On the creation of a pronunciation dictionary for Hungarian

Stephen Grimes
stgrimes@indiana.edu
Midwest Computational Linguistics Colloquium
UIUC, May 20-21, 2006

Pronunciation Dictionary?

Consider the word *szabadság ‘liberty; freedom’*

- The ‘d’ is pronounced [t] due to voicing assimilation.
- The ‘ds’ consonant cluster is actually pronounced [cc] due to affrication.
- ‘sz’ is a digraph
- ‘á’ uses a diacritic; there are character encoding issues

Outline for today:

- Motivation for a pronunciation dictionary
- A strange kind of phonology? The relationship between spoken and written Hungarian
- Error identification, deliberate omissions, and future additions
- Applications of a pronunciation dictionary in cross-linguistic study and within Hungarian linguistics

Motivation: what is it used for?

- Studying properties of the “mental lexicon”, phonological neighborhoods, Neighborhood Activation Model
- Phonological complexity
- Phonotactics, phonostatistics
- Statistical models of sonority
- Establishing a markedness hierarchy
- Used in acoustic model for speech recognition
- Interesting when viewed as a phonology problem

- What it’s not used for!
- For native speakers or foreigners seeking such as resource, the “Magyar kiejtési szótár” already exists (Fekete, 1995).

Why Hungarian?

- Agglutinative language with high morpheme::word ratio.
- More inflectional morphemes than English.
- Several computational tools are already available (Kornai, 1986; Halácsy et al, 2004)
- Relatively close relationship between writing and pronunciation allows for this.
- Studying Hungarian is fun!

The Hoosier Mental Lexicon

- HML (Nusbaum et al., 1984): developed at the Psychology Department at Indiana University
- For ~20,000 English words, HML gives written form, a broad phonetic transcription, and the corpus frequency of the word.
- Additionally, the HML contains data on word familiarity ratings, judged by subjects on a scale of one to seven.

Goals for Hungarian dictionary

- Correct the majority of sound/symbol discrepancies.
- One-to-one mapping of sound to symbol
- Use ASCII-based alphabet for portability
- Phonetic forms should represent the idealized standard dialect, present day...
- Large vocabulary: 67,000 words
- Include word frequencies

Orthography	OGOB7 (Szigetvári)
cs	C
ch	H
dzs	D
dz	F
gy	G
ly	L
ny	N
sz	S
ty	T
zs	Z
á	A
é	E
í	I
ó	O
ú	U
ö	w
o	W
ü	y
u	Y

Outline for today:

- Motivation for a pronunciation dictionary
- A strange kind of phonology? The relationship between spoken and written Hungarian
- Error identification, deliberate omissions, and future additions
- Applications of a pronunciation dictionary in cross-linguistic study and within Hungarian linguistics

Changes needed to create dictionary

- historical spelling variants
- digraphs and trigraphs
- phonological rules
  - vowel, consonant length alternations
  - several types of assimilation
  - glide insertion
- morphophonological rules
  - imperative forms

digraphs and trigraphs

Fortunately digraph ambiguity occurs only in compound words.

[zs]
rézsun ‘on the slope’ (rézsu ’slope’, uncommon)
rézsün ‘copper hedgehog’ (réz ‘copper’, sün ‘hedgehog’)

[szs]
sertészsír ‘pork grease’ (sertés ’pig’, zsír ‘grease’)
kertészsír ‘gardener’s grave’ (kertész ’gardener’, sír ’grave’)

[cs]
lécsín ‘liquid beauty’ or ‘slat track’

[tty]
hattyúk ‘six hens’ or ‘swans’

Are such examples widespread or isolated???

Phonology: assimilation

- Voicing assimilation
- Palatalization
- Affrication
- Nasal place assimilation

NB. Some types of assimilation are marked in the orthography.

/ember + vel/ -> [emberrel]

Phonology: regressive voicing assimilation

- The voicing feature for a consonant must agree with the voicing feature of a following consonant

Exceptions:

- 1. do not undergo assimilation: h, j, m, n, ny, r
- 2. do not cause assimilation: v

Phonology: regressive voicing assimilation

Written Form	Pronounced Form	Gloss
abszolút	apszolút	’absolute’
joghurt	jokhurt	’jogurt’
olvasd el	olvazsd el	’read it’
népdal	nébdal	’folksong’
kútban	kúdban	‘in the well’
húsdaráló	húzsdaráló	‘meatgrinder’
kerékgyártó	keréggyártó	’wheelmaker’

Phonology: nasal place assimilation

- A nasal consonant must agree with the place of articulation feature of a following consonant.

