Gramática para el análisis del diccionario VOX

M. Antònia Martí
Universidad de Barcelona

Irene Castellón
Universidad Politécnica de Cataluña

Abstract

La presente comunicación expone una parte del trabajo realizado en el marco del proyecto Esprit Acquilex, y trata de manera específica de la gramática realizada para el análisis del diccionario VOX.

Este proyecto consiste en un programa de investigación para el desarrollo de técnicas y metodologías que extraen la información contenida en diccionarios de uso público que disponen de una versión en soporte magnético (MRDs), con el objetivo de construir el componente léxico de sistemas de procesamiento del lenguaje natural.

No se dispone todavía de un sistema completo que produzca diccionarios computacionales a gran escala aprovechando los diccionarios ya existentes. Se trata en definitiva del desarrollo de herramientas y propuestas metodológicas orientadas a este fin.

Distinguiremos diversas etapas en la constitución de bases de conocimiento léxico a partir de MRDs:

a)- Tratamiento del MRD para volcar la información que contiene en una estructura de base de datos (Base de Datos Léxica, BDL).

b)- Especificación e implementación de métodos automáticos y semi-automáticos para extraer la información implícita y explícita contenida en los diferentes campos de la BDL.

c)- Definición de una estructura común a todos los diccionarios fuente: C.L.E. (Common Lexical Entry)

d)- Definición y construcción de la Base de conocimiento léxico (BCL) a partir de la información obtenida en b).

La presente comunicación se centra en el apartado a y trataremos también algunos aspectos de b.

Gramática para el análisis del diccionario VOX

M. Antònia Martí
Universidad de Barcelona

Irene Castellón
Universidad Politécnica de Cataluña

0.- Introducción

La presente comunicación expone una parte del trabajo realizado en el marco del proyecto esprit Acquilex, y trata de manera específica de la gramática realizada para el análisis del diccionario VOX.

Este proyecto consiste en un programa de investigación para el desarrollo de técnicas y metodologías para extraer la información contenida en diccionarios de uso público que disponen de una versión en soporte magnético (MRDs), con el objetivo de construir el componente léxico de sistemas de procesamiento del lenguaje natural (Boguraev & Briscoe 1987 - 89).

No se dispone todavía de un sistema completo que produzca diccionarios computacionales a gran escala aprovechando este tipo de diccionarios. De momento se trata, en definitiva, del desarrollo de herramientas y propuestas metodológicas orientadas a este fin.

herramientas en diferentes áreas:

- lexicografica: desarrollo de una macroestructura (templeta) resultado del análisis y comparación de diferentes Bases de Datos de diccionarios: se trata de una base de datos común a diferentes diccionarios, no de un diccionario concreto (contexto multilinge).
- Computacional: desarrollo de analizadores, de entornos para la adquisición de conocimiento léxico, etc.
- lingística: Formalizacion del contenido, elaboración de la gramática de las definiciones, representación de su contenido semántico y de otros campos relevantes, etc.

Las principales características del proyecto son:

- contexto multilinge y múltiples fuentes de conocimiento. (LDOCE, COLLINS, COBUILD, GARZANTI, DMI DATABASE, VAN DALE, VOX)
- Diccionarios bilinges: transferencia de información léxica
- derivación de una estructura conceptual común
- representación de los datos léxicos en forma de estructuras jerárquicas.

Distinguiremos diversas etapas en la constitución de bases de conocimiento léxico a partir de MRDs:

a)- Tratamiento del MRD para volcar la información que contiene en una estructura de base de datos léxica: cuerpo

léxico de un amplio abanico de aplicaciones prácticas del LN.

b)- Especificación e implementación de métodos computacionales y semi-automáticos para extraer la información implícita y explícita contenida en los diferentes campos de la BDL.

c)- Definición y construcción de la Base de conocimiento léxico (BCL) a partir de la información obtenida en b).

La presente comunicación se centra en el apartado a y trataremos también algunos aspectos de b.

