FROM SCIENCE TO DATA SCIENCE

LUC BORUTA, THUNKEN INC.

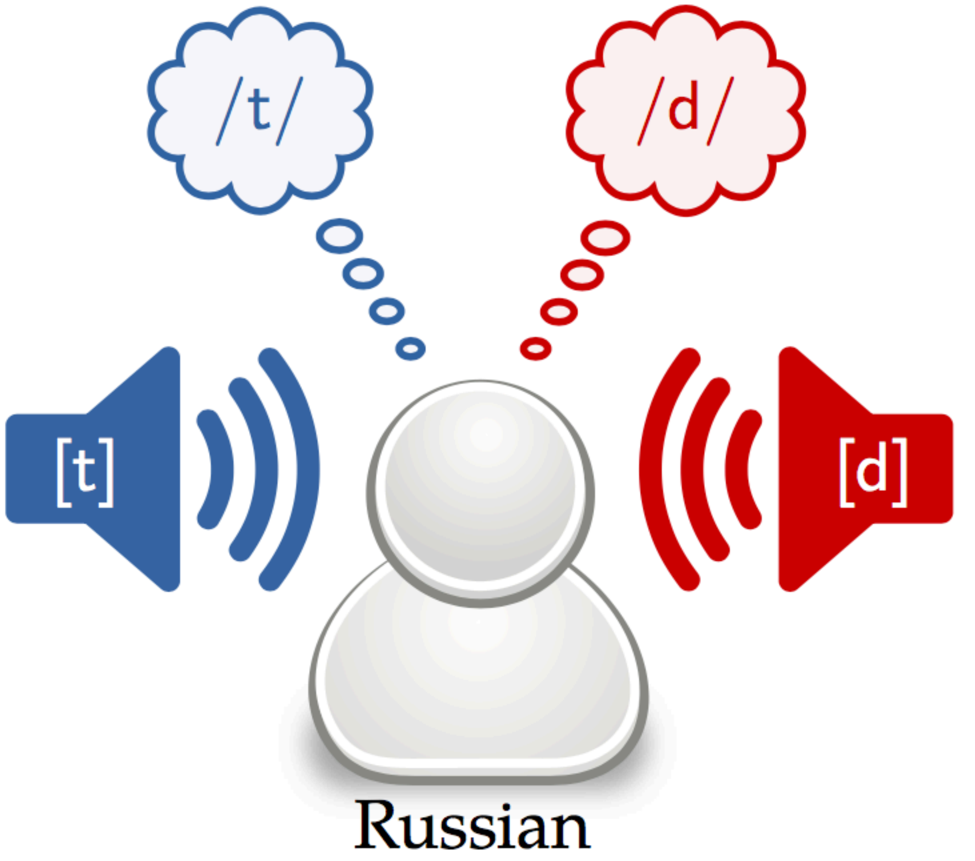
FROM SOFT SCIENCE TO HARD SCIENCE TO DATA SCIENCE

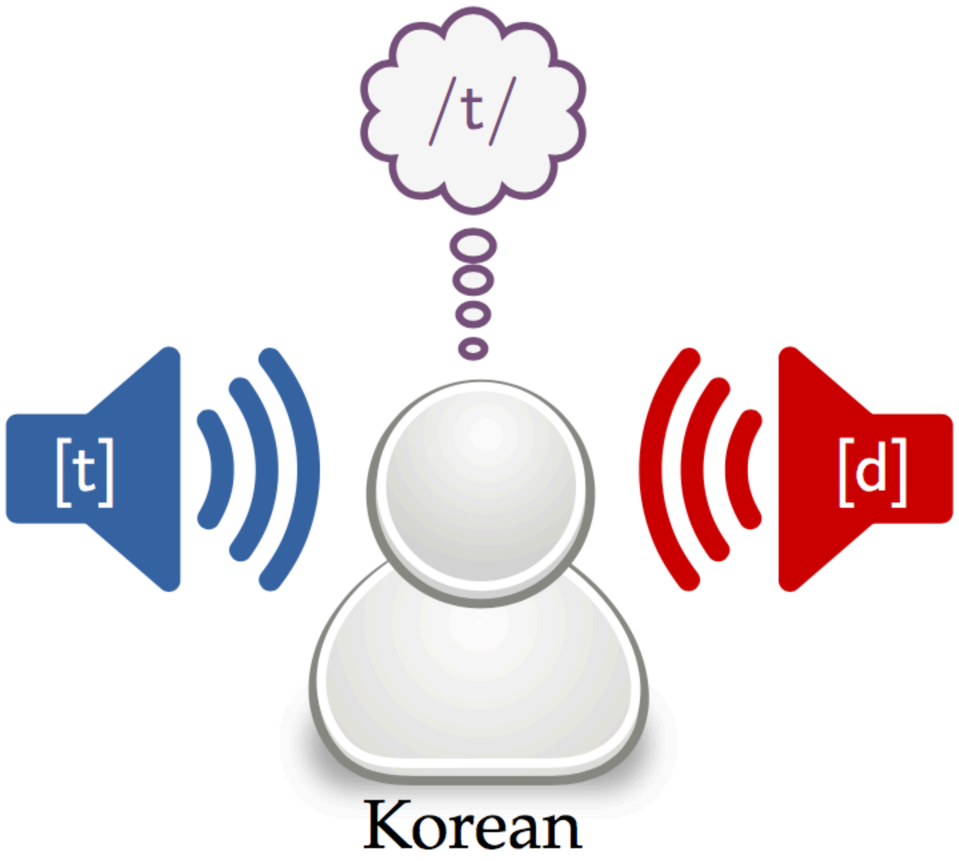
- DEUG, Theoretical Linguistics @ Toulouse II
- Master, Computational Linguistics @ Paris 7
- Ph.D., Computational Linguistics @ Paris 7
- Sr. NLU Developer @ Nuance, Montréal
- R&D Director @ MyScienceWork, Luxembourg
- CEO @ Thunken, Washington D.C. / Luxembourg

PH.D. @ PARIS 7 (2009–2012)

- Labs: Alpage (P7/INRIA) + LSCP (ENS/EHESS/CNRS)
- Grad school: Frontières du Vivant (P5/P7/FBS)
- Project:
  - Computational models of early phonological acquisition
  - “Psychocomputational linguistics”
  - Unsupervised machine learning (fun!)
- Contribution: 100% of negative results!

