

Acercando la lingüística teórica y el PLN

VI Seminario de investigación TIC-ETL
UNED. 13 de diciembre de 2013

Dra. Rocío Jiménez Briones
rocio.jimenez@uam.es



GUIÓN DE LA PRESENTACIÓN

- 1. Breve historia de un (des)encuentro.**
- 2. FunGramKB: un proyecto de PLN fundamentado en la lingüística teórica.**
- 3. Conclusiones.**

La lingüística y el PLN

- El PLN es una disciplina a medio camino entre la **lingüística aplicada** y la **inteligencia artificial** que se centra en el estudio y diseño de sistemas de computación efectivos para **facilitar la comunicación entre personas o entre personas y máquinas** de forma fluida y flexible (Periñán Pascual 2012).

La lingüística y el PLN

1940+1950

→ Traducción automática y desciframiento criptográfico del lenguaje (Warren Weaver).

1960

→ Gramática Generativa de Chomsky (1957, 1965) y la Inteligencia Artificial.

1970

→ Representación semántica y teorías funcionales: Gramática Sistémico-Funcional de Halliday (1961, 1967).

1980

→ Teorías lexicistas (Kaplan y Bresnan 1982; Pollard y Sag 1987) y la lingüística de corpus.

La lingüística y el PLN

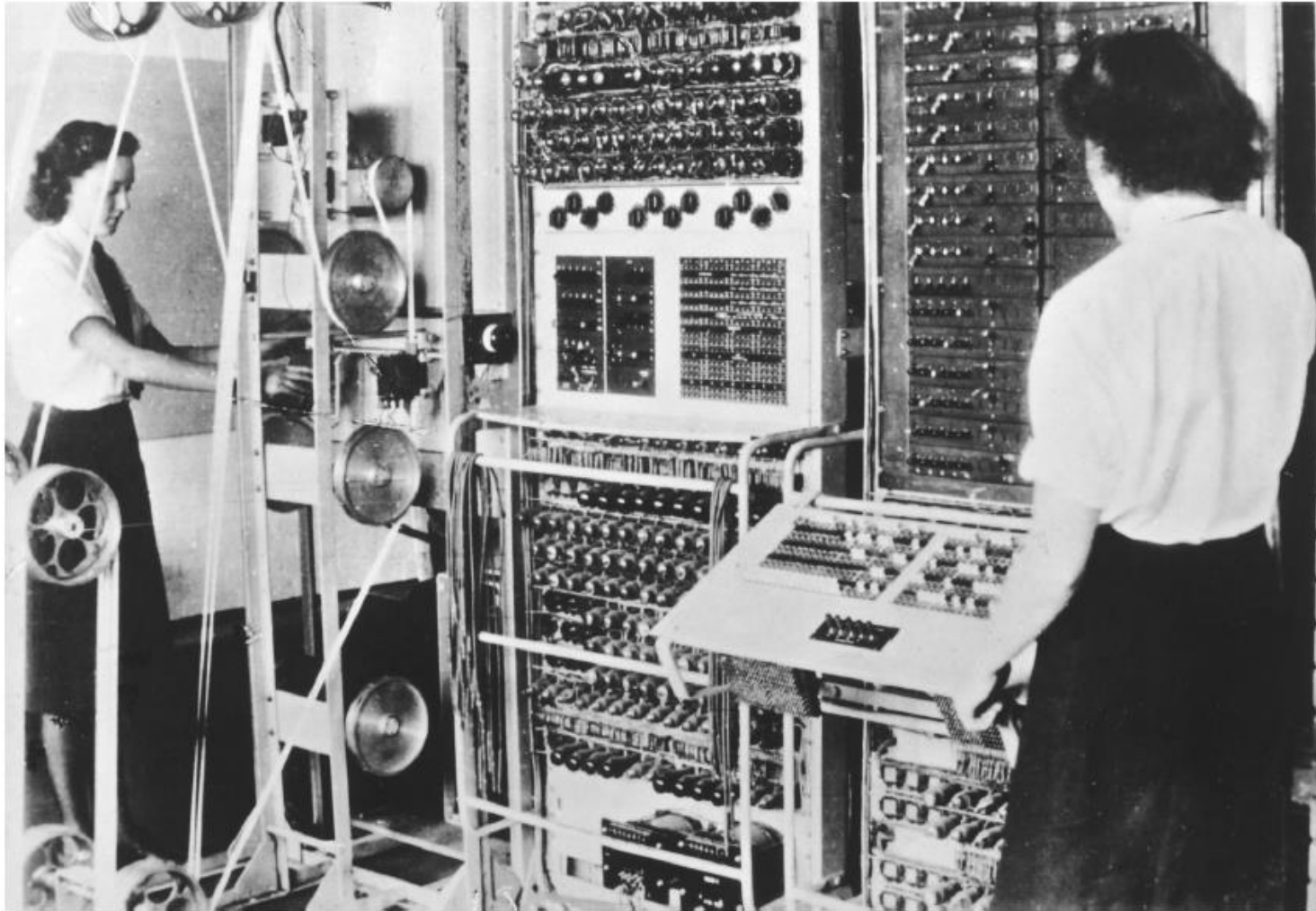


Figura 1: Colossus (www.britannica.com)

La lingüística y el PLN

1990-
actualidad



Supremacía de métodos estadísticos.

¿Por qué?

