Measuring Linguistic Contamination

John Nerbonne
Rijksuniversiteit Groningen

In collaboration with Wybo Wiersema, Groningen

Measuring Linguistic Contamination

Thesis:

We can measure linguistic contamination

- ... in Syntax
- ... assuming some tools from computational linguistics

This talk

- Language contact as current research problem
- Why measure contamination?
- Computational linguistic fundamentals
  - Excursus measurement theorie
- Statistics: Permutation tests
- Data: English of Finnish emigrants
- Resultats
- Further steps

Language contact is a current research problem

- Mobility is large, growing
- Multilinguality is the norm
- Languages in contact influence one another
  - first languages influence second languages
  - and *vice versa*
- What are the factors, how important are they?
  - Experience, attitude, instruction, relation between source and target language

Current Research Problem

"No easy way of measuring or characterizing the total impact of one language on another in the speech of bilinguals has been, or probably can be devised. The only possible procedure is to describe the various forms of interference and to tabulate their frequency."
U. Weinreich, *Languages in Contact*, 1968, p.63

- Poplack & Sankoff, *Linguistics* 1984, Borrowing
- Poplack, Sankoff & Miller, *Linguistics* 1988 Lexical Borrowing

Language contact: Dialectometric Approach?

Measure distance using dialectometric techniques?

- ... another situtation vis-à-vis data: no atlases, no large body of analysed, comparable material

Idea

- Goal: **detect** lots of syntactic differences
- Material: Corpora of language use in contact situations
- Mark syntactic categories of words with (POS TAGS)
- Collect and analyse trigrams of tags

Computational Linguistics Basis

- Tagger: *Trigrams 'n Tags* (TnT) (Th. Brants, Saarland)
- TnT 96.7% correct for Penn Treebank
- Spoken material (500K words) of the *International Corpus of English* (Great Britain) used als training material
- 270 Tags, 195 used
- 87% correct in spoken material (checked by hand)
- 74% correct bigrams, 65% correct trigrams

Tagging

Oh	that	’s	a	just	a	fun
INT	PRON	COP	ART	EXCL	ART	N-COM
	in	a	’	Helsinki		
	PREP	ART	PAUSE	N-PROP		

Tag-Trigrams: INT-PRON-COP, PRON-COP-ART,... auch ##-#-INT, #-INT-PRON, PAUSE-NPROP-#, NPROP-#-##

Cf. Jan Aarts & Sylviane Granger “Tag Sequences in Learner Corpora” in: S.Granger (ed.) *Learner English on Computer*. London: Longman. 1998.

Indirect Measurement

- We aim to observe differences in syntactic use
  - including overuse and underuse, not just “errors”
- Lexical categories mirror syntactic analysis (projection principle, headedness)
- We *assume* that syntactic differences correlate strongly with the distribution of tag-trigrams
- ... even if POS information does not determine syntax 100%!

Excursus: Syntactic Footprints

- We *assume* that syntactic differences correlate strongly with the distribution of tag-trigrams
- Technique is therefore *indirect*—but complete analysis of this material in large quantities is unthinkable
- The history of the development of measurement techniques is encouraging!

  —Temperature, east/west longitude (first) measured indirectly.
- Likewise linguistic measures, e.g. MLU as indication of children's linguistic maturity

Statistical Significance

- Aarts & Granger examined tag-trigrams, but did not subject their collections to statistical analysis
- We wish to compare histograms of about 10^4 elements (of $\approx 10^6 = 100^3$ possible combinations)
- Only 13,784 trigrams actually occur
- Solution: permutation test

Permutation Tests: Basic Ideas

1. Determine difference between samples, e.g. with cosine

$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \cdot |\vec{y}|}$$

2. Check whether the differences are due to chance

3. "Shuffle" the data, draw two similarly sized samples, and measure the difference

4. Repeat step (3) e.g. 10,000 times, and then check whether the original sample is among the most extreme

5. Estimation of stat. significance, i.e., the probability that the original samples were due to chance (p-value).

Permutation Tests

Question: Do $\{A, B, C, \ldots\}$ and $\{D, E, F, \ldots\}$ differ wrt. X at the level $p < p'$?

