Topic expressions in Spanish: contrasting corpus and questionnaire data in the analysis of prepositional synonymy

12th International Cognitive Linguistics Conference
University of Alberta, Edmonton, 23-28 June 2013

Anton Granvik
HANKEN School of Economics, Helsinki
anton.granvik@hanken.fi

1. Background

Spanish prepositions

- 17 basic prepositions, hundreds of compound prepositions (*NGLE*)
- The compound prepositions have appeared (historically) in order to specify certain meanings and relieve the basic prepositions of excessive usage work load
  - The topic meaning ('about') is one case in point (Halliday 1967)
- Basic prepositions *de* and *sobre* have always been used with this meaning (DE CIVITATE DEI, SED HAC SUPER RE NIMIS DIXI) (Bassols de Climent 1967, I)
- *Acerca de* and *en torno a* appear later as alternatives
- Further expressions: *en cuanto a*, (*con*) *respecto a/de*, *en relación a/con*, *en lo tocante a*,...

The four prepositional topic expressions

de ('of, from')	polysemous, topic only one out of dozens of senses
sobre ('on, over')	polysemous, topic only one out of a dozen or so senses
acerca de ('about')	monosemous, formally related to locative *cerca de* ('near, close to')
en torno a ('around')	polysemous: locative, approximative vs. abstract (=topic) meaning

Aim of the study

- Address the differences in usage between four near-synonymous prepositions
- Are the expressions synonymous or not? To what degree?
- Compare different analyses (Arppe 2006, 2008; Liu 2013, Vanhatalo 2003)
  - What can corpus analysis tell us about synonymy
  - Compare the results of corpus analysis with questionnaire data
  - What other information might be used and/or needed in order to reach a "full" account of this issue?

2. Methodology and corpus

Corpus analysis

- Annotation of examples
- Individual characterization of each expression
- Quantitative analysis: multinomial logistic regression (SPSS)

Questionnaire data

- Quantitative and qualitative data analysis
- Comparison of the results

Corpus del español (Davies 2002-)

- Large diachronic corpus of Spanish (100 million words)
  - Comparable to the COCA-corpus
- 4 x 100 examples of the four topic expressions (TE)
- Manual annotation of the 400 examples
- 15 syntactic and four semantic factors

Syntactic factors

HEAD/GOV word class (N, V, A, 0 / N, pron, V, conj)
HEAD/GOV complexity (NP(1-3) vs. N)
HEAD/GOV number (sg vs. pl)
HEAD/GOV definiteness (def vs. indef)
HEAD/GOV determinacy (det vs. indet)
HEAD/GOV modification (modif vs. unmodif)
HEAD/GOV attribute (attr vs. no attr)
HEAD before/after the TE

Semantic factors

Word class (of HEAD and GOV element) = Communication/cognition/action/general
Animacy (of HEAD / GOV element) = Human/Non-human/collective/Unspecified
Presence of other Topic expression (e.g. *de* and *sobre*) in the same clause
Abstract/figurative vs. concrete reading

Ag_anim	uniq	uniq+	H_fig	H_wc	H_wc+	H_lex	HEAD	H_num	H_def	H_det	H_modif	H_attr	H_compl	H_pos	Nr. 2	Nr.	Source	Genre	Marker	EXAMPLE	G_lex	GOV	G_num	G_def	G_det	G_modif	G_attr	G_compl	G_fig	G_wc
col	1	1	fig	gral	m_com	informe	N	sg	0	1	0	0	1	pre	1	2	Bolivia:ERBOL:	N	sobre	la Comisión de	acciones	N	sg	1	1	1	1	3	1	act
hum	1	1	fig	com	com	conferencias	N	pl	1	1	0	1	2	pre	2	5	Enc: Crítica liter	AC	sobre	Ralph Waldo E	principio	N	sg	1	1	1	0	2	1	abs_obj
hum	1	1	conc	com	com	platicar	V	sg	act	inf	0	1	1	pre	12	40	Entrevista (PA	OR	sobre	y en las platica	tema	N	sg	1	1	0	0	1	1	abs_obj
							O								128	88	Enc: José Gaos	AC	en torno a	(El pensamient	filosofía	N	sg	1	1	1	0	2	1	abs_obj
0	0	de	fig	gral	act	girar	V	sg	act	inf	0	0	0	pre	139	111	Enc: Enseñanza	AC	en torno a	como tal, lo cie	idea	N	sg	1	1	0	1	2	1	cog
hum	1	1	conc	com	com	cuestionamie	N	pl	0	0	1	1	2	pre	186	249	Entrevista (PA	OR	en torno a	consenso total	lo que	pron	sg	1	1	0	1	2	1	prod_cog
unsp	1	1	fig	cog	cog	dudas	N	pl	0	0	0	0	0	pre	216	87	Arg:Prensa:21_	N	acerca de	El ministro de	que	conj	sg	0	0	0	1	1	1	abs_obj
hum	0	de	conc	com	com	hablar pestes	V	sg	act	ind	1	0	1	pre	273	117	Tántalo en el tr	F	acerca de	el cuñado de A	modales	N	pl	1	1	0	1	2	1	abs_obj
hum	1	1	fig	gral	act	hacer	V	pl	act	inf	1	0	1	pre	274	83	Entrevista (Zed	OR	acerca de	muchos años y	esto	pron	sg	0	0	0	0	0	1	abs_obj
hum	1	1	fig	cog	cog	contento	A	sg	1	0	0	0	0	pre	301	574	Entrevista (ABC	19-OR	de	, tiene en la ca	tener	V	sg	act	inf			0	0	abs_obj
unsp	1	1	fig	cog	cog	eso	pron	sg	0	1	0	0	1	pre	326	1450	Habla Culta: Mi	19-OR	de	No, no, no, no,	conservar	V	sg	act	inf			0	0	act
hum	1	1	conc	com	com	hablar	V	sg	act	ind	0	0	0	pre	365	154	Habla Culta: La	19-OR	de	cristiano. Esto	esto mism	pron	sg	0	0	1	0	1	1	abs_obj

