http://dx.doi.org/10.12795/PH.1997.v11.i02.07

CODIFICACIÓN EN EL LEXICÓN DE LAS RELACIONES DE CONCURRENCIA

Mª Ángeles Zarco Tejada

Lexical selections are not predictable from semantic information since they obey to arbitrary reasons. Thus, co-occurrence restrictions should be codified if high-quality translations count among one of our aims. We claim here that Lexical Functions seem to be a type of interlingual, expressive and monotonic representation suitable for a computational interpretation and that they should be included in the lexicon as part of the information given for every lexical entry. Using such type of formalism we avoid using bilingual context-dependent rules.

1. EL LEXICÓN: CARACTERIZACIÓN Y EJEMPLOS.

Los datos lingüísticos en un sistema de Traducción Automática (TA) se dividen en datos gramaticales y datos léxicos. Los primeros reponden a las gramáticas que se utilizan en las rutinas de análisis y de generación, los segundos, a la información específica que cada elemeto léxico de la lengua en cuestión tiene.

El Lexicón de una lengua lista todos los elementos léxicos que pertenecen a dicha lengua correspondientes a cada categoría. Evidentemente, es una parte muy importante de un sistema de TA ya que contiene la información asociada a las palabras que luego habrá de traducir el sistema. Este no es sino el diccionario de consulta del sistema, que presenta unas características peculiares derivadas del entorno de aplicación al que pertenece.

El tipo de información que se codifica en el lexicón está íntimamente relacionado con las decisiones de diseño que tomemos con repecto al lexicón que queremos construir. Básicamente englobaría los siguientes aspectos: **categoría gramatical** (N, V, Adj., etc); **información morfológica**; **rasgos de subcategorización** (tr, int, masc, fem); **información valencial**; **rasgos de casos**; **información semántica** (+animado); así como información sobre **restricciones de selección** (V: "reír": sujeto [+animado]). Es por todo ello que se prefiere el término lexicón al de diccionario.

Si tuviéramos que decir en qué consiste la principal diferencia entre los diccionarios de los sistemas y los de uso humano, diríamos que los diccionarios de los sistemas no pueden evitar lo obvio, sino que deben codificar absolutamente toda la información de manera explícita. Además de esto, dicha información ha de estar codificada de manera que el ordenador la pueda usar. Por ello, en los diccionarios bilingües de TA se debe también

incluir las condiciones bajo las cuales una palabra en la lengua de partida se puede traducir por otra equivalente en la lengua de llegada, ya sean condiciones gramaticales, semánticas o estilísticas.

A manera de ilustración enumeramos a continuación la base de datos léxicos de SYSTRAN (Hutchins & Somers, 1992) (sistema de la Comisión de las Comunidades Europeas en Luxemburgo):

1. Un diccionario de las formas principales.
2. Un diccionario bilingüe de entradas individuales.
3. Varios diccionarios contextuales.

En el diccionario de las formas principales se da una descripción morfológica, sintáctica y semántica completa: categoría gramatical, rección, valencia, concordancias, transitividad, tipo de N (abstracto, contable), marcadores semánticos y también una traducción de la base en una palabra equivalente de llegada, acompañada de la información gramatical necesaria para su generación. Se hace una distinción para los homógrafos con diferente categoría gramatical, que tienen diversas entradas, y homógrafos de la misma categoría que se manejan como polisemias por los diccionarios contextuales.

Los diccionarios contextuales se derivan en SYSTRAN automáticamente de un sólo diccionario de partida:

1) Diccionario idiomático, designado para tratar con las expresiones fijas.
2) Diccionario de semántica limitada, define el ámbito de las relaciones sintácticas en los Sintagmas Nominales.
3) Diccionario de homógrafos, lista la información contextual sintáctica requerida para la resolución de algunos homógrafos.
4) Diccionarios analíticos, contienen las excepciones a las reglas sintácticas generales que se aplican a palabras particulares: **nor+inverted subj. N and V** *"nor could he see the difficulties"*.
5) Diccionario de semántica condicional, interviene en el tramo de *transfer* para hacer la última selección léxica del término de la lengua de llegada.

