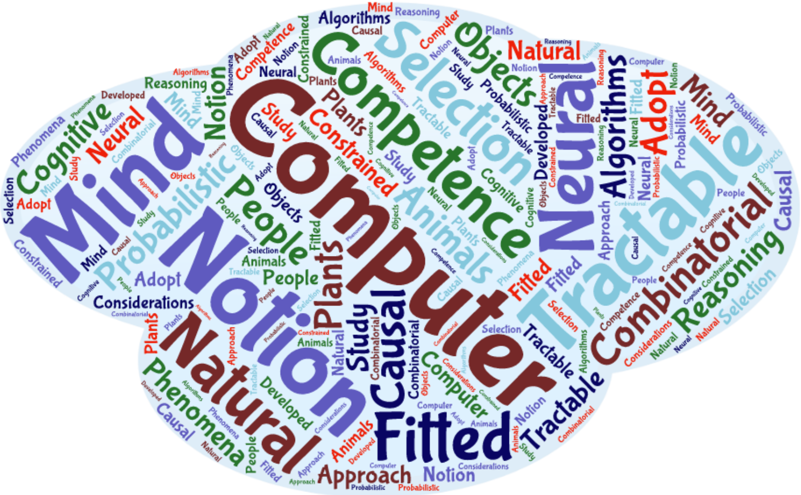
Semantic Complexity and Corpus Analysis

Camilo Thorne

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
camilo.thorne@ims.uni-stuttgart.de

CoSaQ Worskshop, Amsterdam, 28–29.9.17


Computer
Notion
Mind
Tractable
Neural
Competence
Selection
Natural
Fitted
Causal
Combinatorial
Reasoning
Objects
Algorithms
Cognitive
Probabilistic
People
Plants
Animals
Study
Constrained
Phenomena
Approach
Adopt
Developed
Considerations

Outline

Motivation

- Complexity in Language
- Research Question(s)

Quantifier Distribution

- Aristotelian, Counting and Proportional Quantifiers
- Corpus Analysis

Fragment Distribution

- The Fragments of English
- Corpus Analysis

Summary

Motivation

Motivation – Cognitive Complexity

> *"The mind is a neural computer, fitted by natural selection with combinatorial algorithms for causal and probabilistic reasoning about plants, animals, objects, and people."*

Steven Pinker, 1997, *How the Mind Works*

> *"(...) if we adopt a computational approach to the study of cognitive phenomena, (...) a notion of tractable competence can be developed, that is, a notion of competence constrained by considerations on computational tractability."*

Frixione, 2011, *Tractable Competence*

Motivation – Complexity in Language

- Cognitive complexity is mirrored by complexity in natural language(s)

Motivation – Complexity in Language

- Cognitive complexity is mirrored by complexity in natural language(s)

    1. surface form/syntactic complexity ⇒ length of utterance, size of parse tree, etc.

Motivation – Complexity in Language

- Cognitive complexity is mirrored by complexity in natural language(s)

  1. surface form/syntactic complexity ⇒
     length of utterance, size of parse tree, etc.
  2. Kolmogorov complexity ⇒
     size of smallest Turing machine

Motivation – Complexity in Language

- Cognitive complexity is mirrored by complexity in natural language(s)

  1. surface form/syntactic complexity ⇒

     length of utterance, size of parse tree, etc.

  2. Kolmogorov complexity ⇒

     size of smallest Turing machine

  ...

Motivation – Complexity in Language

- Cognitive complexity is mirrored by complexity in natural language(s)

1. surface form/syntactic complexity ⇒
   length of utterance, size of parse tree, etc.
2. Kolmogorov complexity ⇒
   size of smallest Turing machine

. . .

✓ semantic complexity ⇒
complexity of verifying truth conditions as function of (finite) model size

Motivation – Complexity in Language

- Cognitive complexity is mirrored by complexity in natural language(s)

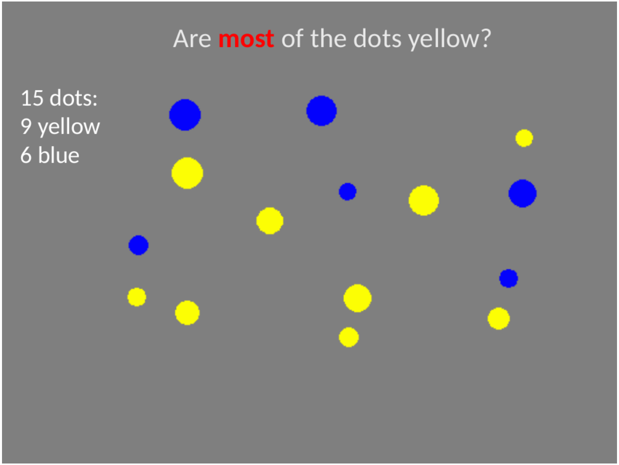
  1. surface form/syntactic complexity ⇒
     length of utterance, size of parse tree, etc.
  2. Kolmogorov complexity ⇒
     size of smallest Turing machine

  . . .

