From bits and numbers to explanations – doing research on Internet-based big data

Veronika Laippala

Digital linguistics

School of languages and translation studies

University of Turku

WIKIPEDIA
Encyclopedia

BBC NEWS

Google books
NGRAM VIEWER

Τι κάνετε;

WIKIPEDIA
The Free Encyclopedia

BBC NEWS

Google books

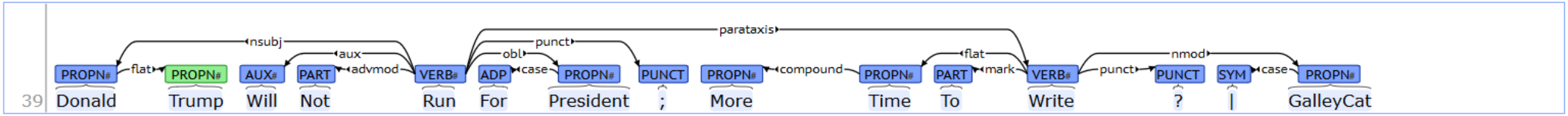
[Turku NLP Group]

English parsebank max 2M | Trump | Search | Case sensitive: ☑ | Hits per page 50

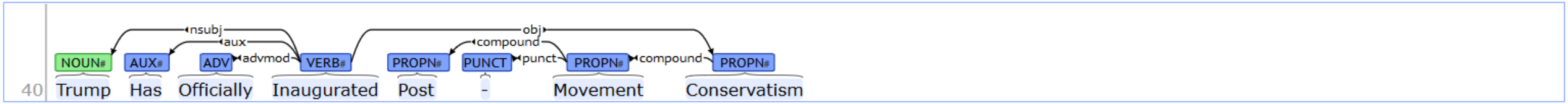
[Link to this query] [Download data] [Query Language]

[Hits in other datasets]

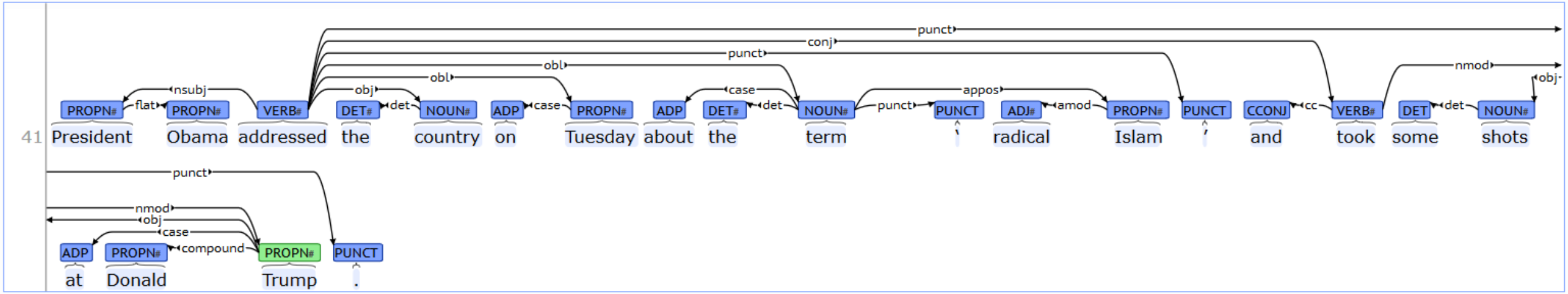
[context] [conllu]



[context] [conllu]



[context] [conllu]



		CONTEXT ALL FORMS (SAMPLE): 100 200 500	FREQ	
1	☐	TRUMP	1309524	

☐ [?]

A	B	C	to get to the bottom of what is at stake in this election. The **Trump** campaign might want to consider bearing down hard on thes
A	B	C	players to watch at the college level, Russ in game athleticism seems to vastly **trump** his combine numbers. Nonetheless, measu
A	B	C	of this argument, in the first presidential debate, when discussing NAFTA, Mr. **Trump** said: " When we sell into Mexico, there's a t
A	B	C	. # Something to think about, in just the second full week of the **Trump** presidency. Download the PDF: qwx24188 qwx24203 qw
A	B	C	the June referendum result hit the US in November with the surprising victory of Donald **Trump** in the presidential election. Afte
A	B	C	the presidential election. After the result, investor focus quickly shifted to what a **Trump** administration will seek to achieve in th
A	B	C	driven by concerns over the stronger US dollar and the potential policy implications of a **Trump** administration. First, dollar stren
A	B	C	, EM investors are worried about the potential for a reversal in globalisation. President-elect **Trump** has talked extensively about

WHAT DO THESE DOCS REPRESENT?

Article Talk

Borgio Verezzi

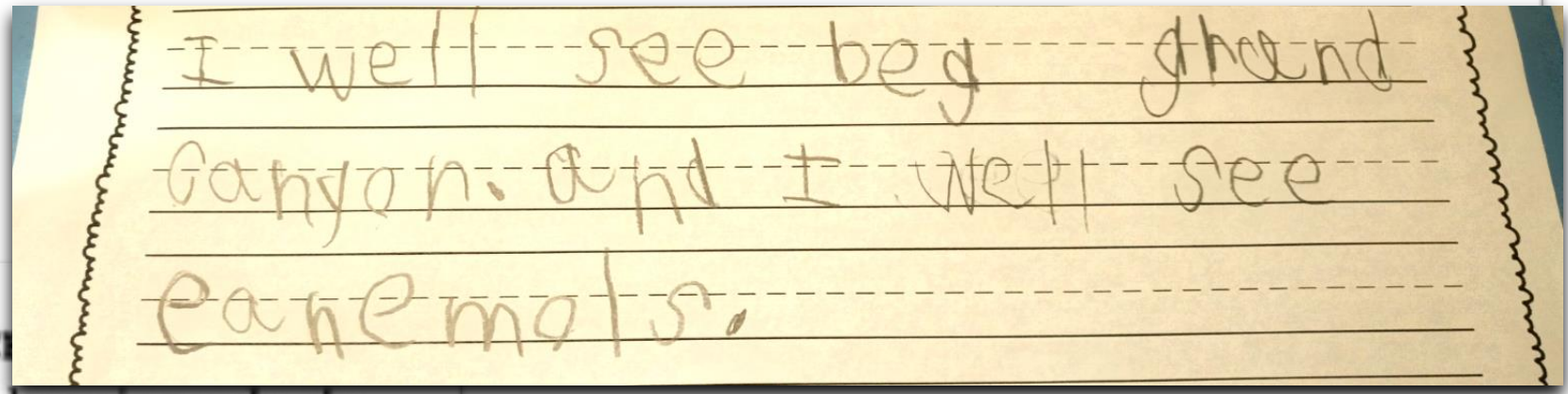
From Wikipedia, the free encyclopedia

Borgio Verezzi (Ligurian: *Bòrzi Veresso*) is a *comune* (municipality) in the Province of Savona in the Italian region Liguria, located about 60 kilometres (37 mi) southwest of Genoa and about 20 kilometres (12 mi) southwest of Savona.

