Zipfian discrimination

Jim Blevins
University of Cambridge
jpb39@cam.ac.uk

Petar Milin
University of Novi Sad
Eberhard Karls Universität Tübingen
petar.milin@uni-tuebingen.de

Michael Ramscar
Eberhard Karls Universität Tübingen
michael.ramscar@uni-tuebingen.de

This talk outlines how form variation can be modelled in terms of equilibria between two dominant communicative pressures. The pressure to **discriminate** forms of a language enhances differences between expressions. Unchecked, this pressure can in principle lead to suppletion of the kind reported in languages such as Yélî Dnye (Henderson 1995). However, in most languages, the pressure towards maximally discriminative expressions is countered by the need to **extrapolate** from sparse input. It has long been known that corpora provide only a partial coverage of the forms of a language (inflectional and derivational). This talk presents evidence that the shortfall is far greater and far more systematic than previously appreciated, and that the coverage of the form variation remains sparse in corpora of up to one billion words. The sampling reported in this talk suggests that the forms in a corpus or encountered by a speaker exhibit a Zipfian distribution at all sample sizes.

The interaction of these pressures also accounts for the role of lexical neighbourhoods. Since most paradigms will be only partially attested, the organization of paradigms into neighbourhoods provides an analogical base for extrapolation.

The status of regularity

It is usually assumed that regularity in a linguistic system is desirable or normative and that suppletion and other irregularities represent deviations from the uniform patterns that systems (or their speakers) strive to maintain. From a discriminative perspective, the situation is exactly reversed. To the extent that patterns like suppletion enhance the discriminability of forms, they contribute to the communicative efficiency of a language. In a discriminative model, such as that of Ramscar et al. (2013), the only difference between overtly suppletive forms such as *mouse*/*mice* and more regular forms such as *rat*/*rats* is that the former serve to accelerate the rate at which a speakers' representation of a specific form/meaning contrast becomes discriminated from the form classes that express similar contrasts. Thus all learning serves to increase the level of suppletion in form-meaning mappings.

Moreover, standard cases of 'suppletion' are merely extreme instances of discriminative contrasts that seem ubiquitous at the sub-phonemic level. In the domain of word formation, Davis et al. (2002) found suggestive differences in duration and fundamental frequency between a word like *captain* and a morphologically unrelated onset word such as *cap*. Of more direct relevance are studies of inflectional formations. Baayen et al. (2003) found that a sample of speakers produced Dutch nouns with a longer mean duration when they occurred as singulars than as when they occurred as the stem of the corresponding plural. In a follow-up study, Kemps et al. (2005) tested speakers' sensitivity to prosodic differences, and concluded that "acoustic differences exist between uninflected and inflected forms and that listeners are sensitive to them" (Kemps et al. 2005: 441). Recent studies by Plag et al. (2014) find similar contrasts between phonemically identical affixes in English.

The role of discriminability

From a discriminative perspective, it is **regularity** that stands in need of explanation. Learning models offer a solution here as well. Unlike derivational processes, inflectional processes are traditionally assumed to be highly productive, defining uniform paradigms within a given class. Lemma size is thus not expected to vary, except where forms are unavailable due to paradigm 'gaps' or 'defectiveness'. Yet corpus studies suggest that this expectation is an idealization. Many potentially available inflected forms are unattested in corpora. As corpora increase in size, they do not converge on uniformly populated paradigms. Instead, they reinforce previously attested forms and classes while introducing progressively fewer new units. As shown in

In Vito Pirrelli, Claudia Marzi, Marcello Ferro (eds.): **Word Structure and Word Usage**. Proceedings of the NetWordS Final Conference, Pisa, March 30-April 1, 2015, published at http://ceur-ws.org

Figure 1, the number of attested inflected noun variants decreases in all random samples, ranging from 1-million to 15-million hits, at which point the 850-million word StdeWaC corpus is essentially exhausted. As sample size increases, there is a marked attenuation in the steepness of the slope steepness, though it never becomes completely flat. This trend is extracted and presented in Figure 2, which plots number of attested forms on the X-axis and slopes of six trends from Figure 1 on the Y-axis. From this relationship we can infer that even if the corpus size were increased to infinity, it would never contain all possible inflected forms of every German noun. As shown in Figure 3, the forms of a language obey Zipf's law at all sample sizes. Speakers must be able to extrapolate from a partial – often sparse – sample of their language, and regular patterns subserve this need.

It takes a neighbourhood

In order for a collection of partial samples to allow the generation of unattested forms, the forms that speakers do know must be organized into systematic structures that collectively enable the scope of possible variations to be realized. These structures correspond to lexical neigbourhoods, whose effects have been investigated in a wide range of psycholinguistic studies (Baayen et al. 2006; Gahl et al. 2011). From the present perspective, neighbourhoods are not independent dimensions of lexical organization but, rather, constitute the creative engine of the morphological system, permitting the extrapolation of the full system from partial patterns. Interesting support for this perspective comes from the study reported in Milin et al. (2011). In this study, analogical extrapolation from a small set of nearest neighbors allowed a system to model the choice of masculine instrumental singular allomorph by Serbian speakers presented with nonce words. Regular paradigms thus enable language users to generate previously unencountered forms, not because they are the product of an explicit rule, or of any kind of explicit grammatical knowledge, but rather they are implicit in the distribution of forms and semantics in the language as a system, much as suggested by Hockett (1967: 221).

