Charles Yang, Kyle Gorman & Jennifer Preys **Saturday 3:00**
(University of Pennsylvania)

Productivity and Paradigmatic Gaps

Summary As young children vividly illustrate (the famous *wug* test; Berko 1958), productivity, the ability to extend observed patterns to novel items, is a key design feature of language. It is thus puzzling that there are corners of the grammar where productivity unexpectedly fails. In a classic paper, Halle (1973) draws attention to paradigmatic gaps; for instance, there are around 75 Russian verbs, all belonging to the second conjugation, that lack a first person singular (1sg.) non-past form, a fact which cannot be predicted on phonotactic grounds alone (Sims 2006). In this study, we use a model of productivity in language acquisition (Yang 2005, 2010) to predict paradigmatic gaps in the inflectional systems of English, Spanish, and Russian from quantitative considerations.

Background Even undisputedly productive generalizations only reach productivity after extensive linguistic input. For instance, children acquiring English past tense typically follow a U-shape learning curve (Pinker 1999), whereby irregular verbs are initially inflected early on, but are later subject to overregularization (*hold-*holded*), signaling the onset of the productive "add /-d/" rule. Since regular verbs are on average less frequent than irregulars, it takes time for the child to encounter enough regulars to posit a productive "add /-d/" rule in the face of exceptional irregulars.

The model The question is how many regulars are "enough" to counter exceptions. Yang (2005) approaches the question through online computational efficiency: a productive rule can tolerate a certain number of exceptions only if treating them as exceptions leads to an lower expected real-time processing time than simply listing all items sorted by frequency. This model is motivated by the psycholinguistic findings that idioms (Cutler & Swinney 1979) and irregulars (Clashen et al. 2004, Penke & Krause 2005) are generally processed faster than compositional phrases and regulars, respectively, even after controlling for frequency; a consequence is that too many exceptions may delay average processing time. Under very general assumptions about word frequencies (i.e., Zipf's law), it is possible to show that if a rule is applicable to $\boldsymbol{N}$ lexical items, the maximum number of exceptions it can tolerate while still maintaining productivity, is $\boldsymbol{N} / \ln \boldsymbol{N}$ (see Yang 2005). Consider a straightforward application. There are approximately 150 irregular verbs in English, which means that there need to be 1000 verbs in total, and 850 (= 1000 - 150) regulars, to ensure the productivity of the "add /-d/" rule (since $1000 / \ln 1000 \approx 150$); in fact, there are 1,252 underived verbs that occur at least once per million words in the COBUILD corpus.

A filibuster-proof majority It is easy to see that a rule followed by a mere majority of items within its scope is not necessarily productive. For instance, nearly 90% of English words place primary stress on the initial syllable (Cutler & Carter 1987), though English metrical stress can only be accurately described by taking syllable weight into account; see Legate & Yang (2012) for an application to this productivity model to evaluate theories of metrical stress and its acquisition. Likewise, it is possible *not* to obtain productivity if the lexical items are subject to several alternations which constitute exceptions for each other, with the result that none reach productivity. We argue this is precisely when paradigmatic gaps arise: the lack of a productive rule means that the learner must hear the inflected forms of these lexical stems, but their absence in the input renders them ineffable.

Case study I: English past tense verbs It is well known (e.g., Pinker 1999) that several English irregular verbs lack preterites (*forgo-*forwent/*forgoed/*forgone-forgone*) or past participles (*stride-strode-*strided/*stridden/*strode*), and corpus statistics confirm the reality of these gaps. At the same time, the majority (102 out of 150) of irregular verbs have syncretic preterites and participles (e.g., *keep-kept-kept*, *bring-brought-brought*). But in order for the syncretism to be automatically extendable, there can be no more than 30 ($\approx 150 / \ln 150$) exceptions, when there are in fact 48 (= 150 - 102). Thus, even though the preterite-past participle syncretism holds for more than twice as many items, it fails to reach the productivity threshold. We correctly predict that the learner will be at a loss when in need of a preterite for *forgo* or a past participle for *stride*.