1.- Tratamiento del MRD

El diccionario fuente del que se parte es un texto continuado que contiene dos tipos de información: el texto del diccionario y códigos de dos tipos:

- Códigos correspondientes a la información léxica, que deben preservarse para su edición (p.e.:sin,., fam., dim., etc.) y

- Códigos tipográficos, que no tienen ningún tipo de relevancia para los lectores, pero que son fundamentales en dos sentidos:

  - porque expresan códigos de edición y

  - [illegible] identifican los

entradas, los campos de cada entrada, los ejemplos, etc.

Este tipo de información se incorpora a las reglas de la gramática de segmentación del texto fuente.

El primer paso que se debe realizar para manejar la información del diccionario es la construcción de una gramática que analice este continuum gráfico y lo segmente de manera pertinente por entradas y, cada entrada, por los diferentes campos que contiene.

Para ello es necesario un estudio previo del diccionario fuente: qué tipo de información contiene, cómo se encuentra estructurada, qué criterios se han seguido en la ordenación de homógrafos, etc.

1.1.- Caracterización del diccionario fuente

Los criterios de edición y la información contenida en los diccionarios de uso general difieren de la información que deben contener los léxicos para el procesamiento del lenguaje natural. El estudio del diccionario fuente tiene como objetivo determinar qué información se utilizará, cual se desestimará y, para la primera, qué cambios se le deberán aplicar en su volcado a la BDL. En esta línea es fundamental determinar que criterios se han seguido en el diccionario fuente para el tratamiento de homógrafos

y entradas y la estructura de cada entrada para determinar de
manera se resolveran en la LDB.

Un ejemplo a tener en cuenta es el criterio en la ordenac
de las acepciones. En el diccionario fuente se adopta un crite
històrico y cuando éste no es posible se adopta un crite
didáctico, como por ejemplo agrupar sentidos análogos. En la
conviene tener los sentidos de las entradas ordenados de
general a más específico para lo cual se deben definir métodos p
identificar los sentidos más generales.

1.2.- Descripción de la gramática

Las reglas de la gramática incluyen los códigos tipográficos
que son indicadores de campos), las etiquetas asociada:
determinado tipo de información, la categoría de la entrada,

El resultado de este proceso será el mismo texto del
cionario organizado según un estructura predefinida y etique
por campos (estructura parentizada) del que se han eliminado
caracteres tipográficos que a partir de este momento ya no so
ningún interés.

1.2.1.- Estructura de las entradas

La construcción de la gramática exige un estudio previo

Considerando la información asociada a cada una de las entradas del diccionario y el orden en que se expone, hemos dividido la estructura de una entrada en ocho campos que nos proporcionan diferentes tipos de información. Se trata de campos delimitados mediante códigos tipográficos y códigos del diccionario de uso común.

ENTRADA: nos da su forma y flexión.

ETIMOLOGIA: informa sobre la procedencia del término.

CATEGORIA: consta de una serie de terminales definidos por el diccionario donde aparecen tanto categorías gramaticales como rasgos morfológicos, (p.e.: adj., fem., m,pl.,etc.)

CONTEXTO: incluye la información geográfica y el área temática.

USO: analiza los códigos de usos particulares (figurado, familiar, etc.)

DEFINICION: se compone del texto de la definición y de jemplos y locuciones. En este momento estamos realizando un estudio sobre los diferentes esquemas de definición para nombres, adjetivos y verbos (Alsawi -1989, Meijs -1989).

FORMA: homófonos de la entrada, superlativos, modelos de conjugación verbal, usos y formas incorrectas, plurales y femeninos irregulares, impropiedades y referencias cruzadas.

RELACIONES informa de las relaciones de la entrada con otras entradas del diccionario, las diferentes relaciones son: sinónimos,

contrarios, ejemplos y frases hechas, sentidos figurados, comjugaciones irreguares, locuciones y formas variantes.

