Measuring Mutual Intelligibility:
Phonetic Distance & Conditional Entropy

Kim Heiligenstein

Seminar in Methodology and Statistics

May 2013

Mutual Intelligibility

- Among speakers of languages with same roots
- Elasticity: Difficulty in establishing distances
- Romance languages: Spanish, Portuguese, French, Italian.
- Subjective tests: intelligibility/proficiency tests
  - Hearing tests
- Objective tests: phonetic distances
  - Orthographic distances

Objective Test: Phonetic Distance

- Levenshtein Distance Algorithm
  - Calculates the least expensive cost of transforming one string into another through deletion, insertion or substitution.
  - Symmetric
  - Can be normalized
- Conditional Entropy
  - Measures the difficulty of predicting the outcome of an unknown random variable given a known one.
  - Asymmetric

Data

- Database of word lists from 4 Romance languages
- Cognates: for all 4 languages, words that have same root derivation.

English	French	Italian	Spanish	Portuguese
adjective	adjectif	aggettivo	adjetivo	adjetivo

- Phonetic transcriptions in IPA and X-SAMPA

Transcription	French	Italian	Spanish	Portuguese
IPA	adʒɛktif	adʒetivo	aðxetiβo	adʒetʃivu
X-SAMPA	adZEktif	adZetivo	aDxetiBo	adZetSivu

Levenshtein Distance

- What it can do for us: Compute how different a word is to another based on the pronunciation.
- The experiment:
  - Hypotheses:
    - There is a significant distance from one language to another.
    - Distances are significantly different from pair to pair.
  - Variables:
    - 1 6-leveled independent variable – language pair
    - 1 dependent variable: Normalized LD

Levenshtein Distance

Example: tʃimitɛro | simtjɛʀ

Italian	t	ʃ	i	m	i	t		ɛ	r	o
French		s	i	m		t	j	ɛ	ʀ	
Operation	del	sub			del		ins		sub	del
Cost	1	1			1		1		1	1

Cost of operations = 6

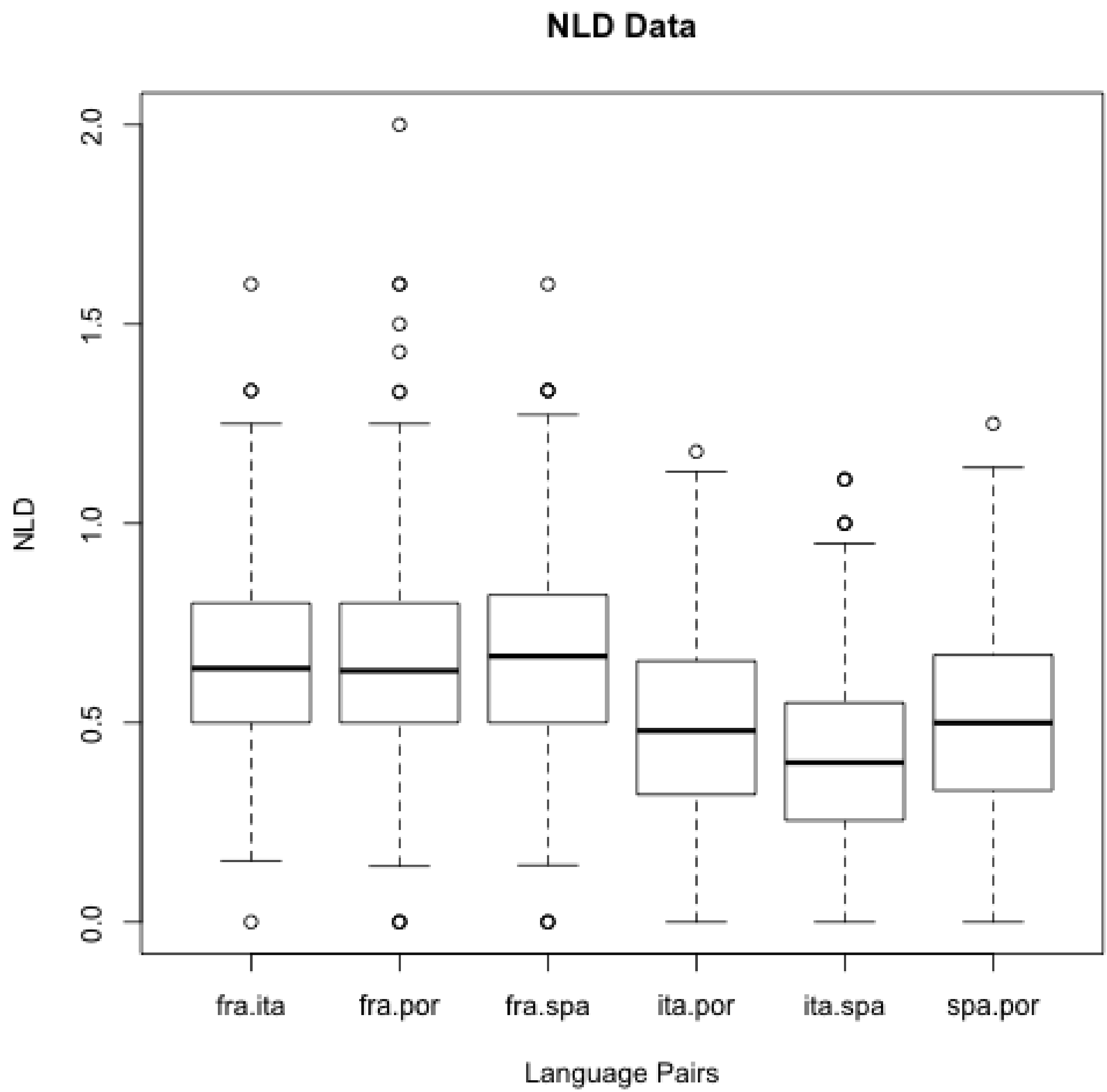
Normalized LD = $\frac{\text{Non-normalized LD}}{\text{Average Length of Both Strings}}$