Contents [hide]

1 Geography
2 Main sights
3 References

WIKIPEDIA
The Free Encyclopedia



"I am no longer surprised at your knowing *only* six accomplished women. I rather wonder now at your knowing *any*."

"Are you so severe upon your own sex as to doubt the possibility of all this?"

"*I* never saw such a woman. *I* never saw such capacity, and taste, and application, and elegance, as you describe, united."

Mrs. Hurst and Miss Bingley both cried out against the injustice of her implied doubt, and were both protesting that they knew many women who answered this description, when Mr. Hurst called them to order, with

Ingredients

- 3/4 cup granulated sugar
- 3/4 cup packed brown sugar
- 1 cup butter, softened
- 1 teaspoon vanilla
- 1 egg
- 2 1/4 cups Gold Medal™ all-purpose flour

WHAT DO THESE DOCS REPRESENT?

Article Talk

Borgio Verezzi

From Wikipedia, the free encyclopedia

Borgio Verezzi (Ligurian: *Bòrzi Veresso*) is a *comune* (municipality) in the Province of Savona in the Italian region Liguria, located about 60 kilometres (37 mi) southwest of Genoa and about 20 kilometres (12 mi) southwest of Savona.

Contents [hide]
1 Geography
2 Main sights
3 References

I well see beg ghand
Canyon. and I well see
canemals.

34

"I am no lo owing *only* six
accomplished now at your
knowing *any*."

"Are you so as to doubt the
possibility of all this

"*I* never saw such a woman. *I* never saw such capacity, and taste, and application, and elegance, as you describe, united."

Mrs. Hurst and Miss Bingley both cried out against the injustice of her implied doubt, and were both protesting that they knew many women who answered this description, when Mr. Hurst called them to order, with

Ingredients

- 3/4 cup granulated sugar
- 3/4 cup packed brown sugar
- 1 cup butter, softened
- 1 teaspoon vanilla
- 1 egg
- 2 1/4 cups Gold Medal™ all-purpose flour

WHAT DO THESE DOCS REPRESENT?

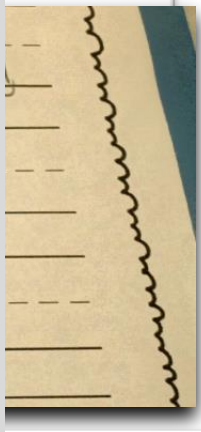
lopedia

òrzi Veresso) is a *comune* (municipality) in the Province of Savona in the Italian region Liguria, located about 60 kilometres (37 mi)
ut 20 kilometres (12 mi) southwest of Savona.

Cont
1 Ge
2 Ma
3 Ref

"a ***register*** is a variety associated with a particular situation of use (including particular communicative purposes)".

(Biber and Conrad 2009: 6)



34

accomplished women. I rather wonder now at your knowing *any*."

"Are you so severe upon your own sex as to doubt the possibility of all this?"

"*I* never saw such a woman. *I* never saw such capacity, and taste, and application, and elegance, as you describe, united."

Mrs. Hurst and Miss Bingley both cried out against the injustice of her implied doubt, and were both protesting that they knew many women who answered this description, when Mr. Hurst called them to order, with

Ingredients

- 3/4 cup granulated sugar
- 3/4 cup packed brown sugar
- 1 cup butter, softened
- 1 teaspoon vanilla
- 1 egg
- 2 1/4 cups Gold Medal™ all-purpose flour

A piece of news, an opinion or something else?
Analyzing and automatically detecting text varieties in the multilingual Internet

EMIL AALTOSEN SÄÄTIÖ

Objectives:

1) describe the full range of registers in the freely accessible Internet
2) develop methods to model and automatically detect the registers
3) apply the methods to detect registers from Universal Parsebanks*

* Universal Parsebanks is a web-crawled data set that includes billions of words in dozens of languages developed in previous projects of our research group. It is freely usable online!

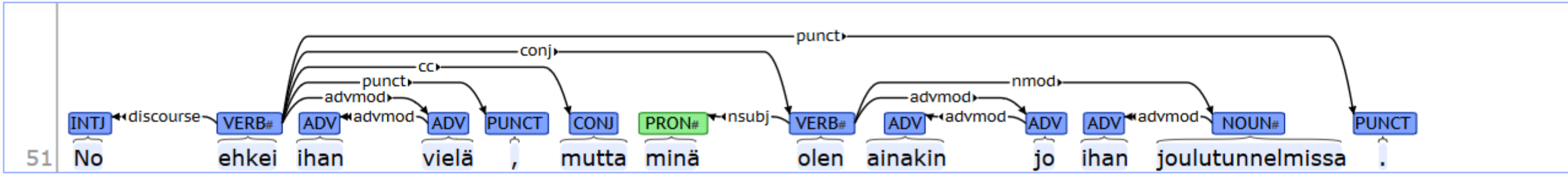
[Turku NLP Group]

Finnish Internet Parsebank Sample | minä | Search | **Case sensitive:** ☑ **Hits per page** 50