1.2.2.- Diccionario

Para el análisis de las entradas ha sido necesaria la creación de un diccionario (tabla Hash) que se compone de categorías y valores, siendo los valores aquellos códigos que vienen ya definidos por el diccionario: categoría, códigos de uso, lugar geográfico y área temática, etiquetas de diferentes relaciones o formas, y otros: números de acepciones y de homógrafos, signos de puntuación y códigos tipográficos.

El resto de categorías que aparecen en la gramática son partes de texto (entrada, flexión, texto de definición, comentarios de formas y relaciones, etc.) que categorizamos como <*TX> y se distinguirán internamente. Pondremos algunos ejemplos de categorías y valores:

CATG: adj.,adj.-f,adv.neg.,com.,f.,m.,etc.

USO: fam., fest., fig., hum., iron.,etc.

TEMA: aeron.,taurom., quím., etc.

GEO: P.Rico, Sal., Al., S.Dom.,etc.

1.2.3.- Análisis gramatical

Para el análisis del MRD se ha utilizado un analizador descenciente, izquierda-derecha e independiente del contexto. Está implementado en LISP.

Las primera regla de la gramática define la estructura general de las entradas, su forma es:

```
DICCION    = ENTRADA [ETIMOLGIA] ACCEPCIONES FORMAS
             RELACIONES '.'.
```

La información referente al uso, área temática, lugar geográfico y categoría la incluimos en el nodo ACEPCIONES. Como ejemplo mostraremos el desarrollo del nodo ENTRADA:

```
ENTRADA     = *NEGR [*NH *PUNG] ( LEXS/ PREFIJO /
SUFIJO) [*COMA] *RED.
```

Donde LEXS es un nodo recursivo que analiza tanto entradas simples como compuestas (con guión intercalado) junto a su flexión si ello es pertinente (con el término "entrada" nos referimos aquí a la palabra que encabeza un artículo de diccionario, en inglés "headword"):

```
LEXS       = LEX [*COMA LEX].
LEX        = *ENTR FLEXS.
FLEXS      = *COMA [*PUNG] *FLEX FLEXS/ *NULL.
```

PREFIJO y SUFIJO son los nodos que desarrollan el análisis de prefijos y sufijos.

PREFIJO = PREF PREFIJO/*NULL.

PREF = *PRF*PUNG [*COMA].

SUFIJO = SUF SUFIJO/ *NULL.

SUF = *PUNG *SFJ [*COMA].[1]

Estas tres reglas como se puede observar son recursivas puesto que en el diccionario aparecen casos de entradas encadenadas por comas.

La recursividad es una propiedad bastante común en los diversos campos contenidos en la información de cada entrada. Los campos recursivos son: entrada(lex, prefijo y sufijo), flexión, geos, tema, usos, relaciones y formas.

La recursividad, sin embargo, no se realiza de igual modo en todos los campos. El campo TEMA realiza la recursividad de forma encadenada entre dos códigos tipográficos p.e.: 'DER.ASTRON.' del diccionario editado se traduce en el diccionario en cinta en 'VERSAL DER. ASTRON. CANVERSAL' siendo VERSAL y CANVERSAL códigos tipográficos y DER. y ASTRON. terminales de la categoría TEMA (derecho y astronomía).

Otros campos, como GEO, necesitan la coordinación para encadenarse. Esto representa que la coordinación no aparece en el mismo código que la información relevante, p.e.: 'Argent. y Chile' en el diccionario en cinta como

CUR Argent RED y CUR Chile. Esto implica un estudio detallado del tipo de recursividad que sigue cada campo.

A continuación veremos de un modo muy general el resto de la gramática.

ETIMOLOGIA: campo referente a la información sobre el origen etimológico de la entrada. Está delimitado por paréntesis.

ETIMOLOGIA = *RED *PUNPA *ETIM *PUNPC.

ACEPCIONES: Es el campo más complejo, se expresa con una regla recursiva debido a que puede aparecer numerosas veces pero con alguna variante por lo que se duplica en dos reglas de distinto nombre:

ACEPCIONES = ACEPCION1 (ACEPCIONES2/*NULL).
ACEPCIONES2 = ACEPCION (ACEPCIONES2/*NULL).

que se reescriben:

ACEPCION1 = *CUR [*CATG]
GEO [INFLEX] TEXT.