- La disponibilidad de extensos **corpus** textuales
- Los avances en **hardware**
- **Internet** (Liddy 2001)

CONSECUENCIA:

Las teorías lingüísticas empiezan a desaparecer en las investigaciones en PLN

La lingüística y el PLN

- “Anytime a linguist leaves the group the recognition rate goes up”

Fred Jelinek (*apud* Periñán Pascual 2012:14)

- “Linguistics, as a discipline, went astray: (...) its theory became so obscure, so baroque, and so self-centered, that it became virtually impenetrable to researchers from other disciplines”

(Wintner 2009: 642)

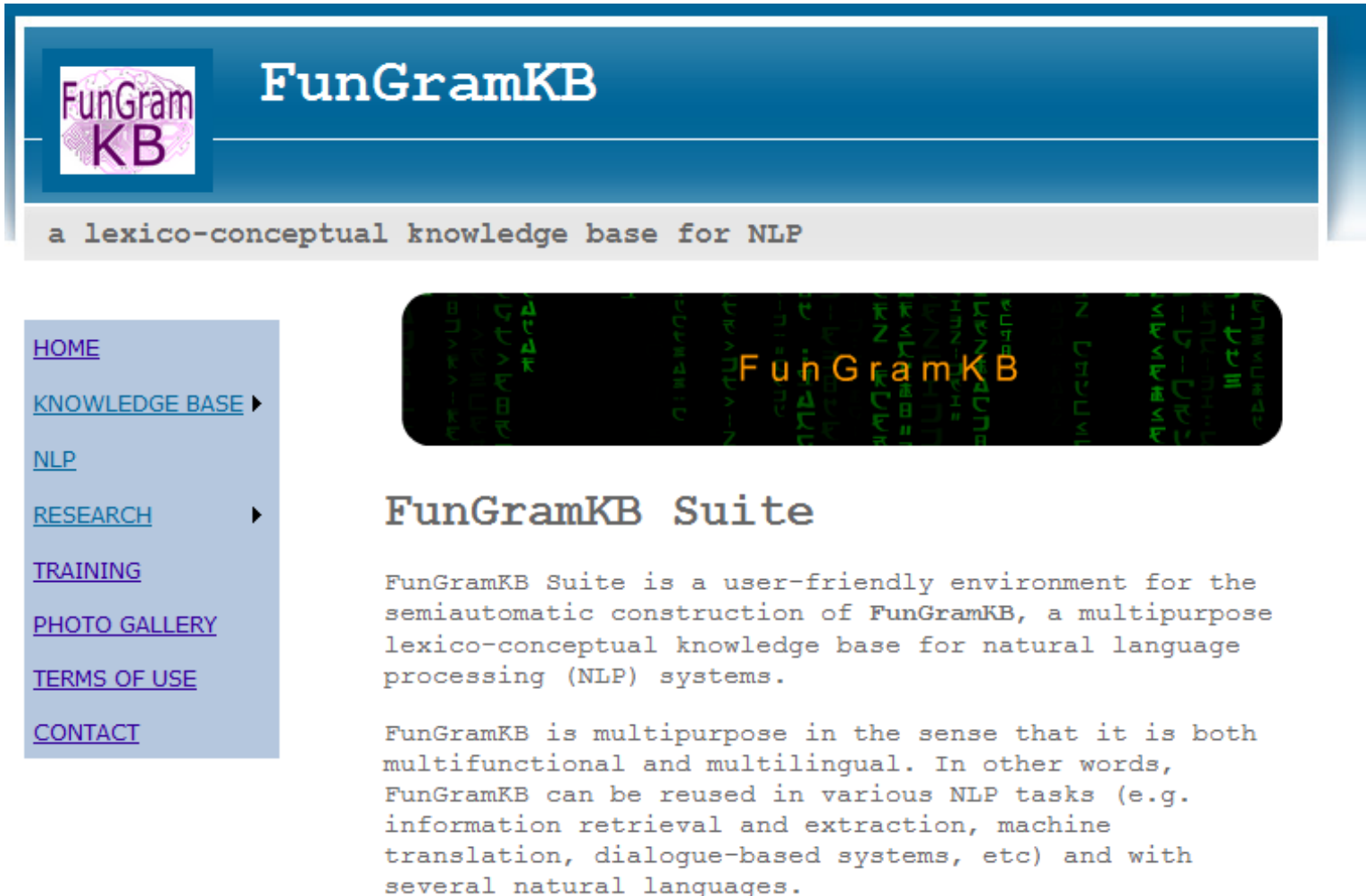
La lingüística y el PLN

- **Ingeniería aplicada:**
 - Ingeniería química => química, física, matemáticas.
 - Ingeniería biomédica => biología y medicina.
 - Ingeniería genética => genética y medicina.
 - ...
 - **Ingeniería lingüística**
 - **Lingüística computacional**
- } lingüística???