Case study II: Spanish mid-vowel diphthongization In some but not all Spanish verbs, an unstressed *e* or *o* in the final root syllable becomes *ie* [je] and *ue* [we] under primary stress (e.g., *negar-niego* 'deny',

aprobar-apruebar 'approve of', but cf. non-alternating *pegar-pego* 'stick onto', *robar-robo* 'steal'). This alternation is found in all three conjugations, but at different type frequencies (verbs occurring at least once per millon tokens in LEXESP; Sebastián et al. 2000). Following Harris (1969), we assume a single rule of dipthongization accounts for *e-ie* and *o-ue* alternations, but that the *i-e* alternation in words like *pedir-pido* derives from underlying /i/, and thus outside the scope of diphthongization.

conjugation		***e-e, o-o***	**e-*ie*, *o-ue***	**productivity threshold**	
first	(*-ar*)	1050	125	166	(yes)
second	(*-er*)	189	29	43	(yes)
third	(*-ir*)	19	33	13	(no)

The model predicts "no change" to be productive in the first and second conjugations; this is supported by children's frequent underapplication of diphthongization in these conjugations (Clahsen et al. 2002), a type of over-regularization. By contrast, the model assigns no productive rule to the third conjugation, and that is exactly where one finds inflectional gaps when primary stress falls in a root mid vowel (e.g., *abolir-*abolo/*abuelo* 'abolish', *colorir-*coloro/*coluero* 'colorize'; Real Academia Española 1992). We compare this prediction to those of Albright's (2003) model of Spanish diphthongization gaps, which incorrectly predicts gaps in other conjugations and fails to reliably predict the gaps recorded in the 3rd conjugation.

Case study III: Russian 1sg non-past verbs Returning to the classic case of Russian, we note that the root-final *t* of many verbs of the Russian second conjugation is realized as *č* in the 1sg. non-past (e.g., *metit'-meču* 'mark') but many verbs instead mutate to *šj* (e.g., *smutit'-smušju* 'confuse') or have a gap in the 1sg. non-past (e.g., *očutit'sja-*očučus'/*očušjus'* 'find oneself'; Halle 1973, Sims 2006). We use the Zaliznjak (1977) morphological dictionary to count each outcome, combining verbs which share roots but which take different prefixes or the reflexive *-sja* (for these always share the same mutation—*č* or *šj*—or are gapped; Pesetsky 1977) and eliminating verb roots not occurring at least once per million tokens in the Russian Reference Corpus (Sharoff 2005). There are 59 verbs belonging to this class, and19 of these verbs follow the minority (*t-šj*) pattern, exceeding the threshold of 14 (≈ 59 / ln 59) needed for productivity. Thus, inflectional gaps in Russian are entirely predictable as a consequence of productivity detection in language acquisition.

Conclusion Paradigmatic gaps arise when productivity fails (Halle 1973), and this is correctly predicted by a formally precise model of language acquisition (Yang 2005, 2010).

References: Albright. A. 2003. In *Proceedings of WCCFL 22*, 1-14. **Berko, J. 1958**. *Word* 14: 150-177. **Clahsen, H., Aveledo, F. & Roca, I. 2002.** *J. Child Lg.* 29: 591-622. **Clahsen, H., M. Hadler, & H. Weyerts**. **2004**. *J. Child Lg.* 31: 683-712. **Cutler, A., & D. Swinney.** 1979. *J. Verbal Learning & Verbal Behavior* 18: 523-534. **Cutler, A., & D. Carter. 1987.** *Computer Speech & Language* 2: 133-142. **Halle, M. 1973**. *LI* 1: 3-16. **Harris, J**. 1969. *Spanish phonology*. Cambridge: MIT Press. **Legate, J. & C. Yang. 2012**. In M. Piatelli-Palmarini & B. Berwick (eds.) *Rich languages from poor inputs.* Oxford: OUP. **Penke, M., & M. Krause. 2002**. *Brain and Language* 81: 301-311. **Pesetsky, D. 1977**. Russian morphology and lexical theory. Ms., MIT. **Pinker, S. 1999.** *Words and rules.* New York: Basic Books. **Real Academia Española. 1992.** *Diccionario de la lengua española*. Madrid: Real Academia Española. **Sebastián, N., F. Cuetos, A. Martí, & M. Carreiras. 2000.** *LEXESP: Léxico informatizado del español*. **Sharoff, S. 2005.** In D. Archer, A. Wilson, & P. Rayson (eds.), *Corpus linguistics around the world*, 167-180. Amsterdam: Rodopi. **Sims, A. 2006.** Doctoral dissertation, Ohio State University. **Yang, C. 2005**. *Linguistic Variation Yearbook* 5: 333-370. **Yang, C**. 2010. *Lingua* 120: 1160-1177. **Zaliznjak, A. 1977.** *Grammatičeskij slovar' russkogo jazyka*. Moskva: Russkij jazyk.