Written Form	Pronounced Form	Gloss
szénpor	szémpor	’coal dust’
különben	különmben	’otherwise’
szenved	szemved	’suffer’
mondja	monygya	’she says’

Phonology: palatalization

- A coronal stop is palatalized, often in imperative or 3rd singular forms.
- {t, d, n} + j -> {tty, ggy, nny}

Written Form	Pronounced Form	Gloss
látja	láttya	‘he sees it’
adjuk	aggyuk	‘we give it’
menjen	mennyen	‘he should go’

Phonology: affrication

- A plosive and a following sibilant coalesce into an affricate of the appropriate place of articulation. The resulting affricate is usually a long consonant, unless reduced due to being adjacent to another consonant.
- {t, ty} + s -> ccs {t, ty} + sz -> cc

Written Form	**Pronounced Form**	**Gloss**
váltson	válcson	‘it should change’
szabadság	szabaccság	‘freedom’
egyszer	eccer	‘once’
maradsz	maracc	‘you stay’

Phonology: glide insertion

A glide consonant [j] is inserted to interrupt hiatus between two vowels whenever one of the vowels is i, í, e, or é.

Written Form	Pronounced Form	Gloss
tea	teja	’tea’
szia	szija	’hello’
hiába	hijába	‘in vain’
nénié	nénijé	‘the aunt’s’
dió	dijó [gyó]?	‘walnut’
kiöl	kijöl	‘extinguish’

Consonant length alternations

- Shortening – Long consonants are produced short before or after consonants
  - orrhang [orhang] ‘nasal’
- Lengthening (treat these as exceptions?)
  - egy [eggy], egyet [eggyet]
  - lesz [lessz]
  - edz [eddz], bridzs [briddzs] (affects all dz, dzs in coda position)
- Deletion in triconsonantal clusters
  - -middle consonant in tri-consonantal sequence can be elided
  - -t and d are particularly susceptible to this
  - mindnyájan [minnyájan], kezdhetjük [keszhettyük] / [keszthettyük]

Vowel length alternations

Written Form	Pronounced Form	Gloss
ors	örs	patrol
gyujt	gyüjt	collect
ródli	ródli	sled
csúzli	csúzli	slingshot

Vowel length phonotactics

- **Mid vowels in word-final position**
  - Typically marked in the orthography
  - Exceptions are function words: no (well, [interjection]), ö ([ahh])
  - Include phonotactic rule to apply to any foreign words

- **High vowels in word-final position**
  - High vowel exceptions (how are they pronounced)?
  - Exceptions?:
    - hamu, Pittyu, falu, kenu, kapu, daru, áru, anyu, apu, saru
    - menü, eskü

Some implementation issues

- Coding issues
  - Each phoneme needs to be coded for articulatory features and sonority value
  - Need morpheme boundaries
  - Know word POS (function vs. content words)
- Care must be taken to apply rules in order
- Currently implemented in Perl using regular expressions

Outline for today:

- Motivation for a pronunciation dictionary
- A strange kind of phonology? The relationship between spoken and written Hungarian
- Error identification, deliberate omissions, and future additions
- Applications of a pronunciation dictionary in cross-linguistic study and within Hungarian linguistics

Omissions from current version

- Correctly identifying compounds: rézsün, sertészsír [possibly not hard to correct]
- Phonological rules applying only in presence or absence of a morpheme boundary or only to certain parts of speech
- “Slang” pronunciations / truncations:
  - asszem (azt hiszem) “I believe (that)”
  - nemtom (nem tudom) “I don’t know”
  - valszeg (valószínuleg) “probably”

Error assessment

- Two types of errors
  - Need to have some notion of precision and recall:

    due to overlooked cases; rules applied incorrectly
  - variability in the language; lack of genuine language standard creates
- How to assign value to correctness?
  - random sample of words to two speakers, see how often they agree on correctness?
- Find more crucial examples to check:
  - Words containing digraphs, words edited by one of my phonological rules
  - egyszer → eGSer → eTSer → eccer

