A Review of Corpus-based Statistical Models of Language Variation

Yao Yao
Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
Hong Kong
ctyaoyao@polyu.edu.hk

Abstract

This paper is a brief review of the research on language variation using corpus data and statistical modeling methods. The variation phenomena covered in this review include phonetic variation (in spontaneous speech) and syntactic variation, with a focus on studies of English and Chinese. The goal of this paper is to demonstrate the use of corpus-driven statistical models in the study of language variation, and discuss the contribution and future directions of this line of research.

1 Introduction

Human language is inevitably variable. The same meaning may be wrapped in different sentence forms without losing the semantic content; the same word or the same sound could be pronounced slightly differently by different speakers, or even by the same speaker but in different linguistic or non-linguistic contexts. Sometimes we can come up with an explanation for the observed differences (e.g. men and women talk differently), but more often than not, variation seems so ubiquitous and random. In fact, variation used to be considered as noise in the signal – something that needs to be filtered out before the signal can be processed. In recent years, however, the value of 'random' variation has been gradually uncovered in linguistic research.

What has changed to cause the rising interest in variation? In our view, the change is largely due to the availability of large-scale linguistic datasets – often extracted from big corpora – and sophisticated statistical tools that allow researchers to look for patterns in a sea of seemingly random and unpredictable data. Thus, variation is no longer viewed as noise but a gold mine of information about how language is produced and used in communication. For instance, examining patterns of pronunciation variation in spontaneous speech can help us understand what factors (e.g. word frequency, contextual predictability, information status) may play a role in the speech production process, what is the relative importance of these factors, and how they interact with each other. Furthermore, a variation model also makes it possible to examine the effect of some particular factor by statistically controlling for other factors that are also active. By comparison, in an experimental study, it is often hard to completely balance all relevant factors when creating experimental stimuli and conditions.

In the remaining of this paper, we will first introduce the general methodology of building corpus-based statistical models of language variation; we will then briefly discuss several previous studies on phonetic variation and syntactic variation that cover a few different languages (English, French, Chinese). Finally, we will briefly discuss the contribution and future directions of this line of research.

29th Pacific Asia Conference on Language, Information and Computation pages 11 - 15
Shanghai, China, October 30 - November 1, 2015

2 General Methodology

The general methodology of a corpus-based variation study consists of two major stages: dataset compilation and model building. A dataset contains observations of the linguistic phenomenon under investigation (e.g. pronunciation of function words in English). The observations are extracted from some corpora and are annotated with a set of linguistic properties. To use the modeling approach, it is necessary that the linguistic variation under investigation is encoded in some quantifiable (or categorical) measures. For instance, variation in word pronunciation may be encoded in the duration of a word, which is a quantitative measure. Such measures will be used as the outcome variable in the statistical model. Furthermore, each observation will be annotated – either manually or automatically – with a number of features that are hypothesized to be predictors of the linguistic variation (e.g. usage frequency of a word might predict the duration of a word in natural production). Variation models typically include thousands or tens of thousands of observations, in order to ensure enough statistical power. Thus, it is critical to choose an appropriate data source that contains enough relevant observations and adequate representation of the predictor variables.

After the dataset is prepared, it will be fed into the statistical model. Currently, the most popular and widely used model in the field is the mixed-effects regression model (Baayen et al., 2008). Compared to a simple regression model, mixed-effects models have the advantage of allowing two levels of predictors: random-effects predictors and fixed-effects predictors. The inclusion of random-effects predictors is particularly useful for modeling linguistic variation, because we know that part of the variation will be truly random and cannot be predicted by any annotated feature. For example, different speakers will pronounce the word *to* slightly differently, and ultimately, some individual differences are beyond the predicting power of speaker sex, age, height, weight, etc. and will have to be random. Similarly, the differences among individual words (e.g. *to* and *too*) could also be idiosyncratic and unpredictable. In a mixed-effects model, random effects may co-exist with fixed-effects, which means that, for example, both gender differences (i.e. sex as a fixed-effects predictor) and true individual differences (i.e. speaker as a random-effects predictor) may both be represented in a model of pronunciation variation.

