The Role of Information Theory in Gap-Filler Dependencies

Gregory Kobele
University of Leipzig
gkobele@uni-leipzig.de

Linyang He
Fudan University
lyhe15@fudan.edu.cn

Ming Xiang
University of Chicago
mxiang@uchicago.edu

1 Introduction

Filler-gap dependencies are computationally expensive, motivating formally richer operations than constituency formation. Many studies investigate the nature of online sentence processing when the filler is encountered before the gap. Here the difficulty is where a gap should be posited. Comparatively few studies investigate the reverse situation, where the gap is encountered before the filler. This is presumably due to the fact that this is not a natural class of dependencies in English, as it arises only in cases of remnant movement, or rightward movement, the analysis of which is shakier and more theory laden than the converse. In languages with *wh*-in-situ constructions, like Chinese, the gap-filler construction is systematic, and natural. Sentences (1) and (2) are declarative and matrix/embedded wh-questions respectively in Mandarin Chinese.

(1) LiuBei zhidao CaoCao ai LuBu
LiuBei know CaoCao love LuBu

'LiuBei knows that CaoCao loves LuBu.'

(2) LiuBei zhidao CaoCao ai shei
LiuBei know CaoCao love who

'Who does LiuBei know that CaoCao love?'
Or: 'LiuBei knows who CaoCao loves.'

Although sentences (1) and (2) have similar word order on the surface, in (2) the in-situ wh-phrase *who* takes scope either over the entire sentence (i.e. the matrix question parse) or at the embedded clause (i.e. the embedded question parse). The scope positions precede the wh-phrase, giving rise to the gap-filler dependencies. Gap-filler constructions raise different problems than do filler-gap ones. In the latter, an item is encountered, which needs to satisfy other (to-be-encountered) dependencies to be licensed. There is no uncertainty *that* a gap must be postulated, only *where* it should be postulated. In gap-filler constructions, a dependency is postulated before the item entering into it appears. In contrast to the filler-gap dependency type, gap-filler dependencies do not require more formal power from the syntax; they can (given a finite upper bound on their number) be analyzed with GPSG-style slash-feature percolation and are thus context-free. In systems with (covert) syntactic movement, the *wh*-mover is predictably silent, and could be optimized away (into the context-free backbone of the derivation tree). The motivation for the postulation of a syntactic dependency is to streamline the account of sentence processing; while a purely semantic scope taking account could be implemented (e.g. using continuations), the role and resolution of semantic information during parsing is not as well understood.

Our goal is to understand the role that information theoretic complexity metrics [3] can play in the analysis of Chinese-like *wh*-in-situ constructions. In particular, whether humans' use of probabilistic cues about the presence of a gap can be modeled using the metrics of surprisal and/or entropy reduction. To this end, we identified a sentence processing data set where such cues were manipulated, wrote a Chinese grammar fragment deriving the stimuli, estimated probabilities from the Penn Chinese Treebank 9.0 [8] (using the Stanford NLP Tregrex [4]), and calculated (using the Cornell Conditional Probability Calculator [1]) surprisal and entropy reduction values at each word. Our results show that complexity metrics computed over abstract syntactic structures are significant predictors of processing cost.

2 The data set

We used a data set from an existing eye-tracking reading experiment (Experiment 1 in [7]). The original experiment consisted of 8 different conditions, which were largely designed to create different scoping possibilities for the in-situ wh-word. We implemented the structural properties of these conditions into our grammatical analysis in section 3, such that every condition could be derived by our grammar. An example of half of the original conditions is given in (3a – 3d).

(3a) Matrix Verb Non-predictive; Lower Verb $+Q$

jizhemen zhidao shizhang toulu-le
Reporter know mayor reveal-perf
shizhengfu yancheng-le naxie-guanyuan
city-council punish which-CL-official

"The reporters knew which officials the mayor revealed that the city council punished."
OR "The reporters knew the mayor revealed which officials that the city council punished."

(3b) Matrix Verb Non-predictive; Lower Verb $-Q$

jizhemen zhidao shizhang huangcheng
Reporter know mayor lie
shizhengfu yancheng-le naxie-guanyuan
city-council punish which-CL-official

"The reporters knew which officials the mayor untruthfully claimed that the city council punished."

(3c) Matrix Verb Predictive; Lower Verb $+Q$

jizhemen xiang-zhidao shizhang toulu-le
Reporter wonder mayor reveal-perf
shizhengfu yancheng-le naxie-guanyuan
city-council punish which-CL-official

"The reporters wondered which officials the mayor revealed that the city council punished."