Future developments

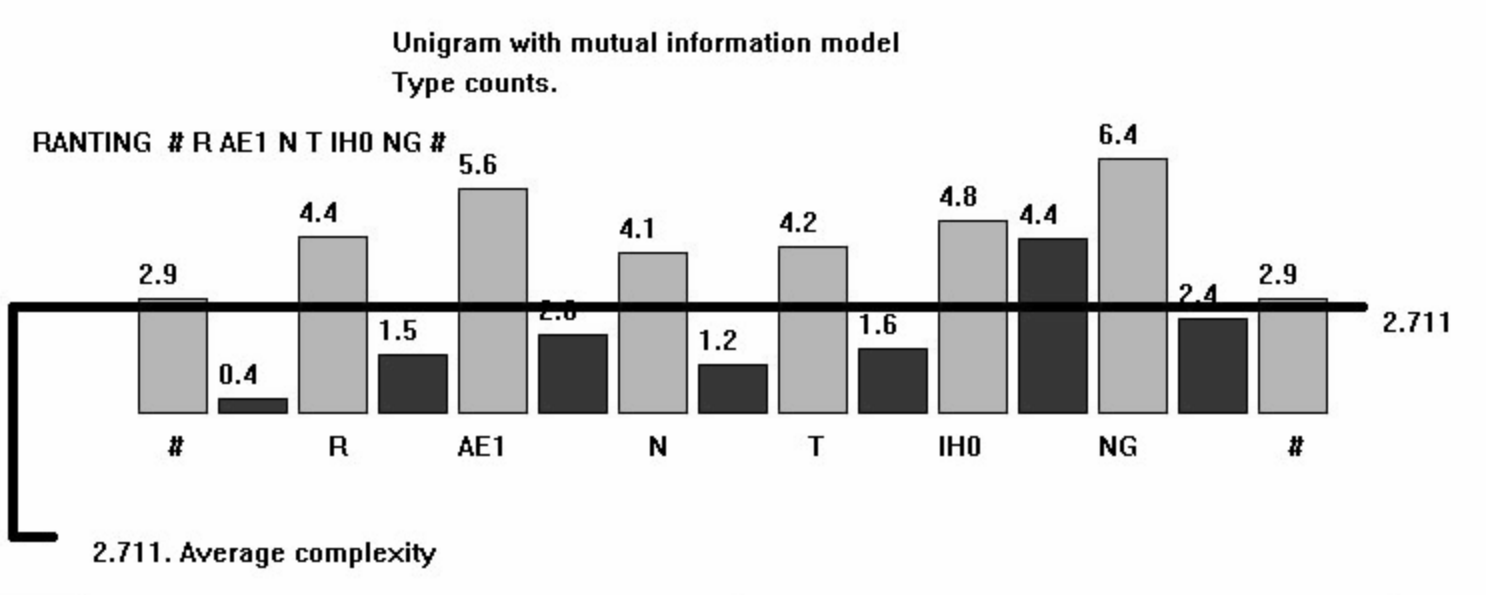
- Will there be standards for phonological lexicons / pronunciation dictionaries??
- Add age of acquisition information or familiarity ratings for each word
- Include subsets of pronounced forms, such as only vowels or only consonants and examine bigrams of these structures
- Include syllable boundaries, CV skeletal structure
- Support for some types of dialectical variation
- Perceptual information: confusion matrices