La lingüística y el PLN

- “De hecho, existen programas del PLN que funcionan perfectamente sin necesidad de estar basados en alguna teoría lingüística, pero se trata en realidad de programas *engañosamente inteligentes*. Por el contrario, las aplicaciones informáticas robustas requieren una base teórica que pueda servir de fundamento al comportamiento deseado (Halvorsen, 1988)” (Periñán Pascual 2012: 15)

La lingüística y el PLN



The image shows a screenshot of the FunGramKB website. At the top, there is a blue header with the FunGramKB logo on the left and the text "FunGramKB" in white. Below the header, a grey bar contains the text "a lexico-conceptual knowledge base for NLP". On the left side, there is a vertical navigation menu with links: HOME, KNOWLEDGE BASE (with a right-pointing arrow), NLP, RESEARCH (with a right-pointing arrow), TRAINING, PHOTO GALLERY, TERMS OF USE, and CONTACT. In the center, there is a black rounded rectangle with a green and yellow digital code background and the text "FunGramKB" in yellow. Below this, the heading "FunGramKB Suite" is followed by two paragraphs of text describing the suite as a user-friendly environment for constructing a lexico-conceptual knowledge base for NLP systems, and as a multipurpose tool for various NLP tasks in multiple languages.

FunGramKB

a lexico-conceptual knowledge base for NLP

- [HOME](#)
- [KNOWLEDGE BASE](#) ▶
- [NLP](#)
- [RESEARCH](#) ▶
- [TRAINING](#)
- [PHOTO GALLERY](#)
- [TERMS OF USE](#)
- [CONTACT](#)

FunGramKB Suite

FunGramKB Suite is a user-friendly environment for the semiautomatic construction of FunGramKB, a multipurpose lexico-conceptual knowledge base for natural language processing (NLP) systems.

FunGramKB is multipurpose in the sense that it is both multifunctional and multilingual. In other words, FunGramKB can be reused in various NLP tasks (e.g. information retrieval and extraction, machine translation, dialogue-based systems, etc) and with several natural languages.

Figura 2: <http://www.fungramkb.com/>

¿Qué es FunGramKB?

- “El diseño de una interfaz lingüístico-conceptual para el análisis de textos en inglés y español. Estudio de su aplicación en los sistemas de recuperación de la información en entornos multilingües”. Proyecto coordinado entre la **UNED y la Universidad de La Laguna, 2012-2014. IP: Ricardo Mairal Usón, UNED (España).**
- “Elaboración de una subontología terminológica en un contexto multilingüe (español, inglés e italiano) a partir de la base de conocimiento FunGramKB en el ámbito de la cooperación internacional en materia penal: terrorismo y crimen organizado” (FFI2010-15983), **2011-2013. IP: Ángel Felices Lago, Universidad de Granada (España).**
- “Fundamentos para la construcción del nivel gramatical en el entorno de una base de conocimiento léxico-conceptual” (FFI2010-17610), **2011-2013. IP: Francisco J. Ruiz de Mendoza Ibáñez, Universidad de La Rioja (España).**

¿Qué es FunGramKB?

Ireland's economic woes

Tiger, tiger, burning dull

Emerald isle in the red

Nov 11th 2010 | from PRINT EDITION

Tweet 6 Me gusta

When the Luck of the Irish Ran Out: The World's Most Resilient Country and its Struggle to Rise Again. By David Lynch. Palgrave Macmillan; 248 pages; \$26 and £16.99. Buy from [Amazon.com](#), [Amazon.co.uk](#)

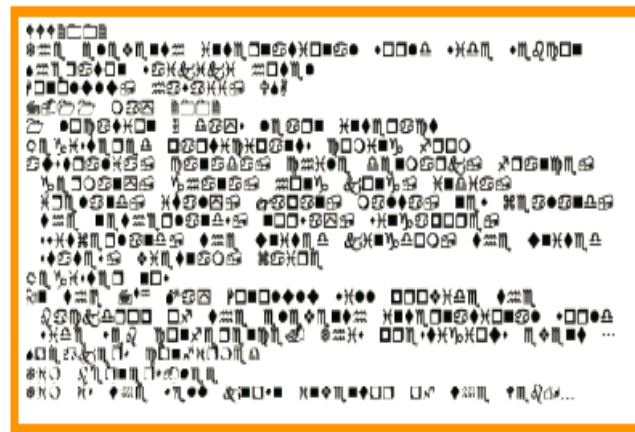
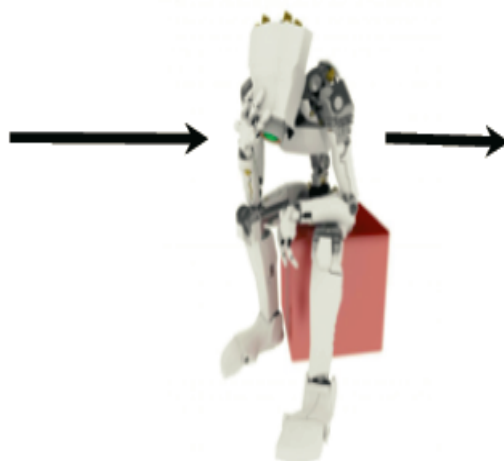
EUROPE has seen many spectacular stories in the past 20 years, but few can match Ireland's rise and fall. A depressed and indebted country in the late 1980s suddenly became the bubbly Celtic Tiger of the 1990s. Then the bubble just as suddenly burst, an experience that could cost Ireland as much as one-fifth of GDP annually in years to come. In 2010 the budget deficit will be at least 32% of GDP and the public debt is almost 100% of GDP. After a brief interlude of net immigration, the Irish are emigrating once again.

Was the Celtic Tiger all an illusion? Fintan O'Toole, a veteran leftist writer, seems to think it was. His previous book, "Ship of Fools", examined the Irish property bubble, how it inflated and deflated, and seduced politicians, bankers and officials alike. His angry new book enlarges on the theme, denouncing not only the political parties (especially ruling Fianna Fail), regulators and civil servants, but also the Roman Catholic church and even the education and health-care systems. Ireland, he concludes, needs to start all over again and create what he calls a new republic.

