Grammatical classes

5 June 2018

Placing words into grammatical classes

Here's a sample of automatically-discovered grammatical classes, using the '`ortho-merge`' strategy from '`gram-class.scm`'. I seem to have lost/corrupted a previous, larger dataset, so this was remade from scratch the last few days. Source dataset is '`en_pairs_cfive_class`'.

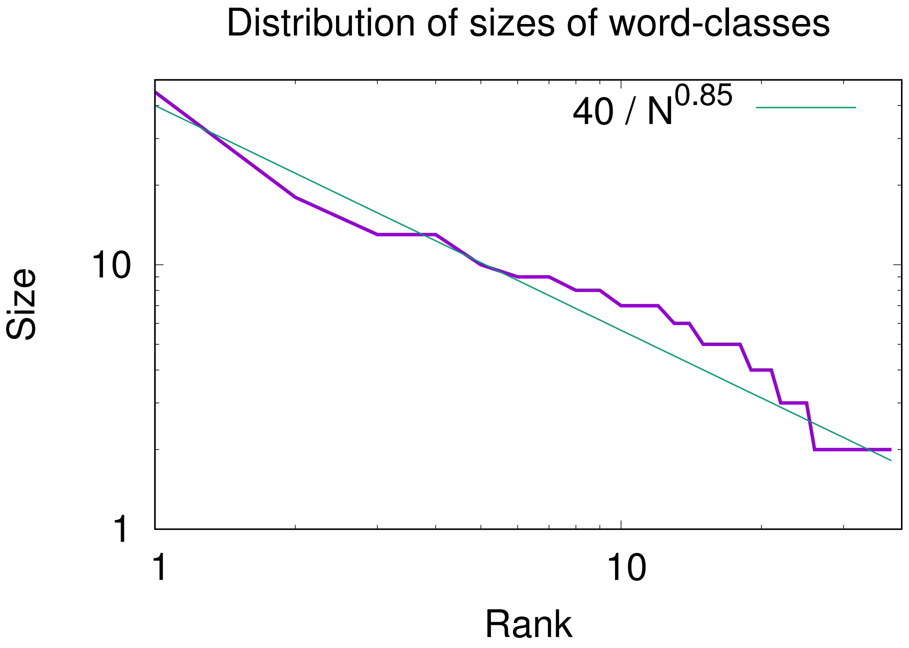
This run took 48 hours (its very far from done, this is a snapshot), it found 230 words that it could classify into 38 classes. (Seven more words got classified as I prepared this, so the counts may be off.) The table below lists them exhaustively. Note that some words appear multiple times: for example, "French English" appear as common nouns, but also in thier own class. Similarly, "mother father" appear in multiple classes. Some words are clearly mis-classified, but there are not many of those. The classes are clearly semantic in nature; for example, there are two distinct classes of prepositions. The semantics is entertainingly insightful: "voice mother hands heart head father mind face feet" are parts of oneself, with some unexpected members: "mother, father" are not normally considered to be body parts, but are, in some sense, deeply, "parts of oneself". Similarly, "wife arm daughter friend mouth friends brother" are mostly relatives and relationships, yet "arm mouth" are not. Perhaps the arm and mouth have a mind of thier own, functioning a bit independently from the true self?

I'll try to run this a few more days, and present a newer report. While reviving this old code, I realized that the classification algorithm being used here has multiple faults and is a bit crude. I'm writing a nicer algo right now. I don't really know how to compare the quality of the algos, at this point.

Meanwhile, you shuld be able to get similar results, by applying the code in '`gram-class.scm`' to a dataset that contains disjuncts dervied from MST parses.

Size	Members	Comments
43	village city question subject sea town girl public land French English fire King war boy air morning words others poor best second world door book heart body case night room whole light country people house children last present ground water family first other	Nouns, mostly
18	for in from at on by of with all towards within near against under through over upon into	Prepositions
12	help hear keep leave take find get make see give say go	Personal verbs
12	fine word large moment certain small woman new good man great little	Adjectives, mostly
10	fall action history character state position sense force knowledge pleasure	
9	full nature part death power most some out one	
9	voice mother hands heart head father mind face feet	Body-parts
8	till whether since because until where if when	Time
7	will would might should may can must could	Imperatives
7	or but perhaps nor though And while	Conjunctions
7	wife arm daughter friend mouth friends brother	Relatives
6	feel believe myself am know think	Beingness
6	rest end body name side power	
6	heard taken given already done seen	Past perfect - action
5	really always still also now	
5	year place same day way	
5	kept held called made found	Possesive verbs
4	son arms own eyes	
4	her me him us	Anaphora
4	heard felt knew saw	Simple past - action
3	making such like	
3	our its their	Possesives - plural
3	five three four	
3	during between among	Prepositions
2	once least	
2	thus sometimes	Deduction
2	therefore indeed	Deduction
2	is was	to be - Singular
2	are were	to be - Plural
2	! ?	Sentence end
2	, ;	Punct
2	And The	Sentence start
2	they we	Anaphora
2	cannot shall	Imperatives
2	mother father	
2	sort number	
2	France England	

Here's a graph of the above distribution. Its on a log-log scale. It looks to be approximately Zipfian. That's no surprise.



The End