Language Sciences 60 (2017) 80–93

Contents lists available at ScienceDirect

Language Sciences

journal homepage: www.elsevier.com/locate/langsci

Exploring the relation between semantic complexity and quantifier distribution in large corpora

Jakub Szymanik [a,*], Camilo Thorne [b]

[a] Institute for Logic, Language and Computation, University of Amsterdam, P.O. Box 94242, 1090 GE Amsterdam, The Netherlands
[b] Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Pfaffenwaldring 5b, D-70569 Stuttgart, Germany

ARTICLE INFO

Article history:
Available online 7 March 2017

Keywords:
Generalized quantifiers
Semantic complexity
Corpus analysis
Generalized linear regression models
Analysis of deviance

ABSTRACT

In this paper we study if semantic complexity can influence the distribution of generalized quantifiers in a large English corpus derived from Wikipedia. We consider the minimal computational device recognizing a generalized quantifier as the core measure of its semantic complexity. We regard quantifiers that belong to three increasingly more complex classes: Aristotelian (recognizable by 2-state acyclic finite automata), counting ($k+2$-state finite automata), and proportional quantifiers (pushdown automata). Using regression analysis we show that semantic complexity is a statistically significant factor explaining 27.29% of frequency variation. We compare this impact to that of other known sources of complexity, both semantic (quantifier monotonicity and the comparative/superlative distinction) and superficial (e.g., the length of quantifier surface forms). In general, we observe that the more complex a quantifier, the less frequent it is.

1. Introduction

Linguists and philosophers have been searching for various ways to estimate the complexity and expressiveness of natural language. One important debate pivots around the Equivalent Complexity Thesis (see Miestamo et al., 2008); that is, the question whether all languages of the world are equally complex or can express equally complex concepts. It is not surprising that such questions can sparkle lively discussion, after all, a proper answer would involve integrating many aspects of linguistics, e.g., grammatical complexity, cognitive difficulty, cultural diversity, etc. As Sampson et al. (2009) puts it:

> Linguists and non-linguists alike agree in seeing human language as the clearest mirror we have of the activities of the human mind, and as a specially important of human culture, because it underpins most of the other components. Thus, if there is serious disagreement about whether language complexity is a universal constant or an evolving variable, that is surely a question which merits careful scrutiny. There cannot be many current topics of academic debate which have greater general human importance than this one.

These endeavors are usually driven by different (but often related) questions: What are the semantic bounds of natural languages or, in other words, what is the conceptual expressiveness of natural language (see, e.g., Szymanik, 2016)? What is the 'natural class of concepts' expressible in a given language and how to delimit it (see, e.g., Barwise and Cooper, 1981; Piantadosi,

* Corresponding author.
E-mail addresses: J.K.Szymanik@uva.nl (J. Szymanik), camilo.thorne@ims.uni-stuttgart.de (C. Thorne).

http://dx.doi.org/10.1016/j.langsci.2017.01.006
0388-0001/

2011)? Are there differences between various languages with respect to semantic complexity (see, e.g., Everett, 2005)? Or from a more methodological perspective: how powerful must be our linguistic theories in order to minimally describe semantic phenomena (see, e.g., Ristad, 1993)? A similar question can be also asked from a cognitive angle: are some natural language concepts harder to process for humans than others (see, e.g., Feldman, 2000; Szymanik and Zajenkowski, 2010)?

The outcomes of such debates heavily depend on the underlying operationalization of the complexity notion, hence, we propose a measure of semantic complexity (see Szymanik, 2016). It focuses on the meaning of the quantifiers abstracting away from many grammatical details as opposed to, for example, typological (cf. McWhorter, 2001) or information-theoretic approaches (cf. Juola, 1998) known from the literature. The goal of this paper is to give a proof-of-concept that such an abstract notion of semantic complexity can be used (together with other linguistic factors) to explain or predict the distribution of quantifiers in natural language textual data.

In order to contribute to the above outlined debate we focus on one aspect of natural language: its ability to express quantities by using the wide repertoire of quantifier expressions, like 'most', 'at least five', or 'all' (see, e.g., Keenan and Paperno, 2012). We restrict ourselves to study generalized quantifiers (GQs) as described by Barwise and Cooper (1981), plus some of the counting and proportional quantifier forms[1] from Szymanik (2016), see Table 1. We identify their main surface forms or lexical variants with high precision, rather than trying to cover all—a considerable challenge given the size of our corpora and thus beyond the scope of this paper. In general, we observe the following:

Table 1
Quantifiers and their semantic complexity. Note: we assume *few* to be the dual of *most*, see also footnote 1. Also, we distinguish between $>1/2$, $<1/2$ and other proportional quantifiers given that they constitute, arguably, the most common such quantifiers.

Class	Examples	Quantifier	Complexity
Aristotelian	'every', 'some'	All, some	2-State acyclic FA
Counting	'more than 4', 'at most 5'	>k, <k	$k+2$-State FA
Proportional	'most', 'less than half'	Most, <1/2,	
	'few', 'more than half'	few, >1/2,	PDA
	'less than three-fifths'	<p/k,	
	'more than two-thirds'	>p/k	

(H) There is a relation between GQ distribution and semantic complexity; more precisely, the more complex a GQ, the less frequent it tends to be.

In order to test **(H)** we leverage on *multiple factor regression models* to reasonably quantify the predictive value of all such factors vis-à-vis quantifier frequency (cf. Gries, 2010).

Observation **(H)** is consistent with *the principle of least effort in communication*: speakers tend to minimize the communication effort by generating so-called "simple" messages. We take this result as an argument in favor of the claim, for instance defended by Szymanik (2016), that abstract semantic complexity measures may enrich the methodological toolbox of the language complexity debate.

We also considered whether semantic complexity can be clearly distinguished from syntactic or surface-form complexity. Clearly, semantics is not the only potential source of complexity in language (cf. Miestamo et al., 2008): surface-form length, syntax (e.g., nesting levels of subordinated clauses or parse tree depth), morphology (e.g., complex word inflection and derivation), monotonicity, and such, also all play a role (cf. Castello, 2008).

2. Semantic complexity of quantifiers

2.1. Quantifiers

What are the numerical expressions (generalized quantifiers, GQs) we are going to talk about? Intuitively, on the semantic level, quantifiers are expressions that appear to be descriptions of quantity, e.g., 'all', 'not quite all', 'nearly all', 'an awful lot', 'a lot', 'a comfortable majority', 'most', 'many', 'more than k', 'less than k', 'quite a few', 'quite a lot', 'several', 'not a lot', 'not many', 'only a few', 'few', 'a few', 'hardly any', 'one', 'two', 'three', etc. To concisely capture the semantics (meaning) of the quantifiers we should consider them in the sentential context, for instance:

(1) More than seven students are smart.
(2) Fewer than eight students received good marks.
(3) More than half of the students danced nude on the table.
(4) Fewer than half of the students saw a ghost.

[1] Thus, we ignore quantifiers as 'not all' or the distributive reading of 'each'.