Corpus Linguistics 2015

Abstract Book

Edited by

Federica Formato and Andrew Hardie

Lancaster: UCREL

Online edition 21st July 2015, with errata corrected

Table of contents

Plenaries

Papers

Ortega, L. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. Applied Linguistics 24/4: 492-518.

Taguchi, N., Crawford, W., and Wetzel, D. 2013. What Linguistic Features Are Indicative of Writing Quality? A Casse of Argumentative Essays in a College Composition Program. TESOL Quarterly, 47 (2), 420 430.

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

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

1 Introduction

Exactly what can we measure with today's corpora? Can corpora act as a proxy to specifically-designed datasets across a variety of contexts? As argued in Ross (2014), linguistic research tends to be biased toward high-frequency phenomena, meaning that we tend to only understand the most common features in languages rather than exploring the full capacity of the human language faculty. Corpus methods are one strategy to address this issue.

On the one hand, corpora are optimal for research on low-frequency phenomena because they provide direct empirical evidence in the form of millions or even billions of words. On the other hand, such large corpora are available only for a small range of linguistic varieties: usually English, usually the standard written variety, and usually contemporary usage by normal adult speakers. Sufficient corpus size is necessary for both finding relevant data and making statistical generalizations.

Therefore here I discuss the difficulties and possibilities associated with researching a particular low-frequency construction in corpora representing historical, dialectal and acquisition data for English, with implications for other languages as well. The outlook is optimistic, with such research just barely possible with the modern corpora available today.

2 The *try-and*-V construction

The *try-and*-V construction is a particular instance of a general control-verb pseudocoordination construction in English. Although several other subject control verbs such as *be sure* and *remember* can appear as the first verb in the construction (Ross 2014:211), they are too infrequent, especially in written usage, to be thoroughly investigated and statistically analyzed in most corpora. However, we can reasonably investigate the construction through its usage with *try*: the verb *try* is the 127th most frequent word in the *Corpus of Contemporary American English* (COCA: Davies 2008), with about 10 instances of *try-and*-V per million words.

Pseudocoordination has caught the attention of a number of researchers because it appears to be a mismatch between syntax (coordination) and semantics (subordination) and displays several unusual morphosyntactic properties (Ross 1967;

Culicover & Jackendoff 1997; Wiklund 2007; among others). In English, there are two types:

(1) He went and saw the movie.

(2) We will try and use corpora effectively.

The former, found with motion verbs, can be used with any morphological inflection as long as that inflection is found on both verbs (cf. Wiklund 2007). The latter, found with control verbs, may only be used in contexts with bare, uninflected verbs (Carden & Pesetsky 1977) such as imperatives, infinitives and the present-tense (except third-person-singular):

(3) We try and use corpora effectively.

(4) *He tries and use(s) corpora effectively.

(5) *We tried and use(d) corpora effectively.

This *Bare Form Condition* (BFC) can be generated by two independent properties (Ross 2013, 2014): that the second verb is a true, bare infinitive; and that the first verb must have parallel morphology to that second, necessarily uninflected verb, analogously to the requirement in motion verb pseudocoordination.

Try-and-V is frequent enough to be studied in corpora of standard English and there have been several successful studies (Lind 1983; Hommerberg & Tottie 2007; Maia 2012), which indicate that the construction is more frequent in spoken English and more frequent in British English than American English. Only in spoken British English is it used more often than *try-to*-V (about 70% of the time). Additionally, the BFC is widespread and consistent.

Below I present three case studies looking at the BFC beyond adult usage of contemporary, standard English, stretching corpora to their limits but with successful results showing that the BFC is robust.

3 Case study 1: Historical development

Although claimed to be a relatively recent phenomenon by some and dismissed as a modern error by prescriptivists, *try-and*-V has a nearly 500 year history in English having developed alongside *try-to*-V (Hommerberg & Tottie 2007; Tottie 2012).

Tottie (2012:210) claims that *try-and*-V predates *try-to*-V with raw frequencies of the sequences *try and [verb]* and *try to [verb]* in the *Early English Books Online* (EEBO) corpus, but this claim is problematic when the data is manually filtered.

The first task in research for this time period is finding a corpus with enough data; EEBO is sufficient, but without part of speech tagging this potentially ambiguous construction is challenging. The raw sequence *try and [verb]* might be normal coordination (*try and fail*), not complementation via pseudocoordination (*try and[=to] win*), with this ambiguity being the source of the construction:

(6) I will aduenture, or trie and seeke my fortune. (Baret 1573; Tottie 2012:207)

An automated search listed all instances of *try* during the 1500s in EEBO, including spelling variation, which were manually filtered to consider only those instances with *and* or *to* followed by verbs that could potentially appear to be infinitival complements. Of those, many were still ambiguous, as shown in Table 1.

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**

Table1: *try and/to* in EEBO 1500-1600

The results reveal that though both *try-and*-V and *try-to*-V date to the 1500s, there is no conclusive evidence that *try-and*-V is older or was more frequent at first because the majority of its instances were ambiguous during this period. We can only conclude that ambiguous contexts with *and* were more frequent than ambiguous contexts with *to*.

At this time, *try-and*-V was limited to non-finite contexts (infinitives and imperatives); the modern version of the BFC developed during the mid-1800s with present-tense usage (Ross 2013:120).

4 Case study 2: Dialectal variation

Although comparisons have been made between British English and American English, other dialects, where there might be significant variation, are more difficult to explore. The recently released *Corpus of Global Web-based English* (GloWbE: Davies 2013), with 1.9 billion words of informal written English from 20 dialects, provides an appropriate data set. After automated searching with part-of-speech tagging and manual filtering of formally ambiguous results, the BFC is shown to be ubiquitous and nearly exceptionless (Table 2).