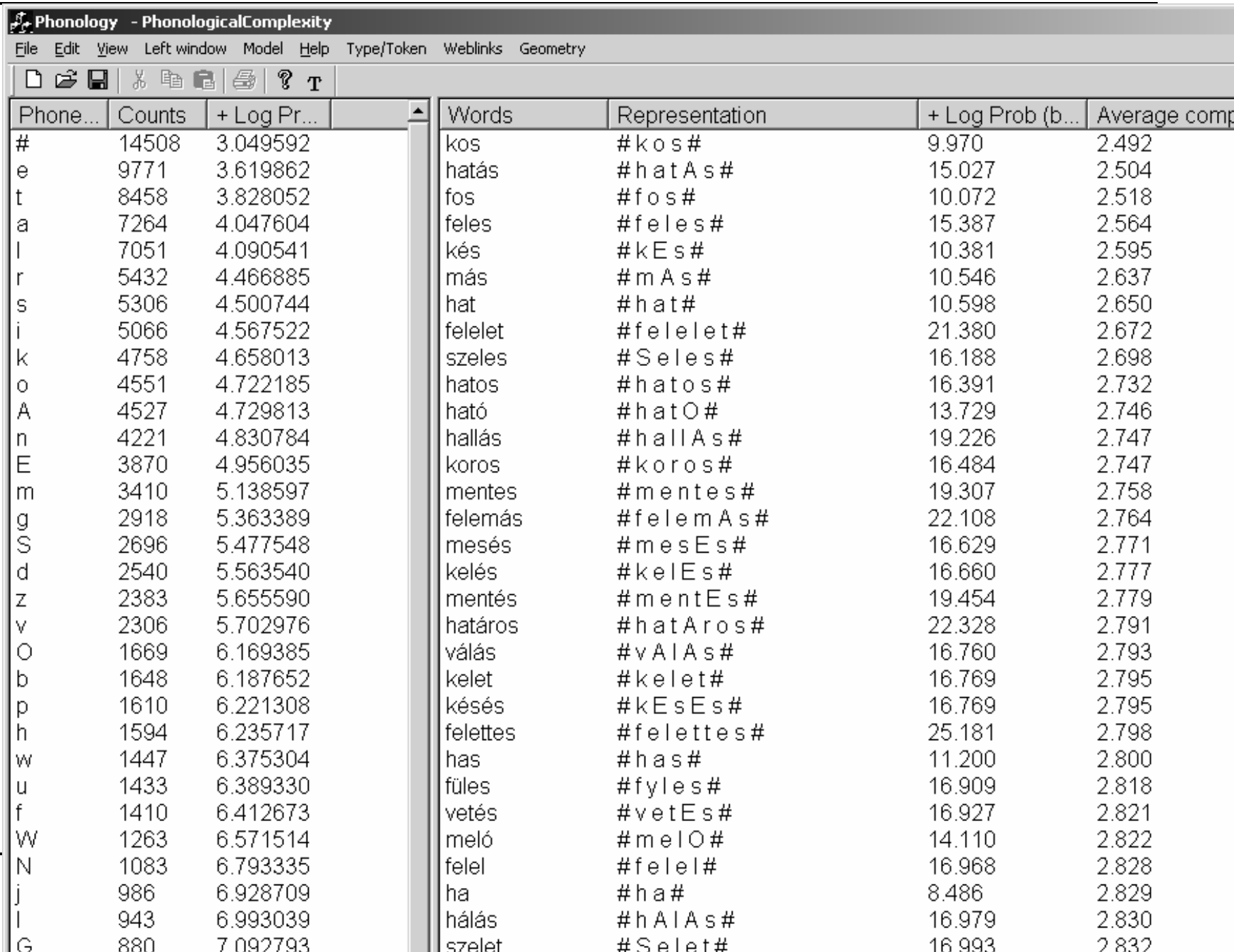
Outline for today:

- Motivation for a pronunciation dictionary
- A strange kind of phonology? The relationship between spoken and written Hungarian
- Error identification, deliberate omissions, and future additions
- Applications of a pronunciation dictionary in cross-linguistic study and within Hungarian linguistics

Phonotactics and probability

Phonological Complexity (Goldsmith, 2002)





Phone...	Counts	+ Log Pr...
#	14508	3.049592
e	9771	3.619862
t	8458	3.828052
a	7264	4.047604
l	7051	4.090541
r	5432	4.466885
s	5306	4.500744
i	5066	4.567522
k	4758	4.658013
o	4551	4.722185
A	4527	4.729813
n	4221	4.830784
E	3870	4.956035
m	3410	5.138597
g	2918	5.363389
S	2696	5.477548
d	2540	5.563540
z	2383	5.655590
v	2306	5.702976
O	1669	6.169385
b	1648	6.187652
p	1610	6.221308
h	1594	6.235717
w	1447	6.375304
u	1433	6.389330
f	1410	6.412673
W	1263	6.571514
N	1083	6.793335
j	986	6.928709
I	943	6.993039
G	880	7.092793

