Mark Davies (Provo)

Advanced research on syntactic and semantic change with the *Corpus del Español*

Historische Korpora der spanischen Sprache wie CORDE und ADMYTE sind für lexikographische Forschungen wertvoll, da ihre Suchmöglichkeiten auf einzelne Wörter und Phrasen ausgerichtet sind. Im Unterschied dazu hat das kürzlich veröffentlichte *Corpus del Español* etliche Eigenschaften, die das Korpus für syntaktische und semantische Forschungen wertvoll machen. Dazu gehören: Grammatikalische Klassifizierung, Lemmatisierung, eine große Synonymen-Datenbank, die Anwendung von Häufigkeitsstatistiken während der Abfrage und die Möglichkeit, benutzerbestimmte syntaktische und semantische Klassen zu kreieren. Der Beitrag bietet zahlreiche anschauliche Beispiele, wie diese Fähigkeiten der Suchmaschine genutzt werden können, um eine breite Auswahl diachronischer syntaktischer und semantischer Phänomenen des Spanischen zu untersuchen.

1. Overview

Within the past five years at least three large corpora of historical Spanish have become available. These include the *Corpus Diacrónico del Español* (CORDE; <http://corpus.rae.es/cordenet.html>), the *Corpus del Español* (<http://www.corpusdelespanol.org>), and more than 300 medieval texts from the Hispanic Seminary of Medieval Studies (HSMS) on CD-ROM (see <http://www.hispanicsociety.org> and a similar collection of texts via ADMYTE: <http://www.admyte.com>; cf. Wanner, in this volume). In total they comprise nearly 200 million words of text from the 1200s to the early 1900s, although there is some degree of overlap between the various corpora.

For researchers interested in examining the historical shifts with any exact words or phrase, any one of the three corpora should suffice, since the search engines in each of the three corpora allow such queries. In terms of research on historical syntax and semantics, however, we find that there are significant differences between the three corpora. With both the HSMS / ADMYTE texts on CD-ROM, as well as CORDE, searches for syntactic constructions are either very difficult or impossible. Because CORDE is not "lemmatized", it is impossible to search for all of the forms of a given verb at one time. For example, a researcher interested in the phrase [HACER merced] (*fizole merçed, hacen merced*, etc) would have to search for each of the many forms of *fazer* individually.

More seriously, neither of these two corpora are "tagged" for part of speech. Therefore a researcher who wants to look at clitic placement with infinitives (e.g. *para lo fazer / cantar / dezir*) would have to compile a list of hundreds or thousands of infinitival forms from Old Spanish, and then perform each of these searches individually. This difficulty would of course be multiplied if the researcher wants to examine a construction comprised of several parts of speech or lemma, such as [pronoun + FAZER + infinitive] (*le fizo pensar*, *nos fazen estar*, etc). In this case, the number of possible strings would be in the tens of thousands, and the searches would have to be carried one by one, involving weeks or months of searching. One option would be to allow "regular expression" to use patterns to try and match verb forms or part of speech (e.g. forms ending in *ar/er/ir/yr/jr* for the Old Spanish infinitive), but the search engine used in CORDE does not allow this type of pattern matching either. Therefore, both of these corpora are of very limited use for syntactic research, unless one obtains the actual text files and then uses a customized concordancing program to search for strings.

2. The *Corpus del Español*

The *Corpus del Español*, on the other hand, was explicitly designed to offer an extremely wide range of queries that allow researchers to examine both syntactic and semantic shifts in the language (see Davies 2003b, 2003c for a technical introduction to the architecture of the database). In terms of its background, I would note that this corpus was funded by the United States National Endowment for the Humanities, and that it was created solely by myself from April 2001–July 2002. The corpus is freely available, and it contains 100 million words of text from the 1200s–1900s, including 20 million words from the 1200s–1400s, 40 million words from the 1500s–1700s, and 40 million words from the 1800s–1900s.

As mentioned, the *Corpus del Español* allows an extremely wide range of queries, beyond those of exact words and phrases. The following sample queries summarize some of these features, and users of the web-based corpus have access to a number of context-sensitive tutorials to guide them through the possibilities.

