Semantic Complexity of Quantifiers and their Distribution in Corpora

Camilo Thorne
IBM CAS Trento - Trento RISE
Povo di Trento, Italy
c.thorne.email@trentorise.eu

Jakub Szymanik
Institute for Logic, Language, and Computation
Amsterdam, The Netherlands
j.k.szymanik@uva.nl

Abstract

The semantic complexity of a quantifier can be defined as the computational complexity of the finite model checking problem induced by its semantics. This paper describes a preliminary study to understand if quantifier distribution in corpora can be to some extent predicted or explained by semantic complexity. We show that corpora distributions for English are significantly skewed towards quantifiers of low complexity and that this bias can be described in some cases by a power law.

1 Introduction

Quantification is an essential feature of natural languages. It is used to specify the (vague) number or quantity of objects satisfying a certain property. Quantifier expressions are built from *noun phrases* (whether definite or indefinite, names or pronouns) and *determiners* resulting in expressions such as "a subject", "more than half of the men", "the queen of England", "John", "some", "five" or "every" (see Peters and Westerståhl (2006) for an overview).

More recently, interest has arisen regarding *semantic complexity*, that is, the complexity of reasoning with (and understanding) fragments of natural language. One model that has been proposed to study natural language semantic complexity is to consider the computational properties that arise from formal semantic analysis, see e.g., Ristad (1993); van Benthem (1987); Kontinen and Szymanik (2008). One could wonder whether speakers (due to their restricted cognitive resources) are naturally biased towards low complexity expressions, see Szymanik and Zajenkowski (2010); Schlotterbeck and Bott (2013). Additionally, related work by Thorne (2012) shows that, when one considers *the satisfiability problem* of specific fragments of English then computationally tractable combinations of constructs occur more frequently than intractable ones.

This paper extends such work by showing that: *(i)* quantifiers can be ranked w.r.t. to their semantic complexity, viz., their computational complexity w.r.t. the *model-checking* problem, and their expressiveness; *(ii)* within a selected set of corpora quantifier distribution is skewed towards computationally easier quantifiers; and *(iii)* such distribution describes a power law.

2 Generalized Quantifiers and Semantic Complexity

Generalized Quantifiers. Generalized quantifiers are usually taken to denote relations holding between subsets of the universe. For instance, in a given model $\mathcal{I} = (\mathbb{D}_\mathcal{I}, \cdot^\mathcal{I})$ the statement "most As are B" says: $\#(A^\mathcal{I} \cap B^\mathcal{I}) > \#(A^\mathcal{I} \setminus B^\mathcal{I})$, where $A^\mathcal{I}, B^\mathcal{I} \subseteq \mathbb{D}_\mathcal{I}$ and $\#(A)$ stands for the cardinality of set

Proceedings of the 11th International Conference on Computational Semantics, pages 64–69,
London, UK, April 15-17 2015.

Table 1: Top: Base FO (Aristotelian and counting) and proportional generalized quantifiers studied in this paper, ranked by semantic complexity; $> k$ and $< k$ comprise by abuse the superlative quantifiers "at least k" and "at most k". Bottom, left: Sample English sentences realizing Ramsey quantifiers; notice the use of the reciprocal "each other". Bottom, right: Semantic complexity of Ramsey quantifiers by quantifier class.

