From position to function: Exploring word distributions within intonation units in American English conversation

Ryan Ka Yau Lai, Lu Liu, Haoran Yan and John W DuBois
University of California, Santa Barbara
{kayaulai, lu20, haoranyan, dubois}@ucsb.edu

1 Introduction

Traditionally, Firthian semantics (Firth 1957) examines meanings of linguistic forms through co-occurrences with other forms. This distributional method has enjoyed tremendous success in computational approaches, yet there has been less attention to how forms are distributed within larger units. Discourse markers' functions are often linked to positions in interactional units like turns and sequences (e.g. Sato 2008, Kim 2022, Fuentes-Rodríguez et al. 2016), but other form classes or prosodic units like the intonation unit (IU; DuBois 1992, Chafe 1994, Wahl 2015) are less frequently investigated. In this study, we examine the length of IUs in which words appear and position of words within IUs in the Santa Barbara Corpus of Spoken American English (DuBois et al. 2000), which is manually annotated for IUs based on acoustic cues (DuBois 1992). We find strong systematicity in word distributions across the lexicon, modellable with simple probabilistic models.

2 Exploring prosodic profiles

We first plot the distribution of words within the IU in heatmaps (Figure 1). Most words display clear tendencies as to where they appear in IUs, with three types of patterns. Firstly, words have different length preferences: Interjections prefer very short IUs and prepositions typically prefer longer ones. Secondly, some distributions are centred around a fixed place value, e.g. subject pronouns tend to be first and auxiliaries second. Finally, some distributions are centred around a fixed value from the *end* of an IU: accusative pronouns tend to come last, while determiners and prepositions are typically 1-2 places from the next IU boundary. Some words display bimodal distributions: conjunctions often have one mode near the front of an IU and another, smaller one near the end.

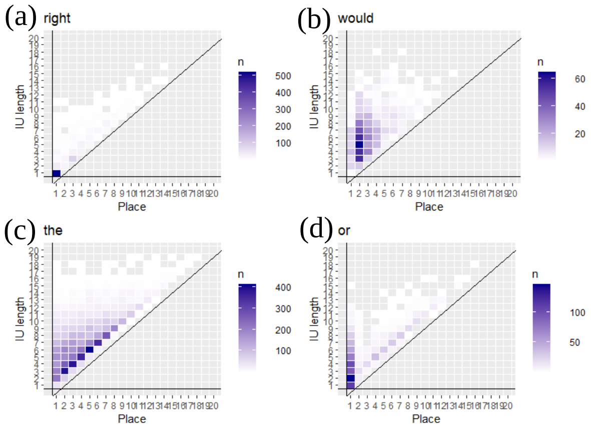

Figure 1: Heatmaps of place and length for the short-biased *right* (a), front-biased *would* (b), end-biased *the* (c) and bimodally distributed *or* (d). The *y*-axis gives the length of the IU where a word appears; the *x*-axis gives the *place*, i.e. sequential position of a word within an IU. The darker a position in the heatmap, the more tokens found in it.

Hierarchical clustering on the joint distributions of the 200 word-types with highest Juilland's *U* (Gries 2008) values, based on Tai & Pham-Gia's (2010) measure of cluster width, reveals syntacto-semantically interpretable clusters. Results at 22 clusters are in the Appendix. Interjections take up two clusters, typically occupying one-word IUs (and occasionally the ends of longer IUs), consistent with their often strong associations with intonation contours (Norrick 2009). At initial positions of longer IUs are conjunctions and other words relating different stretches of discourse, often serving as prefaces (Kim & Kuroshima 2013) to turns. *Wh*-words also tend to come first in an IU and modal-evidential verbs (main and auxiliary) second – words typically described as constituting recognisable turn beginnings in turn-initial position (Schegloff 1996), but the IU-initial tendency remains even in turn-medial positions, e.g. after filtering out uppercase-initial instances. One cluster contains words like *know* and *think* preferring final positions of two-word IUs, reflecting their role in stance-marking chunks like *I think* (Thompson 2002). Words attracted to IU ends include nouns and non-nominative pronouns, projected by words attracted to (ante)penultimate positions like prepositions and determiners.

3 Modelling prosodic profiles

To go beyond exploratory analysis to predictive modelling, we model the words' prosodic profiles with a Bayesian approach, focusing on words with unimodal distribution. We adopt a parametric approach so the distributions can be summarised using a small number of interpretable parameters.

For each word, we first modelled the length of IUs that it appears in using a negative binomial distribution. We use the parametrisation standard in negative binomial regression (Ver Hoef et al. 2007) with the following probability mass function:

$$f(y;\mu,\phi) = \binom{y+\phi-1}{y}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{y}$$

where μ is the mean and ϕ a dispersion parameter; the variance is $\mu(1+\mu/\phi)$. Since 0 places are impossible, we truncated the distribution at 0.

