Semantic Complexity of Quantifiers and their Distribution in Corpora

Camilo Thorne
IBM CAS Trento – Trento RISE, Italy
c.thorne.email@trentorise.eu

Jakub Szymanik
Institute of Logic, Language and Computation – UvA, The Netherlands
J.K.Szymanik@uva.nl

A. Problem

The semantic complexity of a generalized quantifier is a phenomenon arising from their formal semantic representation [8].

Question (1): Can semantic complexity predict the distribution of quantifiers in corpora?

Question (2): Is such distribution power-law skewed towards low complexity quantifiers?

B. Generalized Quantifiers

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.

- ▷ Semantic complexity formally refers to the computational complexity of the induced finite model checking problem: does $\mathcal{I} \in Q$ [8]?
- ▷ We are interested in complexity w.r.t. the size of the finite model $\mathcal{I}$, that is, in data complexity [5].

C. Quantifier Typology

Semantic complexity induces a split into tractable and intractable quantifiers:

(1) Tractable quantifiers come in two flavors:

- First order (FO) quantifiers are FO expressible, have low semantic complexity, and split into Aristotelian and counting quantifiers.
- Proportional quantifiers are used by speakers when referring to collections and go beyond FO.

(2) Intractable quantifiers derivable via Ramseyfication, expressed by the reciprocal "each other" [3].

D. Power Laws

A power law or Zipfean relation between frequency $fr(Q)$ of quantifier Q and its complexity $rk(Q)$ can be described by the equation:

$$fr(Q) = a/rk(Q)^b.$$

- ▷ Power laws are skewed distributions where intuitively the topmost (w.r.t. rank) 20% outcomes of a variable concentrate around 80% of the frequency [1].
- ▷ Parameters a and b are inferred via linear regression [7], as power laws are equivalent to linear models on the log-log scale.

E. Corpus Analysis

We counted generalized quantifiers indirectly, by matching part-of-speech patterns (regular expressions) such as:

(1) for the Aristotelian quantifier *all*, we considered its lexical variants and matched

```
.*( 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) for Ramsey quantifiers we matched at the same time FO, counting or proportional patterns and