Mr O'Toole's writing is splendidly sharp, but his conclusion seems too gloomy. Despite the excesses of Ireland's time as a Celtic Tiger, the country has changed for the better. Real businesses, from pharmaceuticals to computing, continue to flourish. Foreign investment is still being lured in by low corporate-tax rates. The country churns out many good graduates. Public services are patchy but improving, as is infrastructure. Corruption is a problem, but it is surely not as bad as in many other European countries. And, as elsewhere, the church's baleful influence is now hugely diminished.

Related topics

- World politics
- Politics
- Irish politics
- Government and politics
- European politics



No entiendo lo que dice el texto.



Figura 3: Procesamiento de un texto

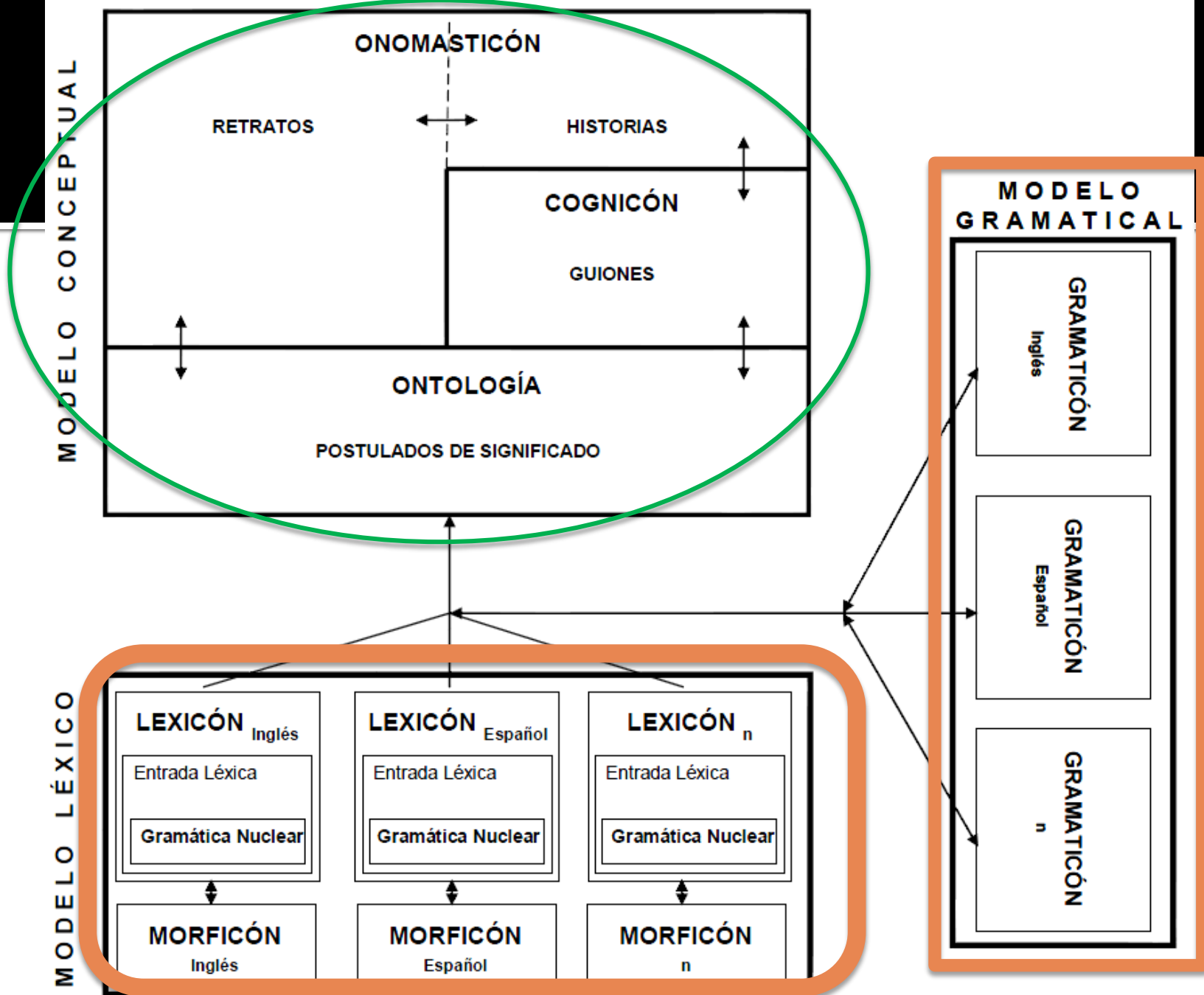


Figura 4: Arquitectura de FunGramKB (Mairal Usón 2012)

Módulo léxico

- Compuesto por:
 - A) **Los morficones**: uno para cada lengua de la base de conocimiento, donde se recogen aspectos de la **morfología flexiva**.
 - B) **Los lexicones**: un lexicón para cada una de las siete lenguas con las que se trabaja: alemán, búlgaro, catalán, francés, inglés, español e italiano. En cada lexicón encontramos la información **morfosintáctica** de las unidades léxicas de cada lengua en particular.