To obtain the joint distribution of place and length, we then modelled the distribution of the place conditional on the length. For the front-biased words, we modelled the place values directly. Since back-biased words tend to be consistently the same number of places from the end of the IU, we model the *back* values of those words by subtracting place from IU length and adding one. The conditional distributions of the place and back values were modelled as Poisson distributions with rate parameter λ, and values below 1 and above the length truncated.

The models were fit in a Bayesian framework in Stan through RStan (Stan Development Team 2023a, 2023b). Priors were set on the parameters as follows: $\lambda \sim Gamma(3,3)$, $\phi,\mu \sim Gamma(1,1)$. The means of the posterior distributions of λ and μ, along with the 'variance' of IU length $\mu(1+\mu/\phi)$, are shown in Table 1 and Table 2 for eight words.

From μ values, which reflect length preferences, clearly *yes* and *right* are much more biased towards short IUs than the rest. This is expected from their functions as interjections: They can function alone to express stance alignment (DuBois 2007) and, for *right*, as backchannels. *Right* has great variance in IU length considering how short the length usually is, reflecting *right*'s secondary use as an adjective.

λ values reveal *yes* and *he* to be most attracted to the edges of IUs, followed by *right*, whereas *the* and *an* are the farthest from IU edges. The interjections' attraction to front edges may allow for early action ascription in the IU, considering their stance alignment functions (cf. Levinson 2012 for similar discussions in the context of turns), and the attraction of *he*, a highly accessible (Ariel 2001) referential expression, to IU beginnings reflects general preferences for producing highly accessible elements first (Levshina 2022). The articles' relatively long distance from the IU edge allows them to project lengthy, inaccessible referential expressions in English.

word	λ	μ	$\mu(1+\mu/\phi)$
yes	1.76	0.22	0.49
he	1.99	6.07	8.80
just	3.20	6.03	11.7
would	3.24	6.65	9.17

Table 1: Parameter estimates for front-biased words. Note that these are not true estimates of means and variances because the distributions are truncated.

word	λ	μ	$\mu(1+\mu/\phi)$
right	2.67	0.58	3.85
an	3.86	6.68	10.23
little	3.73	7.09	10.60
the	4.59	7.01	9.83

Table 2: Parameter estimates for back-biased words.

4 Conclusion and future directions

Words in English conversation reliably pattern as to where they occur in IUs of what length. Some of these distributions can be modelled with simple probability distributions with parameters revealing of the words' functions. This shows location within IUs as a promising avenue for examining linguistic function distributionally, adding to analyses based on collocations and interactional units, perhaps even suggesting refinements of traditional syntax-based word classes like nouns and verbs, while incorporating interjections/discourse markers that do not fit neatly into sentence-based analyses.

We plan to extend these models to account for special words, e.g. those like *'re* or *'m* where initial positions are much less likely than Poisson-like models predict. We also plan to model words with clearly bimodal distributions like *or*. Finally, we hope to compare word distributions within IUs with other units like the turn, turn-constructional unit and sequence, to determine how much additional information IUs capture.

References

Ariel, Mira. 2001. Accessibility theory: An overview. In Ted Sanders, Joost Schilperoord & Wilbert Spooren (eds.), *Text representation: Linguistic and psycholinguistic aspects*, 29–87. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/hcp.8.04ari.

Chafe, Wallace. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press.

DuBois, John W. 1992. Discourse transcription. *Santa Barbara Papers in Linguistics* 4. 1–225.

DuBois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson & Nii Martey. 2000. *Santa Barbara Corpus of Spoken American English*. CD-ROM. Philadelphia: Linguistic Data Consortium. Linguistic Data Consortium.

DuBois, John W. 2007. The stance triangle. In Robert Englebretson (ed.), *Pragmatics & Beyond New Series*, vol. 164, 139–182. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/pbns.164.07du.

Firth, J.R. 1957. A synopsis of linguistic theory, 1930-1955. In J.R. Firth (ed.), *Studies in Linguistic Analysis*. Oxford: Basil Blackwell.

Fuentes-Rodríguez, Catalina, María Elena Placencia & María Palma-Fahey. 2016. Regional pragmatic variation in the use of the discourse marker pues in informal talk among university students in Quito (Ecuador), Santiago (Chile) and Seville (Spain). *Journal of Pragmatics* 97. 74–92. https://doi.org/10.1016/j.pragma.2016.03.006.

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. https://doi.org/10.1075/ijcl.13.4.02gri.