1.1 Simple pattern matching
cans — *cansado, descansar, incansable* — "tired"

1.2 Collocations
borrar.* el/la *.n — *huella, marca, pasado* — "erase the N"

1.3 Word forms (20.000 lemma)
haber. — forms of haber: *ha, ouo, habrá* — "to have (PP)"

1.4 Parts of speech (30+ categories)
*.adv — adverbs

1.5 Synonyms / antonyms (synonym sets for 30.000+ words)
!inteligente — *vivo, capaz, agudo ...*

1.6 Combinations of word, lemma, parts of speech, and synonyms
!mandar.* que *.v_subj_ra — *hicieron que dijera, mandó que volvieran*
"they made her say, he ordered them to return"

1.7 Frequency – limiting and sorting
haber.* // 1500s>5 -1900s — *avia, uvo, obiese*
forms of *to have* (+ PP) that occur at least five times in the 1500s, but do not occur in the 1900s

1.8 Customized, user-defined lists
estar.* tan [gonzález:emociones].*
estaba tan alegre, estoy tan deprimido
any form of estar "to be" + "as" + any form of any word in the [emociones] list created by [gonzález]

Once users have submitted the query via the web interface, they then see a listing of the frequency of each of the matching forms in each century from the 1200s–1900s, as well as three different registers from the 1900s (19L = literature, 19O = oral, and 19M = miscellaneous (mainly non-fiction; newspapers and encyclopedias)). The following table is an example of partial results for the query [pn_obj.* querer.* *.v_inf] – any object pronoun + a form of the verb *querer* "to want" + an infinitive (a construction studied by Davies 1995b, 1998 and others).

#	*Phrase(s)*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*19L*	*19O*	*19M*
1	te quiero decir				17	10	1	8	49	11	38	
4	me quiero ir				32	10	1	2	23	7	16	
19	le quiere dar	9	1	1	7	6	2	6	4		3	1
22	te quiero contar			1	11	3			4		4	
	...											

Table 1. Frequency results

Users can click on words or numbers to see a KWIC (keyword in context) display of a particular phrase in a given century, a particular phrase in all centuries, or a set of phrases in any selected set of centuries. The following is an example of the KWIC display (note that the context here is greatly reduced in order for it to fit in this table). Note that – as with a traditional KWIC display – users can sort by left and right contextual words, as well as selecting a passage in order to see more context.

1/1	Cen	Text	Re-sort by: L-2 L-1	C	R-1 R-2
3	12	Libro de los f..	tiene gela forçada. Et non	le quiere dar	lo que a tomado & en logar
8	15	La Serrana de ..	desdicha el desengaño. No	me quiero casar,	padre, que creo que
14	19L	Follaje en los ..	¡Haré lo que quiera, no	me quiero ir!	Ya soy grande y sé hacer
27	19O	España Oral: C..	a mi madre y a mi padre -	Te quiero decir	que es una cosa que yo - y
	...	...	...	...	...

Table 2. KWIC display on the web

The important point – for the purposes of this paper – is that the *Corpus del Español* allows an extremely wide range of searches – far beyond anything remotely possible with CORDE. In addition, it is also very fast, usually requiring no more than one or two seconds for even the most complex queries of the entire 100 million word corpus. All of this functionality means that via the *Corpus del Español*, researchers of historical Spanish syntax and semantics finally have a tool that allows them to carry out in-depth research on a wide range of phenomena, in ways that could only be imagined three or four years ago.

3. Using the *Corpus del Español* for research on diachronic syntax

In the following sections, we will provide more detailed examples of some of the types of queries on historical Spanish syntax that can be carried out with the *Corpus del Español*. We will consider searches involving lemma, part of speech, synonyms, and customized lists, as well as searches in which we limit the output by the number of occurrences in a particular historical period or a register of Modern Spanish. In the final section of the paper, we will consider ways in which the *Corpus del Español* can be used to research semantic change.

In terms of lemmatization and part of speech annotation, basic searches might include a listing of all forms of a given verb that occurred only in Old Spanish (e.g. *dixiemos*, *dezian*, *dira* for *decir* "to say") or a listing of infinitives that have entered the language during the past 200 years (e.g. *detectar*, *incrementar*, *financiar*). However, it is possible to combine these to create more powerful queries of the syntax. For example, we can combine part of speech categories to search for all construction involving a preposition + a subject pronoun + an infinitive. These would give us examples of the "lexical subject of infinitive construction" that has been studied by Davies (2003a) and others.