  ✓ semantic complexity ⇒
  complexity of verifying truth conditions as function of (finite) model size

✌ Semantic complexity seems to mirror best cognitive complexity

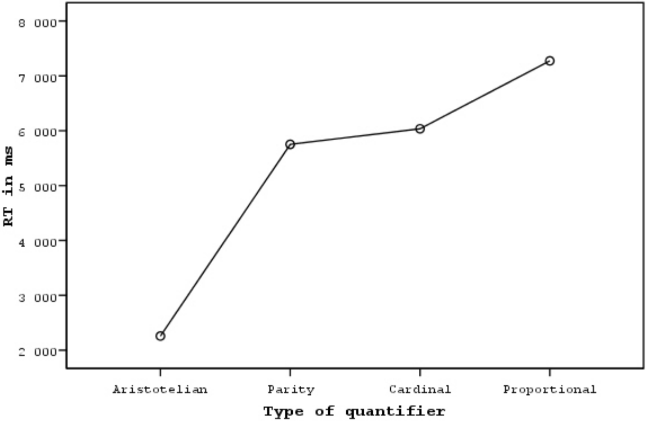
Motivation – Picture Verification Task [Szy09]



Are **most** of the dots yellow?

15 dots:
9 yellow
6 blue

Mtivation – Picture Verification Task



Parity: “exactly 2” *Cardinal*: all the other counting quantifiers

Motivation – Hypothesis

We can approximate the distribution of linguistic structures by analyzing (large) corpora

Motivation – Hypothesis

We can approximate the distribution of linguistic structures by analyzing (large) corpora

Hypothesis 1

Linguistic structures in are distributed w.r.t. semantic complexity

Motivation – Hypothesis

We can approximate the distribution of linguistic structures by analyzing (large) corpora

Hypothesis 1

Linguistic structures in are distributed w.r.t. semantic complexity

Hypothesis 2

The distribution is biased towards low complexity structures

Quantifier Distribution

Aristotelian, Counting and Proportional Quantifiers

$\vdots$

each wordle is a cloud

most wordles contain some kind of message

fewer than $1/2$ of wordles suck

few wordles solve mathematical problems

exactly one wordle is contained in these slides

$\vdots$

Aristotelian, Counting and Proportional Quantifiers

$\vdots$

each wordle is a cloud

most wordles contain some kind of message

fewer than $1/2$ of wordles suck

few wordles solve mathematical problems

exactly one wordle is contained in these slides

$\vdots$

L-Expressibility and Complexity [Szy09]

Definition (L-Expressibility)

Generalized quantifier Q over domain Δ is expressible in logic L iff there exists a formula such that

$$(R_1, \ldots, R_n) \in Q \;\Leftrightarrow\; (\Delta, \cdot^{\mathcal{I}}) \models \overline{Q}(\overline{R}_1, \ldots, \overline{R}_n)$$

L-Expressibility and Complexity [Szy09]

Definition (L-Expressibility)

Generalized quantifier Q over domain Δ is expressible in logic L iff there exists a formula such that

$$(R_1, \ldots, R_n) \in Q \ \Leftrightarrow\ (\Delta, \cdot^{\mathcal{I}}) \models \overline{Q}(\overline{R}_1, \ldots, \overline{R}_n)$$

- The semantic complexity of generalized quantifier Q is the cost of computing

$$(\Delta, \cdot^{\mathcal{I}}) \models \overline{Q}(\overline{R}_1, \ldots, \overline{R}_n)$$

L-Expressibility and Complexity [Szy09]

Definition (L-Expressibility)

Generalized quantifier Q over domain Δ is expressible in logic L iff there exists a formula such that

$$(R_1, \ldots, R_n) \in Q \;\Leftrightarrow\; (\Delta, \cdot^{\mathcal{I}}) \models \overline{Q}(\overline{R}_1, \ldots, \overline{R}_n)$$

- The semantic complexity of generalized quantifier Q is the cost of computing

$$(\Delta, \cdot^{\mathcal{I}}) \models \overline{Q}(\overline{R}_1, \ldots, \overline{R}_n)$$

- We measure computational cost in $\#(\Delta)$: data complexity

L-Expressibility and Complexity [Szy09]

Definition (L-Expressibility)

Generalized quantifier Q over domain Δ is expressible in logic L iff there exists a formula such that

$$(R_1, \ldots, R_n) \in Q \;\Leftrightarrow\; (\Delta, \cdot^{\mathcal{I}}) \models \overline{Q}(\overline{R}_1, \ldots, \overline{R}_n)$$