2. LAS RELACIONES DE CONCURRENCIA Y SU CODIFICACIÓN.

En este artículo hacemos una exposición de la necesidad de codificar como parte de la información del lexicón el Nivel Sintagmático. El objeto lingüístico de nuestro análisis serán las **Colocaciones** y más específicamente los **Predicados Complejos** (PCs).

Cuando hablamos de colocación nos referimos a la relación léxica entre elementos lingüísticos, por tanto nos referimos a una estructura que consta de al menos dos palabras, que suelen o deben aparecer juntas en relación sintáctica directa en el texto para designar un

significado específico (a pesar de que existan otras palabras con significado similar) (Mitchell, 1971; Zuluaga, 1975; Aisenstadt, 1978; Cowie, 1981; Sinclair, 1991).

Principales tipos de construcciones donde tal selección léxica ocurre, según Allerton (1984):

1. N --> prep de control: *on Friday, at Christmas.*
2. N --> prep. de complemento: *attack on, defence of.*
3. V --> prep de complemento: *despair of, hope for.*
4. adj --> prep. de complemento: *free from/of, essential to.*
5. N deverbal --> V general transitivo: *make a suggestion, put a question, give an answer.*

Un Predicado Complejo (n1 5 en Allerton) es una estructura lingüística (colocación) formada por un verbo (colocante) y un nombre (base), que presenta las siguientes características:

1. El verbo no 'suele' tener una gran carga semántica (Cattell, 1984). Utilizando un término de Jespersen, son verbos 'light' en cuanto a que no aportan gran contenido semántico a la acción expresada por la estructura compleja: ***hacer** una oferta, **dar** una patada.*
2. Sin embargo, existen algunos casos donde los verbos ofrecen un contenido específico (Allerton, 1984): ***echar** un vistazo, **ofrecer** disculpas.*
3. Gran parte de dichos Predicados Complejos tienen el mismo contenido semántico que sus correspondientes verbos lexicalizados:

   hacer una oferta — ofertar

   dar un beso — besar

   echar un vistazo — ojear

4. Aún cuando el verbo que forma parte de dichos PCs sea de carácter general, existe una relación de dependencia entre el V y el N, de tal manera que el N rige al V y no otro en tal estructura. El verbo se encuentra léxicamente regido, existiendo un número muy reducido de posibles sinónimos que puedan aparecer en lugar del primero: ***echar/dar/*otorgar** un vistazo.*
5. El significado del todo ha de ser composicional. Es una expresión divisible en constituyentes semánticos a diferencia de las frases idiomáticas (Cruse, 1986).
6. Son estructuras con un mayor o menor grado de cohesión, dependiendo del número de posibles sinónimos que admita, a diferencia de las frases idiomáticas que son inmutables en relación a operaciones de sustitución, trasposición y expansión (Cowie, 1981): *to commit suicide — to commit a crime.*

Existen dos métodos generales para distinguir las colocaciones del discurso libre: el **método lexicográfico** y el **estadístico**, el primero basado en la intuición y la experiencia del especialista, el segundo en la frecuencia. El primero evitaría que se tomen en consideración colocaciones del tipo **el-hombre**; **doctor-enfermera** (Church & Hanks, 1990; Smadja, 1993).

Hemos dicho anteriormente que los PCs presentan en algunas ocasiones correspondientes lexicalizados con el mismo contenido semántico. Sin embargo, puesto que nuestro marco de estudio de los PCs es el de la TA, tenemos que aludir a un segundo tipo de relación Paradigmática, las relaciones de transfer, que relacionan una palabra o frase en una lengua con otra semánticamente equivalente en otra lengua. (Evens, 1988; Tsujii, 1989).

En este sentido, en la traducción de un PC se pueden dar las siguientes posibilidades:

```
PC --> PC

Daniel da un paseo --> Daniel takes a walk

PC --> Verbo

Daniel pone la zancadilla a su madre -->Daniel trips up his
   mother

Verbo --> PC

Daniel se suicidó --> Daniel commited suicide
```

¿Cuál es la relevancia de las colocaciones en nuestro ámbito de aplicación?

