Richard Futrell

Predicting Gaps

1. What is the Accessibility Hierarchy?

Keenan and Comrie (1977) present a typological generalization about relative clause formation in the world's languages. The structures they examine consist of two parts: one that specifies a set of objects (syntactically a head noun), and one part that restricts the interpretation of the head noun to some subset of which a certain sentence is true (the restricting clause). So in (1) below, the italicized head nouns presents a set of possible entities, and the underlined restricting clauses limit the interpretation of the head noun to a subset of those entities (all examples are observations from the Switchboard corpus, unless otherwise noted):

(1) a. I have a lot of *friends* who <u>like to go fishing</u>.

b. the *car* that <u>I bought</u>

c. the *issues* that <u>people really want to know about</u>

The way in which the restricting clause limits the interpretation of the head noun depends on the syntactic position that is missing in the restricting clause. So for instance the restricting clause in (1a), "like to go fishing", lacks a subject, so we know that the noun phrase of (1a) can only refer to "friends who like to go fishing". Similarly, the verb "bought" in (1b) lacks an object that we would otherwise expect, so we know that the noun phrase can only refer to *cars* such that it is true that "I bought the *cars*." We say that the relative clause "relativizes from" the missing position. For instance, (1a) relativizes from subject position, (1b) from object position, and (1c) from a PP adjunct.

Based on a sample of 50 languages, Keenan and Comrie demonstrate that certain limitations apply to the syntactic positions that can be relativized from, defining an "Accessibility Hierarchy" such that:

(2) Any relative clause-forming strategy must apply to a continuous segment of the Accessibility Hierarchy (AH).

(3) Strategies that apply at one point of the AH may in principle cease to apply at any lower point.

The hierarchy itself is found empirically in the language sample:

(4) Sbj > Direct Obj > Indirect Obj > Oblique > Genitive > Object of Comparison (OComp) The hierarchy is an ordered list of syntactic positions within the restricting clause which provide the restriction on the head noun. Essentially, one can read the symbol > in the hierarchy as 'is more accessible to relativization than'. So any language that allows (1b), an object relative, will also allow subject relative (1a), but not necessarily oblique relative (1c); whereas a language that allows (1c) will allow both (1b) and (1a).

The authors provide a processing explanation for the persistence of the AH cross-linguistically. Relativization of OComp is allegedly harder to process than relativization of a subject, so a grammatical strategy powerful enough to handle a relativized subject may not be powerful enough to handle a relativized OComp. Their specific proposals are 1. that elements that are usually mandatory are more likely to relativize and 2. that syntactic positions that require independent reference are easier to relativize from.

2. **Some possible processing explanations**

I will examine some processing explanations for the AH inspired by the treatment of filler-gap dependencies as constructions in Sag (2009), who explores the restrictions on the various kinds of filler-gap dependencies, highlighting (among others) restrictions on what syntactic or semantic category can appear as parts of the different constructions.

A grammatical construction may have both 'hard' and 'sboft' restrictions on what linguistic elements can combine into it. A 'hard' restriction might, for instance, prohibit an adjective from being the subject of a sentence, as in the ungrammatical (5):

(5) *Treated can really make you sick.

'Soft' or probabilistic restrictions are not categorical. For instance, in English, inanimate nouns in subject position, as in (6), are fully acceptable, but relatively rare (making up 4% of subjects in the Switchboard). The typical subject is not inanimate.

(6) Treated wood can really make you sick.

This 'soft' restriction in English surfaces as a 'hard' restriction in other languages, such as the various Pacific Northwest languages that disallow inanimate subjects. In these languages, it would only be grammatical to express (6) as in (7):

(7) You can really be made sick by treated wood.

Overall, one can describe potential occupants of a constructional position using a probability distribution. Elements that are rare in this distribution for one language may be entirely prohibited in the equivalent distribution in another language. Presumably, a construction containing elements that violate probabilistic restrictions will be harder to process than constructions containing only canonical elements; the avoidance of difficulty in processing can then lead to ungrammaticality of those elements.

Turning to the syntactic phenomenon at hand, the relation between head noun and restricting clause can be considered one kind of construction relating a "filler" (the head noun) to a "gap" (the missing

syntactic element in the restricting clause). Critically, every filler-gap construction takes place within the context of some other construction. For instance, the gap in (1a) occurs in the context of the subject construction, and the gap in (1c) occurs in the context of the PP adjunct construction. Both the gap-containing construction and the filler-gap construction impose categorical and probabilistic restrictions on what can constitute them.

If the combination of restrictions from the two constructions is in conflict, then processing difficulty should result, and the combination should be barred in some languages. Suppose for the sake of illustration that relative clause fillers/gaps tend to involve animate nouns, and PP adjunct constructions tend to involve inanimate nouns. Then a relativized PP adjunct, the combination of the two constructions, due to the combination of restrictions, cannot easily accommodate any nouns. It will be exceedingly difficult to process, and we can expect its equivalent construction-combination to be ungrammatical in many languages.

