Lancaster University

July 24th, 2015

Corpus Linguistics 2015

Stretching corpora to their limits: research on low-frequency phenomena

Daniel Ross **djross3@illinois.edu**
University of Illinois at Urbana-Champaign

The size of modern corpora is impressive…

- Google Ngrams: American English **155 billion words**
  http://googlebooks.byu.edu/x.asp

- Corpus of Global Web-based English (GloWbE)
  - **1.9 billion words** across 20 dialects

  http://corpus.byu.edu/glowbe/

Corpora as a tool for theoretical questions…

- Large corpora can provide information about low-frequency phenomena not easily observed otherwise

- For questions of grammaticality or the limits of a language or the human language faculty, we must look beyond frequency:
  - We hope that the data gives us an idea whether a certain expression occurs or not
    - This is a lot to ask of real-world data!

The bias of high-frequency data & phenomena

- As argued in Ross (2014), there is a bias in linguistic research toward that which we observe most often
- This has a negative impact on an often implied goal of understanding the limits of a language and the human language faculty
- Such phenomena are sometimes considered "peripheral", and are often ignored

The bias of high-frequency data & phenomena

- As shown in Ross (2014), though, it is specifically these phenomena that increase the complexity of a grammatical system
  - We assume the language is simpler if we ignore them!

- Today I will present a detailed corpus study of one such phenomenon: *try and* pseudocoordination

Variation is crucial as well…

- A great way to understand the language faculty is to look at how data varies
  - Dialectally, diachronically, and in acquisition

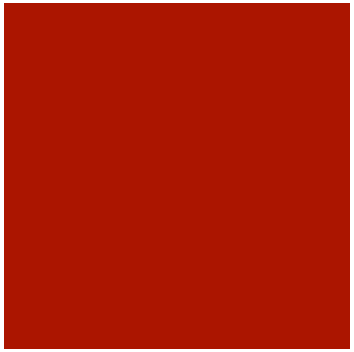
- If data varies, then our linguistic systems must be able to account for this
  - We have positive evidence that such data exists
  - We also know it is not found in all varieties

Variation is crucial as well…

- Corpora can give immediate access to data
  - From dialects, over history, and in acquisition

- But most corpora are based on
  - The standard written variety
  - Contemporary usage by normal adult speakers
  - And usually *English*

Test case: pseudocoordination



- I will try and win the race, but I might not win.
- We try and exercise every day, but we don't always.
- Be sure and proofread articles before submission.
- Remember and brush your teeth!
- Johnny likes to pretend and do his homework.

Always ambiguous

- **I will try and win the race**

- *Reading 1*: I will try, and I will also win!
- *Reading 2*: I will try to win the race; I might not win.

- We are only concerned with the second (as a control verb)

Try specifically…

- Most frequent verb in the construction
  - Properties shared with the other verbs
  - *try* is easiest to study

- *Try* is 127th most frequent word in *COCA*
  - *Corpus of Contemporary American English* (Davies 2008)

- About 10 instances of *try and V* per million words

Another kind of pseudocoordination

- Do you want to go and get something to eat?
  - What do you want to go and eat?
- Yesterday we went and saw a movie.
  - What movie did we go and see?
- Our friend comes and visits us every weekend.
  - Who does he come and visit?

- Note the parallel inflection on both verbs!

Pseudocoordination cross-linguistically

- Pseudocoordination like this is found in a number of languages (Ross 2014, forthcoming)
  - Especially Germanic, Romance and Slavic
  - Also Austronesian, Semitic, Khoisan and more

- A general property is the parallel inflection found on both verbs (e.g., Wiklund 2007)

Morphological restrictions for *try*

- Unlike motion verb pseudocoordination in English and cross-linguistic tendencies, *try* does not appear to be restricted based on parallel forms

- Instead, there seems to be a requirement that neither verb can be inflected (Carden & Pesetsky 1977, inter alia)

- This detail has evaded theoretical analysis

?? Try and use corpora effectively!

?? We will try and use corpora effectively.

?? We try and use corpora effectively.

?? He tries and use corpora effectively.

?? He tries and uses corpora effectively.

?? We tried and use corpora effectively.

?? We tried and used corpora effectively.

?? We try and be good corpus linguists.

?? We try and are good corpus linguists.

?? I try hard and use corpora effectively.

Try and use corpora effectively!

We will try and use corpora effectively.

We try and use corpora effectively.

*He tries and use(s) corpora effectively.

*We tried and use(d) corpora effectively.
We did try and use corpora effectively.

We try and be/*are good corpus linguists.