El uso correcto de combinaciones de lexemas no debe darse por supuesto ni para los hablantes nativos, ya que las colocaciones son generalmente arbitrarias e impredecibles, y, además, varían en gran medida dependiendo del contexto. Podría decirse que el uso correcto de las colocaciones de un hablante nativo es un reflejo de su competencia fraseológica y de un alto dominio del idioma.

Para Mackin (1978), las colocaciones no presentan un problema de comprensión, sino de selección del lexema más usual o natural. Veamos el clásico ejemplo de Mel'...uk:

Inglés: TO ASK a question

Francés: POSER une question

Español: HACER una pregunta

Ruso: ZADAT ('dar') vopros

Sorprendentemente, apenas existen diccionarios de colocaciones en las lenguas. Además, la mayoría de los diccionarios cuentan con colocaciones, pero no léxicas, sino gramaticales.

En Lexicografía Computacional se ha venido haciendo un considerable esfuerzo de investigación sobre las colocaciones en los últimos años y quizá uno de los ejemplos más notables es el proyecto DECIDE (Bárcena & Gerardy, 1995). Sus objetivos son los siguientes: 1) la realización de un estudio exhaustivo de recopilación y evaluación de las herramientas automáticas existentes para la extracción de colocaciones de diccionarios (Cobuild, Collins-Robert) y córpora; 2) el diseño e implementación de una caja de herramientas para este propósito, incorporando y mejorando las herramientas sofisticadas seleccionadas; y, 3) la creación de diccionarios automáticos de colocaciones para inglés, francés y alemán. Existe una clara visión de la necesidad de formalizar el conocimiento semántico de las entradas de las colocaciones.

2.1 Restricciones Léxicas en los Sistemas de TA:

En la Generación Automática, las colocaciones juegan un papel muy importante, siendo uno de los elementos determinantes que pueden llegar a diferenciar una producción de alta calidad de una producción de baja calidad. Así, podemos afirmar que es necesario algún tipo de codificación del nivel léxico. Si además, el sistema es multilingüe, el reto es mayor, puesto que la transferencia léxica ha de ser tratada a cierto nivel de abstracción que sea al menos común a los idiomas que cubre el sistema.

Las entradas de los lexicones de generación suelen estar formados, en su mayor parte, por información que no tiene en cuenta el nivel sintagmático, quizás debido a que éste obedece a razones arbitrarias. Puede decirse que las restricciones léxicas han quedado relegadas a un nivel marginal dentro del esquema general de los sistemas de TA. Wordnet (clasificación semántica universal propuesta por Miller et al. 1993), y las restricciones de selección (Dik, 1978) no sirven para nuestros propósitos, ya que lo que se busca no es información de tipo semántico, sino léxico.

En cuanto a SYSTRAN, hasta 1992 al menos no existía un diccionario específico que se ocupara de los PCs. Estas selecciones léxicas parecen estar distribuidas a lo largo de los diversos tipos de diccionarios: e.g., diccionario de expresiones simples, diccionario analítico o diccionario de semántica condicional.

En Rosetta (Appelo & Landsbergen, 1986), las expresiones básicas no son necesariamente elementos léxicos individuales, sino que una expresión compleja puede ser una expresión básica. Así se han solucionado también problemas de correspondencia estructural, como es el caso del español **madrugar**, y de sus equivalentes en inglés **to get up early**, y el holandés **vroeg opstaan**.

El nivel léxico también se tiene en cuenta en Diogenes (Nirenburg 1988) que, en el apartado de generación del lexicón, incluye información sobre colocaciones, codificada bajo el apartado "Syn-collocations-in".

En FLUSH (Balkan, 1993), Jacobs desarrolla un lexicón flexible que permita la representación de colocaciones predicativas, tanto colocaciones gramaticales como verbo-partícula (**fill out**), marcos de subcategorización y predicados complejos (**give a hug**).

