A Spanish e-dictionary of collocations

María Auxiliadora Barrios

Universidad Complutense de Madrid

Igor Boguslavsky

Universidad Politécnica de Madrid / Russian Academy of Sciences

Diretes – an electronic dictionary of collocations for human users and applications

- Collocation is a special kind of word combinations
- “A collocation **AB** is a semantic phraseme such that its signified ‘X’ is constructed out of the signified of the one of its two constituent lexemes — say, of A — and a signified ‘C’ [‘X’ = ‘A⊕C’] such that the lexeme B expresses ‘C’ contingent on A” (Melčuk 1998).
    - *black coffee* (‘without milk’)
    - *do a favor* (light verb)
    - *heavy smoker* (‘smokes much’)
    - *artesian well*
- Lexical Functions of the Meaning-Text Theory is a formalism for describing collocations in a rigorous and systematic way.
- Human users:
    - phrases that a fluent speaker of the language should know and be able to use
- Applications:
    - idiomatic translation in MT, paraphrasing, disambiguation, etc.

Plan

- Diretes dictionary of Spanish collocations
  - Sources of Diretes: Redes and Práctico
  - Data of Diretes
- A possible application: semantic analysis
  - SemETAP semantic analyzer
  - Adjectival and adverbial Lexical Functions in SemETAP
- Future work

Sources of Diretes data: *Redes* and *Práctico*

- Bosque I. 2004. REDES. Diccionario combinatorio del español contemporáneo. Las palabras en su contexto. Ediciones SM, Madrid.
    - 7,115 entries
- Bosque I. 2006. Diccionario combinatorio PRÁCTICO del español contemporáneo. Las palabras en su contexto. Ediciones SM, Madrid.
    - 14,000 entries
- Carefully selected set of collocations. For each collocation there is a real example of use taken from a corpus of more than 250 millions of words.
- *Redes* is mostly oriented towards research purposes. Combinatorial data are presented by means of lexical classes.
- *Práctico* is conceived as a dictionary for practical purposes. Intended for native speakers, interested in refreshing their mastery of language, for authors, translators and language learners.
- High standard of quality (as opposed to automatically extracted collections of collocations)
- Lack formalization

Diretes

- Electronic dictionaries of collocations within the MTT framework (French, English, Russian, German, Spanish)
- Spanish
    - DiCE: semantic field of emotions (200 entries)
    - DiCoEnviro: semantic field of environment (170 entries)
    - Dicoinfo-ES: semantic field of computer science (1000 terms)
    - Diretes: 664 semantic fields, about 50,000 collocations
        - Among them - 551 adjectival and adverbial collocations beginning with the letter *a*