En este primer caso se contempla el análisis de la primera acepción que se caracteriza por carecer de número.

ACEPCIOn = *CUR [*PUN-] *NUM [*CATG]
GEO [INFLEX] TEXT.

En el segundo caso encontramos en primer lugar el código

tipográfico, seguido opcionalmente de un guión si aparece la información de CATG (categoría), seguido del número de acepción. El nodo CATG indica la categoría de la entrada, entendiendo 'categoría' de una forma amplia ya que nos proporciona tanto información de tipo categorial (nombre, adjetivo, verbo...)como morfológica (género y número).

A continuación el nodo INFLEX proporciona información léxica delimitada por los códigos tipográficos correspondientes:

INFLEX = *VERSAL [(*TEMA)] *CTTX5.

INFLEX nos informa sobre el área temática en la que se emplea el término (TEMA).

TEXT, es el último nodo de acepción, se reescribe del siguiente modo:

TEXT = *RED [*IL] *TP *PUNTO .

TEXT analiza la información léxica referente al ámbito de uso del término (P.e.: fig., fam., rust., iron., etc.), y el texto de la definición, englobado en TP.

Después de ACEPCION encontramos el campo FORMA:

FORMA = ´; PG´ *VERSAL *FORM *CANVERSAL *CUR *TX

Este campo se compone de una etiqueta que indica el tipo de información de que se trata y el texto. Las informaciones que proporciona este campo son:

CONJUG. indica el modelo de conjugación si la entrada se trata de un verbo;
HOMOF. Homófono;
ES INCOR. uso incorrecto;
INCOR. forma incorrecta;
PL. forma plural;
V. referencias cruzadas; etc.

El último campo, RELACIONES, es recursivo ya que puede incluir diferentes relaciones en una misma entrada.

RELACIONES = RELACION RELACIONES / *NULL.
RELACION = *INTRSIN *MAY *REL *CANMAY *CUR *TX.

Informa sobre diferentes relaciones entre la entrada y otras entradas del diccionario. Estas relaciones pueden ser, entre otras:

REL entradas relacionadas;
SIN sinónimos;
CONTR antónimos;
EXPR en frases hechas expresiones; GR o

GRAM.construcciones correctas o especiales de tipo sintáctico.

1.3.- Problemas de implementación de la gramática

Los problemas que plantea la implementación de la gramática son:

- Problemas concernientes a errores tipográficos, que lógicamente aparecen al tratar una obra tan extensa. Conviene resaltar que algunos de estos han sido tipificados e incluidos en la gramática debido a su frecuente aparición.

- Problemas concernientes a las reglas de la gramática: iniciamos el análisis a partir de una muestra del diccionario (397 entradas) y por lo tanto al analizar el diccionario en su totalidad aparecen estructuras o terminales no previstos en la gramática. Para ejemplificar este tipo de problemas detallaremos algunos casos:

  - La aparición de categorías terminales en campos no previstos y sin los códigos tipográficos pertinentes.
  P.e.: la categoría de una entrada aparece al inicio de cada acepción con el código tipográfico correspondiente a la cursiva (adj., prep., etc.). Estas formas pueden aparecer también en el interior de definiciones sin su código tipográfico específico. Debido a esto se ha desarrollado un

el código tipográfico que les corresponde es decir cuando su estado es el correcto.

- La presencia de terminales no contemplados: La documentación del diccionario editado incluye una lista de todas las abreviaturas existentes. En el texto aparecen estas formas sin abreviar, esencialmente en el campo GEO.
P.e.: la forma 'Argent.' nos informa de que un término se utiliza en Argentina, pero en otras entradas aparece 'Argentina'. Para solucionarlo se añaden las variantes, que proporcionan la misma información, en el analizador morfológico.

