The effects of N-gram probabalistic measures on the recognition and production of four-word-sequences.

Eberhard Karls Universität Tübingen
Course: Frequency effects of multi-word sequences
Lecturer: Hendrix-Sun, June , Ph.D.
Student: Johannes Krämer
SS 2017

How to read the title?

- Effects of [...] probabalistic measures [...]

How to read the title?

- Effects of [...] probabalistic measures [...]
  $\rightarrow$ it's about the methods:

How to read the title?

- Effects of [...] probabalistic measures [...]

  $\rightarrow$ it's about the methods:
  - how was it measured?
  - which probabalitic methods were used?
  - which are the best methods?

Which methods were used?

Methods = Predictors

- frequency of occurance
- log probability of occurance
- mutual information

in more detail...

- frequency of occurance

in more detail...

- frequency of occurance
- $\frac{freq-of-event}{all-possible-events}$

in more detail...

- probability of occurance

in more detail...

- probability of occurance
- $P(W_i|W_i-_2, W_i-_1)$

in more detail...

- probability of occurance
- $P(W_i|W_i{-}_2, W_i{-}_1)$
- what is more propable?
  - wit OR with
  - given that previous words are...

in more detail...

- probability of occurance
- $P(W_i|W_i-_2, W_i-_1)$
- what is more propable?
  - wit OR with
  - given that previous words are...
  - I like pizza

in more detail...

- probability of occurance
- $P(W_i|W_i-_2, W_i-_1)$
- what is more propable?
  - wit OR with
  - given that previous words are...
  - I like pizza

(this example is from Mr. Cöltekins SNLP Course)

in more detail...

- probability of occurance
- $P(W_i|W_i-_2, W_i-_1)$
- what is more propable?
  - wit OR with
  - given that previous words are...
  - I like pizza

  (this example is from Mr. Cöltekins SNLP Course)
- log is for scaling

  $\rightarrow$ fitting more on the same plot

in more detail...

- mutual information

in more detail...

- mutual information
- how much do we know about a fact given another fact
- correlation of variables

in more detail...

- mutual information
- how much do we know about a fact given another fact
- correlation of variables
- $I(X;Y)$

in more detail...

- mutual information
- how much do we know about a fact given another fact
- correlation of variables
- $I(X;Y)$
- how much do we now about the gender given the name of a person?

in more detail...

- mutual information
- how much do we know about a fact given another fact
- correlation of variables
- $I(X; Y)$
- how much do we now about the gender given the name of a person?
- how much do we know about FreqAB given FreqA

What is measured?

- onset latency

What is measured?

- onset latency
- production duration

Why do such a experiment in the first place?

What inspired them?

- no one did it before

Why do such a experiment in the first place?

What inspired them?

- no one did it before
- Baayen, Kuperman, Bertram (2010):
  - found interaction between measures of different linguistic units

$\rightarrow$ claim: multiple sources of linguistic information are processed in parallel

This study: Find out the following questions

- whether freq., log prob., MI of 1-4grams affect onset latency and production duration

This study: Find out the following questions

- whether freq., log prob., MI of 1-4grams affect onset latency and production duration
- Which is the better predictor for:
  - onset latency
  - production duration

This study: Find out the following questions

- whether freq., log prob., MI of 1-4grams affect onset latency and production duration
- Which is the better predictor for:
  - onset latency
  - production duration
- Are there (linear) interactions between predictors?

Conclusion spoiler

- MI not very predictive

Conclusion spoiler

- MI not very predictive

  $\rightarrow$ leaves us with:
  - frequency of occurance
  - log probability of occurance

Conclusion spoiler

- MI not very predictive

  $\rightarrow$ leaves us with:
  - frequency of occurance
  - log probability of occurance
- prob. measures up to 4-grams all interact with each other

Conclusion spoiler

- MI not very predictive

  $\rightarrow$ leaves us with:
  - frequency of occurance
  - log probability of occurance
- prob. measures up to 4-grams all interact with each other
  - but how strong?
  - for which N-gram does it work best?
  - what does it mean if they interact?

Setup of the Experiment

- participants:
  - 17 Students (10 female, 7 male)
  - University of Alberta
  - paid: yes

Setup of the Experiment

- participants:
  - 17 Students (10 female, 7 male)
  - University of Alberta
  - paid: yes
- material:
  - extracted 112 most freq. 4grams from BNC (100 mio words)

Setup of the Experiment

- participants:
  - 17 Students (10 female, 7 male)
  - University of Alberta
  - paid: yes
- material:
  - extracted 112 most freq. 4grams from BNC (100 mio words)
    - frequeny: 12 - 117 per million words (?)