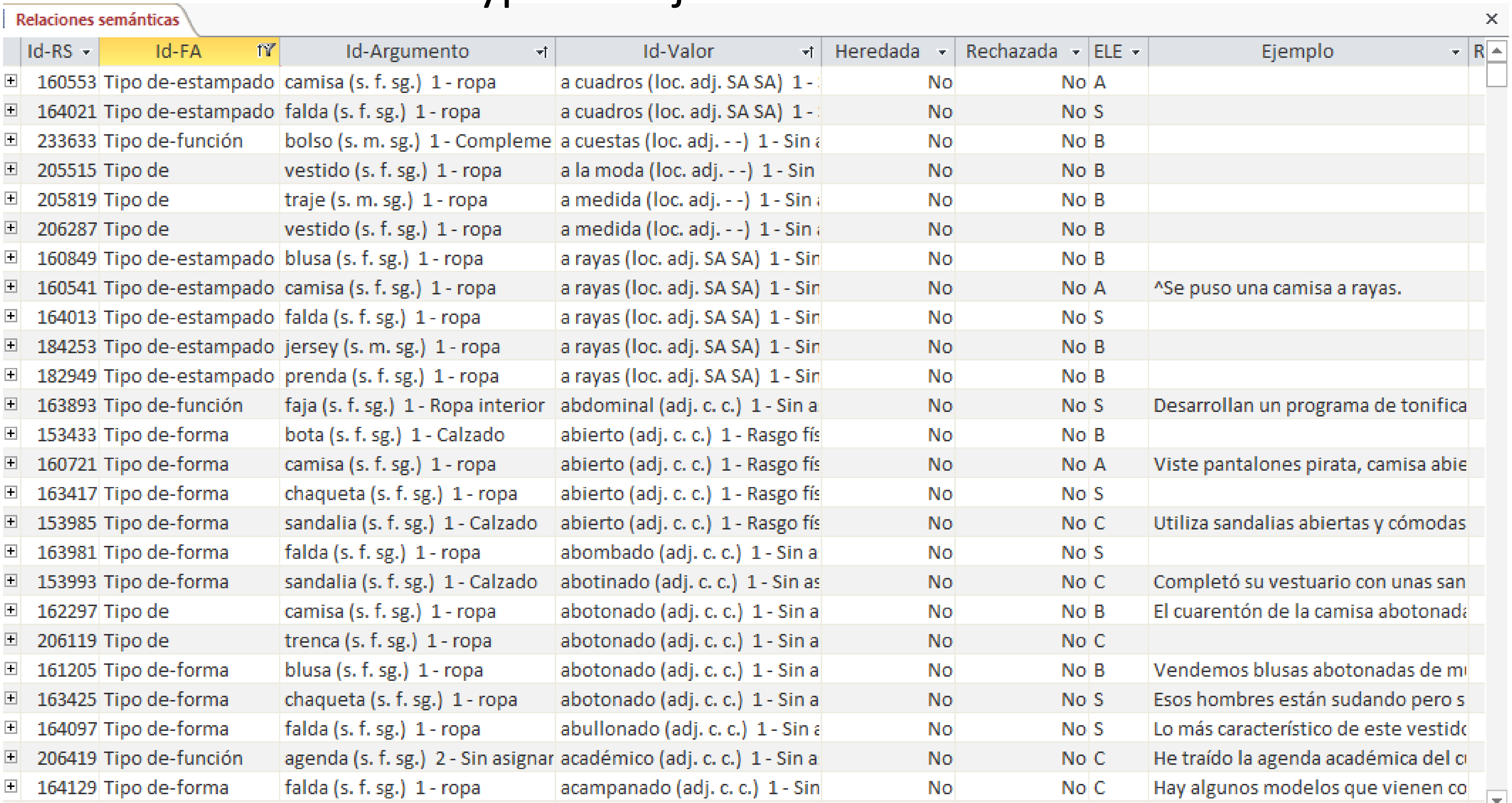
Standard Lexical Functions

- A standard LF satisfies 2 conditions simultaneously:
  - broadness of domain
  - broadness of range
- Adjectives and adverbs can be values of the following standard LFs:
  - Semantic derivatives A_i and Adv_i
  - Magn (‘very, to a high degree’): *infinite patience*
  - Ver (‘such as should be’): *legitimate demand*
  - Bon (‘good’): *fruitful analysis*
  - Pos (‘positive evaluation’): *favourable opinion*
  - Epit (‘redundant clichéd modifier’): *sweet dream*
- Many of them can combine with Anti
- There are many other (non-standard) LFs

TypeOf collocations

- TypeOf (hypernymy, similar to Gener)
- Several semantic variants if TypeOf (examples on the next slide):
    - TypeOf-form
    - TypeOf-function
    - TypeOf-print
    - …