Another way of viewing this hypothesis is in terms of prediction. Once a speaker or hearer is in a PP adjunct construction, she predicts that the coming noun should probably be inanimate. If a gap, which is typically animate, must occur next, this will be cause uncertainty and surprisal about the following word—and countless studies have shown that uncertainty and surprisal are excellent predictors of processing difficulty.

The Accessibility Hierarchy may result, then, from the mismatch between the probability distribution for linguistic elements in relative clause filler-gap constructions, as opposed to the probability distributions for elements in the various constructions in which gaps occur within the restricting clause.

The mismatch between probability distributions is easily quantifiable. But by itself, this measure of processing difficulty for gaps is circular. In the example above, we could claim that PP adjuncts are difficult to relativize for independent reasons, and their divergence of their typical objects from the distribution of typical gaps arises only because PP adjuncts have relativized relatively rarely. There are three ways that I see around this issue. First, we could compare the probability distributions of properties of linguistic elements in certain syntactic positions with some generally applicable property that must apply to all relative clauses. For instance, we may conclude that gaps should occur in positions in restricting clauses where language users expect old information (in terms of information structure): PP adjunct positions may cause language users to expect new information, and thus be an unlikely gap site.

A second method would be to examine differences within the hierarchy: since subjects relativize most easily in the hierarchy, we can expect that divergence from the distribution of subjects might correlate with processing difficulty. As a third way around the circularity issue, we could use not the distribution of

properties of fillers, but rather the distribution of properties of constituents that appear in the same syntactic position as a specific filler. In this paper I will focus on the first method as an exploratory tool, and leave the latter methods for future research.

3. Methods

I will take advantage of the rich annotations of Treebank Switchboard provided by the LINK project, which annotated a hand-parsed fraction of the Switchboard corpus of spontaneous speech for various animacy classes (Zaenen et al., 2004) and for information structure (Calhoun, 2005). Optimally, my dependent variables would include some measure of actual processing difficulty (frequency, reading times, cloze probability, disfluencies, etc.), but for the sake of simplicity, in this exploratory work I will use the AH itself as a dependent variable. Essentially I will be hunting for some measure of processing difficulty that tracks the AH.

I will also only be examining relative clauses of the type that Keenan and Comrie do not call "case-marked." That is, I will not consider pied-piped PPs or case-marked relative pronouns (such as *whose*) in my treatment of obliques, genitives, and objects of comparison. This has the result that all the genitives I examine will be *of*-genitives, as in (8), since *s*-genitives (as in (9)) can only be relativized in a case-marked manner (constructed examples):

(8) The patient *of the doctor* => The doctor who he is a patient of _

(9) The *doctor's* patient => The patient whose doctor _ is John

The LINK annotation also allows me to make a tentative distinction between PP arguments and PP adjuncts.

Nodes for the appropriate syntactic positions were extracted using the tgrep2 queries in Appendix A. All lexically empty nodes were removed. Probability distributions and entropies are calculated from normalized counts in the corpus.

4. Parts of speech

Relative clauses in English modify nouns. While in cases of pied-piping the filler daughter may be a prepositional phrase, the mother node is NP and the constituent being modified is a noun. Since the gap in the restricting clause bears the information of that noun, the gap should appear where speakers expect nouns. I examined the Shannon entropy of the probability distribution over parts of speech for the syntactic positions of the Accessibility Hierarchy. The entropy measures uncertainty about part of speech. Since nouns are the predominant part of speech for each of these positions, we can assume that a low entropy represents low uncertainty that the syntactic position will be occupied by a noun. (I consider pronouns to be nouns for this

specific study.) Other possible parts of speech in these positions are primarily embedded sentences.

Using the entropy of the distributions of parts of speech, we arrive at the hierarchy in (10):

(10)	Hierarchy	SBJ	> IO >	DO >	GEN >	OBL >	OBL(ADJ) >	OC [1]
	Entropy	0.05	0.52	0.74	0.93	1.24	1.43	2.40

The measure of uncertainty about part of speech is wide of the mark for describing the Accessibility Hierarchy. Note that indirect objects come out with less uncertainty than direct objects. Indirect objects are rarely sentences, for instance, rather than nouns. As we will see, indirect objects seem to be the most difficult syntactic position to fit into the correct hierarchy. Where this measure does succeed is in capturing the difficulty of the relativization of objects of comparison. Indeed, the object of comparison is often a number, an adjective, or an adverb; it is only a noun 42% of the time in the Switchboard corpus.

Another approach might be to examine the distribution of kinds of nouns in each syntactic position. Specifically, we might note the prevalence of mandatory resumptive pronouns in languages such as Arabic and the use of resumptive pronouns in some complex relative clauses in English and surmise that gap sites should be in positions where a pronoun is highly likely. Furthermore, pronouns would seem to fit into the information structure of relative clauses nicely; the gap represents an object that has been mentioned and can be seen as having the filler as its antecedent. (11) summarizes the hierarchy of syntactic positions in order of the probability of a pronoun in that position.