Words	Representation	+ Log Prob (b...	Average comp
kos	#k o s#	9.970	2.492
hatás	#h a t A s#	15.027	2.504
fos	#f o s#	10.072	2.518
feles	#f e l e s#	15.387	2.564
kés	#k E s#	10.381	2.595
más	#m A s#	10.546	2.637
hat	#h a t#	10.598	2.650
felelet	#f e l e l e t#	21.380	2.672
szeles	#S e l e s#	16.188	2.698
hatos	#h a t o s#	16.391	2.732
ható	#h a t O#	13.729	2.746
hallás	#h a l l A s#	19.226	2.747
koros	#k o r o s#	16.484	2.747
mentes	#m e n t e s#	19.307	2.758
felemás	#f e l e m A s#	22.108	2.764
mesés	#m e s E s#	16.629	2.771
kelés	#k e l E s#	16.660	2.777
mentés	#m e n t E s#	19.454	2.779
határos	#h a t A r o s#	22.328	2.791
válás	#v A l A s#	16.760	2.793
kelet	#k e l e t#	16.769	2.795
késés	#k E s E s#	16.769	2.795
felettes	#f e l e t t e s#	25.181	2.798
has	#h a s#	11.200	2.800
füles	#f y l e s#	16.909	2.818
vetés	#v e t E s#	16.927	2.821
meló	#m e l O#	14.110	2.822
felel	#f e l e l#	16.968	2.828
ha	#h a#	8.486	2.829
hálás	#h A l A s#	16.979	2.830
szelet	#S e l e t#	16.993	2.832

Phonological lexicon?

- Neighborhood Activation Model (Luce, 1986; Luce and Pisoni, 1998; Barlow, 2000)
- Probabilistic phonotactics (Vitevitch and Luce, 1999)
- Developing accurate models of phonological similarity (e.g. Kapatsinski, to appear)

CALCULATE PHONOTACTIC PROBABILITY

Type or copy and paste your data here. Press [Enter] after each line.

```
cut
k^t
```

Calc your Entry

Clear your Entry

The results of your calculation are displayed here. You may copy and paste results to another program for further analysis.

```
cut
0.0075 0.0221 0.0660
0.0000 0.0027
1.0956 1.0027

k^t
0.0927 0.0392 0.0660
0.0043 0.0024
1.1979 1.0067
```

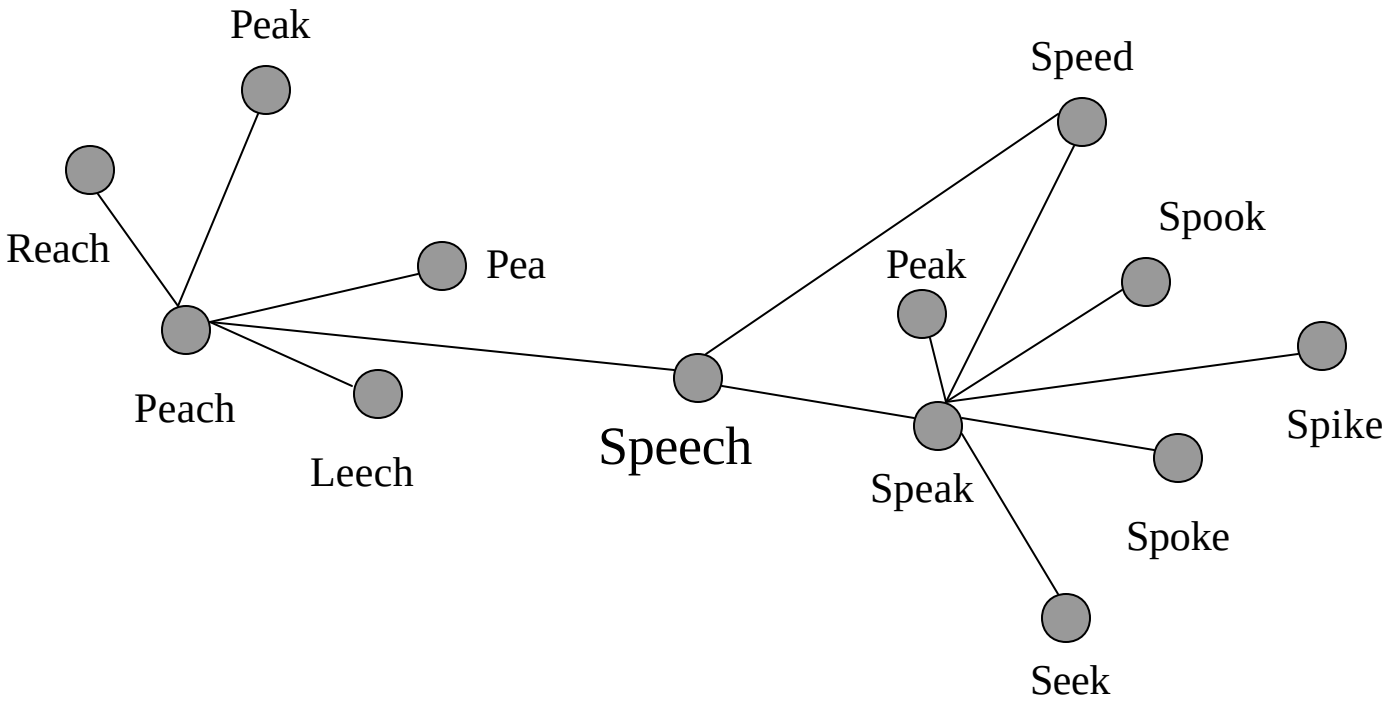
Return to Phonotactic Probability Home Page

