text-to-Speech synthesis", Proceedings IEEE, 1989.

-------- "Inductive learning of pronunciation rules with the DEPES system", Proceedings ICASSP, 1991.

[2] OLABE, J.C., et alia, "Real time text to speech conversion system for Spanish", Proceedings IEEE, 1984.

[3] SANTOS, J.M. y NOMBELA, J.R., "Text-to-Speech conversion in Spanish, a complete rule-based system", Proceedings IEEE, 1982.

[4] ROCA, J.M. et al., "SINCAS: un conversor texto-voz en castellano", Boletín SEPLN, n_5, pp. 112-122, Mayo 1985.

[5] ROMANO, J., "Un sistema automático de síntesis mediante semisílabas", Boletín SEPLN, n_2, pp. 34-45, Mayo, 1984.

[6] "Monográfico sobre reconocimiento y síntesis del habla", Boletín SEPLN, n_6, Junio, 1988.

[7] PARDO, J.M. et al. "Improving Text To Speech Conversion in Spanish:linguistic analysisis and prosody". Proceedings IEEE, 1987, vol.2.

[8] QUILIS, A., "El empleo de los ordenadores en la investigación fonética", Lingüística Española Actual, III, 1981, pp. 197-219.

[9] HART, J.'t y COLLIER, R., "Integrating Different Levels of Intonation Analysis", Journal of Phonetics, vol.3, pp.235-255.