= 0.75

Results

Descriptives

	N	Mean	SD	SE	Min	Max
fra.ita	399	0.65	0.24	0.01	0	1.60
fra.spa	399	0.66	0.25	0.01	0	1.25
fra.por	399	0.66	0.27	0.01	0	2
ita.spa	399	0.42	0.22	0.01	0	1.11
ita.por	399	0.49	0.23	0.01	0	1.18
spa.por	399	0.51	0.23	0.01	0	1.25


NLD Data
NLD
2.0
1.5
1.0
0.5
0.0
fra.ita
fra.por
fra.spa
ita.por
ita.spa
spa.por
Language Pairs

Results

Levene's Test for Equality of Variances

```
> with(nld, leveneTest(NLD,lang.pair))
Levene's Test for Homogeneity of Variance (center = median)
        Df F value Pr(>F)
group    5  1.5629 0.1672
      2388
```

The Levene's test is not significant ($F(5) = 1.56$, $p = 0.17$). Assumption of homogeneity of variance is met.

Results

ANOVA

```
Analysis of Variance Table

Response: NLD
            Df  Sum Sq Mean Sq F value    Pr(>F)
lang.pair    5  22.474  4.4949  77.642 < 2.2e-16 ***
Residuals 2388 138.246  0.0579
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

There is a significant effect of the language pair on the Levenshtein distance ($F(5) = 77.64$, $p < 0.05$).

Results

POST HOC

```
> pairwise.t.test(nld2way$NLD,nld2way$lang1, p.adj="none")

	Pairwise comparisons using t tests with pooled SD

data:  nld2way$NLD and nld2way$lang1

    ita     por
por 1.4e-11 -
spa 9.8e-13 0.3

P value adjustment method: none
> pairwise.t.test(nld2way$NLD,nld2way$lang2, p.adj="none")

