Pausing as Evidence of L2 Proficiency

Finnish Australians Speaking English

Timo Lauttamus, John Nerbonne and Wybo Wiersma

University of Oulu, Rijksuniversiteit Groningen
timo.lauttamus@oulu.fi, j.nerbonne@rug.nl & wybo@logilogi.org

7 August 2008
In this talk we present
A method for finding significant syntactic differences and findings on pausing and proficiency

3 sub-questions about the method
1. What did your corpus look like?
2. What is permutation statistics?
3. How to apply it to syntax?

3 sub-questions about the analysis
1. What general differences did you find?
2. What disfluent speech did you find?
3. What does this tell about pausing?
The Method

- Detect a wide range of syntax differences, and these as
  - significant differences
  - aggregate differences
  - relative differences
- This would enable measuring the syntax part of total impact:

  “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.”
The Method

- Detect a wide range of syntax differences, and these as
  - significant differences
  - aggregate differences
  - relative differences

- This would enable measuring the syntax part of total impact:

  "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."

  - Weinreich, Languages in Contact
The Method

We also want to do this **automatically and computationally**
In order to be able to:

- Mine for differences in syntax between
  - learners versus native speakers
  - speakers of different dialects
  - writers from different discourses
- Test dialectological and other linguistic hypotheses
- Note **over- and under- use** instead of right / wrong
The Method

We did it in four steps:

1. Tag 2 or more collections of comparable material (using an automatic POS-tagger)
2. Take n-grams (2 - 5 grams) of POS-tags
3. Statistically compare their frequencies
4. Sort the significant POS-n-grams by extent of difference

Aarts J. and Granger S. did this without the statistics in: 'Tag sequences in learner corpora: a key to interlanguage grammar and discourse' (1998)
Our Corpus

Origins:

- 20,000 Finns immigrated to Australia
- Working class background, limited education
- 25-40 Years upon arrival

Corpus collected 1995-1998 by Greg Watson:

- of the university of Joensuu, Finland
- two age groups; adults and juveniles
- 350,000 words, 305,000 words free conversation
Our Corpus

Adults:

- over 18 years at arrival, on average 30
- on average 58 at time of interview
- 60 interviews, 65-70 min each (221,000 words)

Juveniles:

- under 16 years at arrival, on average 6
- on average 36 at time of interview
- 30 interviews, 65-70 min each (84,000 words)
Our Corpus

In preparation we Part of Speech-tagged it with:

- **Trigrams ’n ’Tags (TnT) Statistical Part of Speech Tagger**
- made by Thorsten Brants (Universitit des Saarlandes)

It achieves an accuracy of:

- 96.7% on the Penn Treebank
- 85.1% - 90.5% on our spoken material

Accuracy is of course worse for 3-grams:

- 2-grams 74%, 3-grams 65%, 4-grams 58% ...
Permutation Statistics

It is different from parametric (normal) statistics:

- It is about the data, not about the population
  - no need for normality
  - no need for homoscedasticity (eq distrib variances)
  - no absolute need for random sampling
- Still, important for permutation statistics are
  - random assignment and independence of observations
  - in practice no problems for linguistic/dialect data

As a statistical method it is very suitable for linguistics
Pausing as Evidence:
Timo Lauttamus, John Nerbonne and Wybo Wiersma

Introduction
In this Talk

The Method
Our Corpus
Permutation Statistics
Applying it to Syntax

Analysis
General Differences
Disfluent Speech
Pausing

Conclusion
Questions
References

Permutation Statistics
Permutation Statistics

Evidence:

Timo Lauttamus, John Nerbonne and Wybo Wiersma

Introduction

In this Talk

The Method

Our Corpus

Permutation Statistics

Applying it to Syntax

Analysis

General Differences

Disfluent Speech

Pausing

Conclusion

Questions

References
Pausing as Evidence:
Timo Lauttamus, John Nerbonne and Wybo Wiersma

Introduction
In this Talk
The Method
Our Corpus
Permutation Statistics
Applying it to Syntax
Analysis
General Differences
Disfluent Speech
Pausing
Conclusion
Questions
References
Pausing as Evidence:
Timo Lauttamus, John Nerbonne and Wybo Wiersma

Introduction
In this Talk

The Method
Our Corpus

Permutation Statistics
Applying it to Syntax

Analysis
General Differences
Disfluent Speech
Pausing

Conclusion
Questions
References
Permutation Statistics

Evidence:
Timo Lauttamus, John Nerbonne and Wybo Wiersma

Introduction
In this Talk

The Method
Our Corpus

Permutation Statistics
Applying it to Syntax

Analysis
General Differences
Disfluent Speech
Pausing

Conclusion
Questions
References
Permutation Statistics

Evidence:
Timo Lauttamus, John Nerbonne and Wybo Wiersma

Introduction
In this Talk

The Method
Our Corpus

Permutation Statistics
Applying it to Syntax

Analysis
General Differences
Disfluent Speech
Pausing

Conclusion
Questions
References
Permutation Statistics

Introduction
In this Talk
The Method
Our Corpus
Permutation Statistics
Applying it to Syntax
Analysis
General Differences
Disfluent Speech
Pausing
Conclusion
Questions
References
Permutation Statistics