```
.* each/dt other/ap .*
```

We considered large English corpora:

(3) the Brown corpus by [4] ($\sim$60,647 sentences and 1,014,312 word tokens)

(4) a sample of the the ukWaC corpus ($\sim$280,001 sentences and 100,000,000 word tokens) from [2].

We validated the power law models by estimating the goodness-of-fit R^2 coefficient and testing for statistical significance (χ^2 at $p < 0.01$ significance).

11th International Conference on Computational Semantics, London, UK, April 12-15, 2015

E. Semantic Complexity Analysis

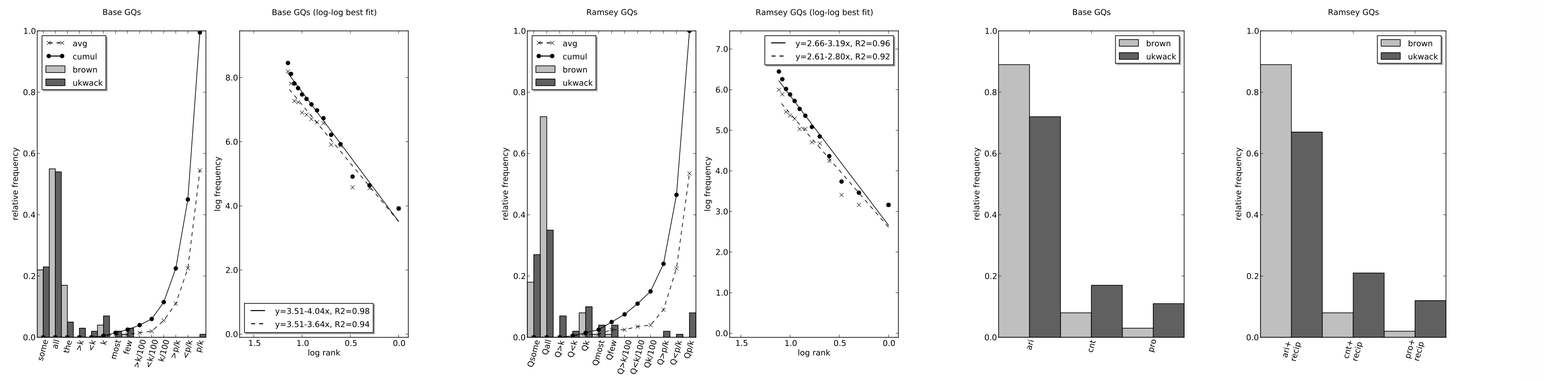
Base FO (Aristotelian and counting) and proportional generalized quantifiers studied in this paper; $> k$ and $< k$ comprise by abuse the superlative quantifiers "at least k" and "at most k" [8]:

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

Left, sample English sentences realizing Ramsey quantifiers; right, semantic complexity of Ramsey quantifiers by quantifier class [9]:

R_Q	Example	Quantifier Class R_Q	S.C.
R_{some}	some children like each other	Aristotelian (*ari+recip*)	$\mathbf{AC}^0$
$R_{>p/k}$	more than $2/3$ of female MPs sit next to each other	counting (*cnt+recip*)	$\mathbf{AC}^0$
R_{most}	most people help each other	proportional (*pro+recip*)	**NP**-complete
$R_{>k}$	at least 2 men married each other in the UK last year		

F. Corpus Experiments



	pro+ recip	*cnt+ recip*	*ari+ recip*	*pro*	*cnt*	*ari*
Brown	4	17	186	20,362	64,846	687,915
ukwack	5,732	10,484	33,107	1,306,971	1,983,694	8,650,650
total	5,736	10,501	33,293	1,327,333	2,048,540	9,338,565

model (base)	$fr(Q) = 3.51/rk(Q)^{3.64}$
model (Ramsey)	$fr(Q) = 2.61/rk(Q)^{2.80}$
R^2 (base)	$R^2 = 0.94$
R^2 (Ramsey)	$R^2 = 0.92$
χ^2-test (both)	$p = 0.0$

G. Conclusions & Further Work

(1) We have showed that semantic complexity can be used to characterize the expressiveness of English quantifiers.

(2) Corpora distributions are significantly skewed towards low complexity quantifiers, and can be in some cases described by a power law.

(3) Our experiments suggest that semantic complexity can be used to predict quantifier distribution.

(4) Related work by the authors suggest that corpora distributions are in general skewed towards "simple" English constructs [11, 10].

(5) In the future, we plan to extend this work by looking at different semantic construction and comparing various languages.

(6) We are also interested in applying the notion of semantic complexity to the linguistic debate on *equivalent complexity thesis*: all natural languages are equally complex (have equal descriptive power) [see, e.g., 6]

H. References

[1] Marco Baroni. Distributions in text. In *Corpus linguistics: An International Handbook*, volume 2, pages 803–821. Mouton de Gruyter, 2009.

[2] Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.

[3] Mary Dalrymple, Makoto Kanazawa, Yookyung. Kim, Sam Mchombo, and Stanley Peters. Reciprocal expressions and the concept of reciprocity. *Linguistics and Philosophy*, 21:159–210, 1998.

[4] W. Nelson Francis and Henry Kucera. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US, 1979.

[5] Neil Immerman. *Descriptive Complexity*. Texts in Computer Science. Springer, New York, NY, 1998.

[6] M. Miestamo, K. Sinnemäki, and F. Karlsson, editors. *Language Complexity: Typology, contact, change*. Studies in Language Companion Series. John Benjamins Publishing Company, February 2008.

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

[8] Jakub Szymanik. *Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language*. PhD thesis, University of Amsterdam, Amsterdam, 2009.

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

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

[11] Camilo Thorne. Studying the distribution of fragments of English using deep semantic annotation. In *Proceedings of the ISA8 Workshop*, 2012.