# “These are a few of my favorite facts”:

## Advances in phonology from new data sources

Bruce Hayes
UCLA

# Topic

- New methods of data gathering in phonology
- What we are learning from them

# Background:
# a view of our research enterprise

- **Intensive inspection of language data**, with discovery of generalizations. This enables...
- **Formal theoretical analysis**, shedding light on the patterns discovered and integrating them into a general phonological theory. We can also pursue:
- **Integration with other areas of cognitive science**, relating our theories to
  - experimental data
  - acquisition
  - learning models
  - processing models

# Phonologists can play a vital role in cognitive science

- because we are:
  - uniquely aware of the complexity and beauty of phonological systems
  - equipped with many good ideas from existing theory

- The data discussed here are particularly important to our role in the cognitive science enterprise, but also bear on some very traditional questions in our field.

# Outline

- Four new kinds of data sources
- Some “favorite facts” obtained from them
- Theoretical consequences of these facts

# Background: data corpora

- Traditional phonological analysis
  - identifies major patterns "by eye"
  - assumes that such patterns are of equal importance for the language learner in constructing her grammar.

# Background: data corpora

- Traditional phonological analysis
  - identifies major patterns "by eye"
  - assumes that such patterns are of equal importance for the language learner in constructing her grammar.
- Experimental evidence (e.g., later in this talk) suggests that this procedure doesn't tell the whole story:
  - the **frequency** of patterns plays a role in the completed grammar

# Background: data corpora

- Traditional phonological analysis
  - identifies major patterns "by eye"
  - assumes that such patterns are of equal importance for the language learner in constructing her grammar.
- Experimental evidence (e.g., later in this talk) suggests that this is wrong:
  - the **frequency** of patterns, particularly where conflicting, plays a role in the completed grammar
- Corpora can give a clear picture of what the language learner confronts when constructing her grammar.

# I. FULL-LEXICON CORPORA

# Technical basis

- It is now possible to gather quantitative data about every word in the dictionary, using the Web.

# Hayes and Londe's (in progress) corpus for Hungarian vowel harmony

- We began with a digital Hungarian dictionary...
- and for each nominal stem formed both possible datives (***-nak*** and ***-nek***, depending on vowel harmony).

# Part of the input list

…

| | |
|---|---|
| bibliofilnak | bibliofilnek |
| bíbornak | bíbornek |
| bíborosnak | bíborosnek |
| bicajnak | bicajnek |
| bicajosnak | bicajosnek |
| bicegnak | bicegnek |
| biciklinak | biciklinek |
| biciklizésnak | biciklizésnek |
| bigottnak | bigottnek |
| bigyónak | bigyónek |
| bikának | bikánek |

*(+ 10,000 more stems)*

…

# Next step:

- Obtain the **Google hit count** for each form, from Hungarian Web pages
- This is done by an Auto-Google program (http://www.linguistics.ucla.edu/people/hayes/), which Googles 20 items/second.
- In all but very common forms, hit counts yield very similar proportions (***-nak*** vs. ***-nek***) to actual frequencies in the language.

# Counts obtained for the stems just listed

| | | | |
|---|---|---|---|
| bibliofilnak | 2 | bibliofilnek | 40 |
| bíbornak | 17 | bíbornek | 0 |
| bíborosnak | 670 | bíborosnek | 0 |
| bicajnak | 5 | bicajnek | 0 |
| bicajosnak | 6 | bicajosnek | 0 |
| bicegnak | 0 | bicegnek | 22 |
| biciklinak | 0 | biciklinek | 83 |
| biciklizésnak | 0 | biciklizésnek | 19 |
| bigottnak | 38 | bigottnek | 0 |
| bigyónak | 44 | bigyónek | 0 |
| bikának | 480 | bikánek | 0 |

# Purpose: verifying proposals made in earlier work

- Earlier work on Hungarian (Szepe 1958, Vágo 1975, Kontra and Ringen 1986, Siptár, and Törkency 2000) has proposed some interesting regularities.
- The stems of interest: those ending in:

| **back vowel** plus one or two **neutral vowels** |
|---|

- In these stems, the frequency of front and back suffix allomorphs depends on two factors.

# The double-neutral effect

- Stems with **back** + **two neutrals** (e.g. *n**o**v**ɛ**mb**ɛ**r*) more frequently take front suffixes than stems with **back** + **one neutral** (e.g. *h**o**t**ɛ**l*)

# The double-neutral effect

- Stems with **back** + **two neutrals** (e.g. *novɛmbɛr*) more frequently take front suffixes than stems with **back** + **one neutral** (e.g. *hotɛl*)