INDICATORS OF ALLOPHONY AND PHONEMEHOOD

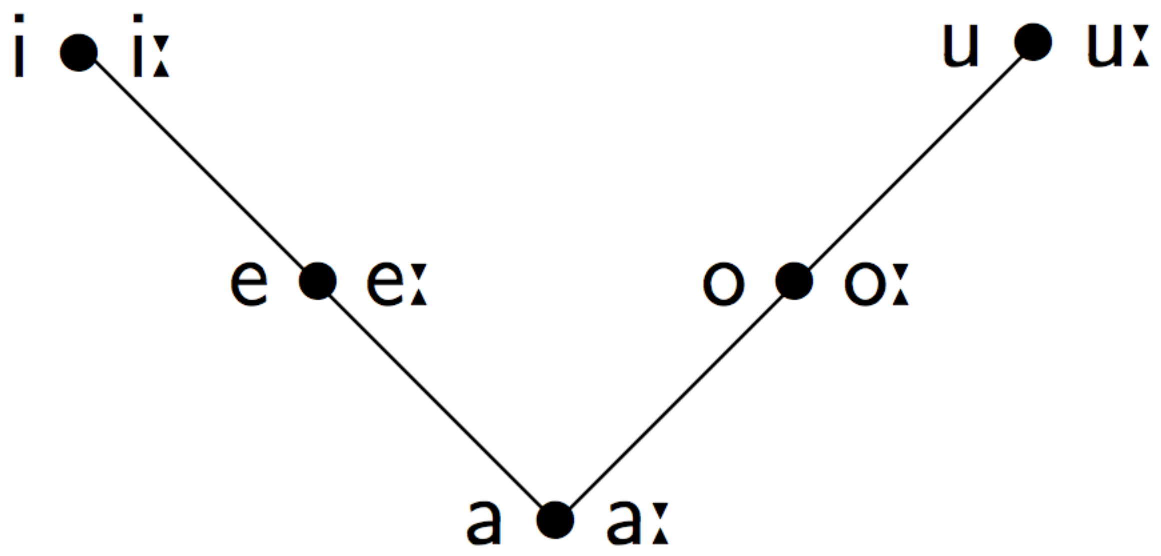




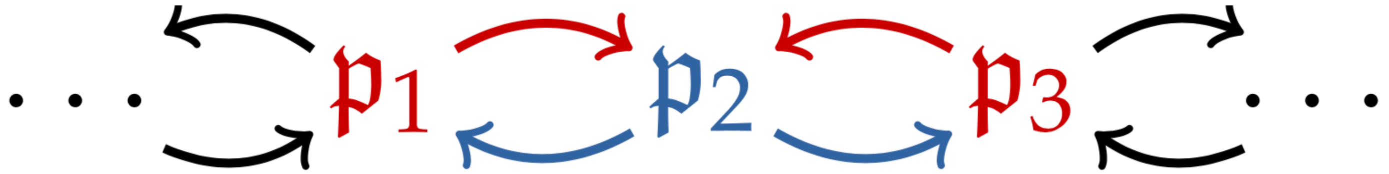
Boruta, 2012, *Indicators of Allophony and Phonemehood.*