Q	**Model Class**	**S. C.**		**Example**
some	$\{\mathcal{I} \mid A^{\mathcal{I}} \cap B^{\mathcal{I}} \neq \emptyset\}$	$\mathbf{AC}^0$	*ari*	some men are happy
all	$\{\mathcal{I} \mid A^{\mathcal{I}} \subseteq B^{\mathcal{I}}\}$	$\mathbf{AC}^0$	*ari*	all humans are mammals
the	$\{\mathcal{I} \mid \#(A^{\mathcal{I}} \cap B^{\mathcal{I}}) = 1\}$	$\mathbf{AC}^0$	*cnt*	the third emperor of Rome was deranged
$> k$	$\{\mathcal{I} \mid \#(A^{\mathcal{I}} \cap B^{\mathcal{I}}) > k\}$	$\mathbf{AC}^0$	*cnt*	more than 5 men are happy
$< k$	$\{\mathcal{I} \mid \#(A^{\mathcal{I}} \cap B^{\mathcal{I}}) < k\}$	$\mathbf{AC}^0$	*cnt*	fewer than 100 violins are Stradivari
k	$\{\mathcal{I} \mid \#(A^{\mathcal{I}} \cap B^{\mathcal{I}}) = k\}$	$\mathbf{AC}^0$	*cnt*	50 MPs voted against the war in Irak
most	$\{\mathcal{I} \mid \#(A^{\mathcal{I}} \cap B^{\mathcal{I}}) > \#(A^{\mathcal{I}} \setminus B^{\mathcal{I}})\}$	$\mathbf{P}$	*pro*	most trains are safe
few	$\{\mathcal{I} \mid \#(A^{\mathcal{I}} \cap B^{\mathcal{I}}) < \#(A^{\mathcal{I}} \setminus B^{\mathcal{I}})\}$	$\mathbf{P}$	*pro*	few people are trustworthy
$> p/k$	$\{\mathcal{I} \mid \#(A^{\mathcal{I}} \cap B^{\mathcal{I}}) > p \cdot (\#(A)/k)\}$	$\mathbf{P}$	*pro*	more than $2/3$ of planets are lifeless
$< p/k$	$\{\mathcal{I} \mid \#(A^{\mathcal{I}} \cap B^{\mathcal{I}}) < p \cdot (\#(A)/k)\}$	$\mathbf{P}$	*pro*	less than $1/3$ of Americans are rich
p/k	$\{\mathcal{I} \mid \#(A^{\mathcal{I}} \cap B^{\mathcal{I}}) = p \cdot (\#(A)/k)\}$	$\mathbf{P}$	*pro*	$1/3$ of Peru's population lives in Lima
$> k\%$	$\{\mathcal{I} \mid \#(A^{\mathcal{I}} \cap B^{\mathcal{I}}) > k \cdot (\#(A)/100)\}$	$\mathbf{P}$	*pro*	more than 10% of Peruvians are poor
$< k\%$	$\{\mathcal{I} \mid \#(A^{\mathcal{I}} \cap B^{\mathcal{I}}) < k \cdot (\#(A)/100)\}$	$\mathbf{P}$	*pro*	less than 5% of the Earth is water
$k\%$	$\{\mathcal{I} \mid \#(A^{\mathcal{I}} \cap B^{\mathcal{I}}) = k \cdot (\#(A)/100)\}$	$\mathbf{P}$	*pro*	15% of Muslims are Shia

R_Q	**Example**
R_{some}	some children like each other
$R_{>p/k}$	more than $2/3$ of female MPs sit next to each other
R_{most}	most people help each other
$R_{>k}$	at least 2 men married each other in the UK last year

Quantifier Class R_Q	**S.C.**
Aristotelian (*ari*+*recip*)	$\mathbf{AC}^0$
counting (*cnt*+*recip*)	$\mathbf{AC}^0$
proportional (*pro*+*recip*)	**NP**-complete

A. Going a step further, we can take a generalized quantifier Q to be a functional relation associating with each model $\mathcal{I}$ a relation between relations on its universe, $\mathbb{D}_{\mathcal{I}}$. This is actually equivalent to their standard Lindström (1966) model-theoretic definition as classes of models:

Definition 2.1 (Generalized Quantifier)**.** Let $t = (n_1, \ldots, n_k)$ be a k-tuple of positive integers. A *generalized quantifier* of type t is a class Q of models of a vocabulary $\tau_t = \{R_1, \ldots, R_k\}$, such that R_i is n_i-ary for $1 \leq i \leq k$, and Q is closed under isomorphisms, i.e. if $\mathcal{I} \in Q$ and $\mathcal{I}'$ is isomorphic to $\mathcal{I}$, then $\mathcal{I}' \in Q$. It gives rise to a *query* $\overline{Q}(R_1, \ldots, R_n)$ such that $\mathcal{I} \models \overline{Q}(R_1, \ldots, R_n)$ iff $\mathcal{I} \in Q$.