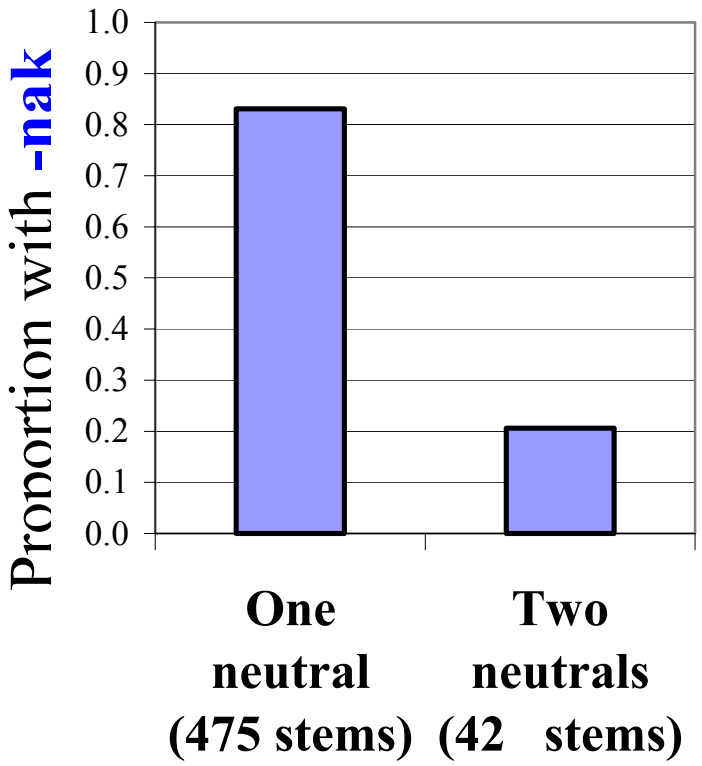
# The height effect

- Stems whose **last neutral vowel** is **lower-mid** (e.g. *hotɛl*) take front suffixes more often than
- stems whose **last neutral vowel** is **mid** (e.g. *kaːveː*);
- which take front suffixes more often than stems whose **last neutral vowel** is **high** (e.g. papiːr).

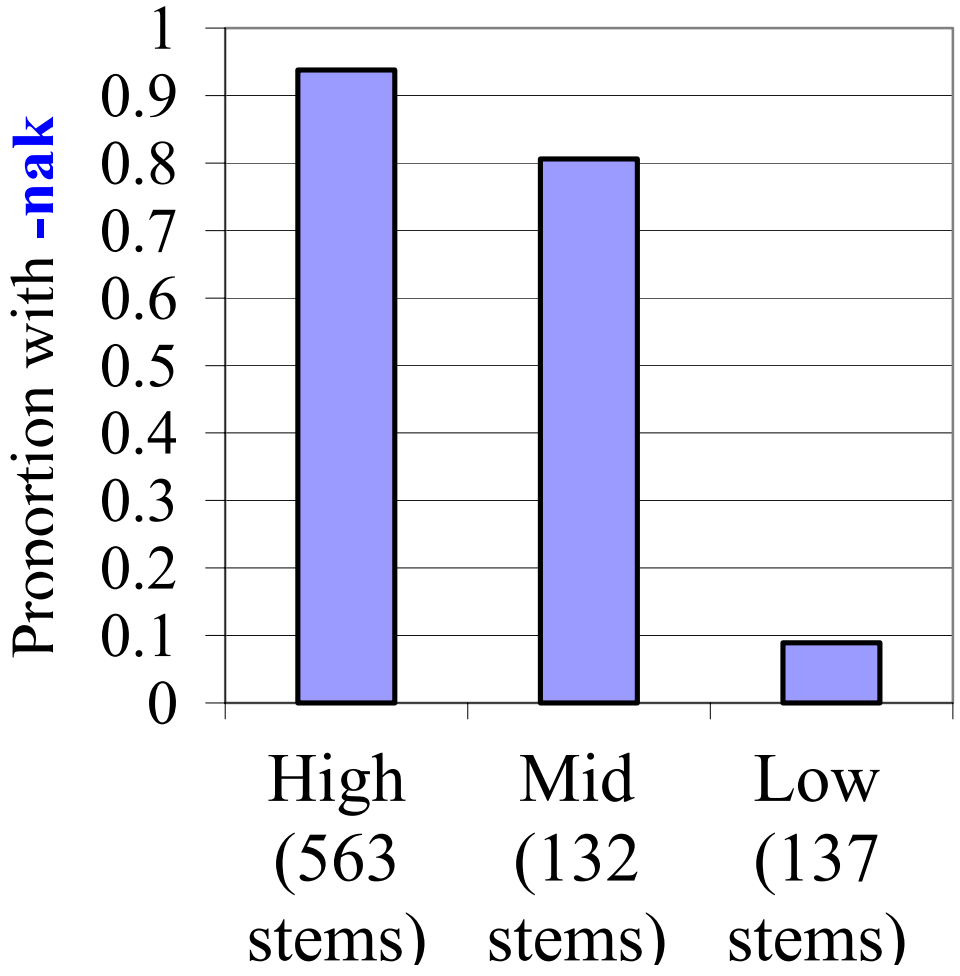
# Word specificity

- These claims are made about the lexicon as a whole.
- Most individual stems always take ***-nak*** or always ***-nek***.
- The effects emerge only when you **aggregate** forms across the phonological category.