1. Measure $\Delta^X(\{A, B, C, \ldots\}, \{D, E, F, \ldots\})$. Call this Δ_0^X.

2. Measure Δ^X for all permutations of $A, B, C, \ldots, D, E, F, \ldots$, i.e.

$$\begin{array}{rcl}
\Delta^X(\{B, C, F, \ldots\}, \{A, D, E, \ldots\}) & = & \Delta_1 \\
\Delta^X(\{A, B, F, \ldots\}, \{C, D, E, \ldots\}) & = & \Delta_2 \\
\vdots & \vdots & \vdots \\
\Delta^X(\{A, C, E, \ldots\}, \{B, D, F, \ldots\}) & = & \Delta_n
\end{array}$$

3. Is Δ_0 among the p' most extreme values?

Normalization

- Permutation tests are insensitive to sample size (in contrast to χ^2), they are based exclusively on relative differences

  No special measures need to be taken to isolate effect size

- Nonetheless we need to normalize to avoid detecting only irrelevant differences

  – we need to permute sentences, not trigrams to avoid measuring only the effect of syntactic coherence
  – but average sentence length differs: 24 wd/sentence vs. 16 wd/sentence
  – so we need to normalize for sentence length

Normalization, More Exactly

We obtain from the tagger a histogram, i.e., a series of counts of all the trigrams of e.g. the young group vs. the older group (more on groups later). We need to keep track of the sums per trigram:

$$\begin{array}{lcll} \mathbf{c^y} & = & < c_1^y, c_2^y, ..., c_n^y > & N^y = \sum_{i=1}^{n} c_i^y \\ \mathbf{c^o} & = & < c_1^o, c_2^o, ..., c_n^o > & N^o = \sum_{i=1}^{n} c_i^o \\ & & & N(= N^y + N^o) \end{array}$$

The relative histograms are the most important and protect us from overemphasizing sheer quantity:

$$\begin{array}{lcll} \mathbf{f^y} & = & < ..., f_i^y (= c_i^y / N^y), ... > & \sum_{i=1}^{n} f_i^y = 1 \\ \mathbf{f^o} & = & < ..., f_i^o (= c_i^o / N^o), ... > & \sum_{i=1}^{n} f_i^o = 1 \end{array}$$

More exactly, p.2

We weight these relative histograms on the basis of the distributions in the aggregated categories:

$$\begin{aligned} \mathbf{p^y} &= < ..., p_i^y (= f_i^y / (f_i^y + f_i^o)), ... > \\ \mathbf{p^o} &= < ..., p_i^o (= f_i^o / (f_i^y + f_i^o)), ... > \end{aligned}$$

and finally via category:

$$\begin{aligned} \mathbf{w^y} &= < ..., p_i^y \cdot c_i, ... > \\ \mathbf{w^o} &= < ..., p_i^o \cdot c_i, ... > \end{aligned}$$

where $c_i = c_i^y + c_i^o$.

Conceptually, these are the values that are compared.

Example

	Group y		Group o		Group y'		Group o'	
	T1	T2	T1	T2	T1	T2	T1	T2
counts $\mathbf{c}$	**15**	**10**	90	10	10	10	17	0
rel. freq. $\mathbf{f}$	0.6	0.4	0.9	0.1	0.5	0.5	1	0
norm. prop. $\mathbf{p}$	0.4	0.8	0.6	0.2	0.33	1	0.67	0
trigram $\mathbf{c}_i$	105	20	105	20	27	10	27	10
redistrib. $\mathbf{C}$	42	16	63	4	9	10	18	0

The normalizations aim at a weighted representation reflecting both the relations within the sample (the different T1, T2) and also the relations within the tag sort (T1 in groups y and o),

A More Readable Representation

In view of the last step:

$$\begin{aligned} \mathbf{w^y} &= < ..., p_i^y \cdot c_i, ... > \\ \mathbf{w^o} &= < ..., p_i^o \cdot c_i, ... > \end{aligned}$$

$\propto$ multiplication by N/n ($\bar{c_i} = N/n$), so that we can read the relative proportions directly. We scale these numbers back by dividing by N. Then we correct by a factor of $2n$, so that we obtain an average value of 1 (pro Trigram):

$$\begin{aligned} \mathcal{C}^{\mathbf{y}} &= < ..., w_i^y \cdot 2n/N, ... > \\ \mathcal{C}^{\mathbf{o}} &= < ..., w_i^o \cdot 2n/N, ... > \end{aligned}$$

Categories have the average value of 1.