(1) [word/phrase] *.prep *.pn_subj *.v_inf

phrase	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*19L*	*19O*	*19M*
de yo haber	0	0	0	0	0	0	2	3	0	3	0
para él ser	0	0	0	1	0	0	1	3	0	3	0
de él haber	0	0	0	1	1	0	0	2	0	2	0
para él tener	0	0	0	1	0	0	1	2	1	1	0
para yo tener	0	0	0	0	0	0	0	2	0	1	1
para nosotros lograr	0	0	0	0	0	0	0	2	0	2	0
para nosotros tener	0	0	0	0	0	0	0	2	0	2	0

Table 3. Multiple part of speech tags

Likewise, we can combine lemma and part of speech categories. For example, the following query will retrieve examples of the "experiencer" subject raising construction discussed in Davies (1997a, 1997b):

(2) [word/phrase] le/les parecer.* *.v_inf [limits] +1600s

phrase	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*19L*	*19O*	*19M*
le pareció ser	0	0	0	16	10	4	0	1	1	0	0
le pareciese convenir	0	0	0	2	3	0	0	0	0	0	0
le pareció estar	0	0	0	5	3	1	0	1	1	0	0
le pareció haber	0	0	0	4	3	2	2	1	1	0	0
le parecía ser	0	0	0	11	3	2	5	2	2	0	0
les pareció ser	0	0	0	9	3	1	0	0	0	0	0
le pareció avisar	0	0	0	0	2	0	0	0	0	0	0
le pareció aguardar	0	0	0	0	2	0	0	0	0	0	0

Table 4. Lemma and part of speech (subject raising)

As mentioned, it is also possible to use synonym information as part of the search. There are more than 30.000 synonym sets in the corpus database, and these can be used directly as part of the query. For example, the following query shows the aggregate frequency of [se] + any form of any synonym of *romper* "to break", which is often used as a decausative verb (see Davies, this volume, for an overview of the historical shifts with *se*). In the table of results, the occurrences are grouped by lemma, and the highlighted entries show those verbs with which there has been a significant increase during the past 100–200 years.

(3) [word/phrase] se !romper.*

se +	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*19L*	*19O*	*19M*
romper	5	5	63	96	103	114	241	295	123	109	63
quebrar	2	9	35	102	65	26	52	65	47	9	9
violar	0	0	0	1	0	9	8	23	2	6	15
destrozar	0	0	0	0	0	5	14	17	12	4	1
despedazar	0	0	0	11	7	7	28	13	10	0	3
divider	0	0	0	0	3	1	3	7	2	3	2
rajar	0	1	0	1	0	3	3	13	7	4	2

Table 5. Synonyms (*se* with synonyms of *romper*)

Another example of the use of synonyms and part of speech is the following query, which searches for any synonym of *difícil* "difficult" + *de* + an infinitive. This is the "object to subject raising" construction investigated by Davies (2002) and others:

(4) [word/phrase] !difícil.* de *.v_inf [limits] +1800s

	12	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*19L*	*19O*	*19M*
duro de pelar	0	0	0	0	0	0	14	1	0	1	0
difícil de explicar	0	0	0	0	0	2	13	26	14	9	3
imposible de describir	0	0	0	0	0	0	12	3	2	0	1
difícil de hacer	0	0	0	3	0	2	10	3	1	1	1
difícil de comprender	0	0	0	1	2	4	9	8	3	2	3
difícil de vencer	0	0	0	1	2	2	9	5	3	1	1
imposible de realizar	0	0	0	0	0	0	9	0	0	0	0

Table 6. Synonyms (object to subject raising)

More complex queries can be formed, which involve synonyms, lemma, and part of speech. For example, the following query will find all of the phrases containing an object pronoun + any form of any synonym of *mandar* "to order / command" + an infinitive. This would be an example of the causative construction studied in Davies (1995a, 1996, 2000). Once again, the results are grouped by lemma.

(5) [word/phrase] le/les !mandar.* *.v_inf[limit] +1800s

.	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*19L*	*19O*	*19M*
le HACER ver	0	0	1	6	6	23	48	21	17	3	1
le HACER comprender	0	0	0	0	0	0	36	5	5	0	0
le HACER dar	0	0	1	41	23	10	28	7	6	0	1
le HACER creer	0	0	0	7	8	9	26	15	10	3	2
le HACER caer	0	0	2	13	3	8	25	4	3	1	0
le HACER temblar	0	0	0	1	3	1	23	6	6	0	0
le MANDAR dar	6	5	9	70	22	7	22	1	1	0	0

Table 7. Synonyms and part of speech (causatives)