TypeOf Adjectival collocations



Relaciones semánticas

Id-RS	Id-FA	Id-Argumento	Id-Valor	Heredada	Rechazada	ELE	Ejemplo	R
160553	Tipo de-estampado	camisa (s. f. sg.) 1 - ropa	a cuadros (loc. adj. SA SA) 1 -	No	No	A		
164021	Tipo de-estampado	falda (s. f. sg.) 1 - ropa	a cuadros (loc. adj. SA SA) 1 -	No	No	S		
233633	Tipo de-función	bolso (s. m. sg.) 1 - Compleme	a cuestas (loc. adj. - -) 1 - Sin a	No	No	B		
205515	Tipo de	vestido (s. f. sg.) 1 - ropa	a la moda (loc. adj. - -) 1 - Sin	No	No	B		
205819	Tipo de	traje (s. m. sg.) 1 - ropa	a medida (loc. adj. - -) 1 - Sin a	No	No	B		
206287	Tipo de	vestido (s. f. sg.) 1 - ropa	a medida (loc. adj. - -) 1 - Sin a	No	No	B		
160849	Tipo de-estampado	blusa (s. f. sg.) 1 - ropa	a rayas (loc. adj. SA SA) 1 - Sin	No	No	B		
160541	Tipo de-estampado	camisa (s. f. sg.) 1 - ropa	a rayas (loc. adj. SA SA) 1 - Sin	No	No	A	^Se puso una camisa a rayas.	
164013	Tipo de-estampado	falda (s. f. sg.) 1 - ropa	a rayas (loc. adj. SA SA) 1 - Sin	No	No	S		
184253	Tipo de-estampado	jersey (s. m. sg.) 1 - ropa	a rayas (loc. adj. SA SA) 1 - Sin	No	No	B		
182949	Tipo de-estampado	prenda (s. f. sg.) 1 - ropa	a rayas (loc. adj. SA SA) 1 - Sin	No	No	B		
163893	Tipo de-función	faja (s. f. sg.) 1 - Ropa interior	abdominal (adj. c. c.) 1 - Sin a	No	No	S	Desarrollan un programa de tonifica	
153433	Tipo de-forma	bota (s. f. sg.) 1 - Calzado	abierto (adj. c. c.) 1 - Rasgo fís	No	No	B		
160721	Tipo de-forma	camisa (s. f. sg.) 1 - ropa	abierto (adj. c. c.) 1 - Rasgo fís	No	No	A	Viste pantalones pirata, camisa abie	
163417	Tipo de-forma	chaqueta (s. f. sg.) 1 - ropa	abierto (adj. c. c.) 1 - Rasgo fís	No	No	S		
153985	Tipo de-forma	sandalia (s. f. sg.) 1 - Calzado	abierto (adj. c. c.) 1 - Rasgo fís	No	No	C	Utiliza sandalias abiertas y cómodas	
163981	Tipo de-forma	falda (s. f. sg.) 1 - ropa	abombado (adj. c. c.) 1 - Sin a	No	No	S		
153993	Tipo de-forma	sandalia (s. f. sg.) 1 - Calzado	abotinado (adj. c. c.) 1 - Sin as	No	No	C	Completó su vestuario con unas san	
162297	Tipo de	camisa (s. f. sg.) 1 - ropa	abotonado (adj. c. c.) 1 - Sin a	No	No	B	El cuarentón de la camisa abotonada	
206119	Tipo de	trenca (s. f. sg.) 1 - ropa	abotonado (adj. c. c.) 1 - Sin a	No	No	C		
161205	Tipo de-forma	blusa (s. f. sg.) 1 - ropa	abotonado (adj. c. c.) 1 - Sin a	No	No	B	Vendemos blusas abotonadas de mu	
163425	Tipo de-forma	chaqueta (s. f. sg.) 1 - ropa	abotonado (adj. c. c.) 1 - Sin a	No	No	S	Esos hombres están sudando pero s	
164097	Tipo de-forma	falda (s. f. sg.) 1 - ropa	abullonado (adj. c. c.) 1 - Sin a	No	No	S	Lo más característico de este vestido	
206419	Tipo de-función	agenda (s. f. sg.) 2 - Sin asignar	académico (adj. c. c.) 1 - Sin a	No	No	C	He traído la agenda académica del cu	
164129	Tipo de-forma	falda (s. f. sg.) 1 - ropa	acampanado (adj. c. c.) 1 - Sin	No	No	C	Hay algunos modelos que vienen co	

Non-standard LFs

- Classified by means of productive semantic features:
  - Material – *tierra abonada* ‘potting soil’
  - Appearance – *mente abierta* ‘open mind’
  - Place – *tráfico aéreo* ‘air traffic’
  - Manner – *decir algo a boca jarro* ‘to say something bluntly’
  - Cause – *sol abrasador* ‘blazing sun’
  - AbleTo – *lugar accessible* ‘accessible place’
  - Quantity – *dividir a partes iguales* ‘divide in equal parts’
  - Time – *convocatoria anual* ‘annual call’
  - Recurrence – *orador asiduo* ‘regular guest speaker’
  - Speed – *trabajar a toda máquina* ‘to work at full speed’