# Verifying the double-neutral effect



# Verifying the height effect



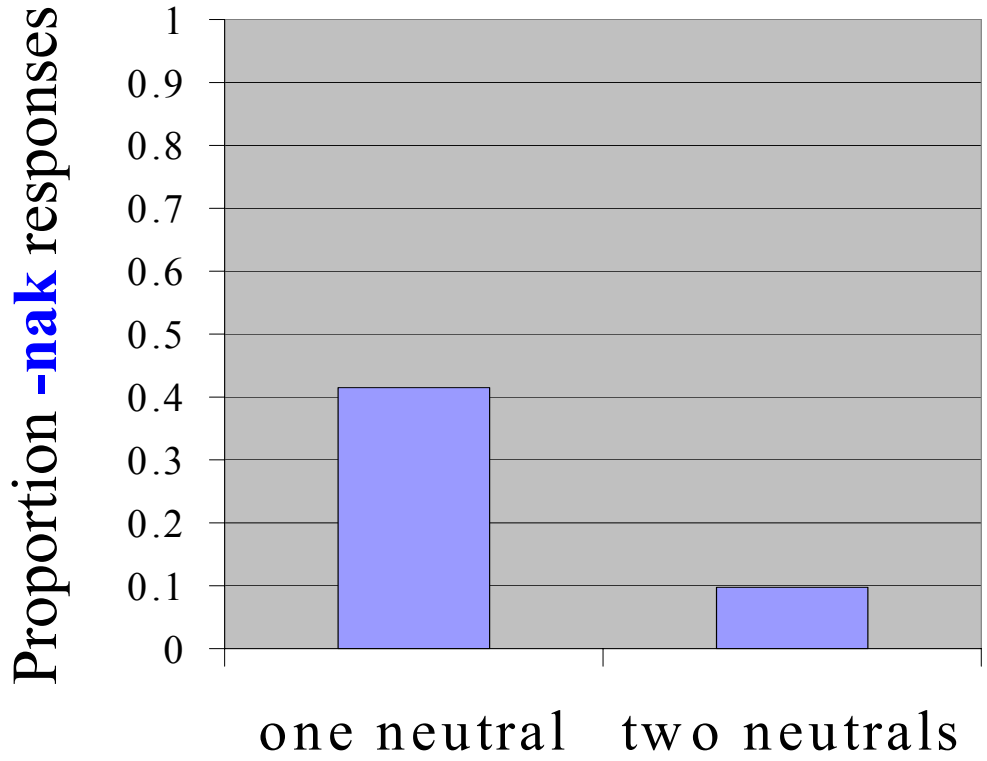
# Why would such data matter?

- Claim: these patterns are not lost on the Hungarian language learner; they are apprehended and internalized.
- Evidence comes from an experiment: given novel ("wug") forms of the relevant phonological shape, speakers behave stochastically, generating ***-nak*** and ***-nek*** forms in proportions that match the lexical statistics of Hungarian.