- La aparición de terminales homónimos es bastante frecuente y se suele resolver mediante el programa de estados.
P.e.: **com.** puede aparecer como 'comercio' en TEMA y como 'nombre común' en CATG.

2.- Estructura parentizada

El programa lisp que realiza el análisis, una vez identificados todos los campos constitutivos de cada entrada, da como resultado una estructura parentizada, donde la información

2.1.- Clasificación de la información

El criterio que hemos seguido para definir la estructura parentizada ha sido considerar como propia de la entrada la información común a todas las acepciones y como propia de cada acepción aquella información que varía de una acepción a otra. Así consideraremos que hay un nivel general de **entrada** y otro más específico de **acepción**.

La información de carácter obligatorio que depende de cada acepción (la categoría), se asume que es la misma para todas las acepciones a no ser que aparezca un nuevo valor para este campo.

Los campos que corresponden al nivel de la entrada son: ENT, palabra de entrada (headword); ETIM, etimologia ; NH, número de homógrafo; FLEX, flexion; FORM, formas especiales ; REL, entradas relacionadas.

Los campos correspondientes al nivel de la acepción son: NA, número de la acepción; CA, categoría; TEMA, área temática; GEO, lugar geográfico; USO, ámbitos de uso; DEF, definición.

Los campos, a su vez, están clasificacdos en campos de información sintáctica, campos de información semántica, campos de

forma y campos que muestran relaciones de la entrada. A continuación damos una relación de esta clasificación:

- campos de información sintáctica:
  - CA: categoria de la acepción
- campos de información semántica:
  - USO: ámbitos de uso
  - TEXT: definición de la acepción
- campos de información contextual
  - GEO
  - TEMA
- campos que aportan información sobre la forma:
  - FLE: flexion de la entrada
  - FORM: formas especiales de la entrada
  - ETIM: etimología
- campos de relación:
  - REL: entradas relacionadas

El esquema de la estructura parentizada es el siguiente:

(ENT: *entrada*
(FLEX: *flexión*
(NH: *n de homógrafo*
(ETIM: *etimología*
(SENSE
(NA: *n de acepción*
(CAT: *categoría*
(GEO: *información geográfica*
(TEMA: *materia o área temática*
(USO: *ámbito de uso*
(DEF: *texto de definición y ejemplos*

(TXF: *texto de forma*
(TIPOR: *etiqueta de relación*
(TXR: *texto de relación*

A continuación presentamos un ejemplo donde podemos apreciar el resultado de la aplicación de los procesos descritos a las entradas del diccionario.

entrada:

> I) **cacho** (l.*calculu*, piedrecita) *m.* fam. Pedazo pequeño de alguna cosa. 2. Cierto juego de naipes. 3. *Méj* y *P.Rico*. Participación pequeña en un número de la lotería. SIN. 1.V. **Pedazo.**

E.Parentizada:

```
(ENT: cacho
(FLEX:
(NH: I
(ETIM: l.calculu. piedrecita
(SENSE
     (NA: 1
     (CAT: m.
     (USO: fam.
     (DEF: pedazo pequeño de alguna cosa
(SENSE
     (NA: 2
     (CAT: m.
     (DEF: cierto juego de naipes
(SENSE
     (NA: 3
     (CAT: m.
     (GEO: Méj y P.Rico
     (DEF: participación pequeña en un número de la
     lotería
(TIPOR: sin
(TXR: v. 1.pedazo.
```

A partir de esta estructura, los datos deben ser almacenados en una templeta, otro tipo de estructura que guarda la información ordenadamente por conceptos y mantiene el texto fuente por partes de manera íntegra. Esta estructura es común a todos los grupos del proyecto y se ha definido a partir de la información contenida en todos los diccionario sobre los que se trabaja.

La templeta se compone de grupos de información morfológica, sintáctica, semántica,sobre variantes, relaciones, etc. De este modo se almacena toda la información clasificada y acompañada del texto fuente.