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

Table2: Infinitive complements of *try* in GloWbE

Across all dialects there are only 64 instances of inflected *try* in the construction. Of these, 46 had a bare second verb, possibly by analogy to *try-to*-V. No dialect frequently uses inflected *try-and*-V. In other, smaller dialects there may still be room for variation, especially in those with non-standard present-tense paradigms (Faarlund & Trudgill 1999) or known exceptions to the requirement for parallel inflection in motion verb pseudocoordination (Rosen 2014). Larger corpora are needed for these dialects.

5 Case study 3: Acquisition in children

The BFC is widespread and historically stable, but is it easily and consistently acquired by children? The corpora available in the CHILDES database (MacWhinney 2000) reveal that it is. No instances of inflected *try-and*-V were found in CHILDES. However, to test this statistically, a single corpus with sufficient tokens of *try-and*-V is required. Most of the corpora contained no more than two instances, but two were identified that were just large enough for this study. Both were samples of British English, where the construction is especially frequent.

First, the Fletcher corpus (Fletcher & Garman 1988; Johnson 1986) was examined, with cross-sectional data from 72 children ages 3, 5 and 7. As shown in Table 3, not only did the children not violate the BFC (statistically significant by Fisher's exact test at p<.05 for 5-7 years), but may have even acquired a categorical difference: *try-and*-V is uninflected, and *try-to*-V is inflected.

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)

Table3: *try and/to* in the Fletcher corpus
(By child, with total instances in parentheses.)

Then the Thomas corpus (Lieven, Salomo & Tomasello 2009) shows that the BFC is acquired early and consistent by a single child, recorded weekly at age 2, then monthly for ages 3 and 4. There are no violations of the BFC, and the lack of inflected *try-and*-V for ages 3 and 4, shown in Table 4, is statistically significant (p<.001).

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

Table4: *try and/to* in the Thomas corpus

This evidence supports the *grammatical conservativity* hypothesis (Sugisaki & Snyder 2013), which states that children will make errors of omission, but few of *comission* (producing elements not found in adults speech).

6 Outlook

Research on a specific syntactic construction based on data from only a single, though frequent, verb is possible but difficult in English. But for other languages resources are needed: for example, Faroese *royna-og*-V (counterpart to *try-and*-V) may exhibit the BFC (Heycock & Petersen 2012:274), but available corpora are limited, such as *Føroyskt TekstaSavn* (about 4 million words) with only 10 instances (9 imperatives and 1 infinitive).

References

Carden, G. & Pesetsky, D. 1977. Double-Verb Constructions, Markedness, and a Fake Co-ordination. *Chicago Linguistics Society* 13: 82–82.

Culicover, P.W. & Jackendoff, R. 1997. Semantic subordination despite syntactic coordination. *Linguistic Inquiry* 28(2): 195–217.

Davies, M. 2008. The *Corpus of Contemporary American English*: 450 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

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

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

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

Fletcher, P. & Garman, M. 1988. Normal language development and language impairment: Syntax and beyond. *Clinical Linguistics & Phonetics* 2(2): 97–113.

Føroyskt TekstaSavn. Faroese text collection by Språkbanken and Fróðskaparsetur Føroya. Available online at http://spraakbanken.gu.se/FTS/search.phtml. (Accessed January 13th, 2015.)

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

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

Johnson, M.G. 1986. *A computer-based approach to the analysis of child language data*. Unpublished PhD thesis, University of Reading.

Lieven, E., Salomo, D. & Tomasello, M. 2009. Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics* 20(3): 481-507.

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

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

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

Rosen, A. 2014. *Grammatical variation and change in Jersey English*. Amsterdam: John Benjamins.

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

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

Ross, J.R. 1967. *Constraints on Variables in Syntax*. Unpublished PhD thesis, Massachusetts Institute of Technology.

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

Tottie, G. 2012. On the History of *try* with Verbal Complements. In S. Chevalier & T. 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.

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

Investigating the Great Complement Shift: a Case Study with Data from COHA

Juhani Rudanko
University of Tampere

Consider the sentences in (1a-b), both from COHA, the Corpus of Historical American English:

(1) a. Would you object to leave home? (1890, FIC)
b. I object to signing such an order. (1891, FIC)

In (1a) the matrix verb *object* selects a *to* infinitive complement, and in (1b) the sentential complement of the same matrix verb is what may be termed a *to -ing* complement, consisting of the preposition *to* and a following gerund. While the examples from COHA show that both types of complements were found in fairly recent English, the infinitival variant has become very rare, or even unacceptable, in current English.

The purpose of the paper is to investigate sentential complements of the matrix verb *object* during the entire time span of COHA, in order to shed light on the two types non-finite complements. To set the stage, the theoretical distinction between the two types of constructions, illustrated in (1a-b), is discussed first. Both constructions involve the word *to*, but it is argued, contrary to Duffley (2000), that only the *to* that precededs a gerund is a preposition. For its part, the *to* in *to* infinitival constructions is under the Infl node, corresponding to the Aux node in more traditional terminology. While some scholars have taken the *to* of *to* infinitives to be a semantically empty element, it is argued that this *to*, similarly to other elements under Infl, may carry a meaning.

A first objective in the empirical part of the study is to provide frequency information on the incidence of the two types of complement, as selected by the matrix verb *object*, in the last two centuries, that is, during the entire time span of COHA, up to 2009. The research tasks here are to find out how long the two complements coexisted side by side and what their frequencies were in each decade. A further task is to identify the period when the gerundial pattern came to prevail over the infinitival pattern.

A second objective is to inquire into the factors that may have played a role in favoring either type of complement during the time when both were found in reasonable numbers in the language. Questions to be investigated include the possibility of semantic differentiation of the two patterns, in the