Setup of the Experiment

- participants:
  - 17 Students (10 female, 7 male)
  - University of Alberta
  - paid: yes
- material:
  - extracted 112 most freq. 4grams from BNC (100 mio words)
    - frequeny: 12 - 117 per million words (?)
  - random selection of 320 4grams:
    - frequeny: 0.3 - 11 per million words

Setup of the Experiment

- participants:
  - 17 Students (10 female, 7 male)
  - University of Alberta
  - paid: yes
- material:
  - extracted 112 most freq. 4grams from BNC (100 mio words)
    - frequeny: 12 - 117 per million words (?)
  - random selection of 320 4grams:
    - frequeny: 0.3 - 11 per million words
  - total: 432 4grams

Where did the data come from?

- most freq. items from BNC

Where did the data come from?

- most freq. items from BNC
- frequencies of the items taken from COCA
  - COCA = The Corpus of Contemporary American and English
  - COCA size: 385 mio. words

Where did the data come from?

- most freq. items from BNC
- frequencies of the items taken from COCA
  - COCA = The Corpus of Contemporary American and English
  - COCA size: 385 mio. words
- don't, you've, wasn't → one word
- data was randomized for every participant

What was the technological setup for the experiment?

- 4gram preceded by a fixation cross

What was the technological setup for the experiment?

- 4gram preceded by a fixation cross
- cross for 500ms $\rightarrow$ black screen 20ms $\rightarrow$ 4gram 1500ms

What was the technological setup for the experiment?

- 4gram preceded by a fixation cross
- cross for 500ms $\rightarrow$ black screen 20ms $\rightarrow$ 4gram 1500ms
- interstimulus 1000ms

What was the technological setup for the experiment?

- 4gram preceded by a fixation cross
- cross for 500ms $\rightarrow$ black screen 20ms $\rightarrow$ 4gram 1500ms
- interstimulus 1000ms
- read out loud aqap ;)

What was the technological setup for the experiment?

- 4gram preceded by a fixation cross
- cross for 500ms $\rightarrow$ black screen 20ms $\rightarrow$ 4gram 1500ms
- interstimulus 1000ms
- read out loud aqap ;)
- 2 possible breaks (most of the participants didn't make use of it)

What was the technological setup for the experiment?

- 4gram preceded by a fixation cross
- cross for 500ms $\rightarrow$ black screen 20ms $\rightarrow$ 4gram 1500ms
- interstimulus 1000ms
- read out loud aqap ;)
- 2 possible breaks (most of the participants didn't make use of it)
- 2 microphones

What was the technological setup for the experiment?

- 4gram preceded by a fixation cross
- cross for 500ms $\rightarrow$ black screen 20ms $\rightarrow$ 4gram 1500ms
- interstimulus 1000ms
- read out loud aqap ;)
- 2 possible breaks (most of the participants didn't make use of it)
- 2 microphones
  - onset latency recording
  - speech recording

The problem of collinearity

- W is our outcome

The problem of collinearity

- W is our outcome
- X and Z are predictors

The problem of collinearity

- W is our outcome
- X and Z are predictors
- if X,Z correlated $\rightarrow$ can't tell wether W is due to X or to Z

The problem of collinearity

- W is our outcome
- X and Z are predictors
- if X,Z correlated $\rightarrow$ can't tell wether W is due to X or to Z
- draw on blackboard!

How to handle Collinearity?

- detect collinearity

How to handle Collinearity?

- detect collinearity
- reduce collinearity

How to handle Collinearity?

- detect collinearity
- reduce collinearity
    - centering

How to handle Collinearity?

- detect collinearity
- reduce collinearity
  - centering
  - residualization

Remove bad data points

- total: 7344 data points (17 x 432)

Remove bad data points

- total: 7344 data points (17 x 432)
- removed: 2251 = 30,7%

Remove bad data points

- total: 7344 data points (17 x 432)
- removed: 2251 = 30,7%
- WHY?
  - 5 items had p = 0 in COCA → 85 data points | 1,2%

Remove bad data points

- total: 7344 data points (17 x 432)
- removed: 2251 = 30,7%
- WHY?
  - 5 items had p = 0 in COCA → 85 data points | 1,2%
  - 1 Participant had to many errors → 35 data points | 0.5%