? I try hard and use corpora effectively.

Generalizations for standard English

- These generalizations have been verified by several grammaticality judgment surveys with native speakers
  - No significant variation found in major varieties (American, British, Australian, Canadian, etc.)
- They can also be shown with corpus data

Previous corpus research

- Lind (1983) finds "the main difference between *try and* and *try to* is one of syntax rather than semantics", with *try and* limited to certain syntactic structures and both associated with certain syntactic structures
- Biber et al. (1999) find *try and* is informal and used especially to avoid the sequence *to try to*…
- Hommerberg & Tottie (2007) likewise find no semantic difference, but find it is more common in spoken and British vs. American English. Only in spoken British English is it more frequent than *try to* (about 70% of the time).
  - Tottie (2012): historical development (more soon…)
- Maia (2012): similar results to the other studies

The *Bare Form Condition*

- Explained by 3 grammatical properties (Ross 2014)
  1. The second verb is a bare infinitive
  2. Both verbs must have the same morphological inflection
  3. The first verb must agree with the subject

- This restriction, especially of inflectional parallelism, is unusual
- Shown by previous research to be consistent and widespread
- I want to fully understand it, with data from:
  - Historical development, dialectal variation and acquisition

Cast study 1: Historical development

- Among prescriptivist accounts, there is a recency fallacy that *try and* is new and in increasing use
  - these comments can be found as early as 1864
- More recently, it has been claimed that *try and* developed during the Great Complement Shift, during the 1800s (Rohdenberg 2003, Vosberg 2006)
  - Specifically, this is said to be due to *horror aequi*:

    I want **to** try and/**to** read that book.

    We will sing, **and** try **and**/to dance.
- *Horror aequi* is still a factor today (Biber et al. 1999)

Cast study 1: Historical development

- However, regardless of *horror aequi*, *try and* did not develop during the 1800s!
  - The OED has an example from 1686
  - Hommerberg & Tottie (2007), Tottie & Hoffman (2011) and Tottie (2012) claim *try and* actually predates *try to*
- The earliest examples of *try and* and *try to* are found in the 1500s.
- Tottie (2012) looks at the sequences "try and" and "try to" in the *Early English Books Online* (EEBO) Corpus and finds use of "try and" predates "try to"

Cast study 1: Historical development

- But Tottie's results were not manually filtered and many are not genuine instances of verbal complementation with *try and*, but literal usage, which is also its source:

  This is not to be tried by the Fathers: but it is to trie and examine the Fathers them selues. (EEBO: Stapleton 1566)

- Ambiguous cases like this were sometimes reanalyzed as verbal complementation

- Around the same time, *try to* developed
  - This related to a general shift in semantics for *try*

- In Ross (2013), I found it hard to tell which was first…

Cast study 1: Historical development

- The manually filtered results are still problematic due to ambiguity:
- Between 1500-1600 in the EEBO corpus:
  - 279 instances of *try and [verb]*
  - 47 instances of *try to [verb]*
- But the majority of the *try and* instances are ambiguous
- The first unambiguous examples of each are found 1550-1600, with *try to* found toward the end of that period
- We cannot conclude that *try and* predates *try to*, and even if so, by less than 50 years

try and [verb]	instances
pseudocoordination	5
ambiguous	186
not pseudocoordination	87
total	**279**
try to [verb]	instances
infinitive complement	34
ambiguous	6
not infinitive compl.	6
total	**47**

Cast study 1: Historical development

- Google Ngrams data for *try and* and *try to*, 1600-2000:

Cast study 1: Historical development

- In the earliest usage, *try and* was limited to usage in non-finite contexts (infinitives and imperatives)
- But there was a major development in the mid-1800s

  Do sit down by the fire, whilst I try and get you some breakfast. (Google Ngrams: Gascoigne 1841)
- Examples of this sort are found from the mid-1800s onward and show further grammaticalization
- The change is also supported by prescripivist commentary from that time (e.g., Waddy 1889)