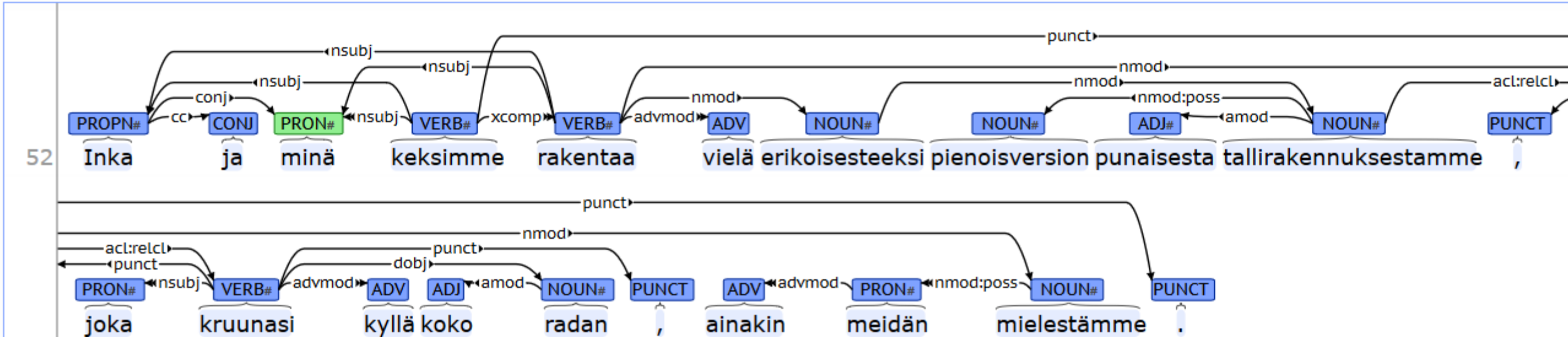
[Link to this query] [Download data] [Query Language]

[Hits in other datasets]

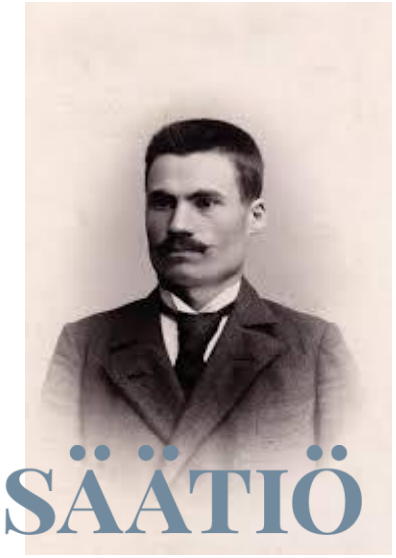
[context] [conllu]



[context] [conllu]



A piece of news, an opinion or something else?
Analyzing and automatically detecting text varieties in the multilingual Internet



EMIL AALTOSEN SÄÄTIÖ

Challenges

- What is the Internet composed of?
- To what extent can registers be automatically identified?
- So much data and so many languages…

Corpus of online registers of English (CORE)
(Egbert, Biber and Davies 2015)

- Unrestricted sample of the searchable web
- 48,571 documents, > 50 million words
- Manually annotated for registers
- 8 main registers functional labels
  - *narrative, informational description, lyrical, informational persuasion, spoken, opinion, how-to, interactive discussion*
- Divided into 27 sub-registers
  - Such as *news, personal blogs, encyclopedia articles, how-to*

- Advice
- Discussion forum
- Description of a person
- Description with intent to sell
- Description of a thing
- Encyclopedia article
- FAQ about information
- Formal speech
- Historical article
- How-to
- Informational blog
- Interview
- News article / news blog
- Opinion blog
- Poem
- Personal blog
- Question / answer
- Research article
- Recipe
- Religious blog / sermon
- Review
- Song lyrics
- Sports report
- Short story
- Travel blog
- Tv subscripts

- Advice
- Discussion forum
- Description of a person
- Description with intent to sell
- Description of a thing
- Encyclopedia article
- FAQ about information
- Formal speech
- Historical article
- How-to
- Informational blog
- Interview
- News article / news blog
- Opinion blog
- Poem
- Personal blog
- Question / answer
- Research article
- Recipe
- Religious blog / sermon
- Review
- Song lyrics
- Sports report
- Short story
- Travel blog
- Tv subscripts

+ Hybrid documents combining several communicative purposes and registers

JPchenet.fi

JPChenet viinit

ETUSIVU YHTEYS

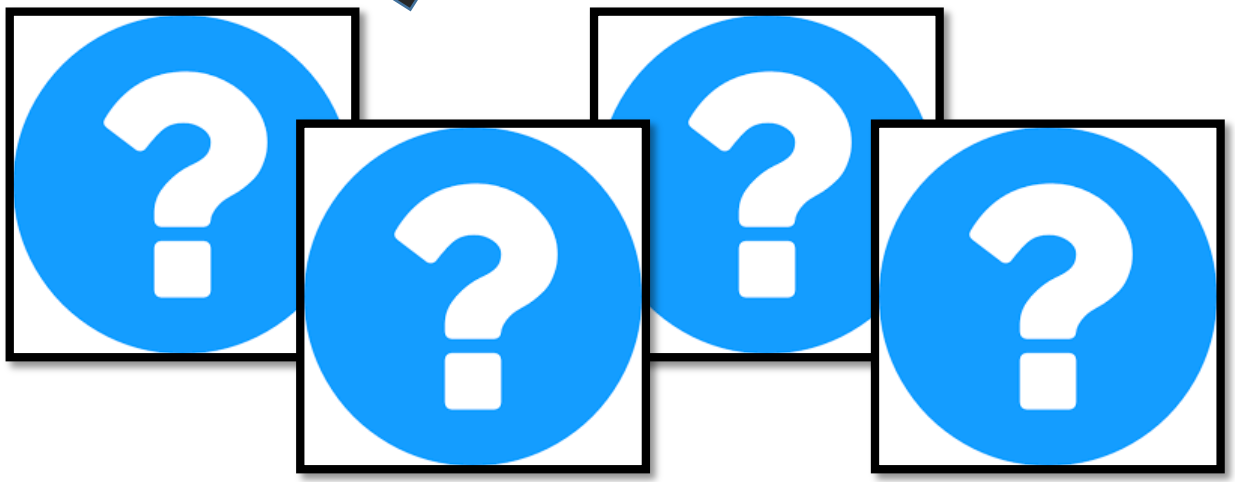
http://jpchenet.fi/etusivu/30-vuotta-suosion-huipulla/

J.P. Chenet: More popular than ever after 30 years!