(3d) Matrix Verb Predictive; Lower Verb $-Q$

jizhemen xiang-zhidao shizhang huangcheng
Reporter wonder mayor lie
shizhengfu yancheng-le naxie-guanyuan
city-council punish which-CL-official

"The reporters wondered which officials the mayor untruthfully claimed that the city council punished."

In the 4 conditions above, the wh-in-situ phrase could either take scope at the highest embedded clause or the lower clause. The matrix verb is manipulated. In the *Matrix Verb Predictive* conditions, the matrix verb *wonder* obligatorily take an interrogative complement clause, and therefore in these conditions the wh-phrase is unambiguously high-scope. In the *Matrix Verb non-predictive* conditions, the matrix verb *know* allows an interrogative complement but does not mandate it. The lower embedding verb is also manipulated. The lower +Q verb, such as *reveal*, is in the same class as *know*; but the lower -Q verb, such as *lie*, blocks the lower scope for the wh-phrase since the verb does not allow an interrogative complement clause. The combination of different matrix and embedding verbs yields the scope-ambiguous 3a , and three unambiguous conditions 3b – 3d.

The original experiment contained four additional conditions, all of which were simpler constructions that only contained one embedded clause. The matrix verb was again either predictive or non-predictive of an upcoming interrogative clause. The embedded clause was either short or long with a control verb predicate. An example is given in (4a – 4d).

(4a/b) Matrix Verb Non-predictive or Predictive; Short

jizhemen (xiang-)zhidao shizhang
Reporter know\wonder mayor
yancheng-le naxie-guanyuan
punish-perf which-CL-official

"The reporters knew\wondered which officials the mayor punished."

(4c/d) Matrix Verb Non-predictive or Predictive; Long

jizhemen (xiang-)zhidao shizhang bangzhu
Reporter know\wonder mayor help
shizhengfu yancheng-le naxie-guanyuan
city-council punish which-CL-official

"The reporters knew\wondered which officials the mayor helped the city council to punish."

In this experiment, participants read sentences silently on a computer screen, and their eye-movements were recorded. The data set consisted of data from fifty native Mandarin speakers, each read 48 critical trials based on the 8 experimental conditions.

3 Grammatical analysis

We use the minimalist grammar (MG) formalism [6] to frame our analysis. This formalism allows for the straightforward and transparent encoding of prominent linguistic ideas into a formal system. The lack of support in the CCPC for covert movement pushed us to adopt a feature movement analysis [2] of the Chinese *wh*-in-situ construction, whereby it is not the *wh*-word itself which moves, but rather just a single (*wh*) feature. This is implemented by deriving a *wh*-word by combining a 'pre-wh-word' with a silent (but otherwise overt) *wh*-moving item. This analysis would allow us to implement the observation that *wh*-words in Chinese can be used as well as indefinites, by relating (derivationally) the *wh*-word and the indefinite, although this did not play a role in our analysis.

The analysis encompasses the four clausal complement selecting verb types in the experimental conditions; control verbs ('help'), declarative complement selecting verbs ('believe' or 'lie'), interrogative complement selecting verbs ('wonder'), and verbs which optionally select either declarative or interrogative complements ('know'). Control structures were analyzed in terms of PRO and null case [5], due to CCPC's lack of support for other alternatives. Verbs selecting interrogative complements selected sentential complements, and then immediately checked a *wh* feature. Verbs which select clausal complements irrespective of their force were given two homophonous lexical entries.

The CCPC forces upon us the (computationally motivated) assumption that only one *wh* feature may be active (i.e. moving) at any given time. Thus upon postulating a *wh* 'gap' (i.e. a covert landing site for *wh*-movement), the parser will categorically rule out the (grammatical im-) possibility that a next word is an interrogative complement selector.

4 Frequency Estimation

The CCPC works by translating MGs to equivalent MCFGs, and then parsing using the MCFG. When multiple rules expand the same non-terminal, we need to assign a (non-unit) weight to these rules. As there is currently no MG (or MCFG) TreeBank for Chinese, we were forced to estimate weights of rules by reasoning about the extant structures in the treebank. Due to the small size of our lexicon, there were only five (non-lexical) non-terminals with multiple rules expanding them.