INDICATORS OF ALLOPHONY AND PHONEMEHOOD

- Phonemes

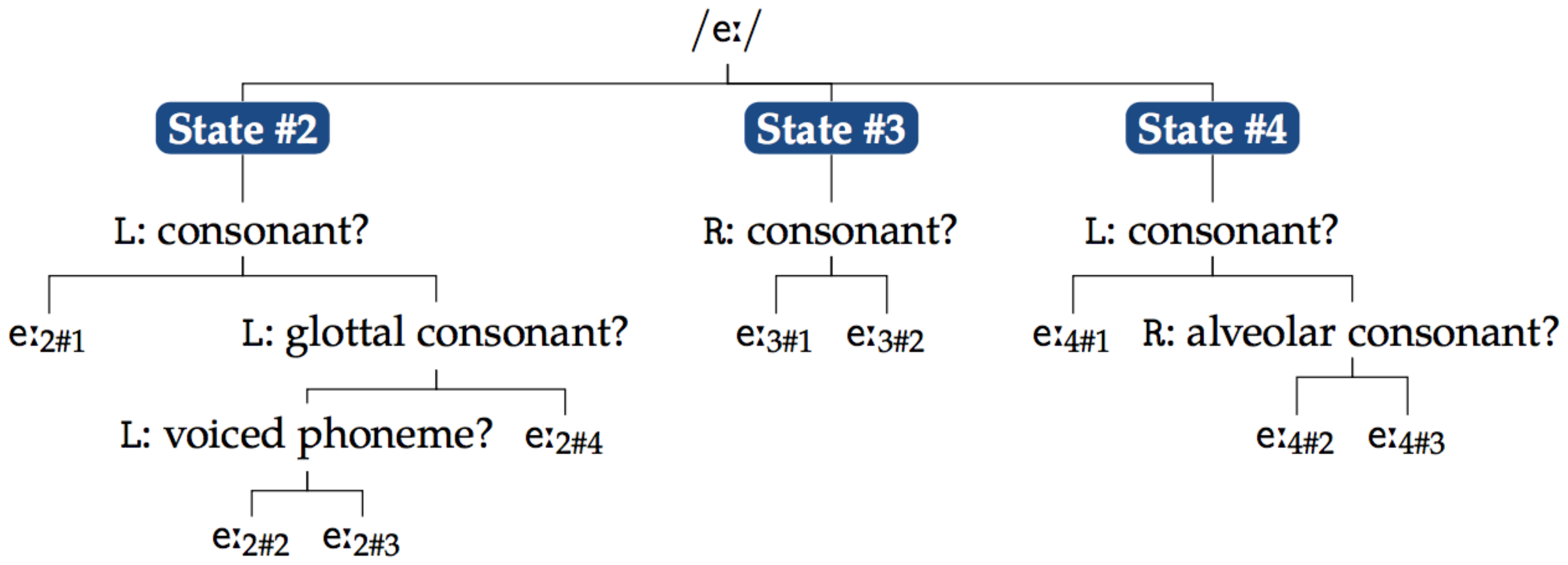


- Allophony