It all started in the beginning of the 1980s, when a young French wine merchant, Joseph Helfrich, had a dream. He noticed that foreigners did not understand French wines. Joseph started to collaborate with Jean-Paul Chanel. Together, they created easy-going wines that could be enjoyed right away, without conservation…. The rest is history….

Manually identified example documents (training data)

New



Classification experiments

- Support vector machine (SVM) in Scikit learn, 80%/20% train / test
- 27 sub-registers in CORE
- Comparison of 3 feature sets generated with the Biber tagger

Laippala, Kyröläinen, Egbert, Biber (submitted): Modeling the jungle of online registers by focusing on register-specific differences.

Type of feature set	Realization
Lexical features	Fishing is a great way for people to enjoy themselves
Grammatical features	^nvbg+++xvbg+ ^vbz+bez+vrb ^at++++ ^jj+atrb+++ ^nn++++ ^in++++ ^nns++++ ^to++++ ^vbi++++ ^ppls+pp3+++
Lexico-grammatical features	Fishing is a great way for people to enjoy themselves ^nvbg+++xvbg+ ^vbz+bez+vrb ^at++++ ^jj+atrb+++ ^nn++++ ^in++++ ^nns++++ ^to++++ ^vbi++++ ^ppls+pp3+++

Results on the sub-register level

Feature set	Precision (%)	Recall (%)	F1-score (%)
Grammatical	64,34	59,14	59,9
Lexical	71,84	70,77	70,8
Lexico-grammatical	**74,48**	**75,13**	**74,5**

Results on the sub-register level

→ Promising

→ What does the model actually grasp?

Feature se			re (%)
Grammatica			9
Lexical	71,84	70,77	70,8
Lexico-grammatical	**74,48**	**75,13**	**74,5**

Interview	Sports news
interview	fight
what	penalty
how	injury
like	win
there	women's
very	jason
did	season
play	game
fashion	won
that	howard

[Turku NLP Group]

Finnish Internet Parsebank Sample

Type in your query

Search

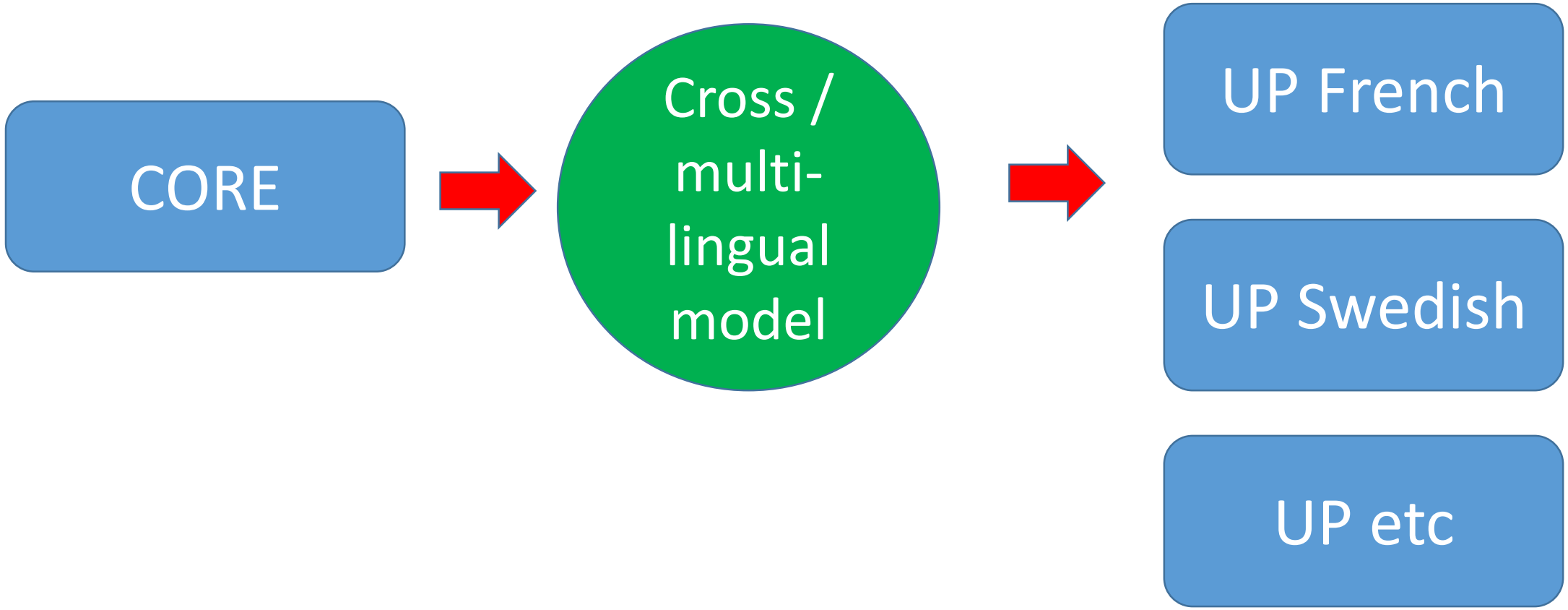
Case sensitive: ☑ **Hits per page** 50

guage]