- The semantic complexity of generalized quantifier Q is the cost of computing

$$(\Delta, \cdot^{\mathcal{I}}) \models \overline{Q}(\overline{R}_1, \ldots, \overline{R}_n)$$

- We measure computational cost in $\#(\Delta)$: data complexity

✓ Example: "each"

1. is $\textsc{Fo}$-expressible as: $(A, B) \in [\![\text{each}]\!] \Leftrightarrow (\Delta, \cdot^{\mathcal{I}}) \models \forall x(\overline{A}(x) \rightarrow \overline{B}(x))$
2. has $\textsc{Space}(\log_2(\Delta))$ semantic complexity

Complexity Ranking [TS15, ST17]

Q	**Semantics** $\subseteq \mathcal{P}(\Delta) \times \mathcal{P}(\Delta)$	**DC**		**Example**
some	$\{(A, B) \mid A \cap B \neq \emptyset\}$	LSPACE	ari	some men are happy
all	$\{(A, B) \mid A \subseteq B\}$	LSPACE		all humans are mammals
no	$\{(A, B) \mid A \cap B = \emptyset\}$	LSPACE		no humans are spiders
$\geq k$	$\{(A, B) \mid \#(A \cap B) > k\}$	LSPACE	cnt	more than 5 men are happy
$\leq k$	$\{(A, B) \mid \#(A \cap B) < k\}$	LSPACE		fewer than 100 violins are Stradivari
most	$\{(A, B) \mid \#(A \cap B) > \#(A \setminus B)\}$	P	pro	most trains are safe
few	$\{(A, B) \mid \#(A \cap B) < \#(A \setminus B)\}$	P		few people are trustworthy
$\geq p/k$	$\{(A, B) \mid \#(A \cap B) > p \cdot (\#(A)/k)\}$	P		more than $2/3$ of planets are lifeless
$\leq p/k$	$\{(A, B) \mid \#(A \cap B) < p \cdot (\#(A)/k)\}$	P		less than $1/3$ of Americans are rich

ari: Aristotelian quantifiers *cnt*: counting quantifiers *pro*: proportional quantifiers

Complexity Ranking [TS15, ST17]

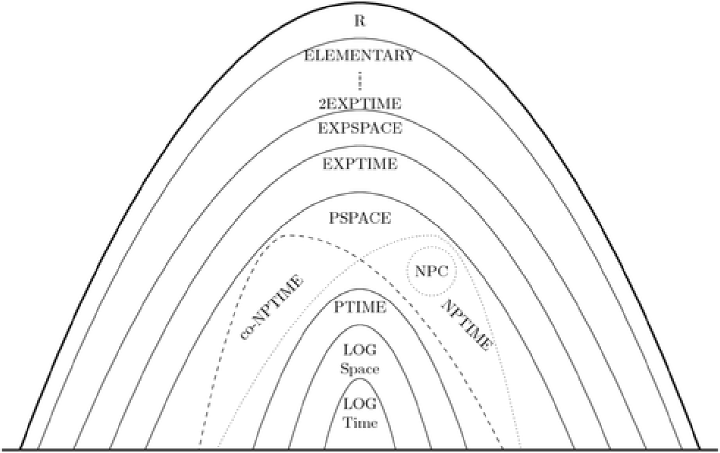
Q	**Semantics** $\subseteq \mathcal{P}(\Delta) \times \mathcal{P}(\Delta)$	**DC**		**Example**
some	$\{(A, B) \mid A \cap B \neq \emptyset\}$	LSPACE	ari	some men are happy
all	$\{(A, B) \mid A \subseteq B\}$	LSPACE		all humans are mammals
no	$\{(A, B) \mid A \cap B = \emptyset\}$	LSPACE		no humans are spiders
$\geq k$	$\{(A, B) \mid \#(A \cap B) > k\}$	LSPACE	cnt	more than 5 men are happy
$\leq k$	$\{(A, B) \mid \#(A \cap B) < k\}$	LSPACE		fewer than 100 violins are Stradivari
most	$\{(A, B) \mid \#(A \cap B) > \#(A \setminus B)\}$	P	pro	most trains are safe
few	$\{(A, B) \mid \#(A \cap B) < \#(A \setminus B)\}$	P		few people are trustworthy
$\geq p/k$	$\{(A, B) \mid \#(A \cap B) > p \cdot (\#(A)/k)\}$	P		more than $2/3$ of planets are lifeless
$\leq p/k$	$\{(A, B) \mid \#(A \cap B) < p \cdot (\#(A)/k)\}$	P		less than $1/3$ of Americans are rich

ari: Aristotelian quantifiers *cnt*: counting quantifiers *pro*: proportional quantifiers

Question 1

Does complexity influence quantifier distribution in (large) corpora?