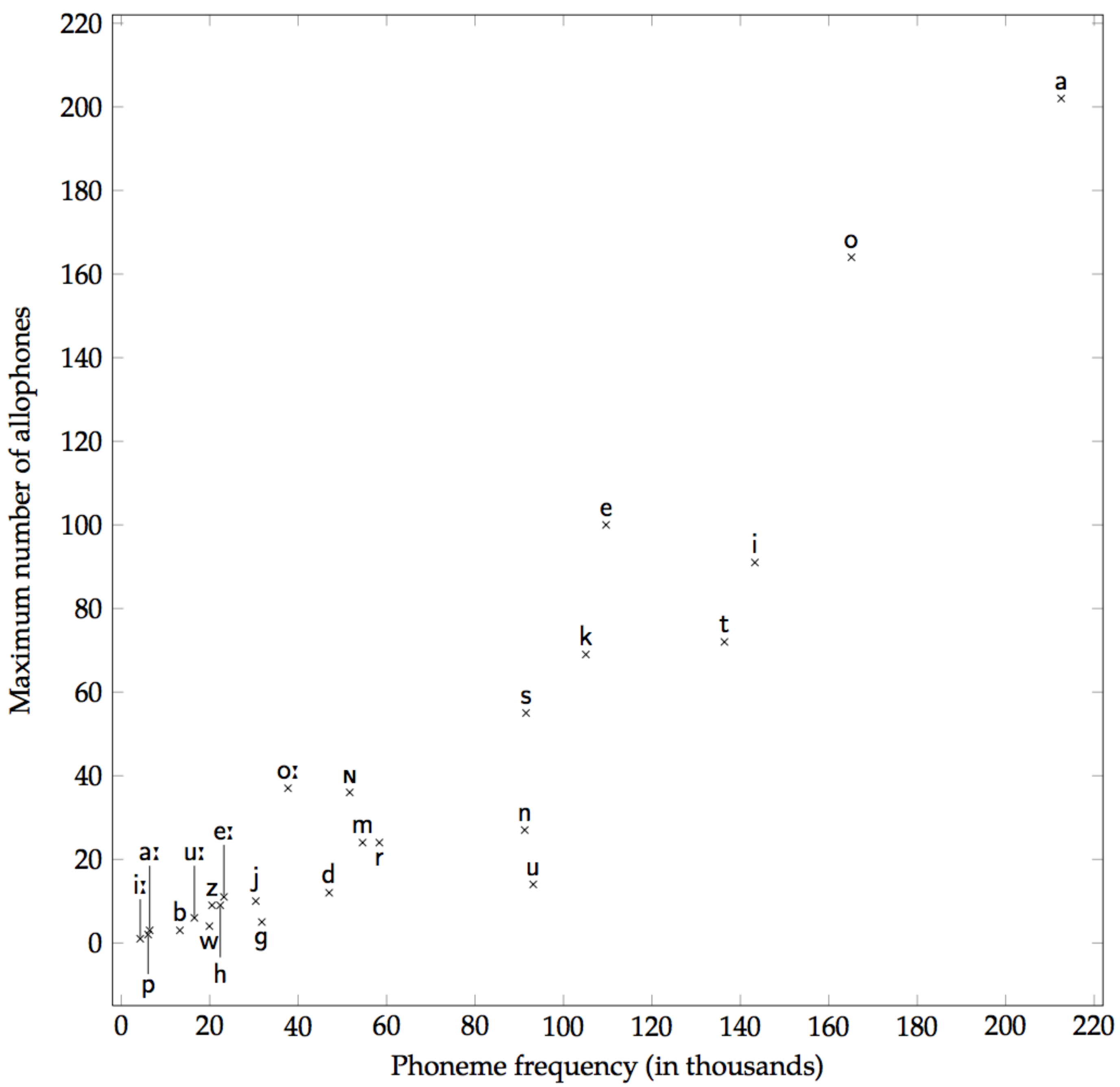
Boruta, 2012, *Indicators of Allophony and Phonemehood.*