Módulo léxico

Concept	+DRY_01
POS	v
Paradigm	Regular
Tense Constraints	no
Reflexivity	Optional
Reciprocity	Never
Aktionsart	CACC
Lexical template: variables	X-Theme, Y-Referent
Lexical template: restrictions	MR2, U = y
Collocations	Y = eyes, hair, hands, tears
Constructions	Caused-motion construction (transitive) Inchoative construction Middle construction Natural force subject construction
Dialect	Standard
Usage	Common
Topic	FACTOTUM
Examples	"Her mouth had dried." (GUE W_fict_prose) "She dried her hands on the kitchen towel, adding, "I worry about Ernest." (FPM W_fict_prose)

Tabla 1: Entrada léxica de *dry* en FunGramKB

Módulo gramatical

- Siguiendo al **Modelo Léxico Construccional** (MLC; Ruiz de Mendoza & Mairal Usón 2008, 2011; Mairal Usón & Ruiz de Mendoza, 2009) y a la **Gramática de Construcciones** (GxC; Goldberg 1995, 2006), se almacenan **las propiedades y características específicas de las construcciones** más significativas de las lenguas de FunGramKB.
- Se compone de varios **GRAMMATICONES**, uno para cada una de las lenguas de la base de conocimiento:



Módulo gramatical

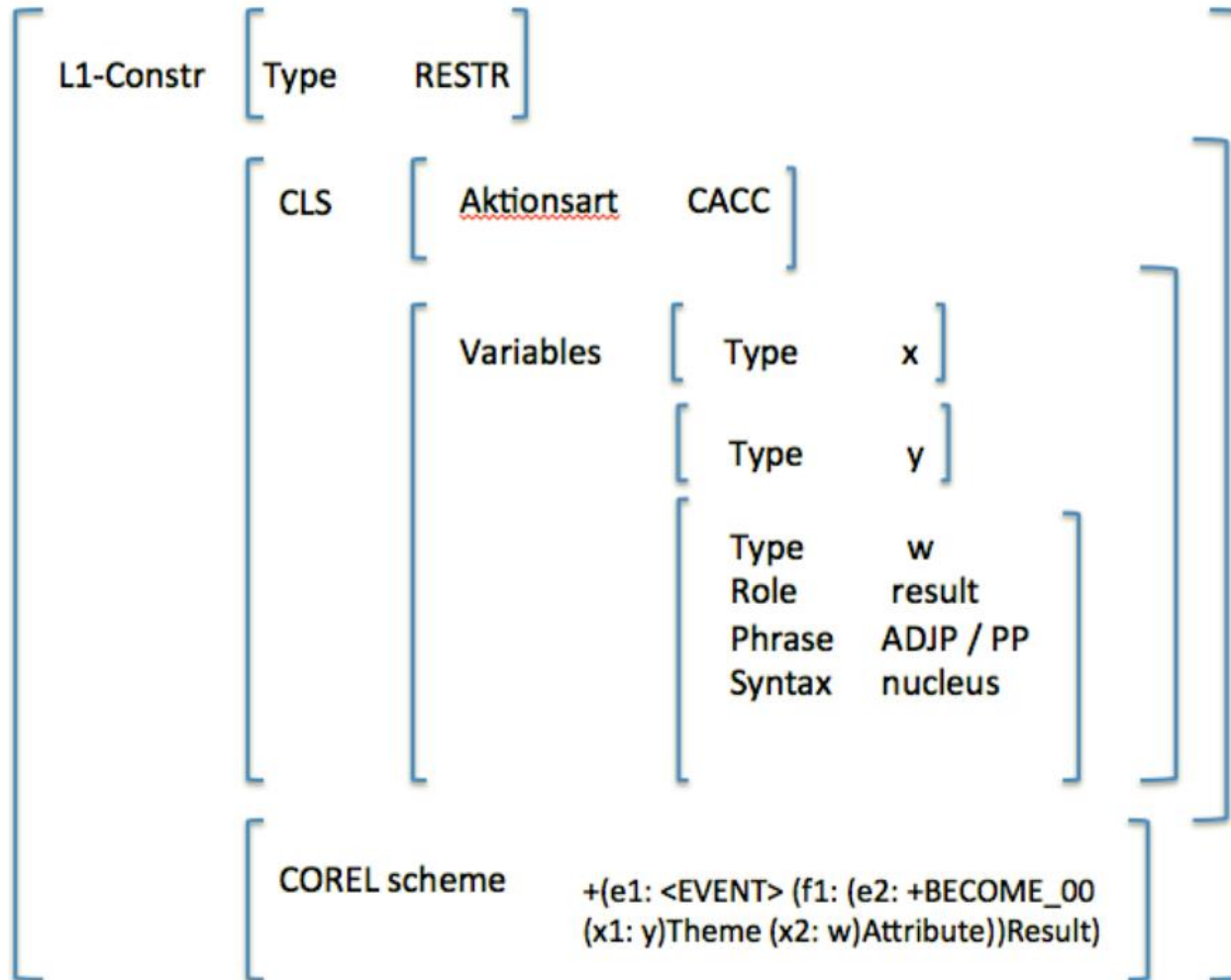


Figura 5: Matriz atributo-valor para la resultativa (Mairal Usón 2012)