Complexity Class Hierarchy (Simplified)



The WaCkY Corpus [BBFZ09]

```
<s>
Flender Flender NP      1       3       VMOD
Werke   Werke   NP      2       3       SBJ
was     be      VBD     3       0       ROOT
a       a       DT      4       7       NMOD
German  German  JJ      5       7       NMOD
shipbuilding    shipbuilding    NN      6       7       NMOD
company company NN      7       3       PRD
,       ,       ,       8       7       P
located locate  VVN     9       7       NMOD
in      in      IN      10      9       ADV
Lubeck  Lubeck  NP      11      10      PMOD
.       .       SENT    12      0       ROOT
</s>
```

	Sentences	**Tokens**	**Source**
WaCkY (Eng)	$\sim$ 43 million	$\sim$ 800 million	Wikipedia (EN, 2008)

Corpus Analysis

- We built a list of simple patterns to identify and count

  1. Aristotelian quantifiers: *all*, *some*, *no*
  2. counting quantifiers: *less (more) than* k
  3. proportional quantifiers: *most*, *few*, *less (more) than* p/k

- Examples:

$$\left\{\begin{array}{lcl} \textit{most} & = & \texttt{most/dt},\ \ \texttt{most/jjs [a-z]}\{\texttt{1,12}\}\texttt{/nns}, \\ & & \texttt{most/rbs [a-z]}\{\texttt{1,12}\}\texttt{/nns}, \\ & & \texttt{more/rbr than/in half/nn}, \\ & & \texttt{more/jjr than/in half/nn} \\ \textit{some} & = & \texttt{some/det} \end{array}\right.$$

Corpus Analysis

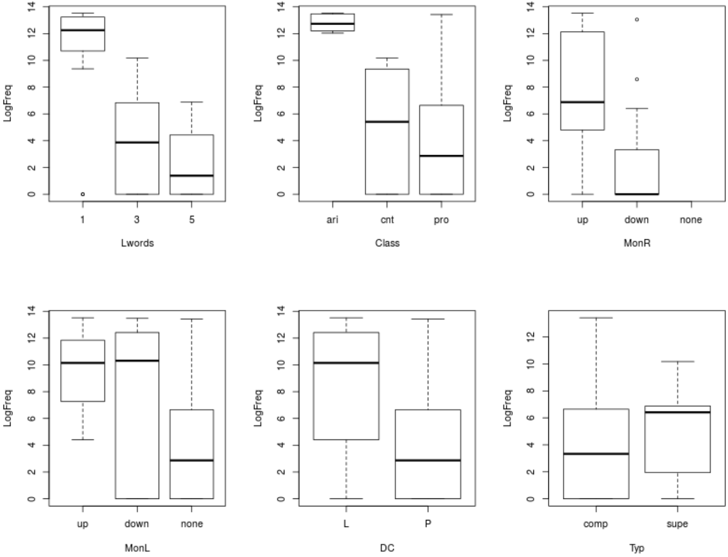
- We built a list of simple patterns to identify and count

  1. Aristotelian quantifiers: *all*, *some*, *no*
  2. counting quantifiers: *less (more) than* k
  3. proportional quantifiers: *most*, *few*, *less (more) than* p/k

- Examples:
$$\left\{\begin{array}{lcl} \textit{most} & = & \texttt{most/dt,}\ \ \texttt{most/jjs [a-z]\{1,12\}/nns,} \\ & & \texttt{most/rbs [a-z]\{1,12\}/nns,} \\ & & \texttt{more/rbr than/in half/nn,} \\ & & \texttt{more/jjr than/in half/nn} \\ \textit{some} & = & \texttt{some/det} \end{array}\right.$$

- We seeked to understand how much their frequency is influenced by semantic complexity (and other features)

Frequency Distribution by Feature [ST17]



GLMs

Definition (Generalized Linear Model (GLM))

A (multifactor) generalized linear model (GLM) has the form

$$f(y^{(j)}) = \theta_1 x_1^{(j)} + \cdots + \theta_k x_k^{(j)} + \theta_{k+1}$$

1. $f\colon \mathbb{R} \to \mathbb{R}$ is a link function (usually: $\ln(\cdot)$)
2. $Y \sim \mathcal{D}$, with $\mathcal{D}$ an arbitrary distribution
3. X_i's can be random effects (mixed model)

✌ **Negative binomial:** assumes that $Y \sim \mathcal{NB}(r, p)$ i.e., a negative binomial GLM will assume that frequency decreases geometrically

Results [ST17]

Analysis 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%	$1.15 \cdot e^{-6}$

- Semantic complexity (GQ class and monotonicity) explain $>$ 50% of error deviance
- If we consider GQs of different lengths, length has also a significant impact

Institut für Maschinelle Sprachverarbeitung

Results [ST17]

Analysis 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%	$1.15 \cdot e^{-6}$