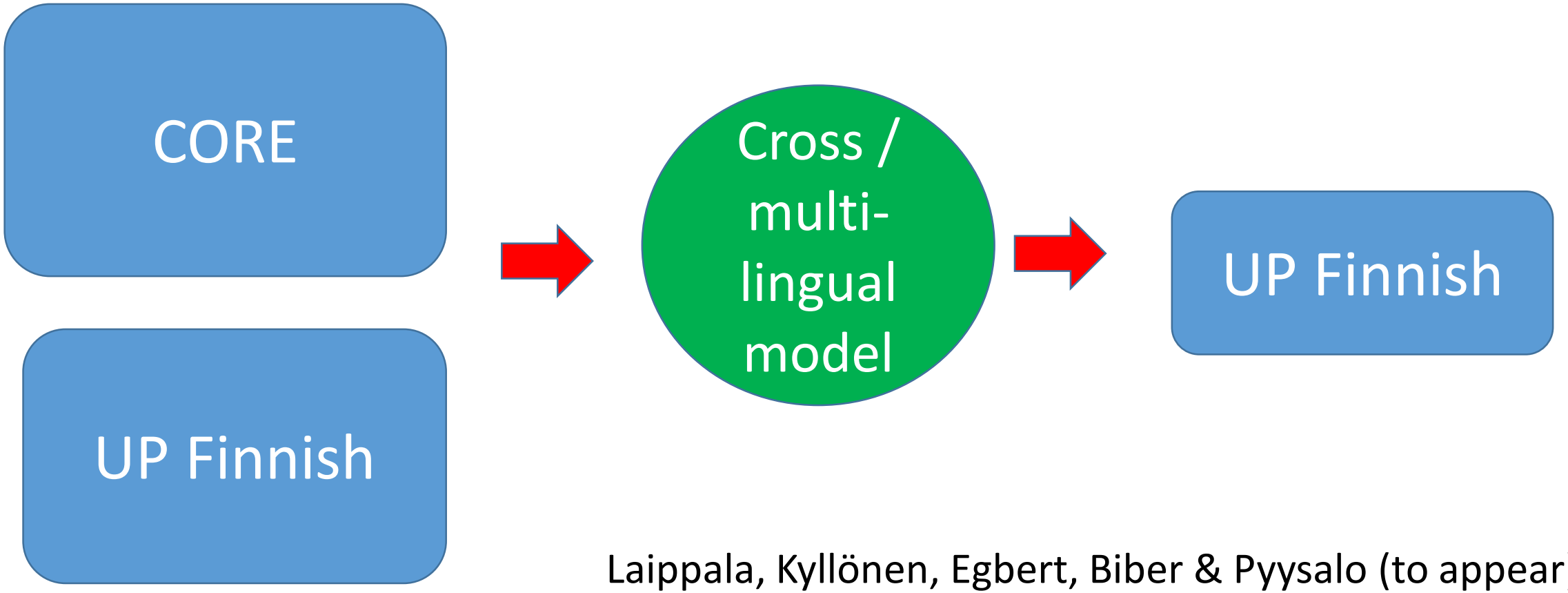
- Finnish parsebank max 2M
- French parsebank max 2M
- Galician parsebank max 2M
- German parsebank max 2M
- Greek parsebank max 2M
- Hebrew parsebank max 2M
- Hindi parsebank max 2M
- Hungarian parsebank max 2M
- Indonesian parsebank max 2M
- Irish parsebank max 2M
- Italian parsebank max 2M
- Japanese parsebank max 2M
- Kazakh parsebank max 2M
- Korean parsebank max 2M
- Latin parsebank max 2M
- Latvian parsebank max 2M
- Norwegian-Bokmaal parsebank max 2M
- Norwegian-Nynorsk parsebank max 2M
- Old_Church_Slavonic parsebank max 2M
- Persian parsebank max 2M

?? What about other languages??

Cross / multilingual learning to help?



Cross / multilingual learning to help?



Laippala, Kyllönen, Egbert, Biber & Pyysalo (to appear): Toward Multilingual Identification of Online Registers. Proceedings of NoDaLiDa, September 2019.

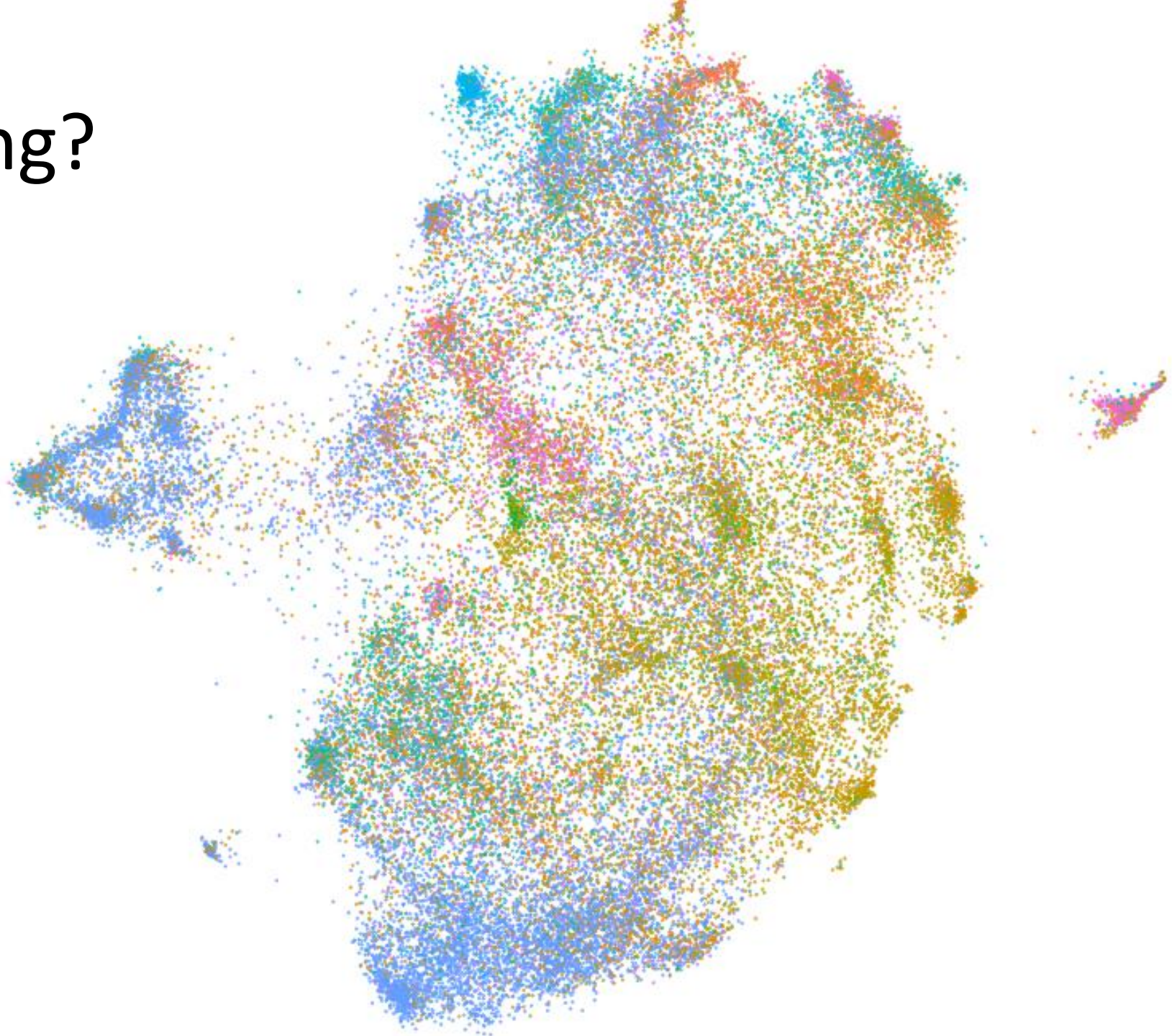
Register	English	Finnish
Narrative	12,541	778
Opinion	3,960	339
Informational description	3,195	379
Discussion	2,697	140
How-to	955	144
Informational persuasion	684	446
All	**24,912**	**2,237**