3. Data

Corpus del Español – overview

PREP	*de*	*sobre*	*acerca de*	*en torno a*	Total
topic uses	100	100	100	100	400
total	1334	255	100	290	1978
percentage	7,5 %	39,2 %	100 %	34,5 %	20,2 %
Total uses	1 163 904	33 572	782	1006	1 199 264
Expected nr of TE uses	87 315	13 165	782	347	242 521

- Characteristics of Head element (HEAD)

HEAD	*de*	*sobre*	*acerca de*	*en torno a*	mean
N	37 %	73 %	51 %	78 %	60 %
V	54 %	27 %	42 %	20 %	36 %
A	9 %	0 %	2 %	0 %	3 %
O	0 %	0 %	5 %	2 %	2 %
Sum	100 %	100 %	100 %	100 %	100 %

- Characteristics of Governed element (GOV)

GOV	*de*	*sobre*	*acerca de*	*en torno a*	mean
N	61 %	91 %	80 %	87 %	79 %
pron	24 %	9 %	15 %	10 %	15 %
V	10 %	0 %	1 %	0 %	3 %
conj	5 %	0 %	4 %	3 %	3 %
sum	100 %	100 %	100 %	100 %	100 %

$p = 0{,}000$, $\chi2 = 67{,}426$ (9 df); $p \leq 0$, $\chi2 = 47{,}466$ (9 df)
N = noun, V = verb, A = adjective, O = no head; N = noun, pron = pronoun, V = verb, conj = (subordinate clause introduced by a) conjunction

Preliminary observations

- *De* and *sobre* are highly frequent, despite relatively low percentages of use as TE
- *Acerca de* only functions as a topic expression
- *En torno a* is the least frequent
- Based on general frequencies *de* and *sobre* are the unmarked pair of TEs, *acerca de* and *en torno de* are stylistic alternatives
- But, only *acerca de* and *en torno a* can be used as sentence-level topicalizers
- Only *de* and *acerca de* are used with adjectival heads and take infinitives as complements; they are also combined with clausal complements
- *Sobre* is most restricted, taking only N and V as heads; and N and pronouns as complements
- *En torno a* is similar *to sobre*, but has a few cases without lexical head and with clausal complements

Multinomial logistic regression data

Case Processing Summary

		N	Marginal Percentage
Marker	de	99	25,3%
	sobre	100	25,5%
	acerca de	95	24,2%
	en torno a	98	25,0%
Valid		392	100,0%
Missing		8	
Total		400	
Subpopulation		65[a]	

a. The dependent variable has only one value observed in 26 (40,0%) subpopulations.

Model Fitting Information

Model	Model Fitting Criteria: -2 Log Likelihood	Likelihood Ratio Tests: Chi-Square	df	Sig.
Intercept Only	594,716			
Final	271,640	323,075	24	,000