Semantic Complexity. An important consequence of this definition is the notion of *semantic complexity*, which refers to the computational complexity[1] of the finite model checking problem that it induces, namely the question: does $\mathcal{I} \models \overline{Q}(R_1, \ldots, R_n)$? When considering such problem we are interested in its complexity w.r.t. the size of the model $\mathcal{I}$, that is, in *data complexity*, see Immerman (1998). Semantic complexity induces a partial ordering (ranking) of quantifiers. Furthermore, it induces a partition into *tractable* and *intractable* generalized quantifiers. Respectively: quantifiers, for which model checking is *at most* **P**, and quantifiers for which model checking is *at least* **NP**-hard.

Tractable Quantifiers. Tractable quantifiers come in two flavors: first-order (FO) and proportional:

1. **FO.** FO quantifiers Q of type t over $\tau_t = \{R_1, \ldots, R_n\}$, are quantifiers which give rise to FO queries (with identity). They are those with the lowest semantic complexity, viz. $\mathbf{AC}^0$ (the data complexity of model checking in FO with identity is in $\mathbf{AC}^0$, see Immerman (1998)). See Table 1, top. They are typically split in the literature into:

[1] Please refer to Papadimitrou (1994) for the basics of computational complexity.

- *Aristotelian* quantifiers (*some*, *all*), which are the quantifiers dealt with in traditional syllogistic logic.
- *Counting* quantifiers (*the*, $< k$, $> k$, k), in which the number of individuals in the domain verifying a given property is specified. Counting quantifiers while sharing the same semantic complexity of Aristotelian quantifiers, are nevertheless more expressive and can be distinguished via their associated language problem (van Benthem (1987)): given a model $\mathcal{I}$ and query $\overline{Q}(R_1, \ldots, R_1)$ one can construct a finite state automaton $A_\mathcal{I}$ for $\mathcal{I}$ and a word w_Q over the alphabet $\{0, 1\}$ s.t. $\mathcal{I} \models \overline{Q}(R_1, \ldots, R_1)$ iff $w_Q \in L(A_\mathcal{I})$ where $L(A_\mathcal{I})$ denotes the language recognized by $A_\mathcal{I}$. The automaton $A_\mathcal{I}$ will have 2 states whenever Q is Aristotelian, and at most $k + 2$ states whenever Q is a counting quantifier[2].

2. **Proportional.** More interesting are *proportional* quantifiers, e.g., *most* ("most men") and $> p/k$ ("more than one third of men"). They are used often by speakers when referring to collections (the denotation of collective or plural nouns) and their quantitative properties. Proportional quantifiers are strictly more expressive than Aristotelian or counting quantifiers. Indeed, as Barwise and Cooper (1980) showed, they are not FO expressible. This is reflected by their higher semantic complexity, in **P** (Szymanik, 2010).

Intractable Quantifiers. Intractable quantifiers can be derived from tractable ones via various model-theoretic operations. One such operation is *Ramseyfication*, that turns a monadic quantifier of type $(1, 1)$ into a polyadic quantifier of type $(1, 2)$. Ramseyfication is expressed by the reciprocal expression "each other" under the default (strong) interpretation of Dalrymple et al. (1998). It intuitively states that the models of the resulting Ramseyfied quantifiers are graphs with connected components. Intractability arises when it is applied to proportional quantifiers, giving rise to so-called "clique" quantifiers (Szymanik, 2010). See Table 1, bottom.

3 Pattern-based Corpus Analysis

Semantic complexity and expressiveness can be leveraged to induce both a partition and a ranking of quantifiers[3], where Aristotelian quantifiers occupy the lowest and Ramseyfied proportional quantifiers the highest end of the spectrum. Such theoretical results are reflected by their distribution in (English) corpora.

Quantifier Patterns. We identified generalized quantifiers indirectly, via part-of-speech (POS) patterns that reasonably approximate their surface forms and lexical variants. The POS tags were required to filter out contexts in which quantifier words do not express a quantifier such as "no" in "you cannot say no" (an interjection) –as opposed to "no" in "no tickets were left" (a determiner). Each such pattern defined a quantifier *type*. This done, we counted the number of times each type is instantiated within a sentence in the corpus, that is, its number of *tokens* (lexical variants). We considered Penn Treebank/Brown corpus POSs (Francis and Kucera, 1979)[4]. We present two such patterns (the others were defined analogously):

1. to identify the Aristotelian quantifier *all*, we considered its lexical variants "all", "everybody", "everything", "every", "each", "everyone" and "the N" (where N is a plural noun), and built the regex: *.*(every/at | Every/at | all/abn | All/abn | the/at .*/nns | The/at ./nns | everything/pn | Everything/pn | everyone/pn | Everyone/pn | everybody/pn | Everybody/pn | each/dt | Each/dt).**;
2. to identify Ramsey quantifiers, we checked for sentences that match *at the same time* the regular expressions of the base (FO, counting or proportional) quantifiers and the following pattern for the reciprocal: *.* each/dt other/ap .**.