- Semantic complexity (GQ class and monotonicity) explain $>$ 50% of error deviance
- If we consider GQs of different lengths, length has also a significant impact

✌ GQ distribution skewed towards cheap and short quantifiers

Discussion

- Preliminary study to determine if semantic complexity influences quantifier distribution
- Examined distribution on large encyclopedic corpora (Wikipedia)
- Results indicate that the cheaper a quantifier, the more frequent it is
- Used generalized regression analysis to quantify impact
- But: the class of quantifiers studied here was quite small

Fragment Distribution

Reasoning Complexity

- Suppose we came across the following argument in some text

  Every Italian loves pasta and football
  Silvio is Italian

  that entails

  Silvio loves pasta

Reasoning Complexity

- Suppose we came across the following argument in some text

  Every Italian loves pasta and football
  Silvio is Italian

  that entails

  Silvio loves pasta

- Common-sense reasoning
- Different constructs give rise to different semantic complexity

Question 2

Does complexity influence construct distribution in (large) corpora?

The Fragments of English [PHT06]

- The fragments of English are linguistically salient, ambiguity-free subsets of English [PHT06]

The Fragments of English [PHT06]

- The fragments of English are linguistically salient, ambiguity-free subsets of English [PHT06]

(a) polynomially translate to $\mathrm{F}\textsc{o}$ meaning representations φ

(b) basic machinery of formal semantics

The Fragments of English [PHT06]

- The fragments of English are linguistically salient, ambiguity-free subsets of English [PHT06]

  (a) polynomially translate to $\textsc{Fo}$ meaning representations φ

  (b) basic machinery of formal semantics

- Defined by constraining syntax, semantics and vocabulary

The Fragments of English [PHT06]

- The fragments of English are linguistically salient, ambiguity-free subsets of English [PHT06]

(a) polynomially translate to FO meaning representations φ

(b) basic machinery of formal semantics

- Defined by constraining syntax, semantics and vocabulary

✌ Define semantic complexity as logical satisfiability!

The Fragments of English [PHT06]

Fragment	Coverage	Fo Operators	
COP($^\neg$)	Copula ("is a"), nouns ("man"), intransitive verbs ("runs"), "every", "some" names ("Joe"), adjectives ("thin") (+"not")	$\{\forall, \exists, (\neg)\}$	P
COP($^\neg$)+TV	COP($^\neg$) +transitive verbs ("loves")	$\{\forall, \exists, (\neg)\}$	
COP($^\neg$)+DTV	COP($^\neg$) +ditransitive verbs ("gives")	$\{\forall, \exists, (\neg)\}$	
COP($^\neg$)+TV +DTV	COP($^\neg$)+TV + ditransitive verbs	$\{\forall, \exists, (\neg)\}$	
COP$^\neg$+Rel	COP$^\neg$+ ("who", "that", "which") "and", intersective adjectives (+"or")	$\{\forall, \exists, \wedge, \neg, \vee\}$	NP-hard
COP$^\neg$+Rel +TV	COP$^\neg$+Rel +transitive verbs	$\{\forall, \exists, \wedge, \neg, \vee\}$	
COP$^\neg$+Rel +DTV	COP$^\neg$+Rel +ditransitive verbs	$\{\forall, \exists, \wedge, \neg, \vee\}$	
COP$^\neg$+Rel +TV+DTV	COP$^\neg$+Rel+TV +ditransitive verbs	$\{\forall, \exists, \wedge, \neg, \vee\}$	

The Fragments of English (Examples)

Fragment	**Example**	**Fo**
COP	Every politician cheats	$\forall x(\mathit{Politician}(x) \rightarrow \mathit{Cheat}(x))$
COP$^{\neg}$	Some philosopher is not trustworthy	$\exists x(\mathit{Philosopher}(x) \wedge \neg \mathit{Trusted}(x))$
COP$^{\neg}$+TV	John does not love Luke	$\neg \mathit{Loves}(\mathrm{John}, \mathrm{Luke})$
COP+TV +DTV	John gives a book to Jane Some man likes every candy	$\exists x \mathit{Book}(x) \wedge$ $\mathit{Gives}(\mathrm{John}, x, \mathrm{Jane})$ $\exists x(\mathit{Man}(x) \wedge$ $\forall y \mathit{Candy}(x) \rightarrow \mathit{Likes}(x, y))$
COP$^{\neg}$ +Rel	Some man who does not cheat is trustworthy	$\forall x(\mathit{Man}(x) \wedge \neg \mathit{Cheat}(x)$ $\rightarrow \mathit{Trusted}(x))$
$\vdots$	$\vdots$	$\vdots$

Semantic Annotation - Boxer [Tho12]

Exploit deep semantic parsers, in particular Boxer 2.0 [Bos08]

Semantic Annotation - Boxer [Tho12]

Exploit deep semantic parsers, in particular Boxer 2.0 [Bos08]

- When parsing Wh-questions from the TREC 2008

  What is one common element of major religions?

