Lectal conditioning of lexical collocations

Jose Tummers[1,2] Dirk Speelman[2] Dirk Geeraerts[2]
[1]Leuven University College [2]University of Leuven
jose.tummers@khleuven.be, {dirk.speelman,dirk.geeraerts}@arts.kuleuven.be

1 Problem statement

The last decade, empirical linguistics focusing on genuine data has largely benefited from theoretical developments in Construction Grammar and from methodological and technical innovations in usage-based linguistics. In both frameworks, there is an obvious interest for lexical selectivity and idiomatic language use as part of the interplay between lexicon and grammar in probabilistic language models (Gries 2008). Lexical preference patterns are modeled along the paradigmatic axis, called collostructions (Stefanowitsch & Gries 2003, 2008), as well as the syntagmatic axis, called collocations (Sinclair 1991; Speelman et al. 2009; Wulff 2008), proving that the instantiation of constructions and constructional slots is at least partially conditioned by lexical selection restrictions.

Less attention has been paid to the lectal dimension of language use, referring to language external sources of variation. However, in a usage-based language model, the properties of the actual usage settings should be taken into account since they influence the language use (Geeraerts 2005). In this respect, Stefanowitsch & Gries (2008) explored the relation between register and collostructions.

In this contribution, we will focus on the lectal conditioning of lexical collocations. First, we will analyze how register and national variety modify the distributional properties of AN collocations in Dutch. Next, we will analyze how those lectal variables alter the impact of lexical collocations on the alternation between two inflectional variants of the adjective in Dutch definite NPs with a singular neuter head noun. In this NP construction, the adjective displays an alternation between the standard inflected form (1) and its marked uninflected counterpart (2):

(1) het vriendelijk-e kind
the friendly-INFL child
(2) het vriendelijk-ø kind
the friendly-ZERO child

Within the intricate network of variables governing this alternation, the lexical collocation strength of the AN pair exerts a major impact on the inflectional realization of the adjective, the use of the uninflected alternative being favored in AN collocations (Tummers 2005). Furthermore, the lectal variables hypothesized to modify the impact of lexical collocations on the adjectival inflection both have a significant effect on the choice of the inflectional alternative, the use of the uninflected adjective being favored by Belgian Dutch as well as informal registers in Belgian Dutch and (highly) formal registers in Netherlandic Dutch.

The following research questions will be addressed to disentangle the relation between lexical collocation strength on the one hand and the lectal variables on the other hand:

1. To what extent is the distribution of AN collocations in Dutch modified by register and national variety?
2. To what extent is the impact of AN collocations on the selection of the adjectival alternative in Dutch altered by register and national variety?

The answers to those questions will shed light on the relation between collocation strength on the one hand and the lectal variables on the other. Is there a consecutive relationship between both, do they both act independently or do they act in mutual interaction?

2 Results and discussion

A database of 4,964 definite NPs with a singular neuter head noun (3,810 inflected and 1,154 uninflected adjectives) was extracted from the Corpus of Spoken Dutch (Oostdijk 2000). That repository of spoken Dutch contains data from Belgian and Netherlandic Dutch, the two national varieties, and various registers ranging from highly informal (colloquial speech) to highly formal (prepared speeches in parliament). The lexical collocation strength between A and N lemmas was computed using the log likelihood ratio, G^2 (Dunning 1993).

In answer to research question 1, figure 1 visualizes the G^2-distributions in the four different registers grouped by national variety, showing differences induced by both register and national variety. Moreover, the distribution shows a strong positive skew, yielding a lot of outliers which are not all included in the boxplots (range(G^2) = [0.00;1782.99]).