	Pairwise comparisons using t tests with pooled SD

data:  nld2way$NLD and nld2way$lang2

    fra     ita
ita < 2e-16 -
spa < 2e-16 0.00022

P value adjustment method: none
```

Levenshtein Distance: Conclusions

- There distances from one language to another are significant.
- The distances are significantly different from one language pair to another.
  - Especially for the Italian-Spanish pair.

Do shorter distances correspond to low entropies ?

Conversely, do longer distances predict high entropies ?

Conditional Entropy

- What it can do for us: Quantify the uncertainty of being able to interpret a word in a foreign language.
- $$H(X|Y) = -\sum_{x \in X, y \in Y} p(x,y) \log_2 p(x|y)$$
- Ability to map phoneme in foreign language (heard conditioning variable) to phoneme in native language (conditioned variable to be identified)

Conditional Entropy

- The experiment:
  - Hypotheses:
    - The conditional entropy of one language given another is significant.
    - The conditional entropies differ significantly from one language to another, and in one direction from another.
  - Variables:
    - Independent variable: foreign (heard language) and native (language to map to)
    - dependent variable: CE

- Example:

Spanish	French
θero θjelo	zeʀo sjɛl

S	Θ	e	ɾ	o	
F	z	e	ʀ	o	
Entropy	$(1/9)\log_2(1/2)$ -0.11	$(1/9)\log_2(1/2)$ -0.11	$(1/9)\log_2(1/1)$ 0	$(1/9)\log_2(1/2)$ -0.11	
S	Θ	j	e	l	o
F	s	j	ɛ	l	-
Entropy	$(1/9)\log_2(1/2)$ -0.11	$(1/9)\log_2(1/1)$ 0	$(1/9)\log_2(1/2)$ -0.11	$(1/9)\log_2(1/1)$ 0	$(1/9)\log_2(1/2)$ -0.11

- $H(F|S) = -(-.11-.11-0-.11-.11-0-.11-0-.11)$

  $= .66$

- Example:

Spanish	French
θero θjelo	zeʀo sjɛl

F	z	e	ʀ	o	
S	Θ	e	ɾ	o	
Entropy	$(f;s)\log_2(s\|f)$	$(f;s)\log_2(s\|f)$	$(f;s)\log_2(s\|f)$	$(f;s)\log_2(s\|f)$	
F	s	j	ɛ	l	-
S	Θ	j	e	l	o
Entropy	$(f;s)\log_2(s\|f)$	$(f;s)\log_2(s\|f)$	$(f;s)\log_2(s\|f)$	$(f;s)\log_2(s\|f)$	$(f;s)\log_2(s\|f)$

- $H(S|F) = -(0+0+0+0+0+0+0+0+0)$

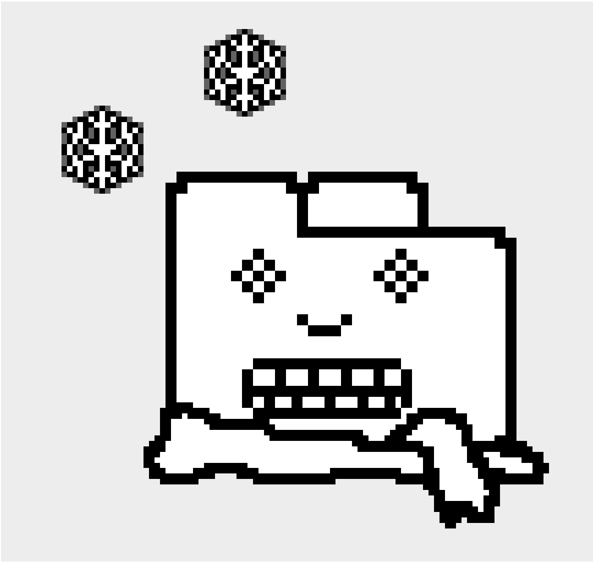
  $= 0$

Certainty in correct mapping is 100%

Based on this example…

- Spanish to French conditional entropy is higher than French to Spanish conditional entropy.
- $H(F|S) > H(S|F)$
- Easier for native speakers of Spanish to understand French than vice versa.

Results



Up Next…

- Finalize the CE data
- Analyze the CE data
- Compare LD to CE for correlation
- Adapt LD algorithm to set different weights depending on pairs
- Compare to subjective data and results