Distribution of Generalized Quantifiers in Large Corpora*

Camilo Thorne
Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart
Pfaffenwaldring 5b, D-70569, Stuttgart, Germany
camilo.thorne@ims.uni-stuttgart.de

Abstract

We describe the results of a large corpus study in which we studied if quantifier distribution (i.e., frequency of quantifier expressions) is influenced by the notion of semantic complexity as defined by generalized quantifier theory –viz., the data complexity and the expressive power of the relations (or higher order functions) that model their meaning. Regression analysis shows that complexity has an statistically significant impact.

Quantifiers in natural languages such as English are multiword expressions, such as "all", "more than", "less than two-thirds" or "most", used to express notions of quantity and number[1].

Generalized quantifier theory, first proposed by Barwise and Cooper (1980) provides a mathematically elegant description of quantifier semantics grounded in (higher order) logic, in the form of relations, constraints or higher order functions. A generalized quantifier Q can be defined as a binary relation $Q(A,B)$ over sets A and B. For instance in the sentence "all men are equal", the determiner "all" states that the set of men is contained in the set of equal beings; More in general, "all" denotes the relation $all = \{A,B \subseteq \Delta \mid A \subseteq B\}$, where Δ is the so-called domain of discourse, or, equivalently, a second order function $\lambda A.\lambda B.\forall x(A(x) \Rightarrow B(x))$.

A question regarding quantifiers that has gained interest lately both in cognitive science and cognitive linguistics is the *complexity* of quantification[2]. Generalized quantifier theory attempts to answer this question via the computational data complexity and the expressiveness of the logical expressions underpinning English quantifiers, a.k.a. *semantic complexity* (Szymanik, 2009). This yields a classification of quantifiers into tractable (**PTime** data complexity) and intractable (**NP**-hard data complexity), and within tractable quantifiers, into so-called Aristotelian, counting and proportional quantifiers (Szymanik and Zajenkowski, 2010).

While several authors have tried to empirically test semantic complexity theory using cognitive experiments, this contribution pursues a different path by considering *large corpora analysis* (Gries, 2010). By observing patterns of multiword expressions denoting Aristotelian, counting and proportional quantifiers over a corpus of close to a billion English words –the WaCky

*Work to be presented at the *QUAD: QUantifiers And Determiners* workshop at the 2017 ESSLLI Summer School, Toulouse, France, between 17.7.2017 and 21.7.2017.

[1]The results in this abstract have been published in a special issue of Language Sciences as joint work with J. Szymanik (Szymanik and Thorne, 2017)

[2]One example is the picture verification task, in which people are asked to verify sentences with quantifiers against pictures with balls of different colors, and both their accuracy and response time (Szymanik and Zajenkowski, 2011; MacLeod et al., 1978).

Table 1: Right: Comparison of regression models. While all models significantly improve on the baseline (Poisson mixed model), the negative binomial model with *fixed* effects shots the best AIC fit score (highlighted in gray). Left: analysis of frequency deviance for the best model (BINOM). Class, length and monotonicity of the quantifier's right argument have an statistically (very strongly) significant impact on frequency. Class (complexity) explains more than one quarter of deviance.

Feature	Deviance	p-value
Length (words)	47.06%	$3.47 \cdot e^{-10}$
Class	27.29%	$5.25 \cdot e^{-7}$
Type	0.02%	0.97
Right mon.	25.65%	$1.15 \cdot e^{-6}$

Model	AIC	p-value
POISSON-MIXED	1446287.3	(baseline)
POISSON	1446000.0	$1.191 \cdot e^{-10}$
BINOM-MIXED	426.7	$< 2.2 \cdot e^{-16}$
BINOM	409.3	$< 2.2 \cdot e^{-16}$

corpus, a semi-automatically curated English corpus extracted from a 2008 Wikipedia dump by Baroni et al. (2009)–, we:

1. Check if their (frequency) distribution is skewed towards low complexity –Aristotelian– quantifiers.

2. Investigate the impact of semantic complexity on frequency vis-à-vis other factors (monotonicity, length in number of words, type –superlative, comparative–).

The main analysis technique we use are **(a)** generalized linear mixed regression models and **(b)** an analysis of deviance[3], appropriate for count and frequency data (Dobson and Barnett, 2008).

Table 1 summarizes our results. The best –negative binomial– model (a geometrically decreasing model) indicates a bias towards low complexity quantifiers. As the reader can see, length (in number of words) explains by itself 47.06% of the deviance, followed by quantifier class (27.29%) and (right) monotonicity (25.65%), and that together account for more than 52% of deviance. In all three cases, such influence is also statistically significant. The impact of length is likely due to the fact that the likelihood of a multiword expression decreases the more tokens it spans (Baroni, 2009). The impact of length (and monotonicity) on the other hand indicate that when we focus on quantifier expression of similar length, semantically simpler quantifiers outnumber more complex ones.

[3]Deviance is a measure analog to variance for linear models; the main difference is that rather than comparing the distribution of standard errors across groups, we compare the goodness of fit of the *reduced* GLMs induced by the groups (Dobson and Barnett, 2008).

References

Baroni, M. (2009). 39 distributions in text. In *Corpus linguistics: An International Handbook*, Volume 2, pp. 803–821. Mouton de Gruyter.

Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta (2009). The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation 43*(3), 209–226.

Barwise, J. and R. Cooper (1980). Generalized quantifiers and natural language. *Linguistics and Philosophy 4*(2), 159–219.

Dobson, A. J. and A. G. Barnett (2008). *An Introduction to Generalized Linear Models*. CRC Press.

Gries, S. T. (2010). Useful statistics for corpus linguistics. In A. Sánchez and M. Almela (Eds.), *A mosaic of corpus linguistics: selected approaches*, pp. 269–291. Peter Lang.

MacLeod, C. M., E. B. Hunt, and N. N. Mathews (1978). Individual differences in the verification of sentence-picture relationships. *Journal of verbal learning and verbal behavior 17*(5), 493–507.

Szymanik, J. (2009). *Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language*. Ph. D. thesis, University of Amsterdam, Amsterdam.

Szymanik, J. and C. Thorne (2017). Exploring the relation between semantic complexity and quantifier distribution in large corpora. *Language Sciences 60*, 80–93.

Szymanik, J. and M. Zajenkowski (2010). Comprehension of simple quantifiers. Empirical evaluation of a computational model. *Cognitive Science: A Multidisciplinary Journal 34*(3), 521–532.

Szymanik, J. and M. Zajenkowski (2011). Contribution of working memory in parity and proportional judgments. *Belgian Journal of Linguistics 25*(1), 176–194.