Depending on the type of the outcome variable, one may use either mixed-effects linear regression model (for numerical outcome variables) or mixed-effects generalized regression model (for categorical outcome variables). Research on modeling language variation

2.1 Modeling phonetic variation

This vein of corpus-based language variation research first started with studies on phonetic variation – probably because phonetic features are readily quantifiable. Some of the pioneering works on English pronunciation variation were completed around the turn of the century (Bell et al. 2009; Fosler-Lussier and Morgan 1999; Gregory, et al. 1999; Jurafsky et al. 1998, 2001a, among others), with phonetic data from the Switchboard corpus of telephone conversations (Godfrey et al. 1992), which contains 240 hours of speech (of which 4 hours are phonetically transcribed and used in the statistical models).

The studies above mostly examined word duration and vowel pronunciation (full vs. reduced) as parameters of pronunciation variation. In addition to describing the general picture of variation, these studies were also deeply interested in the effects of probabilistic factors (e.g. word frequency, contextual probability, etc) on pronunciation variation. The results presented in these studies are cited as empirical support for the general claim that probabilistic relations have profound influence on the representation and production of words in speech (Jurafsky et al., 2001b)

Later on, with the completion of the Buckeye corpus (Pitt et al., 2007), which contains 40 hours of phonetically transcribed conversational speech, another batch of corpus-based phonetic variation studies appeared (Johnson, 2004; Gahl et al., 2012; Yao, 2009, 2011, etc). Since the Buckeye corpus is recorded in a studio, the recording quality is high enough to warrant automatic measurement of VOT (Yao, 2009) and vowel formants (Yao et al., 2010). This allows for modeling of gradient vowel dispersion, measured by the distance between a specific vowel token from the center of the vowel space on a F1-F2 plane (Bradlow et al., 1996).

Furthermore, some of the variation studies based on the Buckeye corpus (Gahl et al., 2012; Yao, 2011) focused on the effects of a particular lexical measure called phonological neighborhood density. Phonological neighborhood density refers to the number of similar-sounding words given a specific target word. Thus, the models built in these studies had one critical predictor (i.e. phonological neighborhood density), and all the other non-neighborhood predictors were included as control variables. Results from these studies revealed the effects of phonological neighborhood structure in word production when all other factors that could also influence word production were statistically controlled.

In addition to English, corpus-based pronunciation variation research has also been conducted in other languages (Dutch: Pluymaekers et al., 2005, among others; French: Meunier and Espesser, 2011; Yao and Meunier, 2014; Taiwan Southern Min: Myers and Li, 2009).

2.2 Modeling syntactic variation

The work on modeling syntactic variation started later than the work on modeling phonetic variation. Most of the pioneering works were done by Bresnan and her colleagues at Stanford (Bresnan, 2007; Bresnan et al., 2007; Bresnan and Ford, 2010; Tily et al., 2009; Wolk et al. 2011, etc) on dative variation (e.g. *I gave John a book* vs. *I gave a book to John*) and genitive variation (e.g. *John's book* vs. *the book of John*) in English. For the American English data, Bresnan and colleagues also used the Switchboard corpus. Since syntactic variation has a discrete set of variants (i.e. different sentence forms), the phenomenon is modelled by generalized regression models. Bresnan and colleagues' work showed that the choice of the surface form under investigation was predictable from a set of factors relating to different components in the local sentence (e.g. semantic type of the verb, NP accessibility, pronominality, definiteness, syntactic complexity, etc) and the context (e.g. presence of parallel structures). When taking all the factors into consideration, Bresnan et al.'s models can correctly predict the surface dative/gentive form in more than 90% of the cases (compare with a baseline accuracy around 79%). Variation patterns revealed in Bresnan et al.'s works were later confirmed in behavioral experiments (e.g. Bresnan and Ford, 2010).

Inspired by Bresnan and colleagues' work on English syntactic variation, there have also been a few studies that apply a similar modeling approach to the study of syntactic variation in Chinese languages (Cantonese: Starr, 2015; Mandarin: Yao, 2014; Yao and Liu, 2010).