En EUROTRA (Balkan, 1993) la definición de los PCs está asociada a la definición de los nombres predicativos. Por ejmplo, un N predicativo (Npred) se define como un **n** que tiene una estructura (Na, Nb, Nc, ...), donde Na=sujeto profundo, Nb=objeto profundo, etc. Así, un Npred puede aparecer en la siguiente estructura:

Det_def Npred [SUBJ_GEN Na][Prepb Nb][Prepc Nc]

The attack of the enemy on the city

Para cada construcción de este tipo existe una relación de paráfrasis con estructuras donde aparecen verbos de apoyo:

John made an attack...

Reproducimos el ejemplo de Balkan respecto a la entrada "**influence**":

{cat=n, predic=yes, is_frame=arg12, pform_of_arg2=over, svneut=have, svincho=gain, svdur=keep, svterm=lose, sviter=none}

donde se especifican los siguientes PCs: have influence, gain influence, keep influence y lose influence.

Danlos y Samvelian (1992) sugieren como metodología para el tratamiento de verbos de apoyo en TA empezar por el sustantivo (el elemento que se considera responsable de la selección del verbo acompañante). Se trataría de evitar las reglas bilingües sensibles al contexto, tales como por ejemplo:

avoir (_habitude) -> be in

Danlos & Samvelian argumentan que el nivel de transferencia debería limitarse a una regla de traducción léxica sencilla **(habitude->habit)** ya que estos elementos no suelen cambiar en la traducción.

Quizás la adaptación más clara de la teoría de Mel'...uk en un sistema de TA sea Heid & Raab (1989), donde, en la parte semántica del diccionario modular, se codifican las restricciones de concurrencia mediante Funciones Léxicas:

```
(problem
(...)
(causfunc(create, pose))
(real(solve, ...)
(...))
```

Para terminar, haremos referencia al trabajo de Smadja & Mckeown (1991), Cook, programa de generación de oraciones que acumula la información colocacional en un lexicón flexible. En éste una entrada está representada por un conjunto de pares de atributos que codifican todas las posibles realizaciones en las que la entrada forma parte de una relación sintagmática restringida para un campo semántico determinado:

Synt_R contiene la palabra o frase y la clase a la que pertenece.

Sem-R representa el significado de la entrada como indicador del campo.

{SV-collocates} lista los colocantes con los que la entrada funciona como sujeto.

{OV-collocates} lista los colocantes verbales con los que la entrada funciona como objeto.

{VO-collocates} lista los objetos colocantes con los que la entrada funciona como verbo.

{NJ-collocates} lista los colocantes adjetivos usados para modificar un N.

{VR-collocates} lista los colocantes adverbios usados para modificar la entrada verbal.

3. LAS FUNCIONES LÉXICAS DE MEL'CUK: MODELO DE CODIFICACIÓN.

Nosotros mostramos a continuación una manera de representar dicho nivel de selección léxica: las Funciones Léxicas.

Una FL es una función en el sentido matemático que representa una idea muy general, tal como 'muy', 'comenzar', 'realizar', etc. o también un cierto rol semántico-sintáctico. Dicha FL asocia con una palabra **P**, denominada su argumento, o **KEY WORD**, el conjunto de palabras o frases que expresan el significado o rol que corresponde a **F**. (Mel'cuk & Zolkovskij, 1988)[1]

Esquemáticamente, la representación de la FL sería:

Nombre de la FL (argumento) = valor

Aunque las funciones se dividen básicamente en dos tipos: **Substitutes** —FLs cuyos valores son sinónimos de los argumentos. (Estos valores son usados "en lugar de" los argumentos: **Syn, Anti, Conv,** entre otras: **Conv(frighten) = to be afraid of**)— y **Semantic parameters** —FLs cuyos valores son expresiones que aparecen en un texto "junto a" su argumento. (Ejemplos de éstas son: **Oper, Func, Caus, Magn**, etc.)—, nosotros definiremos tan sólo aquellas funciones que representan la dependencia léxica que

[1] Nuestra traducción.

se produce en los PCs entre el Nombre y el Verbo —generalmente vacío de contenido— y que son parte de las llamadas "semantic parameters"[2]:

$Oper_1$: La acción básica o relación de D_1 es la de sujeto gramatical sobre la Key word como objeto gramatical:

$Oper_1$(golpe) = dar $Oper_1$(step) = to make

$Oper_1$(medidas) = tomar $Oper_1$(appologies) = to offer

$Oper_2$: La relación básica de D_2 es la de sujeto gramatical de KW como su objeto gramatical:

$Oper_2$(órdenes) = tener $Oper_2$(visit) = to have

$Oper_2$(venta) = estar en $Oper_2$(humiliation) = to undergo

$Func_0$: La acción básica de KW como sujeto gramatical al margen de cualquier otro participante de la situación:

$Func_0$(cambio) = tener lugar $Func_0$(wind) = to blow

$Func_0$(película) = haber $Func_0$(film) = to be on

$Func_1$: La acción básica o relación de KW como sujeto gramatical hacia D_1 como su objeto gramatical:

$Func_1$(remordimiento) = consumir $Func_1$(misfortune) = to befall

$Func_2$: La acción básica de KW como sujeto gramatical sobre D_2 como su objeto gramatical:

$Func_2$(cambio) = afectar $Func_2$(change) = to affect

$Labor_{1,2}$: La acción básica de D_1 como sujeto gramatical sobre D_2 como objeto gramatical, con KW como objeto gramatical marginal:

$Labor_{1,2}$(estima) = tener en $Labor_{1,2}$(oblivion) = to bury in

Pero, además de representar mediante dichas funciones el nivel de dependencia sintagmática, Mel'cuk garantiza la representación de las relaciones paradigmáticas mediante las funciones fusionadas. Así, mediante un mismo tipo de codificación, atendemos a la relación anteriormente citada que se producía entre los PCs y las formas lexicalizadas de la misma lengua:

hacer una oferta — ofertar $Oper_1$(oferta)=hacer//ofertar

---

[2] La definición de los subíndices denominados por nuestro autor como D_1 y D_2 (actantes de estructura sintáctica profunda), así como la determinación de uno nuevo $D_{1\text{-}2}$, son explicados en Zarco, 1994.

dar un beso — besar $Oper_1$(beso)=dar//besar

echar un vistazo — ojear $Oper_1$(vistazo)=echar//ojear

Este tipo de representación cumple además las características que, según Pulman (1991), debe tener todo formulismo gramatical que se intenta tenga una interpretación computacional: Implementable eficientemente, multilingüe, expresivo —debería tener una codificación elegante de conceptos lógicos o lingüísticos—, multiaplicable, facilmente enseñable, reversible —que permita la descripción lingüísitica sea en análisis como en síntesis—, y monotónico —añadiendo una nueva información no afecte la información que ya contiene el sistema.

El interés sobre el uso de las FLs está en construir una estructura de representación para las construcciones colocacionales que son el input y output de correspondencias de transfer y que tienen FLs figurando entre ellas.

La única razón sobre las FLs como representación interlingual reside en el hecho de que las FLs capturan suficiente significado requerido por las bases para que se pueda producir la traducción.Así, las FLs representan una relación importante sintactico-semántica entre la base y colocante.

El potencial combinatorio restringido del lexema colocante se explica listándolo en cada base con la que aparece. Tendremos que crear, pues, en cada entrada un campo dedicado a la información colocacional:

Entrada léxica de "*demonstration*" (Heylen & Verhagen, 1993)[3]:

```
PHON        DEMONSTRATION
SYN         SUBCAT <(PP[2],[PP[3]),[NP[1]>
SEM [5]     RELN DEMONSTRATION
            ARG1 [1]
            ARG2 [2]
            ARG3 [3]
COLLOCATES {give   SEM    REL   OPER1}
                          INST [5]
                   SYN    SUBCAT <NP,[3],[1]>
```

[3] Descripción realizada en HPSG, teoría del lenguaje basada en la unificación. Los elementos en la lista **subcat** de "demonstration" son opcionales, como se indica en la oración de nuestros autores "**John gave a demonstration (of the system) (to the students)**".