INDICATORS OF ALLOPHONY AND PHONEMEHOOD



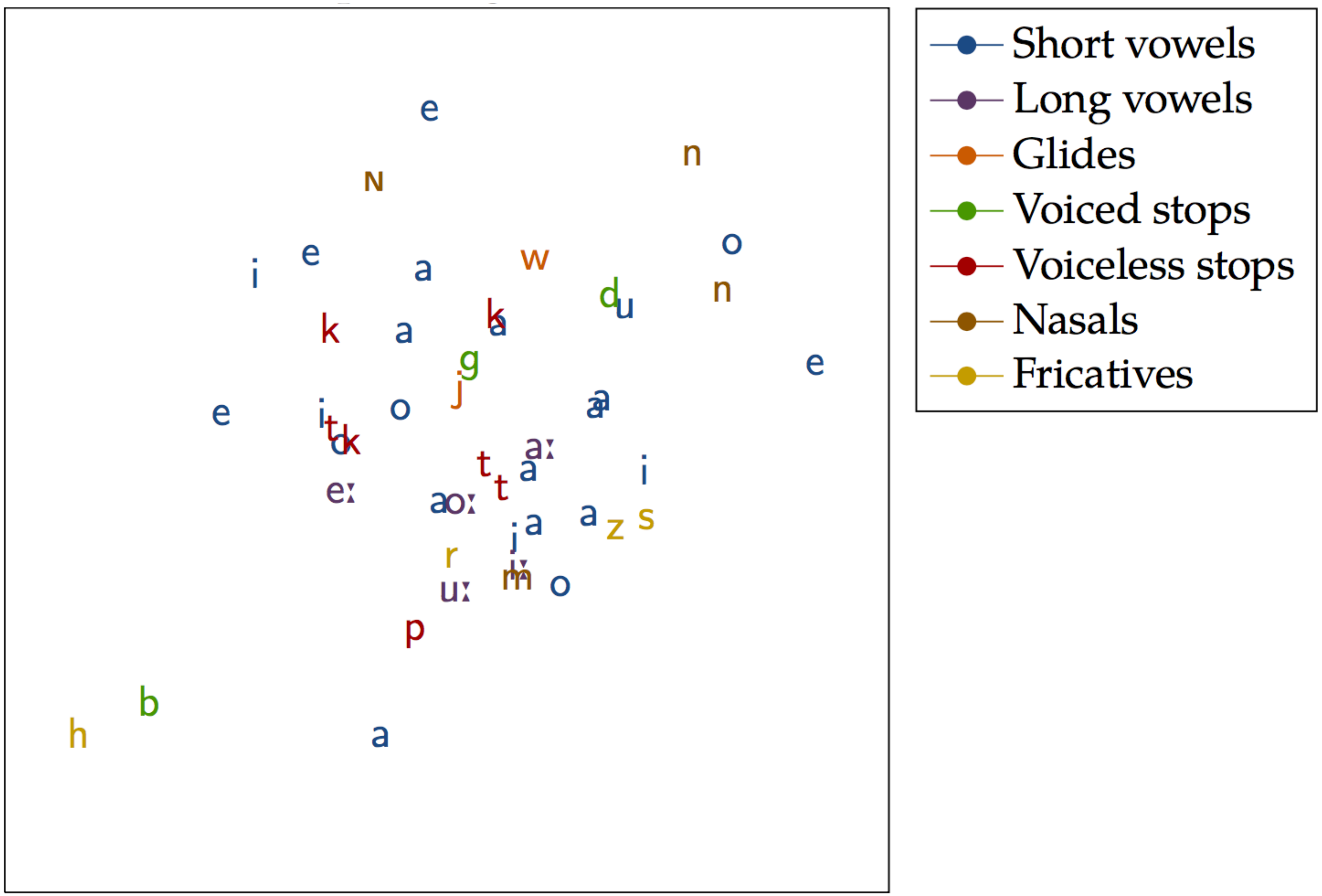
Boruta, 2012, *Indicators of Allophony and Phonemehood.*