Using such patterns we observed the frequency of *(i)* generalized quantifiers, and *(ii)* tractable and intractable Ramsey quantifiers, to see whether such distribution was skewed towards low complexity quan-

[2] More in general, the class REG of regular languages corresponds to the class of quantifiers definable in so-called divisibility logic, see Mostowski (1998).

[3] Note that $\mathbf{AC}^0 \subseteq \mathbf{P} \subseteq \mathbf{NP}$-complete and $\text{REG}_2 \subseteq \text{REG}_{\leq k+2}$.

[4] For the POS tagging, we relied on the NLTK 3-gram tagger, see `http://www.nltk.org/`.

tifiers in the former case, and towards tractable quantifiers in the latter case.

Corpora. To obtain sufficiently large, balanced and representative samples of contemporary English, we considered two corpora covering multiple domains and sentence types (declarative and interrogative). Specifically, we considered the well-known Brown corpus by Francis and Kucera (1979) ($\sim$ 60,647 sentences and 1,014,312 word tokens). We also considered a sizeable sample of a large web corpus, the ukWaC corpus ($\sim$ 280,001 sentences and 100,000,000 word tokens) from Baroni et al. (2009), built from a 2006-2007 crawl of the (complete) .uk domain.

Power Laws. Power laws, first discussed by Zipf (1935), relate the frequency of linguistic tokens to their rank, viz., to the ordering induced by their frequency. They typically predict that frequency is proportional to rank (modulo two real-valued parameters a and b^{-1}), giving rise to non-normal, skewed distributions where the topmost (w.r.t. rank) 20% words in a corpus concentrate around 80% of the probability mass or frequency. They are widespread in natural language data (Baroni, 2009). More recent work (Newman, 2005) has shown that power laws can be variously modified and extended to cover wider spectra of natural language phenomena and rankings. One such possible extension is to consider, as we do in this paper, power laws relating the frequency of a quantifier Q to its semantic complexity rank:

Definition 3.1 (Complexity Power Law)**.** The power law between *quantifier frequency* $fr(Q)$ and *quantifier complexity rank* $rk(Q)$ is described by the equation: $fr(Q) = a/rk(Q)^b$, with $a, b \in \mathbb{R}$.

To approximate the distribution parameters a and b we ran (Newman, 2005) a least squares linear regression, since power laws are equivalent to linear models on the log-log scale[5]. We measured also the ensuing R^2 coefficient, that captures how well the observations fit the inferred power law equation and that ranges from 0 (no fit) to 1 (perfect fit). To validate further our models we ran a χ^2 test (at $p = 0.01$ significance) w.r.t. the uniform distribution as our null hypothesis.

Results and Interpretation. The distributions observed are summarized by Figure 1. The top left corner describes the contingency (raw frequency) tables used for the figures. The top right figures describe relative frequency by quantifier class. The bottom figures, by quantifier: the reader will find to the left the average and cumulative relative frequency plots, and to the right the log-log (power law) regressions.

As expected by the theory, Aristotelian and counting —$\mathbf{AC}^0$— quantifiers occur more frequently than proportional —**P**— quantifiers, and (proportional) Ramsey —**NP**-hard— quantifiers. See Figure 1 (top right). This bias is statistically significant: their distribution significantly differs from uniform or random distributions, as $p < 0.01$.

Furthermore, when we consider separately the distribution of, on the one hand, base quantifiers, and on the other hand, Ramseyfied quantifiers, we can infer power laws skewed towards Aristotelian quantifiers and bounded Ramsey quantifiers: see Figure 1 (bottom). In both cases a high goodness-of-fit coefficient was obtained: $R^2 = 0.94$ for base quantifiers and $R^2 = 0.92$ for Ramseyfied quantifiers (mean distribution).