Cast study 1: Historical development

Summary

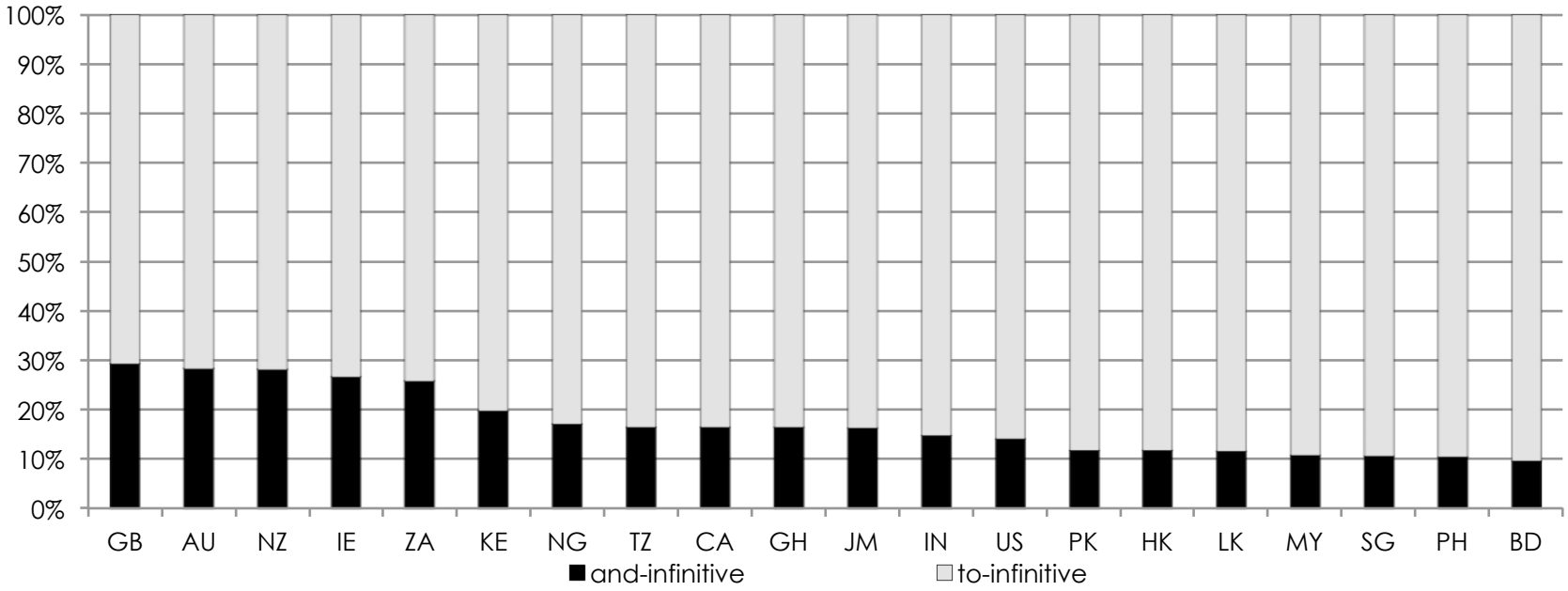
- *Try and* has a nearly 500 year history
- It came about around the same time as *try to* (c.1550)
- It grammaticalized from ambiguous usage
- The *bare form condition*:
  - Originally was associated with non-finite usage
  - Extended during the 1800s to uninflected finite usage

Cast study 2: Dialectal variation

- *Try and* has been in usage since before any major international varieties of English developed, but only later did the modern *bare form condition* arise
- Is this grammatical anomaly found consistently across dialects, or has it further changed?
- We will explore this with the *Corpus of Global Web-based English* (GloWbE: Davies 2013), with 1.9 billion words of informal written English across 20 dialects

Cast study 2: Dialectal variation

- There is variation regarding the frequency of *try and*:



Cast study 2: Dialectal variation

- But the grammatical properties are consistent: the *bare form condition* holds strongly across all of the dialects

	try-and-V	*try-to*-V
Bare	67888 (7%)	282359 (30%)
Inflected	64 (.007%)	595195 (63%)

Cast study 2: Dialectal variation

Summary

- The frequency of *try and* varies by dialect
- The *bare form condition* is consistent across dialects
- Only one confirmed case of dialectal variation
  - Faarlund & Trudgill (1999), in Norwich English, where third-person singular –s is optional: ***He try and see us every day***
  - The *bare form condition* still applies

Cast study 3: Acquisition (L1)

- Given the unusual nature of the *bare form condition*, is it learned early and easily? Lack of dialectal variation suggests they should be able to learn it without difficulty
- How early do children acquire *try and*?
- Do they learn *try and* or *try to* first?
- Do they make mistakes by inflecting *try and*?
  - The *grammatical conservativity* hypothesis predicts they will not overgeneralize: they make errors of omission but not of *comission* (producing elements not found in adult speech) (Snyder 2008; Sugisaki & Snyder 2013)

Cast study 3: Acquisition (L1)

- In the CHILDES database (MacWhinney 2000), I found only two corpora with more than one or two instances of *try and*.
- Both are British English, likely due to the higher frequency
- One is cross-sectional, with 72 children
- One is longitudinal, with one child
- Each barely has enough data for statistical results