(5a) T[**+WH**]

(5b) VP (with and w/o *wh*)

(5c) AgrO (with and w/o *wh*)

The distinctions relevant to the probability distribution over derivations are not always the ones of obvious interest to linguists. For example, there were two MCFG rules for constructing TPs with *wh*-moving subexpressions. Both rules involve checking the case of a subject DP, but differ as to whether this subject DP is itself **+WH** or **-WH** (in which case the TP necessarily contains another *wh*-word). What we counted in the Treebank is the relative frequency of TPs/Ss which contain active *wh*-words[1] where this *wh*-word is the matrix subject, vs a non-matrix subject. Similarly, a VP can be constructed either by merging an object with a lexical verb, or a derived structure (in this case, necessarily a control verb plus infinitival complement clause). Finally, the category 'AgrO' is the category with which the logical subject is merged (sometimes called 'little-v' in the syntactic literature). The relevant distinctions here (for the non-*wh* case) are whether the AgrO is created by a VP checking the case of its object, or by an interrogative sentential complement taking verb checking the *wh*-feature of its complement, or by a declarative sentential complement taking verb combining with its declarative complement. We counted the relative frequency of transitive verbs (including control verbs) vs interrogative sentential complement taking verbs vs declarative sentential complement taking verbs in the corpus. The relevant distinctions in the case of a **+WH** AgrO are different. A **+WH** AgrO can be created by checking the case of the object of a verb if either the object itself, or some other expression in the VP, is itself **+WH**. Alternatively, it can be created by a *declarative sentential complement* taking verb merging with its sentential complement which contains a +WH expression.

[1] A TP contains an active *wh*-word just in case it contains a *wh*-word which takes scope outside the TP.

The other point of grammatical non-determinism involved the lexicon. Given multiple lexical items with the same featural makeup, we needed to assign weights to the rules which realize a syntactic feature bundle as a particular lexeme. As our lexemes represent whole word classes (*help* stands for the class of control verbs), the only real non-determinism here was in the choice of sentential complement taking verbs (both **+WH** and **-WH**). We counted (for the **-WH** case) the relative frequency with which declarative sentential complements are embedded under *reveal*, *believe*, and *know*,[2] and *mutatis mutandis* for the **+WH** verbs, *reveal*, *know* and *wonder*.

5 Results and discussion

We focused on 4 different eye-movement measures. *First pass duration* is the sum of all fixations in a region from the eyes first entering the region until leaving it either to the left or to the right. *Go-past* time is the sum of all fixations from first entering a region until leaving the region to the right, including fixations made during regression to earlier parts of the sentence. *Second pass duration* is the sum of all fixations in a region following the initial first-pass fixations. *Total time* is the overall reading time (all fixations) in a given region. For each eye-movement measure, we computed average reading time (RT), collapsing over participants and trials, for each word region under each condition. Next using the CCPC software, the grammar analysis in section 3 and the frequency estimation in section 4, we generated the entropy reduction (ER) and surprisal predictions for each word region under each condition. We then performed four linear regressions, using ER and surprisal as predictors and the four eye-movement measures as dependent variables.

Neither ER or surprisal are significant predictors for the first pass duration ($ps>.5$). For the go-past time, surprisal is not significant ($p>.2$), but ER is ($p<.05$). However, the model with ER as a predictor accounted for very little of the overall variance in the data (adjusted R^2=0.04). For second-pass and total time RTs, both ER and surprisal are significant (*ps* for ER $<.01$; *ps* for surprisal $<.001$). When both predictors are considered in the same model, R^2=0.23 for the second pass measure and R^2=0.32 for the total time measure. When the two predictors are considered separately, surprisal accounted for more variance in the data than ER (R^2=0.17 surprisal vs. 0.05 ER for the total time; 0.13 vs. 0.03 for the second pass time).

If we consider the four eye-movement measures *first pass*, *go past*, *second pass* and *total time* form a scale to measure effects from the earlier stages of processing to the later stages, we observe that for the current data set information-theoretic complexity metrics such as ER and surprisal seem to mostly explain later measures but not the early ones. With the second pass and total time measures, although ER and surprisal seem to have only accounted for a relatively small amount of variance in the data (with surprisal having a better performance than ER), the current results nonetheless demonstrate the independent effect of abstract structure in parsing, decoupled from effects based on lexical information.

[2] Not the ratio of declaratives vs interrogatives embedded, but, given that a declarative is embedded, how frequently it is embedded under one of these vs the others.

References

[1] Zhong Chen, Tim Hunter, Jiwon Yun, and John Hale. Modeling sentence processing difficulty with a conditional probability calculator. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

[2] Noam Chomsky. *The Minimalist Program*. MIT Press, Cambridge, MA, 1995.

[3] John T. Hale. Information theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412, 2016.

[4] Roger Levy and Galen Andrew. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *LREC*, pages 2231–2234. Citeseer, 2006.

[5] Roger Martin. Null case and the distribution of PRO. *Linguistic Inquiry*, 32(1):141–166, 2001.

[6] Edward Stabler. Derivational minimalism. In *International Conference on Logical Aspects of Computational Linguistics*, pages 68–95. Springer, 1996.

[7] Ming Xiang and Suiping Wang. Locality and expectation in mandarin wh-in-situ dependencies. manuscript under review. 2019.

[8] Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238, 2005.