> in his analogizing … [t]he native user of the language … operates in terms of all sorts of internally stored paradigms, many of them doubtless only partial

References

Baayen, R. H., Feldman, L. B. & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language* **53**, 496–512.

Baayen, R. H., McQueen, J. M., Dijkstra, T. & Schreuder, R. (2003). Frequency effects in regular inflectional morphology: Revisiting Dutch plurals. In Baayen, R. H. & Schreuder, R. (eds.), *Morphological Structure in Language Processing*, Berlin: Mouton de Gruyter, 355–370.

Davis, M., Marslen-Wilson, W. D. & Gaskell, M. (2002). Leading up the lexical garden-path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception & Performance* **28**, 218–244.

Gahl, S., Yao, Y. & Johnson, K. (2011). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language* **66(4)**, 789–806.

Henderson, J. E. (1995). *Phonology and Grammar of Yele, Papua New Guinea*. Pacific Linguistics B-112, Camberra: Pacific Linguistics.

Hockett, C. F. (1967). The Yawelmani basic verb. *Language* **43**, 208–222.

Kemps, J. J. K., Rachèl, Ernestus, M., Schreuder, R. & Baayen, R. H. (2005). Prosodic cues for morphological complexity: The case of Dutch plural nouns. *Memory & Cognition* **33(3)**, 430–446.

Milin, P., Keuleers, E. & Filipović Đurdjević, D. (2011). Allomorphic responses in Serbian pseudo-nouns as a result of analogical learning. *Acta Linguistica Hungarica* **58**, 65–84.

Plag, I., Homan, J. & Kunter, G. (2014). Homophony and morphology: The acoustics of word-final S in English. Ms, Heinrich-Heine-Universität, Düsseldorf.

Ramscar, M., Dye, M. & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of *mouses* in adult speech. *Language* **89(4)**, 760–793.

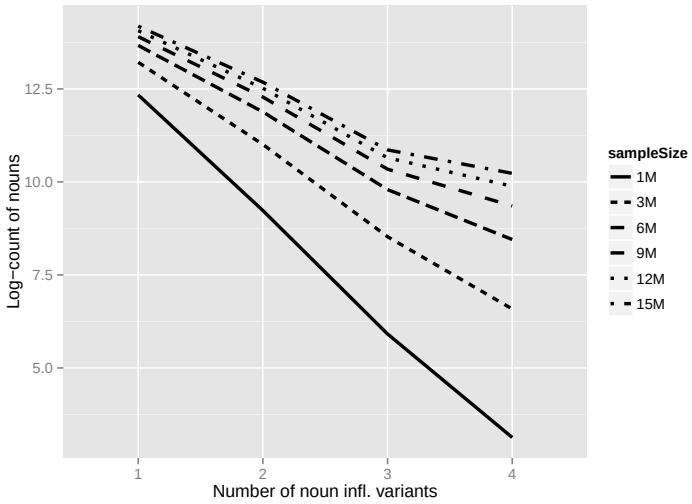

Figure 1: The paradigm non-filling pattern

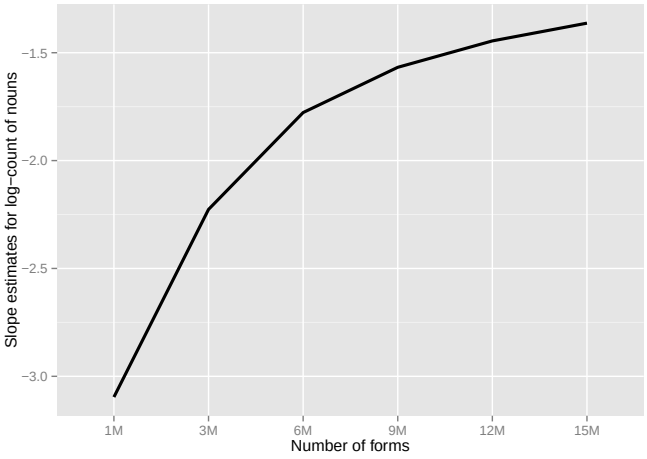

Figure 2: Asymptoting slopes

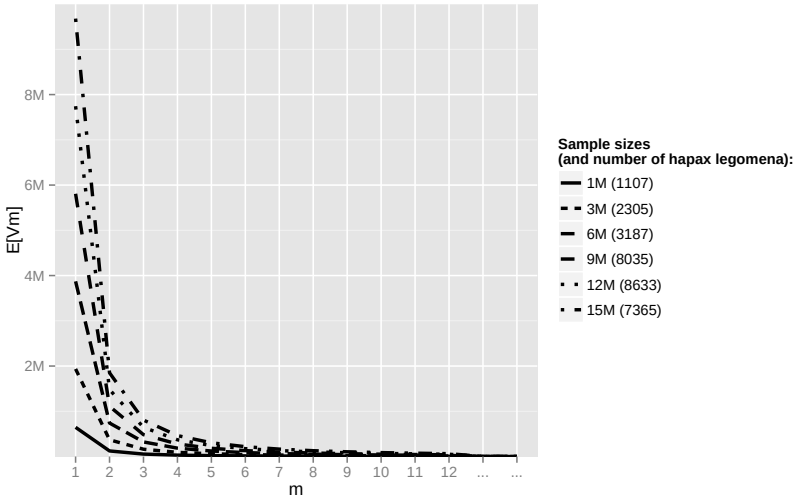

Figure 3: Zipf plot for randomly sampled words