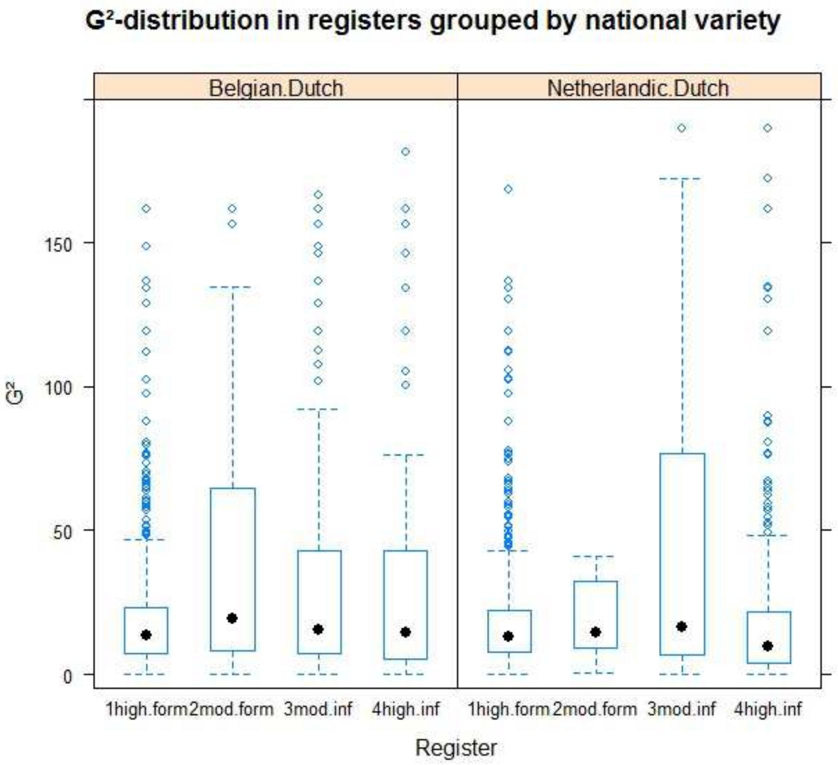

Figure 1: G^2-distribution over AN pairs in registers grouped by national variety

To model the impact of both lectal variables on the lexical collocation strength, viz. G^2, a gamma GLM has been fitted, G^2 displaying a Chi^2-distribution which in turn is a special case of the gamma distribution (Forbes et al. 2011). Table 1 presents the regression coefficients, both lectal variables (`nat.var`, `register`) being dummy coded.

Variable	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.041447	0.003003	13.804	< 2e-16 ***
nat.var=bel	-0.007020	0.003681	-1.907	0.056541 .
register=mod.form	0.009085	0.032179	0.282	0.777704
register=mod.inf	-0.032765	0.003203	-10.231	< 2e-16 ***
register=high.inf	-0.017535	0.003708	-4.729	2.32e-06 ***
nat.var=bel:register=mod.form	-0.028925	0.032297	-0.896	0.370511
nat.var=bel:register=mod.inf	0.016507	0.004426	3.730	0.000194 ***
nat.var=bel:register=high.inf	-0.001795	0.004793	-0.375	0.708046

Table 1: Gamma GLM modeling impact of national variety and register on G^2

Although no significant main effect of the national variety (`nat.var=bel`) is found, there is a significant interaction between register and national variety indicating a different stylistic conditioning of AN collocation patterns in both national varieties of Dutch.

To deal with research question 2, a logistic regression analysis has been performed (`rms` library in R, Harrell 2001) with $\ln(^{P(A.uninflected)}/_{1-P(A.uninflected)})$ as response variable and G^2 (`llr`), national variety (`nat.var`, dummy coding) and register (`register`, dummy coding) as explanatory variables (model statistics: likelihood ratio $Chi^2 = 648.11$, df = 15, $p < 0.0001$, C = 0.732). The regression coefficients (table 2) show an adjustment of the impact of the lexical collocation strength on the inflectional alternation by both lectal variables and their interaction.

Variable	**Coef**	**S.E.**	**Wald Z**	**Pr(>|Z|)**
Intercept	-2.3088	0.1113	-20.75	<0.0001
llr	0.0144	0.0020	7.25	<0.0001
nat.var=bel	0.5516	0.1352	4.08	<0.0001
register=mod.form	2.1823	1.0109	2.16	0.0309
register=mod.inf	1.3902	0.1699	8.18	<0.0001
register=high.inf	0.2149	0.1680	1.28	0.2008
llr:nat.var=bel	-0.0075	0.0023	-3.29	0.0010
llr:register=mod.form	-0.0642	0.0508	-1.26	0.2069
llr:register=mod.inf	-0.0101	0.0021	-4.74	<0.0001
llr:register=high.inf	-0.0033	0.0026	-1.26	0.2059
nat.var=bel:register=mod.form	-1.4334	1.0215	-1.40	0.1605
nat.var=bel:register=mod.inf	-0.0682	0.2170	-0.31	0.7532
nat.var=bel:register=high.inf	1.2169	0.2277	5.35	<0.0001
llr:nat.var=bel:register=mod.form	0.0627	0.0509	1.23	0.2178
llr:nat.var=bel:register=mod.inf	0.0054	0.0026	2.11	0.0350
llr:nat.var=bel:register=high.inf	-0.0001	0.0032	-0.02	0.9811