One other powerful feature of the corpus are the "customized lists" that can be created by end-users. Let us suppose, for example, that a researcher is investigating shifts with causatives and verbs of perception (*hacer*, *dejar*, *ver*, *oír*, etc). The synonym lists would be insufficient, because while they would show the synonyms for a given verb (e.g. *mandar*, as in the example above), they would not list all of the causatives or verbs of perception. However, the end user can create such a list (via the same web interface as the actual search form itself) and then re-use this list in subsequent queries – even weeks or months later. Incidentally, one of the main advantages of exposing the functionality of the customized lists to the end user is that even for lemma and part of speech, the burden of exhaustively annotating the corpus is taken off the corpus creator, because the end users themselves can create customized verb lists or part of speech lists (e.g. temporal adverbs, or a subset of intransitive verbs) where the original annotation is incomplete or insufficient.

As an example of using customized lists, consider the case of a hypothetical user [jones], who creates a list called [perception], containing verbs of perception like *ver*, *oír*, *ver*, *escuchar*, and *sentir*. She could then make reference to this list as part of the normal query syntax, and be able to retrieve a list of matching hits like that shown in the following table (once again, the results are grouped by lemma):

(6) [word/phrase] le/les [jones:perception].* *.v_inf

CAPITALS = *all verb forms*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*19L*	*19O*	*19M*
le OÍR decir	0	0	0	20	17	6	32	23	17	6	0
le VER entrar	0	1	0	4	8	5	26	1	1	0	0
le OÍR hablar	1	0	0	1	2	2	20	3	2	1	0
le VER llegar	0	0	0	3	1	2	17	5	5	0	0
le VER pasar	0	0	0	2	2	0	13	1	1	0	0
le VER salir	3	0	1	6	5	0	12	0	0	0	0
le VER venir	2	1	3	5	3	1	11	0	0	0	0

Table 8. Customized lists (verbs of perception)

Another example is the following one, in which a user [garcía] creates a customized list of decausative verbs (see Davies, this volume), and then uses this list to find all cases of *se* + *le* or *les* + the decausative verb:

(7) [word/phrase] se le/les [garcía:decausative].*

	12	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*19L*	*19O*	*19M*
se le perder	1	0	8	49	26	8	17	34	27	7	0
se me perder	0	0	1	14	13	0	4	28	19	9	0
se le abrir	1	0	0	17	8	5	16	22	15	3	4
se le mover	1	1	1	4	0	2	4	20	19	1	0
se le enredar	0	0	0	0	0	1	10	19	12	6	1
se le hundir	0	0	0	11	0	1	9	16	14	2	0
se le romper	0	0	1	11	4	4	17	16	8	7	1
se me romper	0	0	0	3	0	1	5	16	9	6	1
se me cerrar	0	0	0	4	0	1	5	14	10	4	0
se le cerrar	0	0	2	20	5	3	31	11	4	6	1

Table 9. Customized lists (*se* + pronoun + decausative verb)

4. Using the *Corpus del Español* for research on diachronic semantics

In addition to being very useful for diachronic syntax, the *Corpus del Español* also allows advanced queries that can be used to study diachronic semantics, or changes in meaning over time. Because the frequency information for each one, two, three, and four word string in each century is stored in a relational database (see Davies 2003b, 2003c), this can be exposed to the end user to see and compare related words and phrases. For example, it would be possible to run a query that would find all of the nouns that occur with the adjective *suave* "soft" at least three times in the 1900s, but occur with the adjective *duro* "hard" less than three times in the same period. In less than three seconds we can search the entire 100 million word corpus, and see results like the following:

ADJ	*+suave*	*–duro*	*ADJ*	*+duro*	*–suave*
voz "voice"	22	2	cara "face"	10	0
viento "wind"	10	0	pan "bread"	10	0
músico "music"	6	0	tiempo "time"	9	0
sabor "taste"	5	0	vida "life"	8	0
invierno "winter"	5	1	palabra "word"	8	1
brisa "breeze"	5	1	hombre "man"	7	0
tono "tone"	4	1	ojo "eye"	6	1

Table 10. Comparing word frequencies (nouns with *suave*, *duro*)

This functionality can be used to investigate diachronic shifts in meaning. For example, a modification of the previous query could find all nouns that occurred more than three times with *duro* in the 1500s but less than three times in the 1900s, or vice versa. The range of nouns should give us some insight into the extension (metaphorical, or otherwise) of *duro* in the two periods. The following table shows the nouns that are more common with *duro* in the 1500s (left side) and in the 1900s (right side).