REFERENCIAS

AISENSTADT, E. 1978. "Collocability Restrictions in Dictionaries". En Hartmann R.R.K. Dictionaries and their uses. Papers from the 1978 B.A.A.L. Seminar on Lexicography, 4: 71-74.

ALLERTON, D.J. 1984. "Three (or four) levels of word cooccurrence restriction". En Lingua, 63: 17-40.

APPELO, L. & LANDSBERGEN, J. 1986. "The Machine translation project Rosetta". En Gerhardt (ed) I International Conference on the State of the Art in Machine Translation in America, Asia and Europe: Proceedings of IAI-MT86, IAI EUROTRA-D. Saarbrücken:34-51.

BALKAN, L. 1993. "Review Existing Systems". En Collocations ET-10/75. I: 259-285.

BARCENA, E. & GERARDY, C. 1995. "Recycling resources for the creation of a prototype machine-readable dictionary of collocations". En Résumés de la Journée Linguistique. Universiteit Katholieke Leuven.

CATTELL, R. 1984. "Composite Predicates in English". En Syntax and Semantics, 17. London: Academic Press.

CHURCH, K.W. & HANKS, P. 1990. "Word Association Norms, Mutual Information and Lexicography". En Computational Linguistics, 16: 22-29.

COWIE, A.P. 1981. "The Treatment of Collocations and Idioms in Lerner's Dictionaries". En Applied Linguistics, 3: 223-235.

CRUSE, D.A. 1986. *Lexical Semantics*. Manchester. University of Manchester.

DANLOS, L. & SANVELIAN, P. 1992. "Translation of the Predicative Element of a Sentence: category switching, aspect and diathesis". En *TMI-92: Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*. Montreal:21-34.

DIK, S.C. 1978. *Functional Grammar*. North Holland. Amsterdam.

EVENS, M.W. 1988. (ed.) *Relational Models of the Lexicon. Representing Knowledge in Semantic Networks*. Cambridge: Cambridge University Press.

HEID, U. & RAAB, S. 1989. "Collocations in multilingual generation". En *Fourth Conference of the European Chapter of the ACL*. Manchester:130-135.

HEYLEN, D. & VERHAGEN, M. 1993. "Representation" En *Collocations* ET-10/75, I:119-141.

HUTCHINS, W.J. & SOMERS, H.L. 1992. *An Introduction to Machine Translation*. London, Academic Press.

MACKIN, R. 1978. "On collocations: 'word shall be known by the company they keep'". En *In honour of A.S. Hornby*. Oxford: O.U.P:149-165.

MEL'CUK, I. & ZOLKOVSKIJ, A. 1988. "The Explanatory Combinatorial Dictionary". En Evens (ed) *Relational Models of the Lexicon*. Cambridge University Press.

MILLER, G.A., BECKWITH, R.C., FELLBAUM, C., GROSS, D. & MILLER, H. 1993. *Introduction to Wordnet: An on-line lexical database*. Princetown University.

MITCHELL, T.F. 1971. "Linguistic 'goings on': collocations and other lexical matters arising on the syntagmatic record". En *Archivum Linguisticum* 2: 35-69.

NIRENBURG, S. 1988. "Lexical realization in natural language generation". En *Proceedings of the Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*. University of Pittsburgh.

SINCLAIR, J. 1991. *Corpus, Concordance, Collocations*. Oxford, O.U.P.

SMADJA, F. 1993. "Retrieving Collocations from the Text: Xtract" En *Computational Linguistics*, 19 (1): 143-177.

SMADJA, F. & MCKEOWN, K. 1991. "Using collocations for language generation" En **Comput. Intell**. 7: 229-239.

TSUJII, J. 1989. "Machine Translation in Natural Language Understanding" En *Literary and Linguistic Computing*, 4 (3): 214-217.

ZARCO TEJADA, M. A. 1994. *Las Estructuras Conceptuales y las Funciones Léxicas en el ámbito de la Traducción Automática: elementos relacionables del lexicón*. Tesis Doctoral: Universidad de Cádiz.

ZULUAGA, A. 1975. "La Fijación Fraseológica" En *Thesaurus. Boletín del Instituto Caro y Cuervo*. XXX, 225-248. Bogotá.