Remove bad data points

- total: 7344 data points (17 x 432)
- removed: 2251 = 30,7%
- WHY?
  - 5 items had p = 0 in COCA → 85 data points | 1,2%
  - 1 Participant had to many errors → 35 data points | 0.5%
  - not triggered sound → 869 data points | 11,8%

Remove bad data points

- total: 7344 data points (17 x 432)
- removed: 2251 = 30,7%
- WHY?
  - 5 items had p = 0 in COCA → 85 data points | 1,2%
  - 1 Participant had to many errors → 35 data points | 0.5%
  - not triggered sound → 869 data points | 11,8%
  - research assistants removed → 1258 data points | 17,1%

Other independent variables

- trials 1-432

Other independent variables

- trials 1-432
- Manner
    - stop
    - approximant
    - vowel
    - fricative
    - nasal

Other independent variables

- trials 1-432
- Manner
  - stop
  - approximant
  - vowel
  - fricative
  - nasal
- NumSyll

Other independent variables

- trials 1-432
- Manner
  - stop
  - approximant
  - vowel
  - fricative
  - nasal
- NumSyll
- PhraseABCD

Other independent variables

- trials 1-432
- Manner
    - stop
    - approximant
    - vowel
    - fricative
    - nasal
- NumSyll
- PhraseABCD
    - 117 phrase
    - 310 non-phrase

Interpretation of data

- We will look together at the plots and figger out whats going on there.

Why does FreqB have such an big impact?

- B-gram appeared at position of fixation cross

Why does FreqB have such an big impact?

- B-gram appeared at position of fixation cross
- WordTypeB as influencer?

- look at the results together!

What do the two tables mean?

- onset latency $\rightarrow$ log probability of occurance

What do the two tables mean?

- onset latency $\rightarrow$ log probability of occurance

  $\rightarrow$ 'recognition is mainly underpinned by a mechanism whereby a target N-gram competes with its family members'

What do the two tables mean?

- onset latency → log probability of occurance

  → 'recognition is mainly underpinned by a mechanism whereby a target N-gram competes with its family members'
- production duration → frequency of occurance (amount of experience)

What do the two tables mean?

- onset latency → log probability of occurance

  → 'recognition is mainly underpinned by a mechanism whereby a target N-gram competes with its family members'
- production duration → frequency of occurance (amount of experience)
- most important among: unigrams, bigrams, trigrams, quadgrams

What do the two tables mean?

- onset latency $\rightarrow$ log probability of occurance

  $\rightarrow$ 'recognition is mainly underpinned by a mechanism whereby a target N-gram competes with its family members'
- production duration $\rightarrow$ frequency of occurance (amount of experience)
- most important among: unigrams, bigrams, trigrams, quadgrams
  - onset latency?

What do the two tables mean?

- onset latency → log probability of occurance

  → 'recognition is mainly underpinned by a mechanism whereby a target N-gram competes with its family members'
- production duration → frequency of occurance (amount of experience)
- most important among: unigrams, bigrams, trigrams, quadgrams
  - onset latency?

    → trigrams

What do the two tables mean?

- onset latency → log probability of occurance

  → 'recognition is mainly underpinned by a mechanism whereby a target N-gram competes with its family members'
- production duration → frequency of occurance (amount of experience)
- most important among: unigrams, bigrams, trigrams, quadgrams
  - onset latency?

    → trigrams
  - production duration?

What do the two tables mean?

- onset latency → log probability of occurance

  → 'recognition is mainly underpinned by a mechanism whereby a target N-gram competes with its family members'
- production duration → frequency of occurance (amount of experience)
- most important among: unigrams, bigrams, trigrams, quadgrams
  - onset latency?
    → trigrams
  - production duration?
    → unigrams

What do the two tables mean?

- onset latency → log probability of occurance

  → 'recognition is mainly underpinned by a mechanism whereby a target N-gram competes with its family members'
- production duration → frequency of occurance (amount of experience)
- most important among: unigrams, bigrams, trigrams, quadgrams
  - onset latency?
    → trigrams
  - production duration?
    → unigrams
- BUT they still interact → what does it mean?

The very conclusion

'Finally, the finding that probabilistic measures tied to N-grams up to four-words long **interacted** with each other in the onset latency and production duration analyses suggests that they are **processed in parallel** in both recognition and production.'

Reference:

Antoine Tremblay and Benjamin V. Tucker. 'The effects of N-gram probabilistic measures on the recognition and production of four-word sequences.' IWK Health Center / University of Alberta.