Setting	Monolingual			
Training data	Finnish	English		
Test data	Finnish	English		
AUC	**83.8%**	**93.6%**		

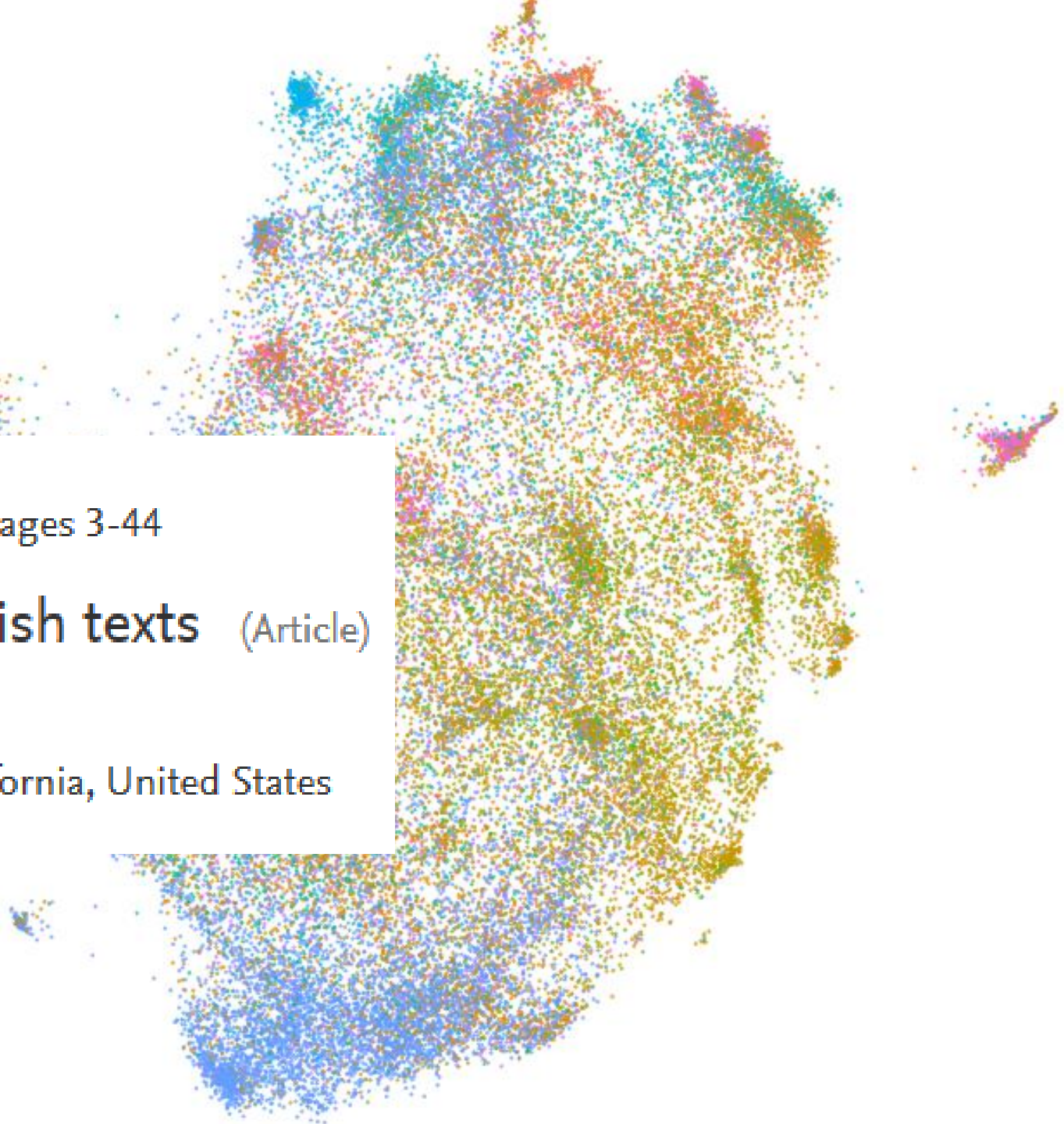
Setting	Monolingual		Cross-/Multilingual	
Training data	Finnish	English	English	En+Fi
Test data	Finnish	English	Finnish	Finnish
AUC	**83.8%**	**93.6%**	**78.6%**	**85.3%**

Other options….?

Clustering?



Clustering?



Linguistics Open Access
Volume 27, Issue 1, 1989, Pages 3-44

A typology of English texts (Article)

Biber, D.