  Boxer outputs

$$\begin{gathered}\exists y \exists z \exists e \exists u (\texttt{card}(y, u) \wedge \texttt{c1num}(u) \\ \wedge\ \texttt{nnumeral1}(u) \wedge \texttt{acommon1}(y) \\ \wedge\ \texttt{nelement1}(y) \wedge \texttt{amajor1}(z) \\ \wedge\ \texttt{nreligions1}(z) \wedge \texttt{nevent1}(e) \\ \wedge\ \texttt{rof1}(y, z))\end{gathered}$$

Semantic Annotation - Boxer [Tho12]

Exploit deep semantic parsers, in particular Boxer 2.0 [Bos08]

- When parsing Wh-questions from the TREC 2008

  What is one common element of major religions?

  Boxer outputs

$$
\begin{gathered}
\exists y \exists z \exists e \exists u (\mathtt{card}(y, u) \wedge \mathtt{c1num}(u) \\
\wedge\ \mathtt{nnumeral1}(u) \wedge \mathtt{acommon1}(y) \\
\wedge\ \mathtt{nelement1}(y) \wedge \mathtt{amajor1}(z) \\
\wedge\ \mathtt{nreligions1}(z) \wedge \mathtt{nevent1}(e) \\
\wedge\ \mathtt{rof1}(y, z))
\end{gathered}
$$

- $\wedge$ and $\exists$ co-occur, but not $\vee$, $\neg$, or $\forall$

Classifying Sentences (Rules)

Classifying Sentences (Rules)

1. Annotate/parse the corpora

Classifying Sentences (Rules)

1. Annotate/parse the corpora
2. Consider sentences expressing operators from classes $c \subseteq \{\wedge, \exists, \forall, \vee, \neg\}$

Classifying Sentences (Rules)

1. Annotate/parse the corpora
2. Consider sentences expressing operators from classes $c \subseteq \{\wedge, \exists, \forall, \vee, \neg\}$
3. Observe the frequency of

Classifying Sentences (Rules)

1. Annotate/parse the corpora
2. Consider sentences expressing operators from classes $c \subseteq \{\wedge, \exists, \forall, \vee, \neg\}$
3. Observe the frequency of

(a) Four P classes: $\{\exists, \wedge\}$, $\{\exists, \wedge, \forall\}$, $\{\exists, \wedge, \vee\}$ and $\{\exists, \wedge, \forall, \vee\}$

(b) Four NP-hard classes: $\{\exists, \wedge, \neg\}$, $\{\exists, \wedge, \neg, \forall\}$, $\{\exists, \wedge, \neg, \forall, \vee\}$ and $\{\neg, \forall\}$

(each class approximates a fragment of English)

Classifying Sentences (Rules)

1. Annotate/parse the corpora
2. Consider sentences expressing operators from classes $c \subseteq \{\wedge, \exists, \forall, \vee, \neg\}$
3. Observe the frequency of

   (a) Four P classes: $\{\exists, \wedge\}$, $\{\exists, \wedge, \forall\}$, $\{\exists, \wedge, \vee\}$ and $\{\exists, \wedge, \forall, \vee\}$

   (b) Four NP-hard classes: $\{\exists, \wedge, \neg\}$, $\{\exists, \wedge, \neg, \forall\}$, $\{\exists, \wedge, \neg, \forall, \vee\}$ and $\{\neg, \forall\}$

   (each class approximates a fragment of English)

4. Study relationships between class frequency *fr*(c) and class rank or expressivity *rk*(c)

Corpora

We considered:

- a subset (A: press articles) of the Brown corpus (http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml)
- a subset (Geoquery880) of the Geoquery corpus (http://www.cs.utexas.edu/users/ml/nldata/geoquery.html)
- a corpus of clinical questions (http://clinques.nlm.nih.gov)
- a sample from the TREC 2008 corpus (http://trec.nist.gov)

Corpus	Size	Domain	Type
Brown	19,741 sent.	Open (news)	Declarative
Geoquery	364 ques.	Geographical	Interrogative
Clinical ques.	12,189 ques.	Clinical	Interrogative
TREC 2008	436 ques.	Open	Interrogative

Motivation – Power Laws

Definition (Power law)

Random variable X follows a power law if the following relation holds

$$\mathit{fr}(x^{(j)}) = \frac{\theta_0}{\mathit{rk}(x^{(j)})^{\theta_1}}$$

Motivation – Power Laws

Definition (Power law)

Random variable X follows a power law if the following relation holds

$$\mathit{fr}(x^{(j)}) = \frac{\theta_0}{\mathit{rk}(x^{(j)})^{\theta_1}}$$

- Zipf posed that power laws arise in natural language data due to the principle of least effort