Kim, Hye Ri Stephanie & Satomi Kuroshima. 2013. Turn beginnings in interaction: An introduction. *Journal of Pragmatics* 57. 267–273. https://doi.org/10.1016/j.pragma.2013.08.026.

Kim, Mary Shin. 2022. Identical linguistic forms in multiple turn and sequence positions in Asian languages. *Journal of Pragmatics* 200. 1–7. https://doi.org/10.1016/j.pragma.2022.06.007.

Levinson, Stephen C. 2012. Action Formation and Ascription. In Jack Sidnell & Tanya Stivers (eds.), *The Handbook of Conversation Analysis*, 101–130. https://doi.org/10.1002/9781118325001.ch6.

Levshina, Natalia. 2022. *Communicative Efficiency: Language Structure and Use*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108887809.

Norrick, Neal R. 2009. Interjections as pragmatic markers. *Journal of Pragmatics* 41(5). 866–891. https://doi.org/10.1016/j.pragma.2008.08.005.

Sato, Shie. 2008. Use of "please" in American and New Zealand English. *Journal of Pragmatics* 40(7). 1249–1278. https://doi.org/10.1016/j.pragma.2007.09.001.

Schegloff, Emanuel A. 1996. Turn organization: one intersection of grammar and interaction. In Elinor Ochs, Emanuel A. Schegloff & Sandra A. Thompson (eds.), *Interaction and Grammar*, 52–133. Cambridge University Press. https://doi.org/10.1017/CBO9780511620874.002.

.Stan Development Team. 2023a. Stan Modeling Language Users Guide and Reference Manual, 2.32. https://mc-stan.org.

Stan Development Team. 2023b. RStan: the R interface to Stan. R package version 2.21.8. https://mc-stan.org/.

Tai, Vo Van & T. Pham-Gia. 2010. Clustering probability distributions. *Journal of Applied Statistics* 37(11). 1891–1910. https://doi.org/10.1080/02664760903186049.

Thompson, Sandra A. 2002. "Object complements" and conversation towards a realistic account. *Studies in Language* 26(1). 125–163. https://doi.org/10.1075/sl.26.1.05tho.

Ver Hoef, Jay M. & Peter L. Boveng. 2007. Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology* 88(11). 2766–2772. https://doi.org/10.1890/07-0043.1.

Wahl, Alexander. 2015. Intonation unit boundaries and the storage of bigrams: Evidence from bidirectional and directional association measures. *Review of Cognitive Linguistics* 13(1). 191–219. https://doi.org/10.1075/rcl.13.1.08wah.

Appendix

Interpretation	Examples	Concentrated in
interjections	hm, oh, right, unhunh	one-word IUs
interjections and vocatives	god, mom, sure, uh, why	one-word IUs and, secondarily, other final positions of shorter IUs
time-/choice-related	after, before, every, or	mostly beginnings of short IUs + sometimes next-to-last positions of longer IUs

conjunction and conjunction-like words	and, so, which, but	strongly initial, well spread across IU sizes
subordinators and modals	how, maybe, what, where	strongly initial, well spread across IU sizes (more short-biased than 10)
modal-evidential verbs	know, mean, think, wanted	second position of two-word IUs
semantically light verbs	came, gon, wan, told	second to third positions of moderate-sized IUs
contractions and modal-evidential verbs	's, goes, guess, should	second positions of short IUs
temporal and modal adverbs	always, just, never, not	third word from the beginning of moderate-sized IUs
semantically light verbs	go, want, went, have	2-4 positions of moderate-sized IUs
light (pro)nouns	day, lot, me, anything	final positions of IUs, well spread out across IU lengths
(diverse)	around, back, time, say	final position across a range of IU lengths
semantically light nouns	everything, something, here	final positions, spread across IU lengths
(diverse)	four, kinda, really, remember	final to penultimate positions of shorter IUs
(diverse)	about, big, long, her	last or penultimate word of moderate-sized IUs
determiners, light content words, some prepositions	an, tell, very, real, call	penultimate position of moderate-sized IUs, highly concentrated
semantically light content words	good, years, great, like	penultimate to antepenultimate positions of short IUs
mostly determiners and prepositions	all, as, by, first, these	penultimate to antepenultimate words of IUs
modal and semantically light verbs	be, even, getting, take	penultimate to fourth-from-last positions of moderate IUs
prepositions and quantitative determiners	any, in, three, through	antepenultimate and penultimate positions of moderate-sized IUs
genitive pronouns and other determiners and semantically light adjectives	another, my, our, than	antepenultimate position across a range of IU lengths
nominative pronouns and modal verbs	are, does, is, it	well spread out or bimodal distribution of positions, short to moderate IUs