Evidence:
Timo Lauttamus, John Nerbonne and Wybo Wiersma

Introduction
In this Talk

The Method
Our Corpus

Permutation Statistics
Applying it to Syntax

Analysis
General Differences
Disfluent Speech
Pausing

Conclusion

Questions

References
Permutation Statistics

Pausing as Evidence:
Timo Lauttamus, John Nerbonne and Wybo Wiersma

Introduction
In this Talk
The Method
Our Corpus
Permutation Statistics
Applying it to Syntax
Analysis
General Differences
Disfluent Speech
Pausing
Conclusion
Questions
References
Permutation Statistics
Pausing as Evidence:
Timo Lauttamus, John Nerbonne and Wybo Wiersma

Introduction
In this Talk
The Method
Our Corpus
Permutation Statistics
Applying it to Syntax
Analysis
General Differences
Disfluent Speech
Pausing
Conclusion
Questions
References

Permutation Statistics

\[
\begin{align*}
group 1 & : 1g, 3b \\
group 2 & : 3g, 1b \\
& + \rightarrow \quad N, 10,000 \text{ times...} \\
& \quad C \\
& C / N = \text{the p value}
\end{align*}
\]
Applying it to Syntax

One firstly needs something to permutate:

- We permutated interviewees
  - more conservative than 3-grams
  - and also easier than sentences (did this earlier)
- For each interview
  - we took 3-grams (N-grams too) of POS-tags
  - we then calculated the 3-gram-promillages for all 3-grams
    (occurrence of 3-gram type per 1000 3-gram tokens)
- These 3-gram-promillage-vectors were then used
  - summed per group after each permutation
Applying it to Syntax

Secondly one needs something to measure extremity:

- Both
  - for the whole group
  - and for each individual POS-3-gram
- We used \( r\)-square and summed \( r\)-square
  - we also tried cosine and summed \( r\)
- \( R\)-square is the square of the difference \( (r)\)
  - for a POS-3-gram-promillage between the 2 groups
- Summed \( r\)-square is the sum of \( r\)-square for all 3-grams
Pausing as Evidence:
Timo Lauttamus,
John Nerbonne and
Wybo Wiersma

Introduction
In this Talk
The Method
Our Corpus
Permutation Statistics
Applying it to Syntax
Analysis
General Differences
Disfluent Speech
Pausing
Conclusion
Questions
References
Applying it to Syntax

Thirdly one can apply normalizations:

- We normalized for frequency
  - divide by half the sum of the POS-3-gram in all vectors
- This increases the weight of less frequent POS-3-grams
  - on the level of the group (summed r-square)
- And it allows one to sort 3-grams based on
  - whether they are more or less typical for each group
  - regardless of frequency
Between Permutations Normalisation

Raw

Normalized
Analysis of the Results

- The following slides summarise some of the material in Lauttamus, Nerbonne, and Wiersma (2007).
- The evidence based on the data of the two groups shows that there are statistically significant syntactic differences between the adult and juveniles groups.
- We argue that some of the significant differences found in the data can be ascribed to the lower level of language proficiency of the adults.
General Differences

Some of the **syntactic differences** found in the data can be described in most general terms as follows (all for the adults group):

1. **Overuse of** hesitation phenomena
2. **Overuse of** parataxis (particularly with and and but) as opposed to hypotaxis
3. **Underuse of** contracted forms
4. **Reduced repertoire of** discourse markers such as you know, I mean
5. **Avoidance of** complex verbal structures
6. **Avoidance of** prepositional and phrasal verbs
7. **Underuse of the** existential there
General Differences

• We therefore argue that the statistical evidence obtained from our data reflects **syntactic distance** between the two varieties of L2 English

• And, consequently, **aggregate effects** of the differences in the two groups English proficiency
Disfluent Speech

- The adults demonstrate typical features of disfluent speech, such as (filled) pauses, repeats, false starts, incomplete or false syntactic structures, arising from difficulties in speech processing, and particularly in lexical access.

- We would like to argue that these disfluencies are automatic reactions to temporal problems in speech planning (cf. Oomen & Postma 2001).
Disfluent Speech

1. skin cancer and /um and uh/ and gene general
2. but /ah I I/never been on
3. clubs spades /hearts and uh/ uh cl oh
4. he was a leading-hand um /leading-hand and ah/ last last
5. as in /a in a/ Finland because especially

As to pausing, note that only filled pauses (vocalised pauses), as in (1), (2), (3), (4), were included and tagged as INTERJEC(tion) in each trigram (here between slashes)
Disfluent Speech

• Features of disfluent speech can occur at any syntactic boundary, at sentence, clause, phrase or word boundary (cf. Paananen-Porkka 2007), and this is what what we would like to argue as well
• They are, of course, characteristic of any kind speech, native and non-native alike, but certainly more frequent in interlanguage or, more generally, in second language acquisition where speakers demonstrate imperfect learning as they study an L2
We applied the computational technique described earlier to examine