INDICATORS OF ALLOPHONY AND PHONEMEHOOD



Boruta, 2012, *Indicators of Allophony and Phonemehood.*

INDICATORS OF ALLOPHONY AND PHONEMEHOOD



Boruta, 2012, *Indicators of Allophony and Phonemehood.*

NLU @ NUANCE (2013–2015)

- Nuance Communications, Montréal
  - ~7k employees, mostly NLP/NLU/CLU
- Natural Language Understanding
  - Personal assistants, knowledge navigators
  - 25+ languages, 120+ countries/territories, 20+ devices
  - Semantic analysis, biiig data

NATURAL LANGUAGE UNDERSTANDING

- Cancel
- Send an email[type] to Ben[recipient] and say what's up[content]
  - Send an email[type] to Ben[recipient] and say… cancel[???]
- Find an Ikea[POI name] store[POI type] near me[relative location]
- Play Vogue[media content] by Madonna[artist]
  - Play the song[type] Vogue[song] by Madonna[singer]
  - Play the music video[type] for Vogue[movie] by Madonna[singer/actress]

NATURAL LANGUAGE UNDERSTANDING

- Science
  - State of the art speech processing, comp. semantics, and multilingual modeling
- Data Science
  - Suboptimal, artificial, and/or unbalanced training data
  - Strong constraints on latency and memory usage
  - End users don’t care if NLU is hard: just make it work!

NLP @ THUNKEN (2016–)

- Thunken, Washington D.C. / Luxembourg
  - NLP consulting
  - Our own products
- IronSift
  - Strategic and competitive intelligence
  - Patents, trademarks, financial statements, clinical trials, drug approvals, government contracts, etc.

NATURAL LANGUAGE PROCESSING + PUBLIC RECORDS

- Science
    - Multilingual named entity linking, comp. semantics, sentiment analysis, etc.
- Data Science
    - Public records are released on a best-effort basis
    - Servers ain’t cheap!
    - End-users expect Google-like search capabilities

FROM SCIENCE TO DATA SCIENCE

- Computational Linguistics vs. Natural Language Processing
  - Same tools, same conferences, etc.
  - CL: scientific study of language, using comp. methods
    - Goal: better linguistic theory
  - NLP: art of solving eng. tasks that use natural language
    - Goal: make it work!

> Big data is like teenage sex:
> everyone talks about it,
> nobody really knows how to do it,
> everyone thinks everyone else is doing it,
> so everyone claims they are doing it...

Dan Ariely, Duke University

FROM LOW-PAYING SCIENCE TO HIGH-PAYING SCIENCE?

- Data science is increasingly popular
  - Named “the sexiest job of the 21st century” by HBR
  - Most recruiters have no clue what data science is (yey!)
- Data science is still science
  - Your work must be rigorous and reproducible
  - …even if it’s just data wrangling for a hackathon

luc@thunken.com