ADJ	*+1500s*	*–1900s*	*ADJ*	*+1900s*	*–1500s*
suerte "luck"	17	0	linea "line"	13	0
hierro "iron"	11	0	disco "disc"	12	0
muerte "death"	10	0	roca "boulder"	12	1
pena "shame"	10	0	cuello "neck"	11	0
bronce "bronze"	6	0	cara "face"	10	0
cuero "leather"	6	1	pan "bread"	10	2

Table 11. Comparing word frequencies across time (nouns with *suave*, *duro*)

The following example is another one that seeks to uncover diachronic shifts in meaning. In this case, we look for all nouns that occur with *fazer* / *hacer* "to do / make" twenty times in the 1200s, but less than three times in the 1900s. As can be seen, many of the phrases with *hacer* refer to pledges and honor, which was a more central aspect of medieval literature (and life) than in the 1900s.

(8) [word/phrase]hacer.* *.n [limits] 1200s> 20 1900s<3

HACER +	*1200s*	*1900s*	*noun*	*1200s*	*1900s*
emienda "reparation"	207	0	limosna "alms"	40	1
adulterio "adultery"	121	0	honra "honor"	34	0
duelo "pain"	77	0	postura "agreement"	31	0
pleito "court case"	76	1	altar "altar"	25	0
engaño "deception"	65	1	adios "goodbye"	23	2

Table 12. Lexical complements (HACER + noun; 1200s / 1900s)

Naturally, one of the most useful tools in examining semantic shift is the thesaurus containing 30.000 synonym sets, which is stored in the relational database and can be accessed directly as part of the query syntax. For example, the following query examines the frequency of the synonyms of *hablar* "to speak" that occur with at least ten times in the 1900s, but less than three times in the 1700s. (Note that the frequency limits are for individual word forms, while the totals shown in the table are the aggregate value for that verb. Note also that the figures only partially take into account possible orthographic changes since Old Spanish.)

(9) [word/phrase]!hablar.* [limits] 1900s>10 1700s<3

verb	12	13	14	15	16	17	18	19
dialogar "to hold talks"	0	0	2	1	0	0	10	150
susurrar "to whisper"	0	0	2	0	2	2	22	21
enunciar "to enunciate"	0	0	0	0	0	2	25	16
cuchichear "to whisper"	0	0	0	0	0	0	4	11
musitar "to murmur"	0	0	0	0	0	2	3	11
balbucir "to stammer"	0	0	0	0	0	0	14	9
disertar "to discourse"	0	0	0	0	1	3	21	9

Table 13. Synonyms (*hablar*, +1900s / –1700s)

It is possible to run even more complicated queries involving synonyms. For example, the following query looks at different ways of expressing "to have a problem (with something)", by looking for any form of any synonym of *tener* "to have" followed by any form of any synonym of *problema*. It shows the phrases that occur more commonly in the 1900s than in the 1800s.

(10) [word/phrase]!tener.* !problema.*[limits] 1900s>10 1700s<3

	17	18	19	19L	19O	19M
tener problema	1	0	296	61	212	23
haber problema	0	4	270	51	203	16
tener difficultad	5	0	43	17	15	11
tener duda	0	3	43	27	13	3
haber duda	0	2	17	7	4	6
haber dificultad	0	2	4	1	2	1

Table 14. Synonyms + lemma (*tener problema*; +1900s / –1800s)

Finally, the customized lists (discussed above) can be invaluable for researching semantic change. Recall that users can create their own lists of any set of words. In the case of semantic change, they might create lists of words relating to a semantic field such as clothing, religious practices, words

expressing anger, or words related to the sea, and then re-use these lists directly as part of the syntax of subsequent queries.

We will briefly provide two examples of the usefulness of such customized lists. In this case, we are investigating the frequency and use of phrases related to "breaking a part of the body", in which [smith] has created a list of [body] parts. The query will produce the results in the following table.