  1. speakers maximize succinctness of utterances conveying a message
  2. hearers have the dual goal

Motivation – Power Laws

Definition (Power law)

Random variable X follows a power law if the following relation holds

$$\mathit{fr}(x^{(j)}) = \frac{\theta_0}{\mathit{rk}(x^{(j)})^{\theta_1}}$$

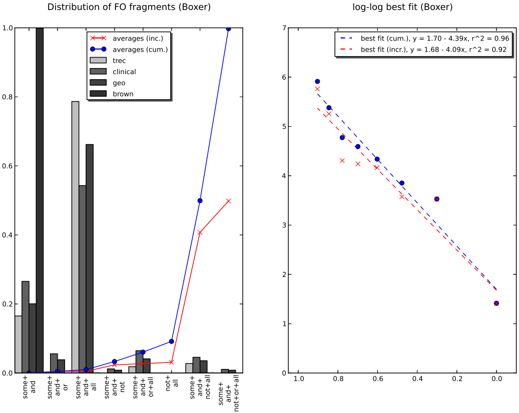
- Zipf posed that power laws arise in natural language data due to the principle of least effort

  1. speakers maximize succinctness of utterances conveying a message
  2. hearers have the dual goal

- N.B. Expressible as a linear model:

$$\mathit{fr}(x^{(j)}) = \frac{\theta_0}{\mathit{rk}(x^{(j)})^{\theta_1}} \quad \Leftrightarrow \quad \ln(\mathit{fr}(x^{(j)})) = \ln(\theta_0) - \theta_1 \ln(\mathit{rk}(x^{(j)}))$$

Power Law Fitting [Tho12]



	(power law)	(R^2)
cum:	$fr(c) = \frac{5.47}{rk(c)^{4.39}}$	0.96
means:	$fr(c) = \frac{5.37}{rk(c)^{4.09}}$	0.92

Discussion

- We have experimented with a methodology based on the Boxer semantic parser
- The distribution obtained may seem to indicate that “non-Boolean-closed” (tractable) fragments occur more often than “Boolean-closed” (intractable) fragments
- Power laws with high R^2 (> 90) could be derived
- But: worked under the assumption that Boxer returns reasonably accurate results, and
  1. Boxer’s precision and recall are not well-known
  2. Boxer blows up the number of existential quantifiers due to Davidsonian event semantics

Summary

Summarizing

- Semantic complexity provides formal, logic-based model of key cognitive dimension of natural language
- Cognitive experiments show that it influences language use
- Our work indicates that it also influences the distribution (frequency) of key language constructs, if
  1. we approximate such distribution by analyzing corpora
  2. we run regression analysis to quantify its impact
- The distributions observed are skewed towards low complexity constructs
- These results are promising, but require methodological refinement (higher accuracy and coverage of semantic analysis)

Thanks!

References I

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.

Johan Bos.
Wide-coverage semantic analysis with Boxer.
In *Proceedings of the 2008 Conference on Semantics in Text Processing (STEP 2008)*, 2008.

Ian Pratt-Hartmann and Allan Third.
More fragments of language.
Notre Dame Journal of Formal Logic, 47(2):151–177, 2006.

Jakub Szymanik and Camilo Thorne.
Exploring the relation between semantic complexity andquantifier distribution in large corpora.
Language Sciences, 60:80–93, 2017.

References II

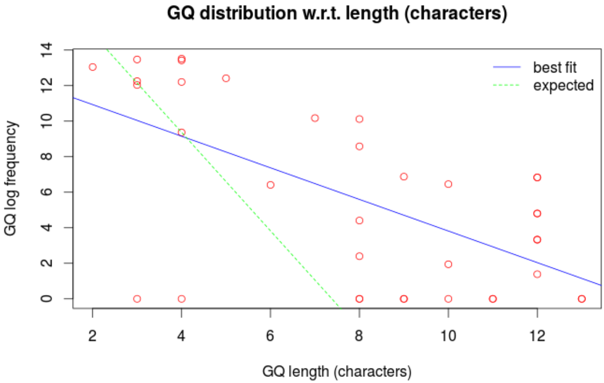
Jakub Szymanik.
Quantifiers in Time and Space.
Institute for Logic, Language and Computation, 2009.

Camilo Thorne.
Studying the distribution of fragments of English using deep semantic annotation.
In *Proceedings of the ISA-9 Workshop*, 2012.

Camilo Thorne and Jakub Szymanik.
Semantic complexity of quantifiers and their distribution in corpora.
In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*, 2015.