Cast study 3: Acquisition (L1)

- The Fletcher corpus (Fletcher & Garman 1988; Johnson 1986) has data from 72 children:
  - Ages: 3, 5 and 7
  - One recording each
  - Informal, unstructured interviews

  - *Statistical test used: Fisher's exact test*

Cast study 3: Acquisition (L1)

- The children use *try to* by age 3 and *try and* by age 5
- The *bare form condition* in fact appears categorical (p<.05)

3 years	*try-and*-V	*try-to*-V
Bare	0	0
Inflected	0	4 (6)
5 years	*try-and*-V	*try-to*-V
Bare	2	0
Inflected	0	6 (10)
7 years	*try-and*-V	*try-to*-V
Bare	4 (8)	0
Inflected	0	6 (12)

(By child, with total instances in parentheses.)

Cast study 3: Acquisition (L1)

- The Thomas corpus (Lieven, Salomo & Tomasello 2009) has data from one child:
  - One recording per week at age 2
  - One recording per month ages 3 and 4
  - Informal, unstructured interviews

  - *Statistical test used: Fisher's exact test*

Cast study 3: Acquisition (L1)

- Thomas uses *try and* and *try to* starting at age 2
- The *bare form condition* is consistent (p<.001 ages 3-4; p<.1 age 2)

2 years	*try-and*-V	*try-to*-V
Bare	2	0
Inflected	0	3
3 years	*try-and*-V	*try-to*-V
Bare	6	5
Inflected	0	35
4 years	*try-and*-V	*try-to*-V
Bare	15	3
Inflected	0	31

Cast study 3: Acquisition (L1)

- Thomas makes two errors, but these actually support the *bare form condition*:

It feels like a crab try and get you.

(try = is trying; 4 years, 7 months)

I trying to concentrate.

(trying = try / am trying; 2 years, 11 months)

Cast study 3: Acquisition (L1)

Summary

- Children learn *try and* and *try to* early
- They learn the *bare form condition* consistently
  - The data suggests they may initially learn a categorical difference between *try and* and *try to* based on inflection
- The results are consistent with *grammatical conservativity*: the children do not produce inflected forms when adults would not

Epilogue: Faroese ‘try and’

- Faroese has a construction equivalent to English *try and*, *royna og* ‘try and’, which has been reported to have similar inflectional restrictions (Heycock & Petersen 2012)
- Faroese corpora are more limited, however
  - Faroese is a North Germanic language spoken by about 55,000 in the Faroe Islands (North Atlantic)
  - Several corpora did not produce any results for *royna og [verb]*
- The best corpus was *Føroyskt TexstaSavn*:
  - about 4 million words
  - The 1998 year of Faroese newspaper *Dimmanlætting*

Epilogue: Faroese 'try and'

- The results from this corpus were inconclusive, however:

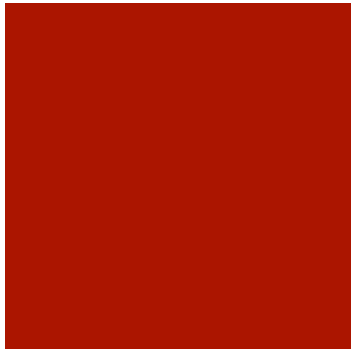
- Only 10 tokens of *royna og [verb]* were identified:
  - 9 were imperatives (both singular and plural)
  - 1 was an infinitive

- This did not provide new evidence for Faroese, and I later did field work and internet surveys to gather data
  - The results reveal a similar grammatical restriction Faroese, but it isn't a complete ban on inflection, rather looking like non-finite forms
  - Grammaticalization in Faroese today is like English in the mid-1800s

Conclusions

- Research on a specific syntactic construction from just a single, though frequent, verb is possible in English
  - Several of the results were difficult to find in available corpora
  - Generally, I felt that I barely managed to find conclusive data when I did
- For other languages, resources are much more limited
- Searching for ambiguous usage in a corpus is challenging; even manual filtering may not resolve it
- Corpus data from a variety of language contexts (dialectal variation, historical development and acquisition) can inform linguistic theory and our knowledge of the human language faculty

Thank you!



Questions?

Daniel Ross
djross3@illinois.edu

References

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Carden, Guy & David Pesetsky. 1977. Double-Verb Constructions, Markedness, and a Fake Co-ordination. *Chicago Linguistics Society* 13. 82–92.

Davies, Mark. 2008. The Corpus of Contemporary American English: 450 million words, 1990-present. http://corpus.byu.edu/coca/.