University of Southern California, United States

Sparse data

$m > 100{,}000$

	able	about	actually	after	… word *m*
Text 1	0	2	0	1	
Text 2	0	4	0	0	
Text 3	0	0	0	0	
Text… N					

$N > 10{,}000$

Idea behind word2vec

Idea behind word2vec

- “You shall know a word by the company it keeps”

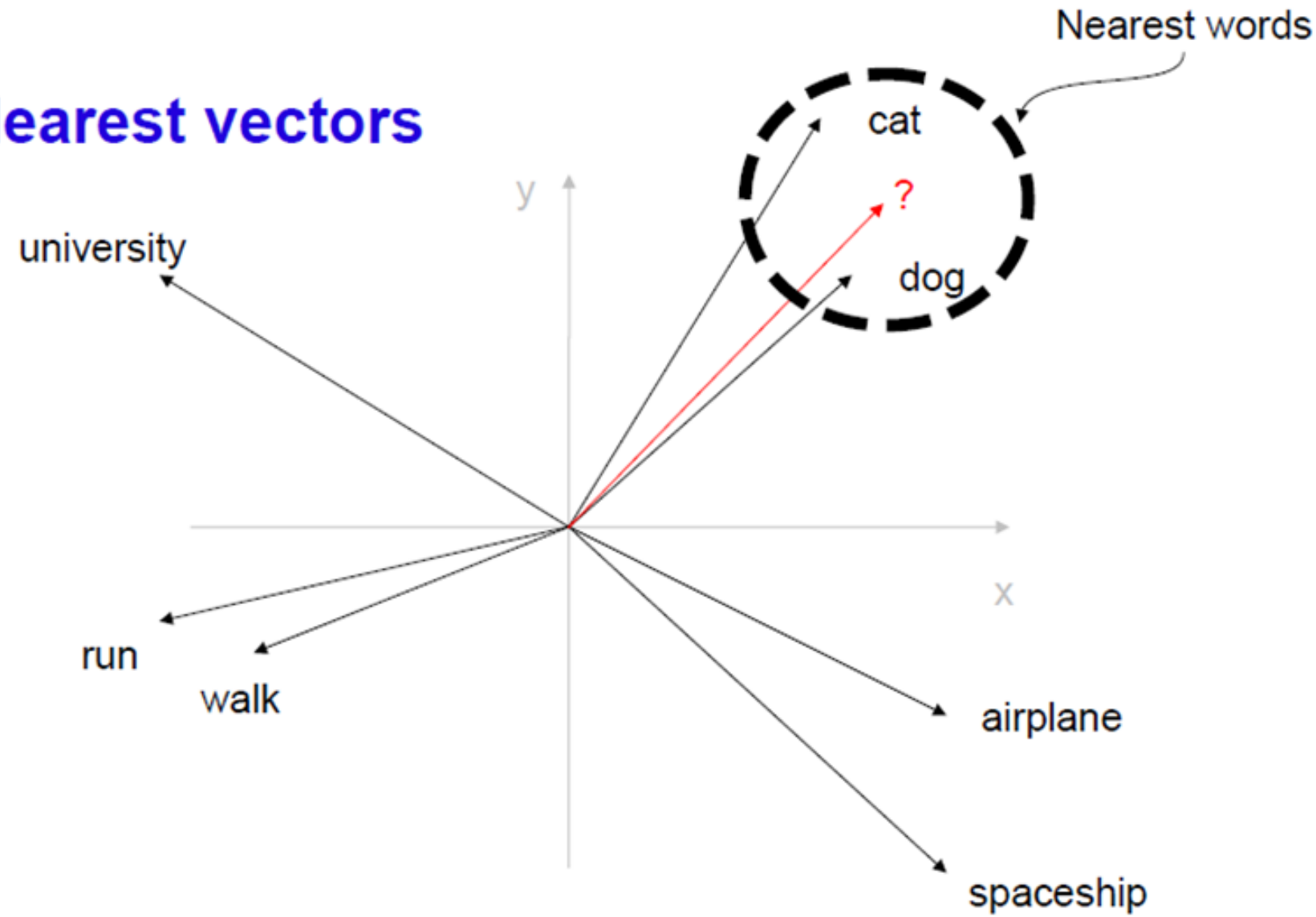
J.R. Firth, 1957

- “We found a cute, hairy wampimuk sleeping behind the tree.”

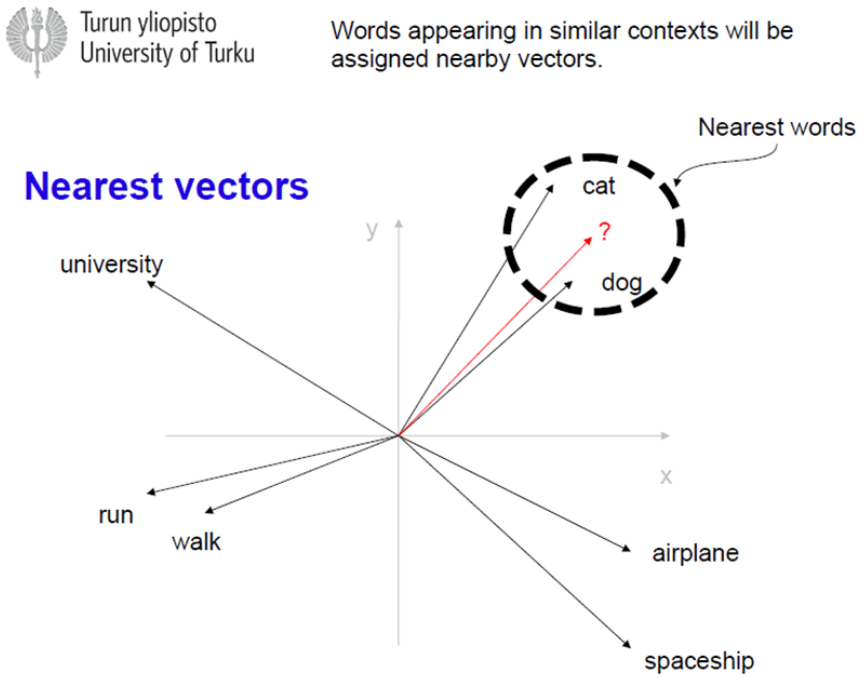
Turun yliopisto
University of Turku

Words appearing in similar contexts will be assigned nearby vectors.