Parameter Estimates

Marker[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B): Lower Bound	Upper Bound
sobre	Intercept	-5,402	1,044	26,762	1	,000			
	G_fig	1,831	,660	7,687	1	,006	6,240	1,710	22,768
	Ag_anim_hum	-1,272	,513	6,140	1	,013	,280	,102	,767
	HEAD_1	2,456	,771	10,160	1	,001	11,660	2,575	52,798
	Genre_News	21,476	,414	2695,315	1	,000	2,123E9	9,438E8	4,776E9
	Genre_Fiction	3,069	,669	21,039	1	,000	21,520	5,799	79,870
	Genre_Academic	21,013	,485	1878,687	1	,000	1,336E9	5,165E8	3,455E9
	G_det_det	,934	,448	4,352	1	,037	2,546	1,058	6,125
	H_def_active	2,380	,842	7,990	1	,005	10,810	2,075	56,323
acerca de	Intercept	-5,439	,997	29,787	1	,000			
	G_fig	2,496	,686	13,223	1	,000	12,134	3,160	46,589
	Ag_anim_hum	-,976	,518	3,555	1	,059	,377	,137	1,039
	HEAD_1	1,058	,665	2,528	1	,112	2,881	,782	10,616
	Genre_News	21,527	,445	2335,827	1	,000	2,234E9	9,332E8	5,348E9
	Genre_Fiction	3,615	,646	31,347	1	,000	37,163	10,482	131,750
	Genre_Academic	21,243	,514	1707,846	1	,000	1,682E9	6,143E8	4,608E9
	G_det_det	1,215	,457	7,059	1	,008	3,369	1,375	8,252
	H_def_active	1,559	,731	4,555	1	,033	4,756	1,136	19,912
en torno a	Intercept	-6,280	1,121	31,369	1	,000			
	G_fig	2,328	,712	10,691	1	,001	10,261	2,541	41,431
	Ag_anim_hum	-2,189	,527	17,226	1	,000	,112	,040	,315
	HEAD_1	2,821	,811	12,103	1	,001	16,794	3,427	82,296
	Genre_News	21,132	,000	.	1	.	1,505E9	1,505E9	1,505E9
	Genre_Fiction	2,400	,736	10,621	1	,001	11,025	2,603	46,692
	Genre_Academic	20,283	,000	.	1	.	6,437E8	6,437E8	6,437E8
	G_det_det	1,826	,485	14,199	1	,000	6,207	2,401	16,043
	H_def_active	2,824	,898	9,891	1	,002	16,848	2,898	97,933

a. The reference category is: de.

Likelihood Ratio Tests

Effect	Model Fitting Criteria: -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests: Chi-Square	df	Sig.
Intercept	342,333	70,692	3	,000
G_fig	290,544	18,903	3	,000
Ag_anim_hum	294,938	23,297	3	,000
HEAD_1	294,173	22,532	3	,000
Genre_News	412,033	140,393	3	,000
Genre_Fiction	320,882	49,241	3	,000
Genre_Academic	339,035	67,395	3	,000
G_det_det	288,290	16,650	3	,001
H_def_active	286,015	14,375	3	,002

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Classification

Observed	Predicted: de	sobre	acerca de	en torno a	Percent Correct
de	90	1	4	4	90,9%
sobre	15	28	20	37	28,0%
acerca de	12	17	37	29	38,9%
en torno a	5	20	13	60	61,2%
Overall Percentage	31,1%	16,8%	18,9%	33,2%	54,8%

Pseudo R-Square

Cox and Snell	,561
Nagelkerke	,599
McFadden	,297

2nd round of observations

G_fig
- If the governed element is used in a metaphorical sense, the odds for using *sobre*, *acerca de* or *en torno a* are six to twelvefold compared to *de*
- I.e. *de* is mostly used with non-metaphorical complements

Ag_anim_hum
- If the agent of the head element (noun/verb/adjective) is human, the odds for using *de* are manyfold compared to the other TEs

Genre (News, Fiction and Academic)
- The odds for using *sobre*, *acerca de* or *en torno a* compared to *de* are manyfold in the Genres other than Oral
- These TEs are clearly preferred in more formal registers
- Only Fiction is significant with regards to *en torno a*

HEAD_1 = Head element is a noun
- If the head is a noun, the odds for using *sobre* and *en torno a* are over tenfold compared to *de*

H_def_active
- If the head element is a verb in active voice (compared to an impersonal or passive voice), the odds for using *sobre*, *acerca de* or *en torno a* are five to sixteenfold compared to *de*

G_det_det
- If the complement (governed element) is a noun with a determiner, such as definite or indefinite article *el/la*, *un/una* the odds for using *sobre*, *acerca de* or *en torno a* are three to six times higher than for *de*
- *De* is used with more variable and unspecified complements, e.g. bare nouns, indefinite and relative pronouns, etc.