(11) [word/phrase]romper el/la/los/las [smith:body].*

romper +	12	13	14	15	16	17	18	19	19L	19O	19M
cabeza	0	0	0	4	15	24	39	21	15	6	0
pierna	0	0	0	0	1	2	2	8	6	2	0
boca	0	0	0	1	1	0	0	6	5	1	0
nariz	0	0	0	0	0	2	3	6	4	2	0
dedo	0	0	0	0	0	0	0	4	3	1	0
ojo	0	0	0	0	0	0	0	3	1	2	0
pie	0	0	1	0	0	1	1	2	1	1	0

Table 15. Customized lists (*romper* + body parts)

A second example shows how one query can provide output that might be inputted into a customized list, to provide a useful semantic field for subsequent queries. We will first search for all nouns that occur in the phrase "to die of N" (die of hunger, die of cancer, etc):

(12) [word/phrase]morir.* de *.n

We can further limit the results to those phrases that occur in the 1900s but not the 1800s, and vice versa. The following table shows some of the nouns with which there is the largest increase or decrease between the 1800s and the 1900s:

morir de "die of"	+18	–19	*morir de* "die of"	+19	–18
pena "sadness"	35	7	ganas "desires"	22	0
dolor "pain"	20	5	calor "heat"	9	1
vergüenza "shame"	24	13	cancer "cancer"	9	0
amor "love"	35	14	miedo "fright"	23	11
tristeza "sadness"	13	4	risa "laughing"	30	19
cellos "jealousy"	8	3	pulmonía "pneumonia"	5	0
pesadumbre "grief"	7	0	aburrimiento "boredom"	5	1

Table 16. Customized lists (*morir de*; 1800s / 1900s)

It is interesting to see the difference between the two centuries. In the texts from the 1800s, there was still a strong sense of honor and the tragic sense of life, as the texts referred to death from sadness, pain, shame, a love gone bad, jealousy, and grief. In the 1900s, there were new labels attached to actual

physical ailments (cancer or pneumonia), but in other cases people die from too much laughter, boredom, or desires (of course most likely in a metaphorical sense).

At any rate, once such a list of "mortal threats" is created – perhaps for each of the eight centuries in the corpus – these lists of words could then be input into a customized list [mortal], or something of the sort. Subsequent queries could then be run using these words as part of the phrase – such as which adjectives occur most commonly with each noun in the different centuries, or which verbs besides *morir* "to die" are used with these words.

As one can readily see, the possibilities for research are endless. Due to the extremely rich query syntax of the *Corpus del Español*, users can research virtually any topic of diachronic Spanish syntax or semantics, at a level far beyond that of any other existing corpus of historical Spanish.

References

Davies, Mark 1995a: The evolution of the Spanish causative construction. *Hispanic Review* 63, 57–77.
— 1995b: Analyzing syntactic variation with computer-based corpora: The case of Modern Spanish clitic climbing. *Hispania* 78, 370–380.
— 1996: The diachronic interplay of finite and nonfinite verbal complements in Spanish and Portuguese. *Bulletin of Hispanic Studies* 73, 137–158.
— 1997a: The history of subject raising in Spanish. *Bulletin of Hispanic Studies* 74, 399–411.
— 1997b: A corpus-based analysis of subject raising in Modern Spanish. *Hispanic Linguistics* 9, 33–63.
— 1998: The evolution of Spanish clitic climbing: A corpus-based approach. *Studia Neophilologica* 69, 251–263.
— 2000: Syntactic diffusion in Spanish and Portuguese infinitival complements; in: Dworkin, Steven / Wanner, Dieter (eds.): *New approaches to old problems: Issues in Romance historical linguistics*. Amsterdam / Philadelphia: John Benjamins, 109–127.
— 2002: *Esto es ligero de fazer*: Object to subject raising in Medieval and Early Modern Spanish; in: Lee, James F. *et al.* (eds.): *Structure, meaning, and acquisition of Spanish*. Somerville, MA: Cascadilla Press, 19–31.
— 2003a: Diachronic shifts and register variation with the "Lexical Subject of Infinitive" construction. (*Para yo hacerlo*); in: Montrul, Silvina / Ordóñez, Francisco (eds.): *Linguistic theory and language development in Hispanic languages*. Somerville, MA: Cascadilla Press, 13–29.
— 2003b: Annotation without lexicons: an alternative to the standard bootstrapping approach; in: Rayson, Paul Rayson *et al.* (eds.): *Proceedings from Corpus Linguistics 2003* (Lancaster, England, March 2003), 174–183.
— 2003c: Relational n-gram databases as a basis for unlimited annotation on very large corpora; in: Simov, Kiril (ed.): *Proceedings from the Workshop on Shallow Processing of Large Corpora* (Lancaster, England, March 2003), 23–33.