Data Collection

- Finns who emigrated to Australia
- Workers and farmers, 25-40 years old (with their kids)
- Corpus collected in 1995-98 by Greg Watson (Joensuu) (*ICAME* 20, 1996, 41–70)
- Kids (< 17) — 30 interviews and
- Adults (≥ 17) — 60 interviews
- 350 K words in total
- Thanks to Lisa Lena Opas-Hänninen, Pekka Hirvonen, en Timo Lauttamus (University of Oulu) for material

Results

- Relative difference between young and old emigrants significant ($p < 0.001$)
- Striking patterns (not always categorically wrong)

’	it	’s	very	low	tax	in	here
PAUSE	PRON	COP	INTNS	ADJ	N-COM	PREP	ADV
a	boat	and	I	was	professional	fisherman	
ART	N-COM	CONJ	PRO	COP	ADJ	N-COM	

Most Important Differences

1	roadworks	and	uh
	hill	and	ah
	N	CONJUNC	INTERJEC
2	I	reckon	it
	that	take	lot
	PRON	V	PRON
3	enjoy	to	taking
	my	machine	break
	INTERJEC	PRON	V
4	but	that	’s
	that	I	clean
	CONJUNC	PRON	V
5	I	’m	uh
	it	’s	uh
	PRON	V	INTERJEC

Most Important Differences, p.2

6	now changing CONJUNC	what but INTERJEC	what some PRON
7	said all PRON	it everybody PRON	’s has V
8	bought lead V	that glass PRON	car windows N
9	that I PRON	was was V	different fit ADJ
10	Oh uh INTERJEC	lake money N	lake production N

Problems

- Distribution of syntactic constructions may by confounded by sentence length (needs to be checked *before* applying the analysis)

- In Finnish data sentence length was 50% longer among those who emigrated young.

Trigrams that only appear in long sentences?

Had you been here, my brother would not ...
We require that you be on time
He speaks bravely, as if here were unafraid!

Other Problems

- Pauses (hesitation noises) und false starts dominate the most significant trigrams
  —attempt to filter
- Identification of sources of contamination
  —attempt to predict based on expectations on the basis of the native language
- Unclear, how much data is needed

Further Work

- Application to Old English Corpus, to examine whether syntax in Latin translations is similar to syntax in non-translated documents (Nijmegen data courtesy of Ans van Kemenade)
  Status: MA thesis (Livi Ruffle)

- Application to student essays (foreign language learners, ICE Corpora)
  Status: a bit of progress (Wybo Wiersma)

Measuring Linguistic Contamination

Thesis:

We can measure linguistic contamination

- ... in Syntax
- ... assuming some tools from computational linguistics

See `www.LogiLogi.org/FiAuImEnRe/` (Wybo's site, software)

Permutation Tests (repeated)

Question: Do $\{A, B, C, \ldots\}$ and $\{D, E, F, \ldots\}$ differ wrt. X at the level $p < p'$?

1. Measure $\Delta^X(\{A, B, C, \ldots\}, \{D, E, F, \ldots\})$. Call this Δ_0^X.

2. Measure Δ^X for all permutations of $A, B, C, \ldots, D, E, F, \ldots$, i.e.

$$\begin{array}{rcl}
\Delta^X(\{B, C, F, \ldots\}, \{A, D, E, \ldots\}) & = & \Delta_1 \\
\Delta^X(\{A, B, F, \ldots\}, \{C, D, E, \ldots\}) & = & \Delta_2 \\
\vdots & \vdots & \vdots \\
\Delta^X(\{A, C, E, \ldots\}, \{B, D, F, \ldots\}) & = & \Delta_n
\end{array}$$

3. Is Δ_0 among the p' most extreme values?

Application 2 (repeated)

When is phonetic correspondence convincing evidence of historical relatedness?

Brett Kessler, *The Structure of Word Lists*, Stanford: CSLI, 2001.

1. Create large contingency table of initial phonemes in semantically similar

2. Ask whether the distributions are independent (χ^2 question)
   —but lots of zeros!

3. Apply measure of similarity (R^2, R, or χ^2)

4. Permute many times and keep track of how often original measure is exceeded