Questionnaires

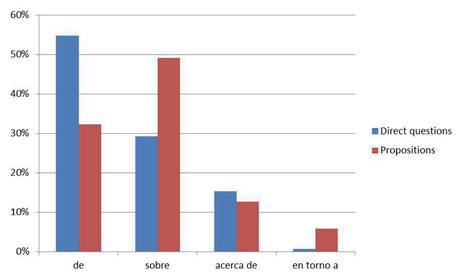
Data from questionnaires A + B

HEAD	*de*	*sobre*	*acerca de*	*en torno a*	Mean
N	59 % / 56 %	48 % / 64 %	54 % / 48 %	56 % / 0 %	54 %
V	41 % / 44 %	52 % / 36 %	46 % / 52 %	44 % / 100 %	46 %
sum	100 %	100 %	100 %	100 %	100 %
Com	43 % / 51 %	43 % / 46 %	46 % / 53 %	24 % / 0 %	39 %
Cog	34 % / 29 %	31 % / 30 %	25 % / 26 %	27 % / 5 %	29 %
Gral	23 % / 20 %	26 % / 24 %	28 % /21 %	49 % / 95 %	32 %
sum	100 %	100 %	100 %	100 %	100 %
N	423 / 452	282 / 349	153 / 85	79 / 21	937 / 907
	30 % / 38 %	45 % / 50 %	16 % / 9 %	8 % / 2 %	100 % / 100 %

- High correlation (r = 0,98) between the two questionnaires
- A: Topic expressions given, 49 TEs to fill in (N = 937); B: no topic expressions given, 53 TEs to fill in (N = 1040)

First two rows (N, V):A) $p \leq 0$, $\chi2 = 12{,}8$ (3 df); B) $p \leq 0$, $\chi2 = 18{,}5$ (df = 3)
Following three rows: A) (Com, Cog, Gral): $p = 0{,}045$, $\chi2 = 12{,}8$ (6 df); B) $p \leq 0$, $\chi2 = 30{,}5$ (df = 6)

TEs in direct questions vs. propositions



3rd round of observations

- *De* and *sobre* stand out as default topic markers
- In questionnaire B, which has no given TEs, the frequencies for *de* and *sobre* go up compared with A: 30 % (*de*) and 45 % (*sobre*) > 38 % and 50 % respectively
- *Acerca de* and *en torno a* show significantly lower frequencies in questionnaire B: 16 % (*acerca de*) and 8 % (*en torno a*) > 9 % and 2 %
- Both *en torno a* and *acerca de* are felt to belong to a formal register
- Only one HEAD expression consistently triggers *en torno a*, namely *girar* 'to circle'

Questions vs. propositions
- In direct questions *de* is the most frequent TE while *sobre* is the most frequent overall

4. Results and discussion

What do the different approaches tell us about the synonymy relations between the four TEs?

- Different approaches yields different kinds of results, i.e. the TEs can be distinguished on different levels, depending on what one looks at and for
  - *De* and *sobre* are the default TEs, *acerca de* and *en torno de* are alternatives
  - *De* is lexically more restricted than *sobre*, but syntactically more independent
- **Multinomial regression analysis** brings forward very detailed information
  - *De* and *en torno a* are best described by the model
  - No significant effect of the semantic classes (com, cog, gral)
- The **questionnaire data** –analyzed on a more holistic level and not annotated– confirm some of the observations made on the corpus data
  - *sobre* is the default TE, followed by *de* and, to a lesser extent,*tent, acerca de*
- But the questionnaires also highlights other aspects
  - Highly marginal use of *en torno a* when not explicitly mentioned
  - *En torno a* is preferredly used with verbs (vs. nouns in log. regr.)
  - *Acerca de* is preferred in more "formal" contexts
- What I haven't done:
  - Detailed collocational/collostructional analysis
  - Annotate the questionnaire data and compare it with the corpus data
  - Thorough analysis of the answers to the open questions included in the questionnaires

References

Arppe, Antti (2006): "Complex phenomena deserve complex explanations – choosing how to think in Finnish", paper presented at the QITL2 Conference, Osnabrück, Germany, 2 June 2006.
Arppe, Antti (2008): *Univariate, bivariate, and multivariate methods in corpus-based lexicography – a study of synonymy*. Publications of the Department of General Linguistics, University of Helsinki.
Bassols de Climent, Mariano (1967): *Sintaxis latina*. 2 vols. Madrid: CSIC.
Davies, Mark (2002-): *Corpus del español. 100 milloin words. 1200s-1900s*. Available on line at http//:www.corpusdelespanol.org.
Halliday, Michael A.K. (1967): "Notes on transitivity and theme in English. Part 2", *Journal of Linguistics*, 3, 199-244.
Liu, Dilin (2013): "Salience and construal in the use of synonymy: A study of two sets of near-synonymous nouns", *Cognitive Linguistics*, 24-1, 67-113.
NGLE = Real Academia Española y AALE (2009): Nueva gramática de la Lengua Española. 2 vols. Madrid: Espasa.
Vanhatalo, Ulla (2003): "Kyselytestit vs. korpuslingvistiikka lähisynonyymien semanttisten sisältöjen arvioinnissa – Mitä vielä keskeisestä ja tärkeästä?", *Virittäjä*, 107:3, pp. 351-369.