The Mental Lexicon as a Graph

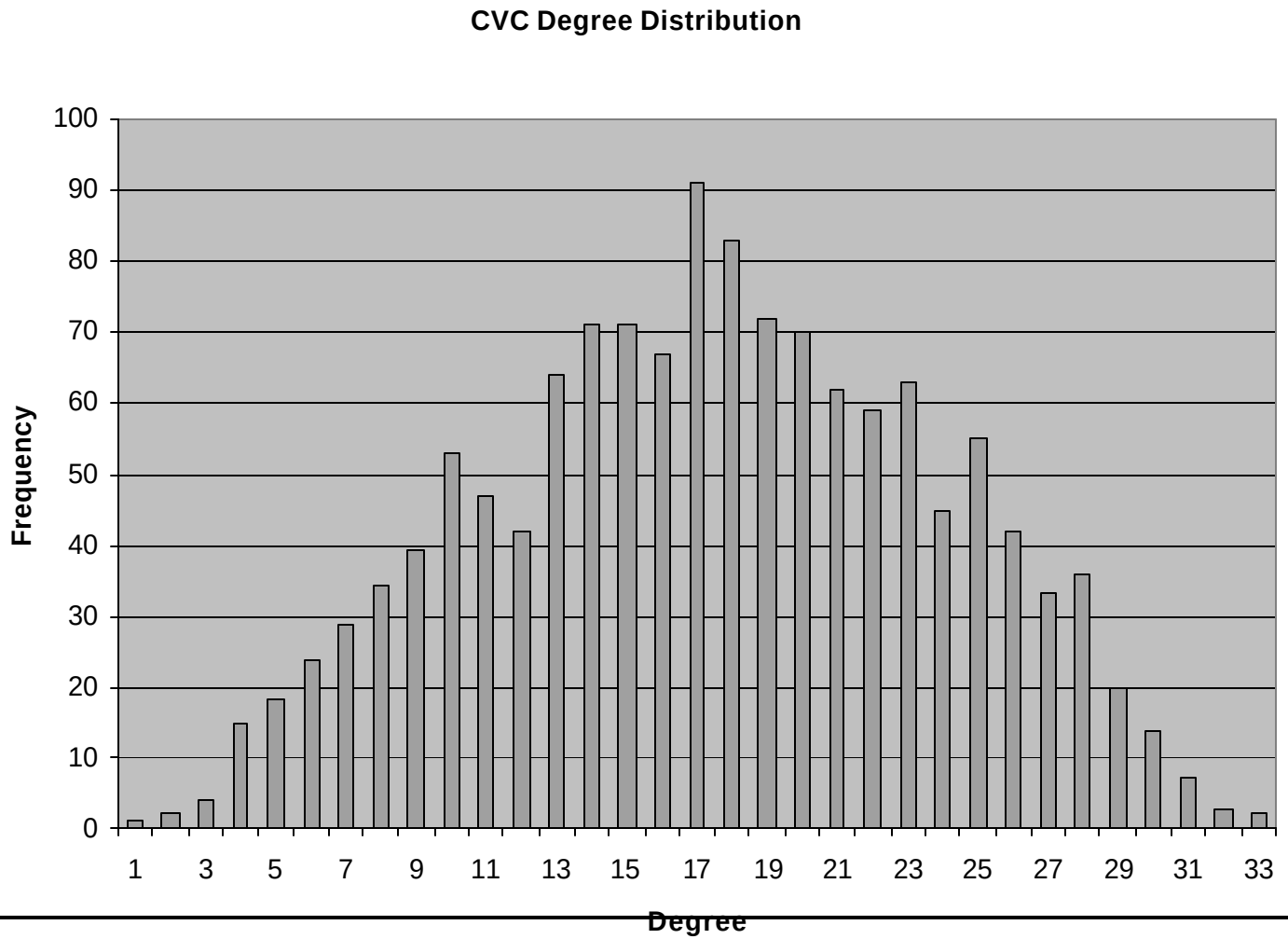
Work by Mike Vitevitch (unpublished) and Gruenenfelder and Pisoni (to appear)