(11)	Hierarchy	SBJ	> IO	>	DO >	OBL >	GEN >	OC >	OBL(ADJ)
	P(Pro|Position)	0.71	0.48		0.31	0.13	0.09	0.06	0.03

The hierarchy generated from pronoun probabilities is slightly better than the first attempt. The only misplaced item is indirect object (and it is not clear where oblique adjuncts should fit into the hierarchy). The problem with indirect objects is again not surprising: dative objects are very often pronouns ("Give me the book!").

5. Information Structure

Perhaps one reason that the probability of pronouns approached the appropriate hierarchy has to do with information structure. Gaps should be given, since contents of the gap are known by the speaker or hearer. As such, gaps are unlikely in positions where speakers expect new, rather than old, information.

---

[1] SBJ = Subject
DO = Direct Object
IO = Indirect Object
OBL = Oblique (all)
OBL(ADJ) = Oblique adjunct
GEN = Genitive
OC = Object of Comparison

Simply examining the probability of a marking of "old" in the LINK corpus in the various syntactic positions, we arrive at the hierarchy in (12):

(12)	Hierarchy	SBJ	>	IO	>	DO	>	OBL	>	GEN	>	OBL(ADJ)	>	OC
	P(Old|Position)	0.299		0.061		0.034		0.031		0.029		0.010		0

The result is similar to (12), except that this method ranks adjunct obliques as more relativizable than objects of comparison. This is likely the case, as (13) certainly sounds more idiomatic than (14), which many speakers find marginal.

(13) The restaurant I ate spaghetti in

(14) The restaurant that Chez Panisse is more expensive than

If we examine the probability of information status that is now "new" (that is, it could be "old" or "mediated," that is, inferable from context), we come closer to Keenan and Comrie's hierarchy:

(15)	Hierarchy	SBJ	>	DO	=	OBL	>	IO	>	GEN	(< OC)	>	OBL(ADJ)
	P(~New|Position)	0.396		0.136		0.136		0.134		0.118	(0.28)		0.073

I do not consider the values for OComp reliable in (12) or in (15), because they are based on 7 data points (as opposed to 245 in the next category by frequency, genitives, and 763 in subjects). Only a fraction of LINK is annotated for information structure, and very few of the annotations touch on objects of comparison.

6. Further research

The methods and approaches here suggest some further research and further hypotheses. The most pressing issue to me seems to be development dependent variables. Here, I reported only whether or not my measures arranged the syntactic positions correctly: but the difference in processability between any two positions is probably variable, so there is a dimension of the data against which my results have not been assessed. One source of a measure of processability of these relative clauses might simply be corpus frequency in a large corpus (it would have to be so to capture relatives based on objects of comparison). Once a suitably rich dependent variable is found, it should be possible to fit the measures I developed here, or some improvement of them, to that dependent variable and establish an actual correlation. Such a model could control for the frequency of the constructions, which likely accounts for much of the processability differences.

All in all, if the hypothesis of probablistic category preferences is correct in explaining some amount of the AH data, then we will have evidence for a thoroughly probabalistic picture of syntax, in which slots in constructions are not filled categorically, but according to probability distributions.

Appendix: tgrep2 Queries

Subject: /SBJ/

Object: /VB/ $. `(/./ !$. /NP/) and `/NP_|SBAR/ > /VP/ $,, /NP_|SBAR/

Indirect object: /NP_|SBAR/ > /VP/ $,, `/NP_|SBAR/

Oblique (PP): /PP/ < (/IN/ !< 'of' !< 'than' $. `/./)

Adjunct oblique: /PP-/ < (/IN/ !< 'of' !< 'than' $. `/./)

Genitive:

@ bad (most|none|much|one|two|all|some|more|less|three|four|five|six|seven|eight|nine|ten|lot|couple|each);

(/PP/ < (/IN/ < of $. `/./) !$ (/N/ << @bad));

Object of comparison:

(/IN/ < 'than' $. `/./)

Works Cited

Calhoun, Sasha, Malvina Nissim, Mark Steedman, and Jason Brenier. 2005. "A framework for annotating information structure in discourse. Proceedings of the Workshop on Frontiers in Corpus Annotations II: 45-52.

Goldsmith, John A. 2007. Probability for linguists. Available at http://humanities.uchicago.edu/faculty/goldsmith/Industrial/Probability.htm

Keenan, Edward and Bernard Comrie, 1977. "Noun Phrase Accessibility and Universal Grammar". *Linguistic Inquiry* 8: 63-99.

Sag, Ivan A. 2009. "English Filler-Gap Constructions". To appear in *Language*.

Zaenen, Annie, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor, and Tom Wasow. 2004. Animacy encoding in English: why and how. In Donna Byron and Bonnie Webber, editors, Proceedings of the ACL Workshop on Discourse Annotation.