Information Theory and Linearization Patterns of English Phrasal Prepositional Verbs

Michael Collins (The Ohio State University)
mxcollins@gmail.com

Research Question

- Phrasal prepositional verb: V + P + P compound (e.g. run out of, put up with).
- Problem: adverb placement within these structures.

a. I (sometimes run out of) ideas.
b. ?I (run sometimes out of) ideas.
c. ?I (run out sometimes of) ideas.
d. *I (run out of sometimes) ideas.

- Phrasal prepositional verbs display a clear preference for adverb placement at the very beginning of the verb phrase. It is interesting to note that this is NOT a general trend of English verbal structure.

a. I sometimes eat.
b. I eat sometimes.
c. I happily went to the store.
d. I went happily to the store.
etc.

- What, then, is the reason for the idiosyncratic behavior of phrasal prepositional verbs?
- Goal: use Information Theory to model our biases against certain permutations (b, c, d) and our preferences for others (a).
- The present examination of phrasal prepositional verbs is a case study in the more general domain of ***syntactic linearization***, the grammatical mechanism that governs word order in sentences.

Theory

- Information content = $-\log_2 P(X=x_i|C)$, where C is some linguistic context (Shannon 1948).
- Levy & Jaeger 2007 argues that Uniform Information Density (UID) is a general principle of rapid, error-free communication. In essence, their theory claims that language users find large "spikes" in information content undesirable.
- This makes sense if we make the reasonable assumption that communication operates within a noisy channel. If a single linguistic item carrying a massive amount of information is somehow lost, then the message becomes very difficult to comprehend. If all items are weighted relatively equally, however, then it is possible to piece the message back together even if there are errors in transmission or reception.
- Prediction: speakers obey the UID principle and choose the adverbial configuration that results in the most evenly distributed information content.

Methods

- The Corpus of Contemporary American English (COCA; http://corpus.byu.edu/coca/) is a useful tool for studying statistical patterns in English speech.
- Eight phrasal prepositional verbs (run out of, catch up with, get up to, keep up with, put up with, end up with, stand up for, hang around with) were selected for testing. COCA's search tools allow us to calculate the information profile of each adverbial configuration (a) – (d). The phrase "put sometimes up with" is given as an example of how this is done.

(i) Count how many times "put" is followed by an Adv, NP, or P in the corpus as a whole.
(ii) Translate the counts into probabilities (bigrams).
(iii) Translate the probabilities into information content.
(iv) Repeat steps (i) – (iii), but use "sometimes" as the base word in the search query instead of "put." Similarly, the next set of searches would use "up" as the base word.

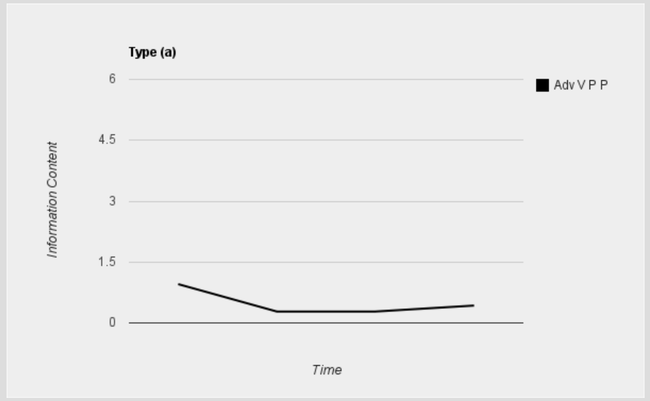
- This process is summarized in the table below.

	Count	Probability	Information Content
put [Adv]	816	0.03	5.06
put [NP]	5145	0.14	2.84
put [P]	29686	0.83	0.27
sometimes [Adv]	3820	0.18	2.47
sometimes [NP]	1887	0.09	3.47
sometimes [P]	4643	0.21	2.25
sometimes [V]	11226	0.52	0.94
up [Adv]			

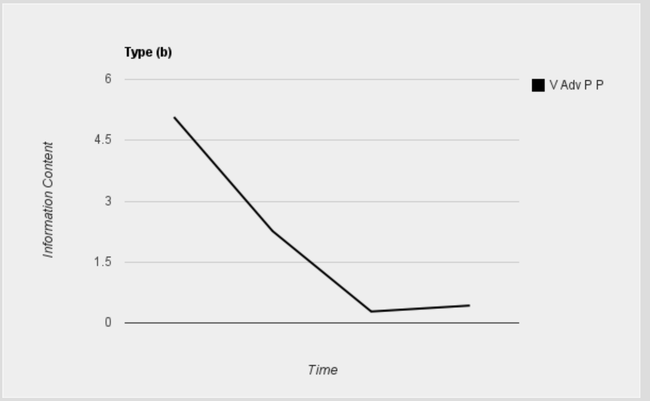
- Note that not all the data in this table will be used in the analysis of this particular phrase. Since, at the moment, we are interested only in "put sometimes up with," the only rows we need are "put [Adv]," "sometimes [P]," "up [P]," and "with [NP]." A different string would, of course, require the use of different data points.