- Each (phonetically distinct) word is a node
- A link is placed between two words if they differ by exactly one phoneme
  - Deletion, addition, or substitution
  - Neighbor, as defined by Luce and Pisoni, Landauer and Streeter, Greenberg and Jenkins
- What are the properties of this graph?

Modeling the Lexicon as a Graph

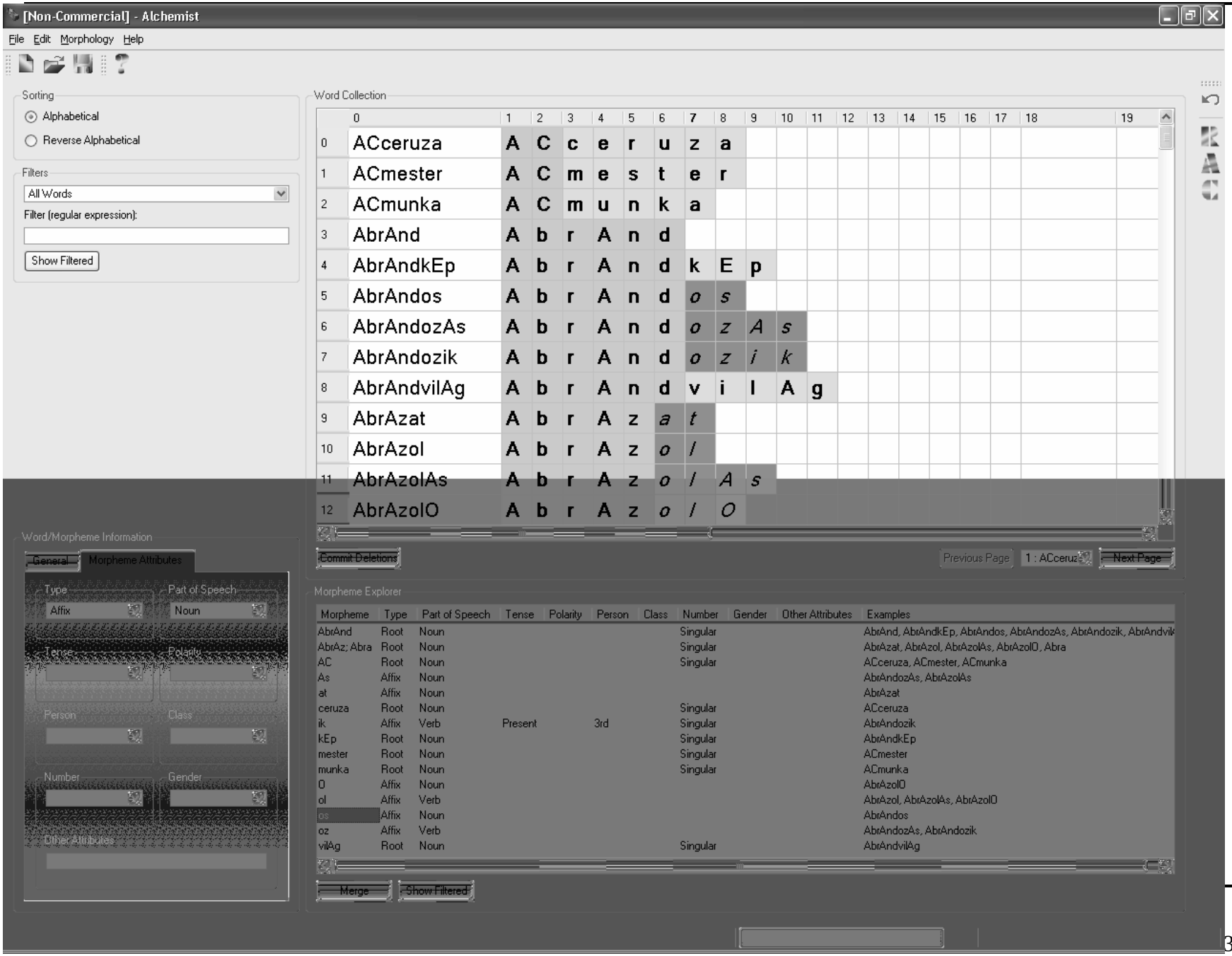


Degree Distribution for CVCs



A basis for further development of resources

- Add further items related to phonological lexicon
- Can serve as a basis for developing a morphologically annotated corpus of Hungarian
  - Morphological parsers for Hungarian exist
  - Using orthographic form, do alignment with the pronunciation dictionary to create a morphologically-annotated pronunciation dictionary.


[Non-Commercial] - Alchemist
File Edit Morphology Help
Sorting
Alphabetical
Reverse Alphabetical
Filters
All Words
Filter (regular expression):
Show Filtered
Word Collection
0 ACceruza A C c e r u z a
1 ACmester A C m e s t e r
2 ACmunka A C m u n k a
3 AbrAnd A b r A n d
4 AbrAndkEp A b r A n d k E p
5 AbrAndos A b r A n d o s
6 AbrAndozAs A b r A n d o z A s
7 AbrAndozik A b r A n d o z i k
8 AbrAndvilAg A b r A n d v i l A g
9 AbrAzat A b r A z a t
10 AbrAzol A b r A z o l
11 AbrAzolAs A b r A z o l A s
12 AbrAzolO A b r A z o l O
Word/Morpheme Information
General
Morpheme Attributes
Type
Affix
Part of Speech
Noun
Tense
Polarity
Person
Class
Number
Gender
Other Attributes
Commit Deletions