Módulo conceptual

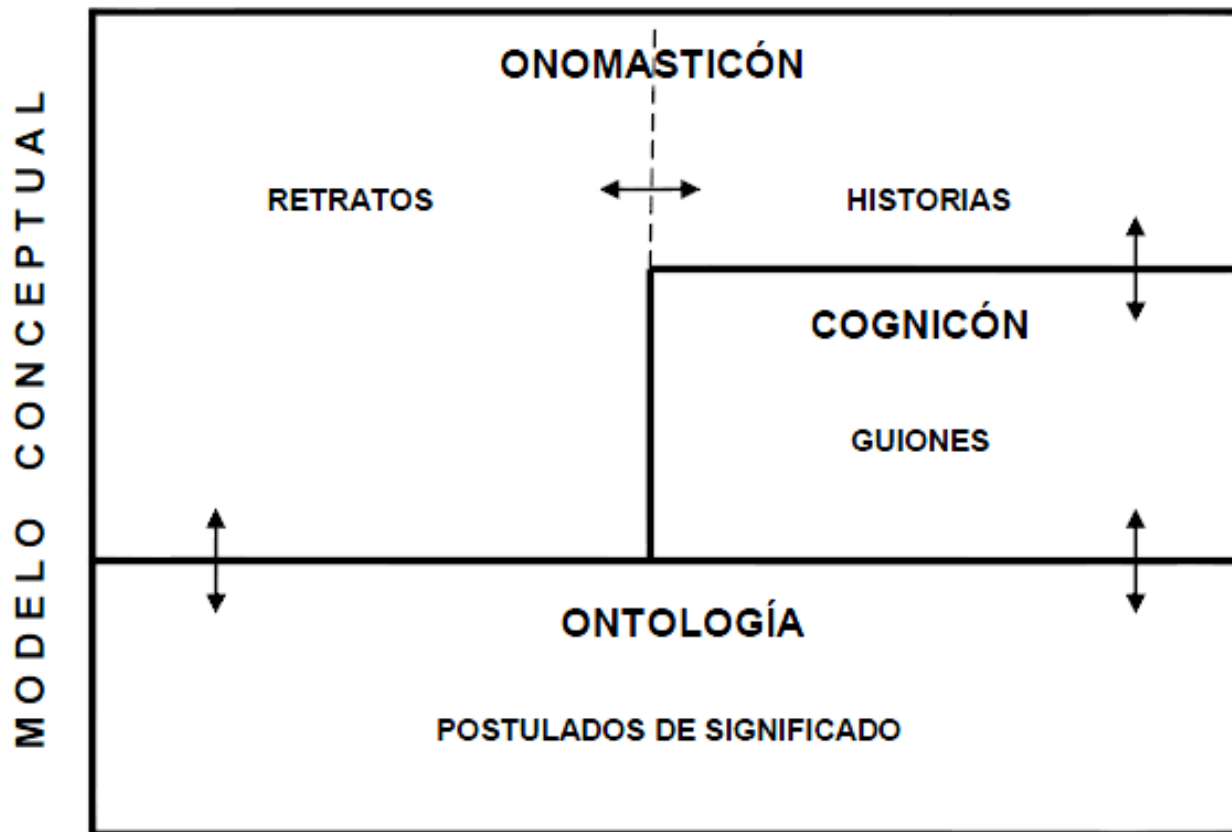


Figura 6: Módulo conceptual de FunGramKB

Módulo conceptual: Onomasticón

- Almacena información relativa a las instancias concretas de **entidades y eventos**, p.e. Javier Bardem, Steve Jobs, La Alhambra, Halloween, etc.
- Este tipo de información se representa mediante estructuras en la forma de **retratos** (*snapshots*) e **historias** (*stories*).

Módulo conceptual: Onomasticón

“París es la capital de Francia”

*(e1: +BE_00 (x1: %PARIS_00)Theme (x2:
\$CAPITAL_00)Referent)
+(e2: +BE_02 (x1)Theme (x3:
%FRANCE_00)Location)

Módulo conceptual: Cognición

- Recoge el **conocimiento de tipo procedimental** (cómo hacer la cama, cómo freír un huevo, etc.).
- Este conocimiento se representa mediante **guiones** (cfr. los *scripts* de Schank y Abelson 1977): esquemas conceptuales en los que se organiza **secuencialmente** una serie de acciones prototípicas siguiendo el modelo temporal de Allen (Allen 1983; Allen & Ferguson 1994).

Módulo conceptual: Cognición

FunGramKB Editor

Cognicon

9 scripts

- @EATING_AT_RESTAURANTS_00
- @GOING_TO_THE_CINEMA_00
- @GOING_TO_THE_DISCO_00
- @PAY_CARD_00
- @PAY_CASH_00
- @READING_00
- @TAKING_A_TAXI_00
- @WATCHING_TELEVISION_00
- @COOKING_OREGANO_LEMON_CHICKEN

New Rename Delete

Conceptual Information:

NAME:	@COOKING_OREGANO_LEMON_CHICKEN_00
SCRIPT:	<pre>+(e1: +HEAT_00 (x1: +HUMAN_00)Theme (x2: \$OVEN_00)Referent (f1: (e2: +BECOME_00(x2)Theme (x3:\$190CENTIGRADEDEGREE_00)Attribute)) Purpose) +(e3: +PUT_00 (x1)Agent (x4: p +OIL_00)Theme (x5)Origin (x6: \$BAKINGDISH_00)Goal) +(e4: +PUT_00 (x1)Agent (x7: \$CHICKENBREAST_00)Theme (x8)Origin (x6)Goal)</pre>
TEMPORAL RELATIONS:	<pre>e1 -> e3 [Before] e3 -> e4 [Before] e4 -> e5 [Before] e5 -> e6 [Before] e6 -> e7 [Before] e7 -> e8 [During]</pre>
DOMAIN:	alimention
CULTURE:	Universal
	Heat oven to 190° C. Grease a baking dish. Place chicken breast