In particular, Yao and colleagues (Yao, 2014; Yao and Liu, 2010) investigated both dative variation and BA-form variation in written Mandarin using data from the Academia Sinica corpus (Chen et al., 1996). Sentence patterns involved in Mandarin dative-variation (e.g. 我送小张一本书 'I gave Xiaozhang a book' vs. 我送一本书给小张 'I gave a book to Xiaozhang' vs. 我把一本书送给小张 'I (BA) a book gave to Xiaozhang') are more complicated than those in English. In addition to the two dative constructions similar to those in English, Mandarin Chinese also allows the direct object to be preposed before the verb. Yao and Liu' work showed that the three-way dative variation in Mandarin Chinese can be modeled by a hierarchy of two models: one on the upper level for the pre-verbal vs. post-verbal distinction and the other on the lower level for the dative vs. double object distinction. Yao and Liu' models raise the prediction accuracy by 27% (upper level) and 7% (lower level) compared to the baseline accuracy levels.

Furthermore, to understand the general properties of the pre-verbal vs. post-verbal word order variation, Yao also built general models on syntactic variation between BA and non-BA sentences. The results from this study showed that the surface word order in Mandarin Chinese is most significantly influenced by the prominence (accessibility, definiteness, etc) and length of the NP, as well as the presence of a similar word order in the nearby context (i.e. parallel structure).

3 Discussion

In this paper, we have briefly reviewed some previous studies that use corpus-based statistical models to investigate language variation phenomena. The focus of this review is on studies of phonetic variation (in spontaneous speech) and syntactic variation in English and Chinese. As discussed above, corpus-based research on linguistic variation is still dominated by studies on

English; by comparison, there is much less research on linguistic variation – especially phonetic variation – in Chinese. One possible reason for the lack of Chinese phonetic variation research is the unavailability of large annotated conversational speech Chinese corpora (to linguists). In our view, the lack of resources may in fact indicate a potential opportunity of collaboration between theoretical linguists and speech engineers (computational linguists). We discuss this in more detail in our next point.

We have observed that so far, the researchers who work on corpus-based language variation studies are mostly linguists who are interested in the general variation patterns or the effects of particular factors that are critical to some linguistic theories. One may say that these researchers are doing 'computational linguistics' in the sense that they use computational (modeling) methods to investigate linguistic questions. In reality, of course, the term 'computational linguistics' refers to the area of study that aims to develop language-related (or text-related) applications in computer science. However, despite the seemingly disparate research interest, we must recognize that these two lines of research do share some common features – mostly in the corpus-based and computational nature of the work – and that people working in these areas may benefit from collaborating with each other. Among other things, computational linguists can help theoretical linguists develop tools for automatically annotating a corpus, and theoretical linguists' work can provide generalizations of variation patterns that may in turn inform computational linguistic applications.

To conclude, while we believe that the research on corpus-based variation research has made significant contribution to the study of language, we are convinced that greater success can be achieved if theoretical and computational linguists will work jointly on these topics.

Acknowledgments

Research reported in this paper was supported by the Early Career Scheme of Hong Kong RGC (Grant No. 558913) and the Newly Recruited Junior Academic Staff funding of the Hong Kong Polytechnic University (Grant Account A-PL27).

References

Baayen, R. H., Davidson, D. J., and D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Bell, A., J. M. Brenier, M. Gregory, C. Girand, and D. Jurafsky. 2009. Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language* 60 (1):92–111.

Bradlow, A. R., G. Torretta, and D. Pisoni. 1996. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20(3–4):255–272.

Bresnan, J. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston and Wolfgang Sternefeld (eds) *Roots: Linguistics in search of its evidential base*, pp 77-96. Series: Studies in Generative Grammar. Berlin: Mouton de Gruyter.

Bresnan, J., Cueni, A., Nikitina, T. and H. Baayen. 2007. Predicting the dative alternation. In G. Boume et al. (eds) *Cognitive foundations of interpretation*, pp 69-94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, J., and M. Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language*, 86(1):168–213.