Previous Page
1 : ACceruz
Next Page
Morpheme Explorer
Morpheme Type Part of Speech Tense Polarity Person Class Number Gender Other Attributes Examples
AbrAnd Root Noun Singular AbrAnd, AbrAndkEp, AbrAndos, AbrAndozAs, AbrAndozik, AbrAndvil
AbrAz; Abra Root Noun Singular AbrAzat, AbrAzol, AbrAzolAs, AbrAzolO, Abra
AC Root Noun Singular ACceruza, ACmester, ACmunka
As Affix Noun AbrAndozAs, AbrAzolAs
at Affix Noun AbrAzat
ceruza Root Noun Singular ACceruza
ik Affix Verb Present 3rd Singular AbrAndozik
kEp Root Noun Singular AbrAndkEp
mester Root Noun Singular ACmester
munka Root Noun Singular ACmunka
O Affix Noun AbrAzolO
ol Affix Verb AbrAzol, AbrAzolAs, AbrAzolO
os Affix Noun AbrAndos
oz Affix Verb AbrAndozAs, AbrAndozik
vilAg Root Noun Singular AbrAndvilAg
Merge
Show Filtered

Applications to the study of Hungarian

- Personal work: representation in phonology
  - Double or single root node (Vago, 1992; Szigetvári, 2001)
  - Complex onsets? (Törkenczy and Siptár, 1999)
  - functional load of segments
  - “Tiers” in language: vowel, consonant, syllable (weight, stress), sibilant projections (Hayes, pc)
- Sonority Hierarchy (Kornai, 1990)
- Vowel length in present Hungarian (Nádasdy and Siptár, 1998)
- Vowel harmony, vacillation in vowel harmony

Acknowledgments

Thanks to Péter Szigetvári, Ádám Nádasdy, Péter Rebrus, Stuart Davis, Ken de Jong, and Damir Cavar.

The dictionary available for download for research purposes at

http://mypage.iu.edu/~stgrimes/dict/