Inheritance of LF values

- Lexical Inheritance Principle (Mel´čuk & Wanner 1996) (aka LF Domain Principle)
- Words sharing a hypernym often develop similar values of LFs.
- CausFunc_0 (‘create, bring into existence’):
  - Building (*house, palace, temple, concert hall*,…) - *to build*
  - Text or music (*poem, novel, essay…, symphony, melody*…) – *to compose*
  - Clothes (*shirt, trousers, coat*,…) – *to make*
- LiquFunc_0 (‘to cause smth not to exist any more’)
- IncepFunc_0 (‘to start existing’)
- FinFunc_0 (‘to finish existing’)
- …

Organization of data in Diretes

- Table 1: assignment of semantic classes (hypernyms) to lemmas:
  - Camisa ‘shirt’ => ‘piece of clothes’
  - Calcetín ´sock´ => ‘underwear’
- Table 2: hierarchy of semantic classes (9 levels):
  - ‘clothing and accessories’ > ‘clothing’, ‘shoes’, ‘accessories’
  - ‘clothing’ > ‘underwear’
- Table 3: inheritance of LF values by semantic subclasses
  - ‘clothing’ inherits some LFs from ‘clothing and accessories’ and has some LFs of its own
  - ‘underwear’ inherits some LFs from ‘clothing’ and has some LFs of its own.
- Table 4: all the collocations (both inherited and added manually)

Statistics for ‘clothing and accessories’

- ‘clothing’ and ‘underwear’: 4989 collocations (2567 inherited and 2422 added manually)
- ‘shoes’: 909 collocations (539 inherited)
- ‘accessories’: 1060 collocations (626 inherited)
- ‘complements’: 987 collocations (151 inherited)

LFs in semantic analysis

- LFs in NLP: idiomatic translation in MT, paraphrasing, generation, disambiguation, corpus annotation.
- Another application: semantic analysis.
- SemETAP
- Task: to represent the meaning of the text in an explicit and unambiguous way.
- SemETAP is an option of the ETAP-4 linguistic processor and reuses its non-semantic modules (morphological analysis, syntactic dependency parsing, and normalization).
- Semantic analysis makes use of linguistic data and extralinguistic information (background knowledge).

More on SemETAP

- Crucial component of SemETAP: inference rules.
- Two levels of semantic structure are distinguished. Basic semantic structure (BSemS) interprets the text in terms of ontological concepts. Enhanced semantic structure (EnSemS) extends BSems by means of a series of inferences.
- LFs are used at two stages:
  - Constructing and normalizing BSemS
  - Drawing inferences of BSemS

Syntactic derivatives (S_i, A_i, Adv_i)

- In BSemS all predicates should be brought to the normalized form, which means that syntactic derivatives should be replaced by their keywords. In case of actantial derivatives, normalization also requires that the i-th actant of the keyword be explicitly established.
- Examples of actantial derivatives:
  - A_1(*fear*) = *fearful1, frightened* (≈ ‘such that fears something’)*,*
  - A_2(*fear*) = *fearsome, fearful2* (≈ ‘such that is feared’);
  - Adv_1(*hurry*) = *hastily* (≈ ‘hurrying’),
  - Adv_2(*permit*) = *with the permission* (≈ ‘being permitted’).

Normalizing operations triggered by these LFs

- A_1: The *child was fearful1 <frightened>* ==> ‘the child feared something’
- A_2: *The consequences were fearsome* ==> ‘one could fear the consequences’
- Adv_1: *He said good bye hastily* ==> ‘he said good bye; while saying it he was hurrying’
- Adv_2: *The evidence was examined by the experts with the permission of the court* ==> ‘the evidence was examined by the experts; the court permitted the experts to examine the evidence’.

Other LFs that trigger inferences

- Real_1(*promise*) = *fulfil - He fulfilled his promise to help me.*

Inference: ‘he helped me’.

- CausFunc_0(*crisis*): *bring about (a crisis).*

Inference: ‘a crisis takes place’.

- LiquFunc_0(*beard*): *shave off (one's beard).*

Inference: ‘the beard exists no longer’.

Conclusions and future work

- A new e-dictionary of Spanish supplied with Lexical Functions and other information (about 50,000 collocations).
  - 20,000 – frequent collocations of peninsular Spanish, that any B2 level student should master
  - 30,000 – domain of the body, body parts, emotions, clothing and accessories.
- Showed a new
- way LFs can be used in NLP applications.
- Goal: 75,000 collocations by the end of 2020.
- Significantly enlarge the set of adjectival and adverbial non-standard LFs.