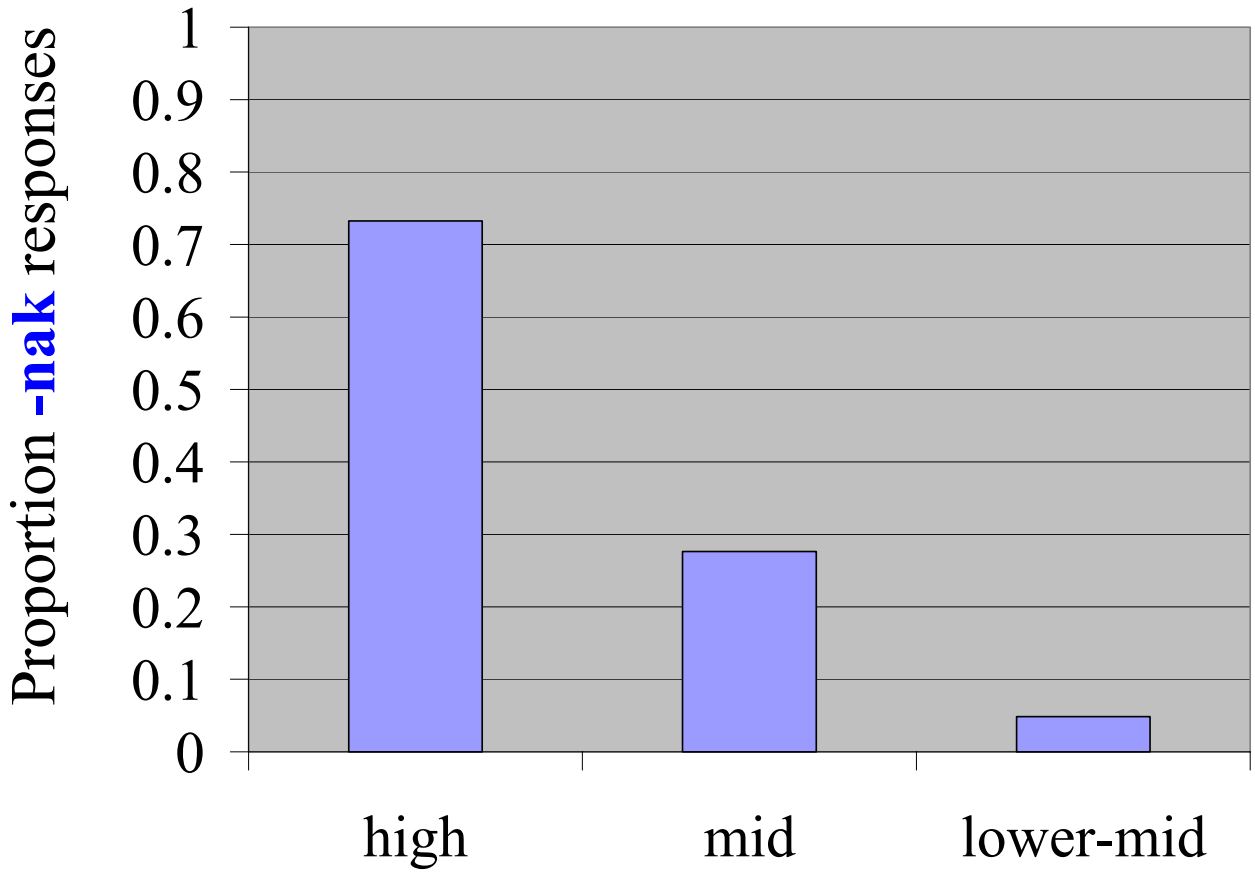
# Details of the Wug test

- Subjects provide the dative form of novel, imaginary Hungarian stems, like ***hádél***, having the relevant vowel sequences.
- We embedded these in sentence frames intended to elicit the dative; either ***hádél-nak*** or ***hádél-nek***
- Count how many ***-nak*** and ***-nak*** responses occur for each wug form (171 native speaker subjects)

# The responses of Hungarian speakers, in the aggregate, show a count effect



# The guesses of Hungarian speakers, in the aggregate, show a height effect



# General picture

- The native speaker possesses a model of the quantitative, as well as qualitative, pattern of the language's paradigms.

# Recent work yielding similar conclusions

- Zuraw (2003 in Bod et al., *Probabilistic Linguistics*)
- Albright and Hayes (*Cognition*, 2003)
- Ernestus and Baayen (*Language*, 2003)
- Pierrehumbert (forthcoming, *LabPhon 8*)

# Should we ignore such data as "extraphonological"?

- I think there are good reasons not to.
  - The data are very systematic.
  - They are based on natural phonological categories, like vowel height.
  - They are eminently **analyzable** — see analysis sections of papers just cited.
  - The analyses use the **normal tools of phonological theory** (features, harmony constraints, ranking, etc.)
- I also think the burden of proof should fall on whoever proposes to ignore data.

# II: MACHINE SEARCHING OF CORPORA FOR GENERALIZATIONS

# Reference


- Albright, Adam and Bruce Hayes (2003) "Rules vs. analogy in English past tenses: a computational/experimental study," *Cognition* 90: 119-161. [http://www.linguistics.ucla.edu/people/hayes/rulesvsanalogy/]

# The project

- **Long term goal**: an automated system that learns the patterns of phonological alternation in paradigms
- **Specific goal**: learn to predict one paradigm member from another.
- **Method**: find the phonological environments of each affix allomorph/segment change, using the algorithm we call *minimal generalization* (Pinker and Prince 1988)
- Application of Albright and Hayes (2003): project **English past tenses** (including irregulars) from their present stems.
- **Test**: can the automated learner's wug-test guesses match those of people?

# Example outputs for Wug verbs

- Past tense of ***spling*** (the model's rating, on a 0-7 scale):

| | |
|---|---|
| ***splung*** | 5.19 |
| ***splinged*** | 5.14 |
| ***splang*** | 4.36 |

- Past tense of ***gezz***:

| | |
|---|---|
| ***gezzed*** | 6.06 |
| ***gozz*** | 3.94 |

# Comparing with Wug test data

- Generally good correlations with native speaker ratings gathered in a Wug test:

| | | |
|---|---|---|
| r = | 0.745 | for regulars |
| | 0.570 | for irregulars |
| | 0.806 | overall |

# Islands of reliability

- We find that *not all regulars are equal*;
- Certain phonological regions (based e.g. on final consonant) are **hyperregular**, in that Wug verbs occupying them are favored even more than usual by native speakers.