Finally Figure 1 (top left) shows that tractable quantifiers occur significantly more often than intractable quantifiers. Furthermore, the same observation applies to FO and proportional quantifiers vis-à-vis their Ramseyfications. Actually, Ramseyfications, whether tractable or intractable, appear to be in general rare ("sparse") in natural language data. We conjecture that this is due to an increase in expressiveness and complexity relatively to proportional and FO quantifiers, that cannot be easily captured through the techniques used to build Table 1, and merits further investigation.

The method described was relatively noisy (the POS tagger had an accuracy of around 80%) and not fully exhaustive (the patterns did not cover all quantifier surface forms). However, we believe that our datasets were large and representative enough, and our rule patterns adequate enough to derive reasonable approximations to the distribution of quantifiers in English.

[5] I.e., $y = a/x^b$ iff $\log(y) = \log(a/x^b) = a - b \cdot \log(x)$.

	pro+ recip	*cnt+ recip*	*ari+ recip*
Brown	4	17	186
ukwack	5,732	10,484	33,107
total	5,736	10,501	33,293

pro	*cnt*	*ari*
20,362	64,846	687,915
1,306,971	1,983,694	8,650,650
1,327,333	2,048,540	9,338,565

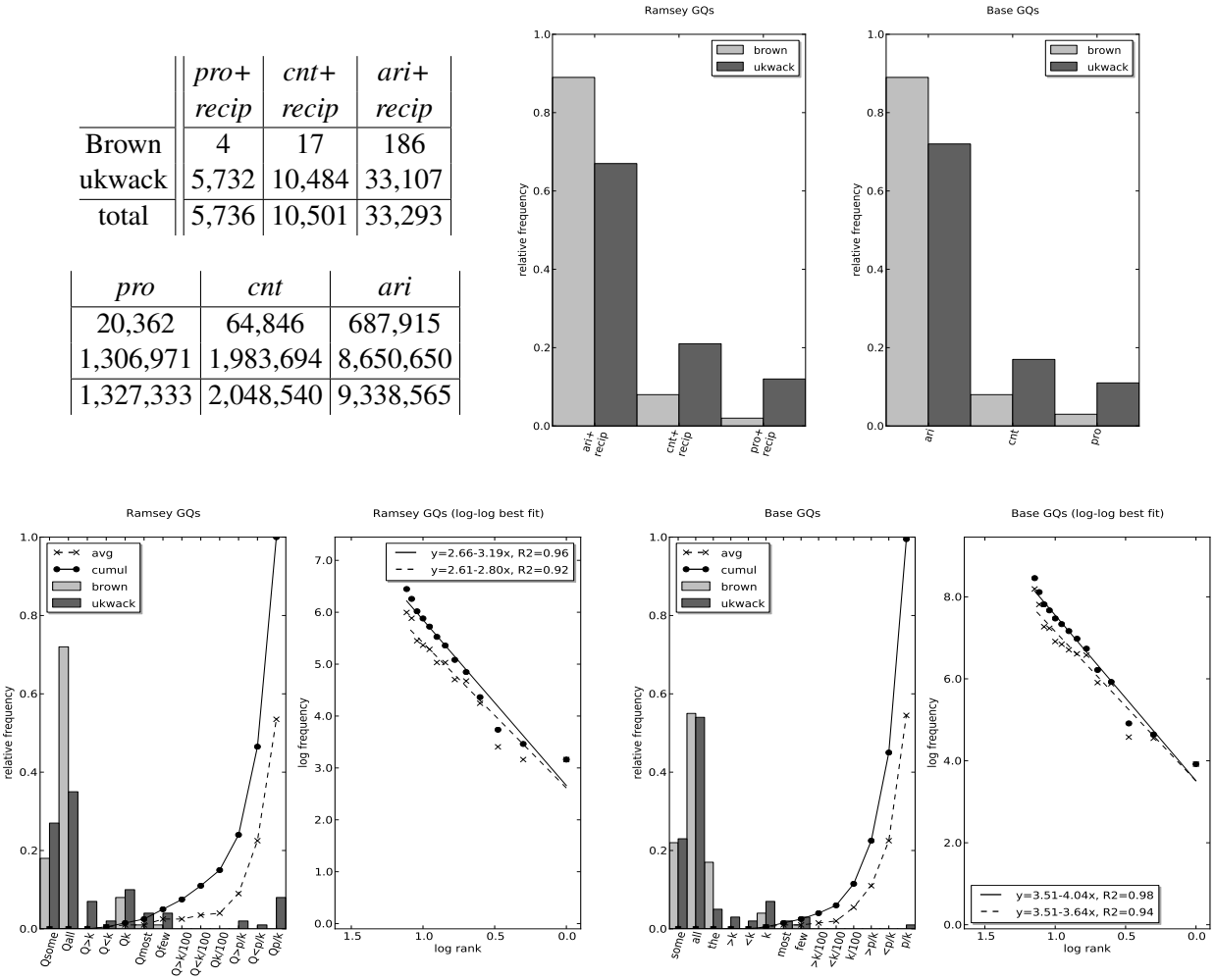

Figure 1: Top, left: Ramsey quantifier (raw) frequencies, and base quantifier (raw) frequencies. Top, right: Ramsey and base quantifier distribution by quantifier class. Bottom, left: Ramsey quantifier distribution and log-log power law regression. Bottom, right: Base quantifier distribution and log-log power law regression.

4 Conclusions

We have studied the semantic complexity and corpora distribution of natural language quantifiers. The computationally easier quantifiers occur more frequently in everyday communication, i.e., their distributions satisfy power laws. Moreover, we have empirically shown—as suggested by Ristad (1993); Mostowski and Szymanik (2012)—that: although everyday English may contain computationally hard constructions, they are infrequent. These results, together with Thorne (2012), suggest that abstract computational properties of natural language expressions can be used to explain their distribution in corpora. Indeed, one of the linguistic reasons to expect power laws in natural language data is the *principle of least effort in communication*: speakers tend to minimize the communication effort by generating "simple" messages (Zipf, 1935).