Information Profiles

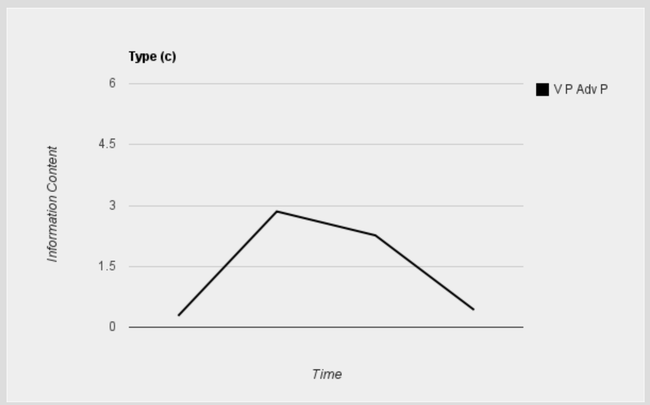
(a) I sometimes put up with nonsense.
Bi: .52 .83 .83 .75
IC: .94 .27 .27 .42



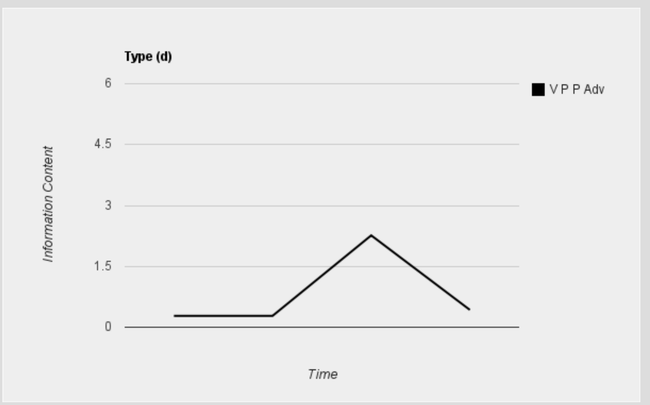
(b) ?I put sometimes up with nonsense.
Bi: .03 .21 .83 .75
IC: 5.06 2.25 .278 .42



(c) ?I put up sometimes with nonsense.
Bi: .83 .14 .21 .75
IC: .27 2.84 2.25 .42



(d) *I put up with sometimes nonsense.
Bi: .83 .83 .25 .75
IC: .27 .27 2.25 .42



Processing Motivations for UID

Preliminaries

- Observe that all eight phrasal prepositional verbs in the test set can "drop" their final preposition and still remain perfectly acceptable lexical items. For example, (put up with) – (with) = (put up).
- Although "put up" does not have the same structural or semantic properties as "put up with," both are still valid English verbs. The same process can be extended further: (put up) – (up) = (put).
- This property forms a crucial part of our explanation for phrasal prepositional verbs' linearization patterns.

Applications

- We hear and parse sentences one word at a time. That is, we do not wait until the speaker is finished with his utterance before analyzing grammatical structures and extracting meaning; comprehension occurs in real-time.
- For example, suppose we encounter the following string: "run sometimes."
- Since an adverb can often occur immediately after a verb, it is natural for us as listeners to parse this string as (V Adv). However, if the utterance continues into "run sometimes out," we are forced to readjust our parse, which results in processing difficulties.

(V Adv) P ??? ??? → (V Adv P)

- However, if the speaker had placed the adverb at the beginning, we as listeners could have "expanded" the verb phrase without undue difficulty.

(sometimes run) in the morning
(sometimes run out) the battery
(sometimes run out of) good ideas

- Note that none of these "expansions" require a dramatic reanalysis of the verbal structure. Thus, comprehension difficulty is reduced when the adverb is in this position.
- In sum, processing considerations provide external motivation for the principle of Uniform Information Density.

Conclusion

- Information Theory and UID can help explain the restricted distribution of adverbs within phrasal prepositional verbs.
- These mathematical tools can also aid in the more general study of syntactic linearization.

References

Levy, R. & F. Jaeger. 2007. Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems* 19: 849-856.

Shannon, C. 1948. A mathematical theory of communication. *Bell Systems Technical Journal* 27: 379-423, 623-656, July, October.

Acknowledgements

I would like to thank Elizabeth Hume for her many helpful comments.