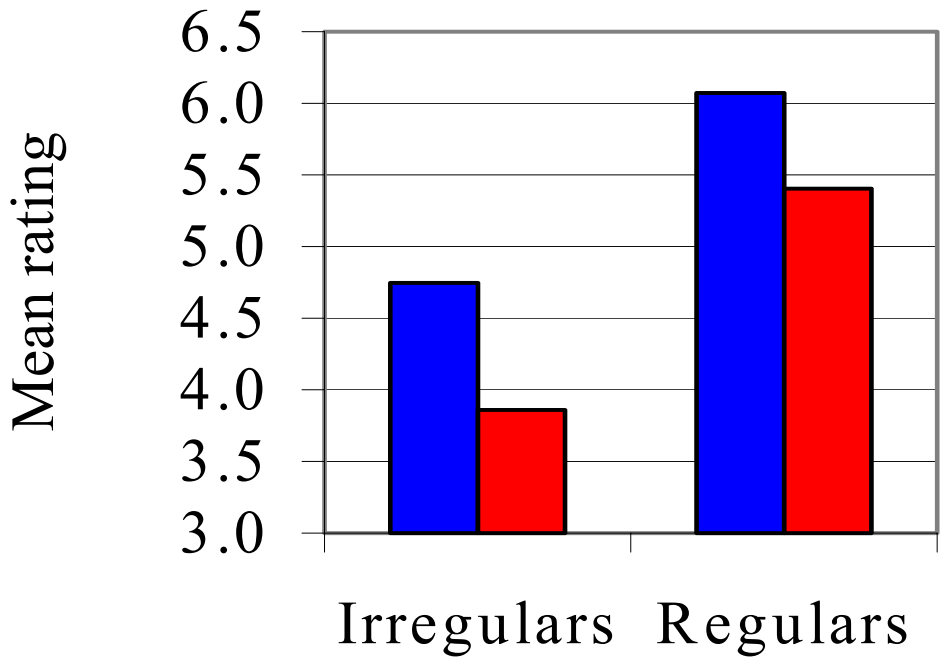
# An example of an island of reliability

- Every verb of English that **ends in a voiceless fricative** ([**f**, **θ**, **s**, **ʃ**]) is regular.
- Our rule-learning system notices this, and thus gives a high score (6.22) to wug verbs ending in voiceless fricatives.
- Speakers tacitly know this as well, as our wug-testees showed by their high ratings for Wug past forms like:

| | | |
|---|---|---|
| ***blafed*** | 6.67 | (scale 1-7) |
| ***wissed*** | 6.28 | |
| ***teshed*** | 6.22 | |

# More generally

- Native speakers rate wug pasts as higher when they occupy an **island of reliability** than when they **do not**.



# Similar results in other languages

- Albright, Adam (2002) Islands of reliability for regular morphology: Evidence from Italian. *Language* 78: 684-709.
- Albright, Adam, Argelia Andrade and Bruce Hayes (2001) Segmental environments of Spanish diphthongization. *UCLA Working Papers in Linguistics* 7, 117-151. [http://www.linguistics.ucla.edu/people/hayes/Segenvspandiph/]

- These studies, like those cited earlier, indicate a richer knowledge of the inflectional pattern than previous research has posited.

# III. THE ARTIFICIAL-LANGUAGE PARADIGM

# Is UG testable?

- The hypothetical question:

> *"Would a language with these properties be learnable?"*

  is common among linguists concerned with questions of Universal Grammar.

- This question is perhaps not as hypothetical as it used to be—due to **artificial-language learning experiments**.

# Form of the experiments

- Construct **miniature languages** that contrast with respect to the relevant properties.
- Give subjects a chance to learn the languages.
- Success/failure, or just relative difficulty, can be informative.

# Colin Wilson's experiment

- Reference:
  - 2003. Experimental investigation of phonological naturalness. In G. Garding and M. Tsujimura (eds.), *West Coast Conference on Formal Linguistics* 22. Cambridge, MA: Cascadilla Press, 533-546.
- Subjects were given one of two artificial languages:
  - the **nasal harmony** language
  - the "**nasals after velars**" language

# The Nasal Harmony language: sample words

| | |
|---|---|
| [du**m**e-**n**a] | [u**k**o-**l**a] |
| [bi**n**u-**n**a] | [di**g**e-**l**a] |
| | [su**t**o-**l**a] |
| | [da**b**u-**l**a] |

- This is a **phonologically natural** language
- Real-life parallels in Lamba, Nyangumarda, Ulithian

# The "Nasals after Velars" language: sample words

| | |
|---|---|
| [u**k**o-**n**a] | [su**t**o-**l**a] |
| [di**g**e-**n**a] | [da**b**u-**l**a] |
| | [du**m**e-**l**a] |
| | [bi**n**u-**l**a] |

- This is a **phonologically unnatural** language, without real-life parallels.

# Training and testing the subjects

- **Training**: 20 items, each presented twice
  - Task: *remember these words*

- **Testing**: 80 items, of which 20 old and 60 new
  - Task: *have you heard this word before*?

- Half the new test items were "grammatical" in the training language; the other half "ungrammatical".

# Results

- **Nasal harmony language**:
  - With significantly greater than chance frequency, subjects were likely to think that new items that were "grammatical" in the training language were words they had heard before.

- **"Nasals after velars" language**:
  - No significant effect

# Wilson's interpretation

> "The results … provide experimental support for the claim, widely held in theoretical phonology, that certain process types have a privileged cognitive status."

- Elaborating: either
  - **phonetic naturalness**, or
  - the basis in a **logical identity relation**

  makes the phonologically "natural" language learnable.
- People are not arbitrary inductive sponges.

# Some other recent artificial language experiments

- Pater and Tessier (2003)
- Nowak et al. (2003)
- Peperkamp and Dupoux (in press)

These vary in whether they found the UG effect they were looking for.