Chen, K.-j., Huang, C.-r., Chang, L.-p., and H.-L. Hsu. 1996. Sinica Corpus: Design Methodology for Balanced Corpora. In B.-S. Park and J.B. Kim (eds) *Proceeding of the 11th Pacific Asia Conference on Language, Information and Computation*, pp.167–176. Seoul: Kyung Hee University.

Fosler-Lussier, E., and N. Morgan. 1999. Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication*, 29(2):137–158.

Gahl, S., Yao, Y., and Johnson, K. 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4):789–806.

J. Godfrey, E. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (*ICASSP-92)*, pp. 517–520.

Gregory, M. L., W. D. Raymond, A. Bell, E. Fosler-Lussier, and D. Jurafsky. 1999. The effects of collocational strength and contextual predictability in

lexical production. In *Proceedings of the Chicago Linguistic Society*, 35, pp. 151–166.

Johnson, K. 2004. Massive reduction in conversational American English. In K. Yoneyama and K. Maekawa (Eds.), *Spontaneous speech: Data and analysis*. *Proceedings of the 1st session of the 10th International Symposium*, pp. 29–54. Tokyo: The National International Institute for Japanese Language.

Jurafsky, D., A. Bell, E. Fosler-Lussier, C. Girand, and W. Raymond. 1998. Reduction of English function words in Switchboard. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP '98)*, Volume 7, pp. 3111–3114. Sydney, Australia.

Jurafsky, D., A. Bell, M. Gregory, and W. Raymond. 2001a. The effect of language model probability on pronunciation reduction. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, Volume 2, pp. 801–804. Salt Lake City, Utah.

Jurafsky, D., A. Bell, M. Gregory, and W. Raymond 2001b. Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee and P. Hopper (Eds.), *Frequency and the Emergence of Linguistic Structure*, pp. 229–254. Amsterdam: John Benjamins.

Meunier, C., and R. Espesser. 2011. Vowel reduction in conversational speech in French: The role of lexical factors. *Journal of Phonetics*, 39(3):271–278.

Myers, J., and Y. Li. 2009. Lexical frequency effects in Taiwan Southern Min syllable contraction. *Journal of Phonetics*, 37 (2):212–230.

Pitt, M. A., L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier. 2007. Buckeye Corpus of Conversational Speech (2nd release). Department of Psychology, Ohio State University. http://www.buckeyecorpus.osu.edu.

Starr, Rebecca L. 2015. Predicting NP Forms in Vernacular Written Cantonese. *Journal of Chinese Linguistics*, 43.1A.

Pluymaekers, M., M. Ernestus, and R. H. Baayen. 2005. Lexical frequency and acoustic reduction in spoken Dutch. *Journal of Acoustical Society of America*, 118(4):2561–2569.

Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., and J. Bresnan. 2009. Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition*, 1(2):147–165.

Wolk, C., Bresnan, J., Rosenbach, A., and B. Szmrecsányi. 2011. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica*, 30(3):382–419.

Yao, Y. 2009. An Exemplar-based approach to automatic burst detection in spontaneous speech. In *Proceedings of the 18th International Congress of Linguists (CIL XVIII)*. Seoul: Korea University.

Yao, Y. 2011. *The effects of phonological neighborhoods on pronunciation variation in conversational speech*. Unpublished PhD dissertation. University of California, Berkeley.

Yao, Y. 2014. Predicting the use of BA construction in Mandarin Chinese discourse: A modeling study with two verbs. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing (PACLIC28)*. December 12-14, 2014. Phuket, Thailand.

Yao, Y., and F.-h. Liu. 2010. A working report on statistically modeling dative variation in Mandarin Chinese. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*. Beijing, China.

Yao, Y., and C. Meunier. 2014. Effects of phonological neighborhood density on phonetic variation: The curious case of French. *The 14th Laboratory Phonology Conference*, Tokyo, July 25-27, 2014.

Yao, Y., Tilsen, S., Sprouse, R.L., and K. Johnson. 2010. Automated measurement of vowel formants in the Buckeye Corpus. *言語研究 Gengo Kenkyu (Journal of the Linguistic Society of Japan)*, 138:99–113.