El paso de la información desde la estructura parentizada a la templeta provoca una serie de problemas. En un diccionario editado el mismo tipo de información puede aparecer en diversos campos y, por el contrario, un campo puede informar de dos cuestiones diferentes. Observemos unos ejemplos:

- En el diccionario fuente, las frases hechas pueden aparecer en el campo DEFINICION o bien en el de RELACION. En la templeta esta información procedente de dos campos distintos debe almacenarse en un único grupo.

- Un caso contrario sería el de CATEGORIA. Como ya hemos dicho el diccionario fuente agrupa en este campo información morfológica y sintáctica; la templeta dispone de dos grupos: el sintáctico y el morfológico por lo que la información procedente de la estructura parentizada se deberá desglosar.

3. Analisis del campo definición

El análisis del campo definición es necesario para la creación de taxonomías, un paso previo a la construcción de la Base de Conocimiento Léxico (BCL).

El primer objetivo de este análisis es la extracción del término genérico de las definiciones de las entradas del diccionario. De este modo se crean unas relaciones jerarquizadas que permiten la clasificación de los conceptos.

Para elanálsis de las definiciones utilizamos el analizador

de H. Alshawi(Alshawi-1989). El análisis precisa el estudio previo de:

a) Categorización de las palabras.
b) Identificación del término genérico para cada categoría.
c) Tipificación de la estructura de las definiciones.

Posteriormente se deberán definir métodos de desambigación de las diversas acepciones para obtener el sentido adecuado en la construcción de la taxonomía.

Trabajos realizados y líneas de investigación iniciadas

La gramática del MRD se ha aplicado ya sobre un total de 10.000 entradas con resultados positivos en un 97% de los casos. Es de suponer, por el volumen de la muestra analizada, que este tanto por ciento variará poco para el resto del diccionario.

Paralelamente al análisis del MRD se realiza el volcado a la LDB.

En la actualidad se ha iniciado el análisis del campo definición. Prar ello se deben determinar los diferentes modelos o esquemas de definición de cada categoría y especificar para cada uno de ellos cual es el término genérico y los modificadores.

La gramática de las definiciones (Alshawi 1989) tiene como objetivo identificar automáticamente estos elementos.

Al mismo tiempo que desarrollamos la gramática definimos diversas estrategias para la desambigación de los términos genéricos obtenidos con el objetivo de construir la estructura taxonómica.

BIBLIOGRAFIA

ALSHAWI, H.A. (1989) "Analysing the Dictionary definitions" in Boguraev, B.- Briscoe, E. eds. Computational Lexicography for Natural Language Processing Longman. Londres.

BOGURAEV, B. - BRISCOE, Ed. (1987) "Large Lexicons For Natural Language Processing: Utilising the grammmar Coding System of LDOCE" Computational Linguistics vol.13 núm. 3-4. págs. 203-218.

BOGURAEV, B.K. - E. BRISCOE (1989) Computational Lexicography for Natural Language Processing Longman, London.

BYRD, Roy- N. CALZORALI- M.S. CHODOROW- J.L. KLAVANS- M.S. NEFF- O.A. RIZK (1987) "Tools and Methods for Computational Linguistics" en Computational Linguistics vol. 13 núm. 3-4 págs. 219-240.

CALZORALI, N. - PICCHI, E. (1988) "Adquisition of Semantic Information from an On-Line Dictionary" en Proceedings of the 12 International Conference on Computational Linguistics, Coling'88, págs. 238-242.

CALZORALI, N. "The Dictionary and the Thesaurus can be combined" (1988) en Relational Models of the Lexicon" M. Walton Evens ed. Cambridge University P. Cambridge.

Diccionario General Ilustrado de la Lengua Española VOX
Ed. Bibliograf S.A.; Barcelona, septiembre 1987

MEIJS, W.- M. van der BROEDER- P. VOSSEN, P. (1989) "Meaning and Structure in dictionary definitions" en Boguraev, B. - E. Briscoe eds. Computational lexicography for Natural Language Processing Longman. London.