# IV. PHONOLOGICAL EXPERIMENTS: INFLECTING THE UNINFLECTABLE

# Reference


- Zuraw, Kie (2005) “Cluster splittability in Tagalog: corpus and survey evidence,” paper given at the 13th Meeting of the Austronesian Formal Linguistics Association, UCLA, Los Angeles, CA.

# Premises of the method

- Borrowings are often **indeclinable**, lacking inflected forms.
- Suppose we persuade speakers to go ahead and inflect them.
- If the borrowings have novel stem shapes, we will see what principles guide speakers in extending their grammar into new territory...

# The puzzle of cluster-splitting infixation

- Example:
  - Tagalog *gradwet* 'graduate' receives the -***um***- infix as either:

    gr-**um**-adwet

    or

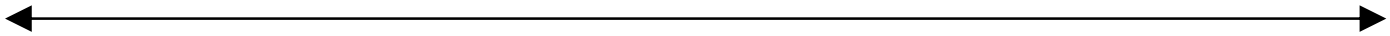
    g-**um**-radwet

- Questions:
  - Why are both outcomes possible?
  - What factors favor the competing outcomes?

# Background: typology of cluster-splitting epenthesis

- Fleischhacker (2002) studied the related phenomenon of **epenthesis in loanword adaptation** (sta → sǝta, ǝsta)
- She found a cross-linguistic **hierarchy of splittability** for *sibilant* + *consonant* clusters:

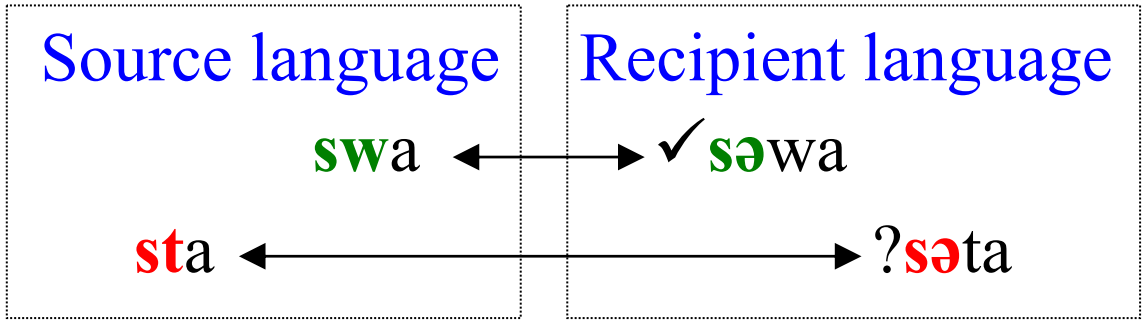
***least splittable*** ST Sm Sn Sl Sr SW ***most splittable***



where
S = sibilant
T = stop
W = glide

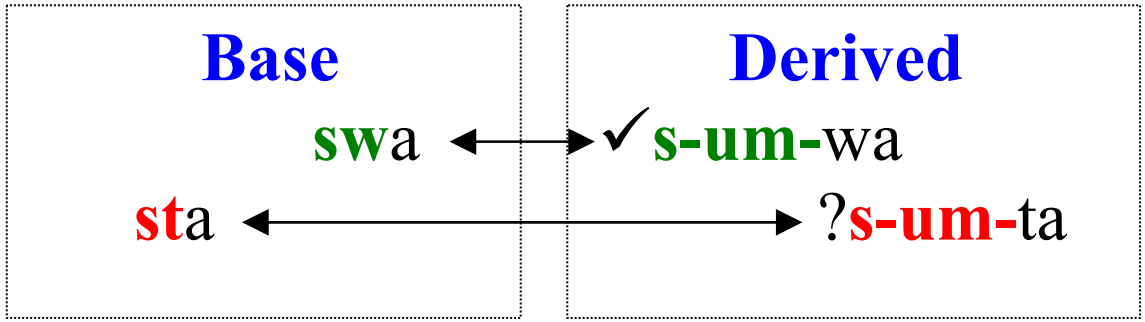
# Fleischhacker's explanation

- Crucial factor is **perceptual similarity**
- Loan adaptation favors maintaining perceptual similarity to the source (Peperkamp 2004)
- Release of S into a **sonorous** consonant is perceptually closer to release into a nonsonorous consonant would be



# Zuraw, adapting Fleischhacker

- Phonological constraints that guide **infixation** are also sensitive to similarity: here, **based-derived** similarity.