As ongoing and further research, we envision several axes. Firstly, to refine our patterns to better cover the lexical variants of the quantifiers considered in this paper. Secondly, to consider much larger corpora. Thirdly, to refine our complexity analysis to explain the frequency gap induced by Ramseyfication. Finally, to run cross-language experiments to address the *equivalent complexity* question: does the complexity of quantification imply a distribution similar accross languages?

References

Baroni, M. (2009). 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.

van Benthem, J. (1987). Towards a computational semantics. In P. Gärdenfors (Ed.), *Generalized Quantifiers*, pp. 31–71. Reidel Publishing Company.

Dalrymple, M., M. Kanazawa, Y. Kim, S. Mchombo, and S. Peters (1998). Reciprocal expressions and the concept of reciprocity. *Linguistics and Philosophy 21*, 159–210.

Francis, W. N. and H. Kucera (1979). Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Immerman, N. (1998). *Descriptive Complexity*. Texts in Computer Science. New York, NY: Springer.

Kontinen, J. and J. Szymanik (2008). A remark on collective quantification. *Journal of Logic, Language and Information 17*(2), 131–140.

Lindström, P. (1966). First order predicate logic with generalized quantifiers. *Theoria 32*, 186–195.

Mostowski, M. (1998). Computational semantics for monadic quantifiers. *Journal of Applied Non-Classical Logics 8*, 107–121.

Mostowski, M. and J. Szymanik (2012). Semantic bounds for everyday language. *Semiotica 188*(1-4), 363–372.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics 46*(5), 323–351.

Papadimitrou, C. (1994). *Computational Complexity*. Addison Wesley - Longman.

Peters, S. and D. Westerståhl (2006). *Quantifiers in Language and Logic*. Oxford: Clarendon Press.

Ristad, E. S. (1993, March). *The Language Complexity Game*. Artificial Intelligence. The MIT Press.

Schlotterbeck, F. and O. Bott (2013). Easy solutions for a hard problem? The computational complexity of reciprocals with quantificational antecedents. *Journal of Logic, Language and Information 22*(4), 363–390.

Szymanik, J. (2010). Computational complexity of polyadic lifts of generalized quantifiers in natural language. *Linguistics and Philosophy 33*(3), 215–250.

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.

Thorne, C. (2012). Studying the distribution of fragments of English using deep semantic annotation. In *Proceedings of the ISA8 Workshop*.

Zipf, G. (1935). *The Psychobiology of Language: An Introduction to Dynamic Philology*. Cambridge, Mass.: M.I.T. Press.