Institut für Maschinelle Sprachverarbeitung

Appendix: Distribution w.r.t. Length



Appendix: Linear Regression (Reminder)

A linear regression model has the form

$$y^{(j)} = \theta_0 x^{(j)} + \theta_1$$

with parameters $\theta = (\theta_0, \theta_1)$ (a gradient and an intercept)

The least squares method infers from training sample $\mathcal{S} = \{(x^{(j)}, y^{(j)})\}_{j \in [1,n]}$ the model whose parameters θ^*

$$\theta^* = \arg\min_{\theta} J(\theta) = \arg\min_{\theta} \sum_{j=1}^{n} [y^{(j)} - (\theta_0 x^{(j)} + \theta_1)]^2$$

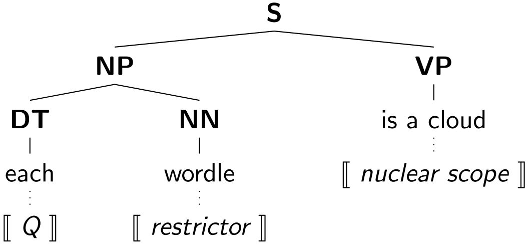
minimize the model's cost $J(\bar{\theta})$ (empirical error)

The R^2 coefficient provides a measure of confidence in $\bar{\theta}^*$

$$R^2 = \frac{Var(X\theta)}{Var(Y)}$$

Appendix: Generalized Quantifiers

Given the domain of discourse $\Delta = \{d_i \mid i \in \mathbb{N}\}$, a generalized quantifier Q of type $(k_1, \ldots, k_n)$ is an n-ary relation $Q \subseteq \mathcal{P}(\Delta^{k_1}) \times \cdots \times \mathcal{P}(\Delta^{k_n})$



S

NP

VP

DT

NN

is a cloud

each

wordle

$[\![$ *nuclear scope* $]\!]$

$[\![$ *Q* $]\!]$

$[\![$ *restrictor* $]\!]$

$$([\![\text{wordle}]\!], [\![\text{is a cloud}]\!]) \in [\![\text{each}]\!] \quad \Leftrightarrow \quad [\![\text{wordle}]\!] \subseteq [\![\text{is a cloud}]\!]$$

Generalized quantifiers describe conditions/constraints to be met by the denotations of sentence subjects (restrictor) and predicates (nuclear scope), viz., its "arguments"

Appendix: GQ Features/Predictors

1. **Class**: Factor encoding GQ class, with three values: "ari" (Aristotelian), "cnt" (counting) and "pro" (proportional)
2. **Data complexity (DC)**: Factor encoding GQ data complexity, with two values: "L" (LSpace) and "P" (P)
3. **Right monotonicity (MonR)**: Factor encoding the monotonicity of subject NP, with three values: "up" (upward monotonic), "down" (downward monotonic) and "neither" (non-monotonic)
4. **Left monotonicity (MonL)**: a factor encoding the monotonicity properties of object NPs
5. **Length in words (Lwords)**: factor that clusters GQ patterns according to the *minimum* number of word tokens
6. **Type (Typ)**: Factor encoding if GQ is superlative or comparative ("comp", "supe")

Appendix: GQ Raw Numbers

Q	**Freq**	**Lwords**	**Class**	**MonL**	**MonR**	**DC**	**FRank**
$\geq k$	51117	3	*cnt*	$\uparrow$	$\uparrow$	LSpace	6
$\leq k$	5953	3	*cnt*	$\uparrow$	$\uparrow$	LSpace	7
$\geq p/k$	1608	4	*pro*	$\uparrow$	$\downarrow$	P	8
$\leq p/k$	16	4	*pro*	$\uparrow$	$\downarrow$	P	13
few	209356	1	*pro*	neither	$\downarrow$	P	5
most	688502	1	*pro*	neither	$\uparrow$	P	3
no	464755	1	*ari*	$\downarrow$	$\downarrow$	LSpace	4
all	1325639	1	*ari*	$\downarrow$	$\downarrow$	LSpace	1
some	742134	1	*ari*	$\downarrow$	$\uparrow$	LSpace	2

- Quantifiers expressed by 32 distinct patterns
- In the analysis we split $\geq k$ and $\geq p/k$ into superlative (= "at least") and comparative quantifiers (= "more than")

Appendix: FOEs & Semantic Complexity [PHT06]

- $\neq$ fragments of English give rise to $\neq$ **semantic complexity**
- Defined as the computational complexity of checking if their $\textsc{Fo}$ meaning representations are **satisfiable** ("Entscheidungsproblem")
- It turns out that in general
  1. fragments that contain **either negation or relatives**, but not both have **tractable** (polynomial) complexity
  2. fragments that cover **both negation and relatives**, have **intractable** (exponential) complexity $\Rightarrow$ encode Boolean satisfiability
- Semantic complexity: measured in terms of the signature of the expressed $\textsc{Fo}$ formulas ($\approx$ names, adjectives, nouns and verbs present in the sentence)