# Prediction: sonority effects on cluster-splitting infixation

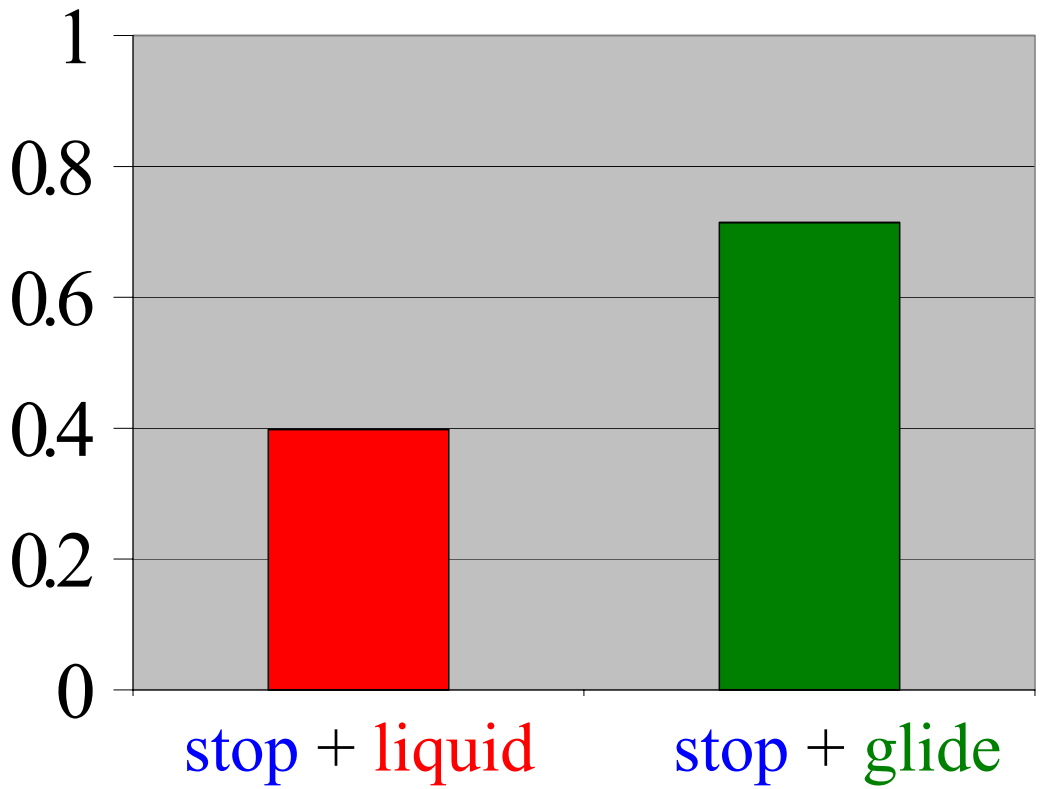
- The higher the sonority of $C_2$ in $C_1C_2$, the more likely infixes should be placed / $C_1$___$C_2$ (and not / $C_1C_2$___).

# Confirmation: corpus study

- Basis: 20 million Tagalog words gathered from the Web
- Next slide: percentage split by infix: ***stop*** + ***liquid*** vs. ***stop*** + ***glide***

*l*

# Percentage split by infix: stop + liquid vs. stop + glide



# Confirmation II: Wug test

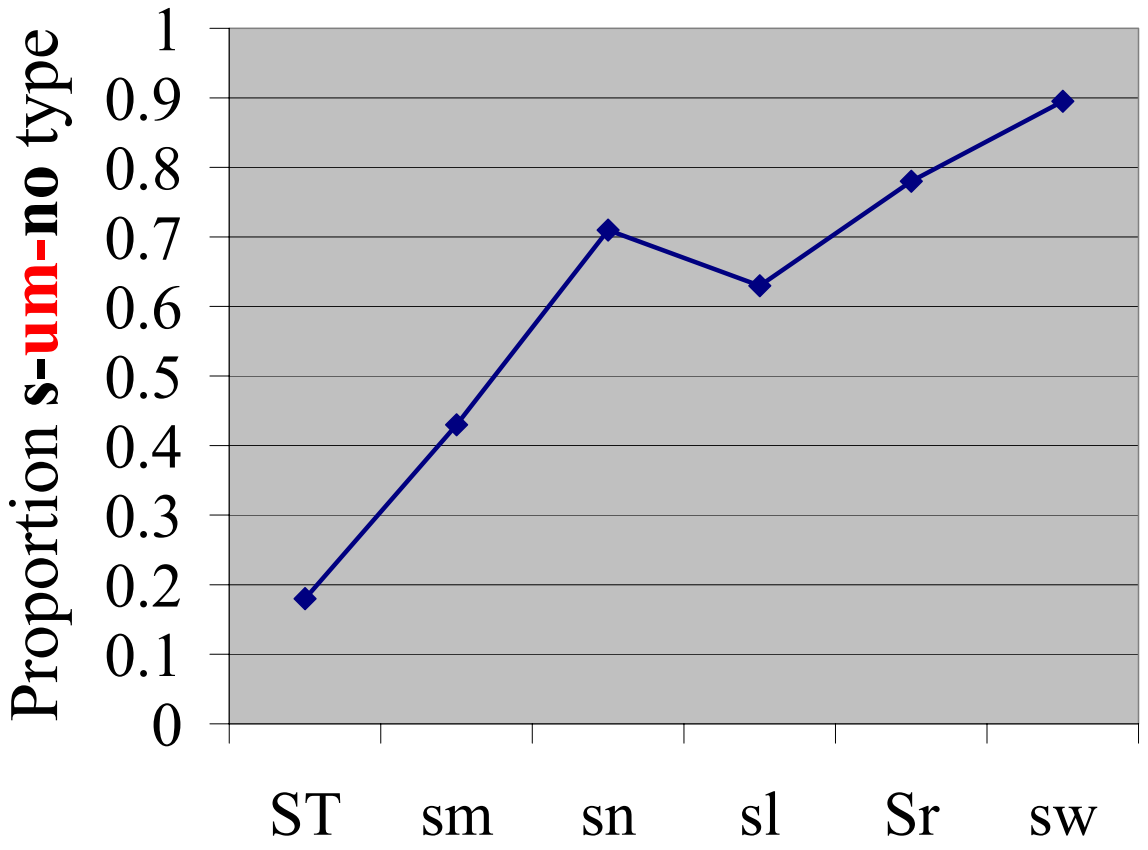
- Tagalog speakers generally treat sC words with a prothetic vowel ([iskul]), however...
- They "sometimes use non-prothesized forms as isolated words, but very rarely with infixation"—i.e. they are largely indeclinable.
- Hence if a speaker is asked to epenthesize into a non-prothesized sC stem, she must **extend her existing grammar** to decide where to put the infix:

sno 'snow' → s-**um**-no, sn-**um**-o

# Testing by long distance over the Web

- Zuraw designed software to administer a Wug test over the Web, eliciting:
  - preference ratings (scale: 1-7) scale for novel forms like **s-um-no** vs. **sn-um-o**.

# Results: preferred epenthesis location by cluster type



# Zuraw's interpretation

- The speakers, acting in novel circumstances, chose the infix location that would **maximize phonetic similarity** of infixed form to base; i.e. when **$C_2$** is more sonorous.

# Zuraw's proposed inference

> "Not all imaginable grammars are equally good from the learner/speaker's point of view"

- Specifically, the principle of phonetic similarity guides native speakers in their active grammatical behavior.
- It cannot be reduced to an "error factor", found only in the diachronic evolution of a language.
- Zuraw cites Ohala (1981), Blevins (2004) as among the works defending a diachronic, error-based approach.

# SUMMARY AND CONCLUSION: THE ROLE OF NEW DATA SOURCES

# What were the "favorite facts"?

- There were four:
  - Speakers **project lexical variation into output variation**, when generating new forms (Hungarian).
  - This is true even when the lexical variation involves **detailed environments** (English past tense islands).
  - A **"natural"** (nasal harmony) language proves learnable in circumstances where a comparable unnatural language is not.
  - Tagalog speakers **follow a principle of phonetic similarity** when they are asked to extend the native pattern of infixation.

# What where might further work along these lines go?

- I would like to see it test **specific formal proposals in phonological theory**.
- The material discussed here mostly bears on very general issues, but with the techniques established it should not be hard to move on to more specific questions.

# Final conclusions

- The proper stance toward new kinds of facts in phonology is **enthusiastic receptivity** (maintaining of course the same standards of rigor we observe elsewhere)
- The "classical" data sources — elicitation, grammars — and "classical" forms of formal analysis will continue to be vital, and central, to our field
- But a broader data perspective is important to the continuing scientific progress of phonology.

# Thank you

For reference list, comments, and any afterthought queries

please send email to bhayes@humnet.ucla.edu

# Some references


- Blevins, Juliette 2004. *Evolutionary Phonology*. Cambridge: Cambridge University Press.
- Ernestus, Miriam and Harald Baayen (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch, *Language* 79, 5-38
- 
- Nowak, Pawel, Anne Pycha, Eurie Shin & Ryan Shosted. 2003. Phonological rule learning and its implications for a theory of vowel harmony. *WCCFL* 22.
- Ohala, John J. 1981. The listener as a source of sound change. In: C. S. Masek, R. A. Hendrick, & M. F. Miller (eds.), *Papers from the Parasession on Language and Behavior*. Chicago: Chicago Ling. Soc. 178 - 203.
- Pater, J. and A.-M. Tessier. 2003. Phonotactic Knowledge and the Acquisition of Alternations. In M.J. Solé, D. Recasens, and J. Romero

(eds.) *Proceedings of the 15th International Congress on Phonetic Sciences, Barcelona*. 1777-1180.

- Peperkamp, Sharon (2004) A psycholinguistic theory of loanword adaptations. *Proceedings of the 30th Annual Meeting of the Berkeley Linguistics Society*.
- Pierrehumbert, J. (forthcoming) An Unnatural Process. *Laboratory Phonology 8*, Mouton de Gruyter.
- Pinker, S. & Prince, A. (1988) On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28, 73-193.
- Wilson, Colin. (2003). Experimental investigation of phonological naturalness. In G. Garding and M. Tsujimura (eds.), *West Coast Conference on Formal Linguistics 22*. Cambridge, MA: Cascadilla Press, 533-546.
- Zuraw, Kie (2005) "Cluster splittability in Tagalog: corpus and survey evidence," paper given at the 13th Meeting of the Austronesian Formal Linguistics Association, UCLA, Los Angeles, CA.