1. if the adults and juveniles show a differential use of pausing (filled pauses), and

2. whether either of the groups behave differently from native speakers of English

Our earlier research (Lauttamus et al. 2007) suggests

- that pausing (filled pauses) may be a statistically significant determinant that distinguishes less proficient learners having acquired an L2 later in life (the adults) from more proficient learners having acquired their L2 at early age (the juveniles).
Pausing

• In line with the evidence discussed in Oomen & Postma (2001), we argue that
  • filled pauses are associated with lexical search phenomena, and that
  • both filled pauses and repetitions may signal problems in constituent construction

• Difficulties in controlling pause duration and placement seem to be common among all second-language learners irrespective of the target language

• Paananen-Porkka (2007), for one, argues that pausing, including filled pauses, seems to be the main source for the anomalies that she found in her study of English speech rhythm by Finnish comprehensive school students
Pausing

• Our goal is now to show in detail if this is the case with the English of Finnish Australians, particularly with the adults

• Our hypothesis is that the adults also pause at inappropriate places compared to the juveniles and native speakers of English

• The syntactic positions of filled pauses will be discussed in detail in our future work
Pausing

- Reminder: We are now dealing with the top 200 trigram types out of a total of 666, which all show significant p-values (at 0.05) for the adults.

- Out of the top 200 POS-trigram types produced by the adults, 37% include at least one filled pause, as in (1) and (2); 6% include at least two filled pauses, as in (3) and (4).

- For the juveniles, there are only 6 (six) trigram types (with significant p-values), of which
  - none contain filled pauses
(1) 
Interj Conj(subord) Art (def) 
politically | uh when the | liberals

(2) 
V(cop,pres,encl) Interj Adv(inten) 
I’ | m ah very | sick

(3) 
Interj Interj Conj(subord) 
and | uh uh because | in the morning

(4) 
Interj Pron(pers, sing) Interj 
and | uh I uh | snow-skied
Conclusion

- The uneven distribution of the filled pauses across the two groups of Finnish Australian English speakers conclusively shows that
  - the juveniles have a much more varied syntactic repertoire (measured in terms of POS-trigrams) than the adults, and
  - that the adults have much more limited and idiosyncratic syntactic patterns at their disposal
Conclusion

• The large number of filled pauses found in the adults speech as opposed to the juveniles is in agreement with the evidence in Paananen-Porkka (2007: 234), who argues that native speakers of Finnish show longer pauses on average in English than in Finnish.

• The statistically significant differential use of filled pauses by the adults can be explained in terms of the adults lesser proficiency (particularly at the level of speech planning) and, consequently, fluency of L2 compared to that of the juveniles.
Concluding Remarks

There is room for fine-tuning the method:

- Find optimum size for data-sets
- Try and evaluate with different measures

The method as is can easily be applied to many data-sets:

- Works on untagged corpora of spoken language
- Can empirically buttress theses with usage data

Software to do it and to pre-process corpora is freely available:

- the ComLinToo http://old.logilogi.org/ComLinToo
Questions

Any questions?
Questions

Any questions?
References

Jan Aarts and Sylviane Granger.
Tag sequences in learner corpora: A key to interlanguage grammar and discourse.

Alan Agresti.
An Introduction to Categorical Data Analysis.

International Phonetics Association.

Thorsten Brants.
Tnt - a statistical part of speech tagger.
References

Eugenio Coseriu.  
*Probleme der kontrastiven Grammatik.*  
Schwann, Düsseldorf, 1970.

Charles Fillmore and Paul Kay.  
Grammatical constructions and linguistic generalizations: the what’s x doing y construction.  

Phillip Good.  
*Permutation Tests.*  

Wander Lowie Kees de Bot and Marjolijn Verspoor.  
*Second Language Acquisition: An Advanced Resource Book.*  
References

Brett Kessler. 
*The Significance of Word Lists.* 

Chris Manning and Hinrich Schütze. 
*Foundations of Statistical Natural Language Processing.* 

A measure of aggregate syntactic distance. 

C.C.E. Oomen and A. Postma. 
Effects of divided attention on the production of filled pauses and repetitions. 
References


References

Peter Sells.  
*Lectures on Contemporary Syntactic Theories.*  
CSLI, Stanford, 1982.

David Sankoff Shana Poplack and Christopher Miller.  
The social correlates and linguistic processes of lexical borrowing and assimilation.  

Sarah Thomason and Terrence Kaufmann.  
*Language Contact, Creolization and Genetic Linguistics.*  

Detecting syntactic contamination in emigrants: The english of finnish australians.  
References

Frans van Coetsem.
Loan phonology and the two transfer types in language contact.

Greg Watson.
The finnish-australian english corpus.
*ICAME Journal: Computers in English Linguistics*, 20, 1996.

Uriel Weinreich.
*Languages in Contact*.
Copyrights Wybo Wiersma, John Nerbonne and Timo Lauttamus, available under the Creative Commons By-Sa license

- Thanks to the OpenClipart archive; Carlitos for the landscape, and unknown authors for the bomb and the frogs.