Davies, Mark. 2013. Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries. http://corpus2.byu.edu/glowbe/.

EEBO TCP. Early English Books Online - Text Creation Partnership. http://www.textcreationpartnership.org/tcp-eebo/.

Faarlund, J. T. & P. Trudgill. 1999. Pseudo-coordination in English: the "try and" problem. *Zeitschrift für Anglistik und Amerikanistik* 47(3). 210–213.

Fletcher, Paul & Michael Garman. 1988. Normal language development and language impairment: Syntax and beyond. *Clinical Linguistics & Phonetics* 2(2). 97–113. doi:10.3109/02699208808985246.

Føroyskt TekstaSavn. Faroese text collection by Språkbanken and Fróðskaparsetur Føroya. http://spraakbanken.gu.se/FTS/ (13 January, 2015).

Google Ngrams. (Michel et al. 2011). http://books.google.com/ngrams/.

Heycock, Caroline & Hjalmar P. Petersen. 2012. Pseudo-coordination in Faroese. In K. Braunmueller & C. Gabriel (eds.), *Multilingual Individuals and Multilingual Societies*, 259–280. Hamburg: John Benjamins.

Hommerberg, Charlotte & Gunnel Tottie. 2007. *Try to* or *try and*? Verb complementation in British and American English. *ICAME Journal* 31. 45–64.

References

Johnson, M. G. 1986. A computer-based approach to the analysis of child language data. Reading, UK: University of Reading Ph.D dissertation.

Lieven, Elena, Dorothé Salomo & Michael Tomasello. 2009. Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics* 20(3). 481–507. doi:10.1515/COGL.2009.022.

Lind, Åge. 1983. The variant forms *try and*/*try to*. *English Studies* 5. 550–563.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for analyzing talk*. Third edition. Mahwah, NJ: Lawrence Erlbaum Associates. http://childes.psy.cmu.edu/.

Maia, Jefferson de Carvalho. 2012. Complementation patterns of the verb *try*. *Revista Virtual dos Estudantes de Letras (ReVeLe)* 4. http://www.periodicos.letras.ufmg.br/index.php/revele/article/view/3945.

Rohdenburg, Günter. 2003. Cognitive complexity and horror aequi as factors determining the use of interrogative clause linkers in English. In Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of Grammatical Variation in English*, 205–249. Berlin: Walter de Gruyter.

Ross, Daniel. 2013. Dialectal variation and diachronic development of *try*-complementation. *Studies in the Linguistic Sciences: Illinois Working papers* 38. 108–147.

Ross, Daniel. 2014. The importance of exhaustive description in measuring linguistic complexity: The case of English *try and* pseudocoordination. In Frederick J. Newmeyer & Laurel B. Preston (eds.), *Measuring Grammatical Complexity*, 202–216. Oxford: Oxford University Press.

References

Snyder, William. 2008. Children's grammatical conversativism: Implications for syntactic theory. In Tetsuya Sano, Mika Endo, Miwa Isobe, Koichi Otaki, Koji Sugisaki & Takeru Suzuki (eds.), *An enterprise in the cognitive science of language: a festschrift for Yukio Otsu*, 41–51. Tokyo: Hituzi Syobo.

Sugisaki, Koji & William Snyder. 2013. Children's Grammatical Conservatism: New evidence. In Misha Becker, John Grinstead & Jason Rothman (eds.), *Language Acquisition and Language Disorders*, 291–308. Amsterdam: John Benjamins.

Tottie, Gunnel. 2012. On the History of *try* with Verbal Complements. In Sarah Chevalier & Thomas Honegger (eds.), *Word, Words, Words: Philology and Beyond: Festschrift for Andreas Fischer on the Occasion of his 65th Birthday*, 199–214. Tübingen: Narr Francke Attempto.

Tottie, Gunnel & Sebastian Hoffmann. 2011. Which came first, *try to* or *try and*? A chicken-and-egg story. Oslo, Norway.

Vosberg, Uwe. 2006. *Die Grosse Komplementverschiebung: aussersemantische Einflüsse auf die Entwicklung satzwertiger Ergänzungen im Neuenglischen*. Tübingen: Gunter Narr Verlag.

Waddy, V. 1889. *Elements of composition and rhetoric with copious exercises in both criticism and construction*. Cincinnati, New York: Eclectic Press, Van Antwerp, Bragg and Company.

Wiklund, Anna-Lena. 2007. *The syntax of tenselessness: tense/mood/aspect-agreeing infinitivals*. Berlin: Mouton de Gruyter.