Homework 1

Richard Littauer, Grady Lamar Payson, Stefan Fischer

February 14, 2012

1 Corpora

The Corpus we've chosen is the Wall Street Journal Corpus.

1.1 X

1. Are passive constructions correlated with *longer* sentence length *in English, in newspapers*?
2. Look for auxiliary verbs (*was, got*) for passive construction sampling (as this is not the only possible way to construct a passive in English) and sentence length.
3. Measure the distribution of auxiliary verbs, and the sentence length differences between sentences with them, and without.
4. N/A.
5. Just do the whole corpus - the corpus is small enough with modern computational power to run the entire thing. If it is too big, it has been pre-segmented into equal sizes, so quota sample from one of the segments.
6. *Was* can be a copula, which here would affect the overall count - in this case, it may have been smarter to look for the passive participle construction on the main verbs, or use a parser. Sentence length is also influenced by the register used in newspaper writers, which means this corpus may affect the outcome. Parsers have trouble with longer sentences, so if this had been used, it might have been deficient for this study. This is a non-trivial research question.

1.2 Y

1. Is there a significant difference with pronoun referral between male and female Named Entities?
2. Pronoun instances, and $n = i$ cases of pronoun/named entity collocations, where i is the distance between the pronoun and the named entity. Also use the Stanford Parser dependency representation along with the distance metric for anaphoric resolution.
3. We would need to measure the optimal distance of i (possibly compared to the parser) as a possibly proxy for anaphoric resolution on a surface level. That distance must be measured, in any event. We would need to measure the male or female occurrences of Named Entities in the text, against some NE dictionary to specify sex, and to make sure that we only select those NE that are persons (not, for instance, *Johns Hopkins University*. We also need to measure cases where the parser or distance proxy fails.
4. N/A

5. Again, this corpus can probably be sampled in its entirety.

6. Parsers are not so great at anaphora resolution; distance may not be a suitable proxy. The dictionary may not have perfect female/male definitions, and in some cases, may fail due to lexical ambiguity (such as the American name *Kelly*, or German *Maria*).

2 Experimental Research

2.1 X

1. Is Red Bull (caffeinated, high-energy drinks) consumption correlated with a higher spoken word per minute count?

2. N/A

3. The base word per minute, and then incrementally the δwpm per Red Bull of participants.

4. Hopefully, you would want to test both stereotypically slow- and fast-talkers, at various times of day. It would be best to test literature, as the best experimental strategy would be to have them read an interesting book (of any sort). Ideally, they should all be native speakers of the experimental language. Undergrads would be suitable for this experiment.

5. Cluster sampling, based on sex and their base wpm (as a slow speaker would be more affected, possibly, than a fast one) would be best.

6. Time of day; personality; reading choice; amount of coffee drunk prior to experiment; amount of sleep the previous night; et cetera.

2.2 Y

1. Do tonal language speakers stay on pitch while singing more than non-tonal language speakers (given a pentatonic scale)?

2. N/A

3. Measure the standard harmonic pitch, the average base f0 levels for each lexical pitch in song by speaker, the deviation from this during song by a speaker, and the speaker's standard f0 levels for each of the lexical tones in their language when spoken. As there is no standard pitch, with f0 being a dynamic feature, the base tone for each targeted word should be measured in relation to the surrounding lexical, syntactic, and intonational environment.

4. A good distribution of male and female singers; gender; age; and professional versus amateur singers. For several languages, both tonal and non-tonal.

5. This should be stratified, as you're comparing different groups, where the measurements are different.

6. Genetic singing ability, feedback mechanism ability (for autocorrection of pitch derivation), environmental differences during recording, smoking and other voice-affecting possibilities, amount of music listened to over lifetime, etc.