Table 2: Logistic regression modeling the impact of G^2, national variety and register on inflectional alternation attributive adjective

First, the impact of the collocation strength on the selection of the uninflected adjective is significantly lower in Belgian than in Netherlandic Dutch (reference value). Next, the effect of the collocation strength on the selection of the uninflected adjective in the moderately informal register (`llr:register=mod.inf`) is significantly lower than for the most formal register (reference value) and the other registers. Finally, the propensity of AN collocations to select the uninflected adjective in the moderately informal register, as compared to the most formal register, is significantly higher in Belgian than in Netherlandic Dutch, as can be inferred from the significant triple interaction (`llr:nat.var=bel:register=mod.inf`).

3 Conclusion

In sum, lexical collocation strength and lectal sensitivity operate in mutual interaction. First, the lexical collocation strength in AN pairs is subject to lectal adjustments. Second, the selection criteria of the adjectival alternatives use lexical collocation strength in a different way depending on the lectal settings, as national variety, register and their interaction significantly constrain the effect of lexical collocation strength on the inflectional variation. Hence, we argue that a comprehensive usage-based language model needs to include a lectal dimension. In this respect, we refer to Cognitive Linguistics, where the recognition of the importance of lectal constraints on language use resulted in Cognitive sociolinguistics (Geeraerts et al. 2010).

References

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19 (1), 61-74.

Forbes, C., M. Evans, N. Hastings, and B. Peacock. 2011. *Statistical Distributions*. Hoboken: John Wiley & Sons.

Geeraerts, D. 2005. Lectal variation and empirical data in Cognitive Linguistics. In: F. J. Ruiz de Mendoza Ibáñez, and M. S. Peña Cervel (eds.), *Cognitive Linguistics and Interdisciplinary Dynamics*, 163-190. Berlin: Mouton de Gruyter.

Geeraerts, D., G. Kristiansen, and Y. Peirsman (eds). 2010. *Advances in cognitive sociolinguistics*. Berlin: Walter de Gruyter.

Gries, S. Th. 2008. Phraseology and linguistic theory: a brief survey. In: S. Granger, and F. Meunier (eds.), *Phraseology: An interdisciplinary perspective*, 3-25. Amsterdam: John Benjamins.

Harrell, F. E. 2001. *Regression Modeling Strategies, with Applications to Linear Models, Survival Analysis and Logistic Regression*. New York: Springer.

Oostdijk, N. 2000. Het Corpus Gesproken Nederlands. *Nederlandse Taalkunde*, 5 (3), 280-284.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Speelman, D., J. Tummers, and D. Geeraerts. 2009. Lexical patterning in a construction grammar. The effect of lexical co-occurrence patterns on the inflectional variation in Dutch attributive adjectives. *Constructions and Frames*, 1 (1), 87-118.

Stefanowitsch, A., and S. Th.Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2). 209-243.

Stefanowitsch, A., and S. Th. Gries. 2008. Channel and constructional meaning: A collostructional case study. In: R. Dirven, and G. Kristiansen (eds.), *Cognitive Sociolinguistics: Language variation, Cultural Models, Social Systems*, 129-152. Berlin: Mouton de Gruyter.

Tummers, J. 2005. *Het naakte adjectief. Kwantitatief-empirisch onderzoek naar de adjectivische buigingsalternantie bij neutra*. PhD dissertation, KULeuven, Faculty of Arts.

Wulff, S. 2008. *Rethinking Idiomaticity: A Usage-based Approach*. London/New York: Continuum Press.