Edit script

Webmaster: Carlos Perrián

Figura 7: Editor del Cognición (García Peral 2013)

Módulo conceptual: Ontología

- “Un inventario del tipo de cosas que existen en un dominio desde la perspectiva de una persona que habla sobre ese dominio” (Sowa 2000).
- La **Ontología Nuclear de FunGramKB** es el **catálogo jerarquizado de los conceptos** que cualquiera de nosotros tenemos en la cabeza cuando hablamos de situaciones cotidianas.
- Refleja el **modelo del mundo** compartido por toda una comunidad.
- Funciona por relaciones de **herencia e inferencia**, siendo la **subsunción** (IS-A) la única relación taxonómica permitida.

Módulo conceptual: Ontología

- Se basa en una **semántica profunda** (= *deep semantics*), a diferencia de la mayoría de ontologías (cfr. EuroWordNet), que siguen un acercamiento de semántica superficial (= *surface semantics*).
- En FunGramKB todos los conceptos están definidos por sí mismos a través de **dos rasgos**: el **marco temático** (*thematic frame*) y el **postulado de significado** (*meaning postulate*).

Módulo conceptual: Ontología

400 events

- +DRY_01
- +FILL_00
- +HARDEN_00
- +HEAT_00
- +IMPROVE_00
- +INCREASE_00
- +INFECT_00
- +LENGTHEN_00
- +MELT_00
- +REPAIR_00
- +SPLIT_00
- +TIGHTEN_00
- +TRANSLATE_00

-BLOW_DRY
+DRY_01

Entities
 Events
 Qualities

Conceptual Information:

CONCEPT:	+DRY_01 <input checked="" type="checkbox"/>
SUPERORDINATE(S):	+CHANGE_00
THEMATIC FRAME:	(x1)Theme (x2: +CORPUSCULAR_00)Referent
MEANING POSTULATE:	+(e1: +CHANGE_00 (x1)Theme (x2)Referent (f1: (e2: n +BECOME_00 (x2)Theme (x3: +WET_00)Attribute))Result)
DESCRIPTION:	remove the moisture from and make dry; "dry clothes"; "dry hair"

Edit structure
Edit properties

English

- air-dry
- blow-dry
- dehumidify
- dehydrate

Spanish

- desecar
- deshidratat
- deshumedecer
- disecar

Italian

- asciugare
- disidratate
- essiccare
- inaridire

3. Conclusiones

- Panorámica del **PLN**: una disciplina “joven” pero con un gran futuro.
- **FunGramKB**: una base de conocimiento cuyo objetivo es implementar sistemas computacionales que simulen el razonamiento humano.
- **FunGramKB** es una realidad gracias a la estrecha colaboración de lingüistas, informáticos, ingenieros del conocimiento, etc.

¡MUCHAS GRACIAS POR VUESTRA
ATENCIÓN!

(podéis leer más en
www.fungramkb.com)

Bibliografía

- Allen, James F., 1983: "Maintaining knowledge about temporal intervals", *Communications of the ACM* 26(11): 832-843.
- Allen, James F. & George Ferguson, 1994: "Events and actions in interval temporal logic", *Journal of logic and computation* 4(5):531-579.
- Chomsky, Noam, 1957: *Syntactic structures*, La Haya: Mouton.
- ---, 1965: *Aspects of the theory of syntax*, Cambridge (Mass.): MIT Press.
- García Peral, Roberto, 2013: *Adquisición y gestión del conocimiento para un sistema de recuperación de información con recetas de cocina*. Trabajo fin de máster. Máster Universitario en Tecnologías de la Información y la Comunicación en la Enseñanza y Tratamiento de Lenguas .Uned.
- Goldberg, Adele, 1995: *Constructions: a Construction Grammar approach to argument structure constructions*, Chicago: University of Chicago Press.
- ---, 2006: *Constructions at work: the nature of generalization in language*, New York: Oxford University Press.
- Grishman, Ralph, 1986: *Computational linguistics: an introduction*, Cambridge: Cambridge University Press.
- Halliday, Michael, 1961: "Categories of the theory of grammar", *Word* 17, 241-92.
- ---, 1967: "Notes on transitivity and theme in English", *Journal of Linguistics* 3, 199-244.
- Halvorsen, Per-Kristian, 1988: "Computer applications of linguistic theory" en Frederick Newmeyer (ed.): *Linguistics: the Cambridge survey II. Linguistic theory: extensions and implications*, 198-219. Cambridge: Cambridge University Press.

- Kaplan, Ronald M. y Joan Bresnan, 1982: “Lexical-Functional Grammar: a formal system for grammatical representation”, en *The mental representation of grammatical relations*, Joan Bresnan (ed.), 173-280. Cambridge (Mass.): MIT Press.
- Levin, Beth, 1993: *English verb classes and alternations: a preliminary investigation*, Chicago/London: University of Chicago Press.
- Liddy, Elizabeth D., 2001: “Natural language processing” en *Encyclopedia of Library and Information Science*, segunda edición, Nueva York: Marcel Decker.
- Mairal Usón, Ricardo, 2012: “La arquitectura de una base de conocimiento léxico conceptual: implicaciones lingüísticas” en *Léxico y sintaxis. Volumen temático de la serie editada por la Sociedad Argentina de Lingüística*, Giammatteo, L., Ferrari, L. y Albano, H. (eds.), 183-210. Buenos Aires: Cuyo, Editorial de la Facultad de Filosofía y Letras– UNCuyo.
- Mairal Usón, Ricardo, Carlos Periñán Pascual & María Beatriz Pérez Cabello de Alba, 2012. “La representación léxica. Hacia un enfoque ontológico”, en *El funcionalismo en la teoría lingüística. La Gramática del Papel y la Referencia. Introducción, avances y aplicaciones*, Ricardo Mairal, Lilián Guerrero & Carlos González (eds), 85-102. Akal: Madrid.
- Mairal Usón, Ricardo & Francisco J. Ruiz de Mendoza, 2009: “Levels of description and explanation in meaning construction”, en *Deconstructing constructions*, Christopher Butler & Javier Martín Arista (eds.), 153-198. Amsterdam: John Benjamins.
- Martí Antonín, Maria Antònia & Joaquim Llisterri, 2001: “La ingeniería lingüística en la sociedad de la información”, *Digithum, Revista Digital d’Humanitats* 3 [http://www.uoc.edu/humfil/articles/esp/listerri-marti/listerri-marti_imp.html]
- Moreno Boronat, Lidia & Antonio Molina Marco, 1999: “Preliminares y tendencias en el Procesamiento del Lenguaje Natural”. *Revista de Inteligencia Artificial* 99 (7).

- Perriñán Pascual, Carlos, 2005: "Procesamiento del lenguaje natural: de lingüista a ingeniero del conocimiento", en *Nuevas tendencias en Lingüística Aplicada*, I. K. Brady, M. Navarro Coy y C. Perriñán Pascual (eds.), 293-317. Murcia: Quaderna.
- ---, 2012: "En defensa del procesamiento del lenguaje natural fundamentado en la lingüística teórica", *Onomázein* 26, 13-48.
- Perriñán Pascual, Carlos & Francisco Arcas Túnez, 2004: "Meaning postulates in a lexico-conceptual knowledge base", en *Proceedings of the 15th International Workshop on Databases and Expert Systems Applications*, 38-42. Los Alamitos: IEEE.
- ---, 2007: "Cognitive modules of an NLP knowledge base for language understanding", *Procesamiento del lenguaje natural* 39, 197-204.
- ---, 2010: "The architecture of FunGramKB" en *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 2667-2674. Malta: ELRA.
- ---, 2011: "Introducción a FunGramKB", *Anglogermánica online* 8, 1-15.
- Perriñán Pascual, Carlos & Ricardo Mairal Usón, 2009: "Bringing Role and Reference Grammar to natural language understanding", *Procesamiento del lenguaje natural* 43, 265-273.
- ---, 2010: "La gramática de COREL: un lenguaje de representación conceptual", *Onomázein* 21, 11-45.
- ---, 2012: "La dimensión computacional de la GPR: la estructura lógica conceptual y su aplicación en el procesamiento del lenguaje natural" en *La Gramática del Papel y la Referencia: introducción, avances y aplicaciones*, R. Mairal Usón, L. Guerrero y C. González Vergara (eds.), 333-348. Madrid: Akal.
- Pollard, Carl J. e Ivan I. Sag, 1987: *Information-based syntax and semantics*, Stanford: CSLI.

- Procter, Paul (ed.), 1978: *Longman dictionary of contemporary English*. Harlow (Essex): Longman.
- Ruiz de Mendoza, Francisco J. & Ricardo Mairal Usón, 2008: "Levels of description and constraining factors in meaning construction: an introduction to the Lexical Constructional Model", *Folia Linguistica* 42 (2): 355-400.
- ---, 2011: "Constraints on syntactic alternation: lexical-constructional subsumption in the Lexical Constructional Model", en *Morphosyntactic alternations in English: functional and cognitive perspectives*, Pilar Guerrero (ed.), 62-82. Equinox Publishing Books.
- Schank, Roger C. & Robert P. Abelson, 1977: *Scripts, plans, goals and understanding: an inquiry into human knowledge structures*. Hillsdale: Lawrence Erlbaum.
- Sowa, John F., 2000: "Ontology, metadata, and semiotics", en *Conceptual structures: logical, linguistics, and computational issues*, Bernhard Ganter y Guy Mineau (eds.), 55-81. Berlín: Springer.
- Tulving, Endel, 1985: "How many memory systems are there?", *American Psychologist* 40: 385-398.
- Van Valin, Robert D. Jr., 2005: *Exploring the Syntax-Semantics Interface*, Cambridge: Cambridge University Press.
- Van Valin, Robert D. Jr. y Randy J. LaPolla, 1997: *Syntax, Structure, Meaning and Function*, Cambridge: Cambridge University Press.
- Velardi, Paola, Michela Fasolo & María T. Pazienza, 1991: "How to encode semantic knowledge: a method for meaning representation and computer aided acquisition", *Computational Linguistics* 17(2): 153-170.
- Wintner, Shuly, 2009: "What science underlies natural language engineering?", *Computational linguistics* 35 (4), 641-644.