Nearest vectors



Clustering registers



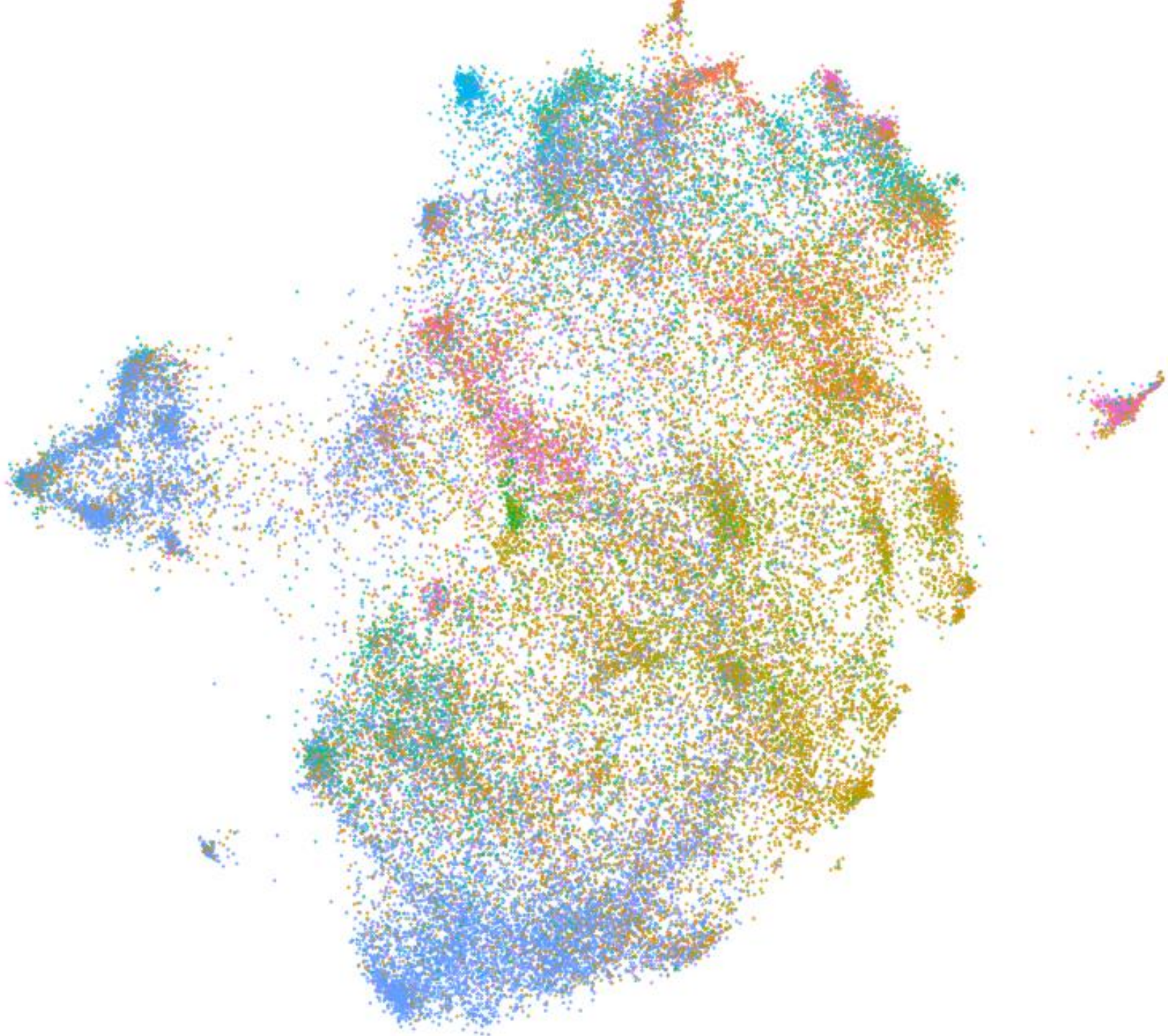
- Original word2vec
  - Context: co-occurrence of words
  - Output: predicted word embedding
  - man _ hamburgers

- From word-level representation to document-level
  - Context: word co-occurrence enriched with syntax
  - Output: predicted document embedding

Document embeddings for online English

- Model: Word2vecf (Levy and Goldberg 2014)
- Training data CORE + documents from Global Web-Based Corpus of English → 120,000 documents
- Contexts: lemmata + syntactic information

	1	2	3	4	… 300
Text 1	0.233608	0.542426	0.376280	0.313996	0.186521
Text 2	0.132363	0.141933	-0.197238	-0.012108	0.046967
Text N					



Jelavic calls for Old Firm to come to the **Premier League**

New Everton signing Nikica Jelavic fears Rangers and Celtic will become second-class clubs if they remain in Scottish football. The Croatian striker moved to Goodison Park on transfer deadline day, with Rangers manager Ally McCoist powerless to prevent the **6m** move. "Rangers and Celtic can't compete with English clubs **financially,** so it would be very important to them if they could join the Premier League one day,"

Jelavic calls for Old Firm to come to the **Premier League**

New Everton signing Nikica Jelavic fears Rangers and Celtic will become second-class clubs if they remain in Scottish football. The Croatian striker moved to Goodison Park on transfer deadline day, with Rangers manager Ally McCoist powerless to prevent the **6m** move. "Rangers and Celtic can't compete with English clubs **financially,** so it would be very important to them if they could join the Premier League one day,"

The Premier League took a big step towards introducing a break-even rule following a meeting in London. There was no formal agreement between the 20 chairmen and chief executives over how to introduce **costs controls**. But clubs agreed to focus on a model similar to the **Financia**l Fair Play **regulations** introduced by Uefa, which require teams to avoid making losses.

Jelavic calls for Old Firm to come to the **Premier League**

New Everton signing Nikica Jelavic fears Rangers and Celtic will become second-class clubs if they remain in Scottish football. The Croatian striker moved to Goodison Park on transfer deadline day, with Rangers manager Ally McCoist powerless to prevent the **6m** move. "Rangers and Celtic can't compete with English clubs **financially,** so it would be very important to them if they could join the Premier League one day,"

The Premier League took a big step towards introducing a break-even rule following a meeting in London. There was no formal agreement between the 20 chairmen and chief executives over how to introduce **costs controls**. But clubs agreed to focus on a model similar to the **Financia**l Fair Play **regulations** introduced by Uefa,

which require teams to avoid making losses**.**

Didier Drogba has told **Chelsea** he does not want further **talks on his future** until the end of the season, leaving it ever more likely that the 34-year-old striker, who has been such a key figure in the club's two semi-final wins over the last six days, will leave for nothing in the summer. Drogba has rejected a one-year **deal** and there have been no further talks to resolve the situation […].

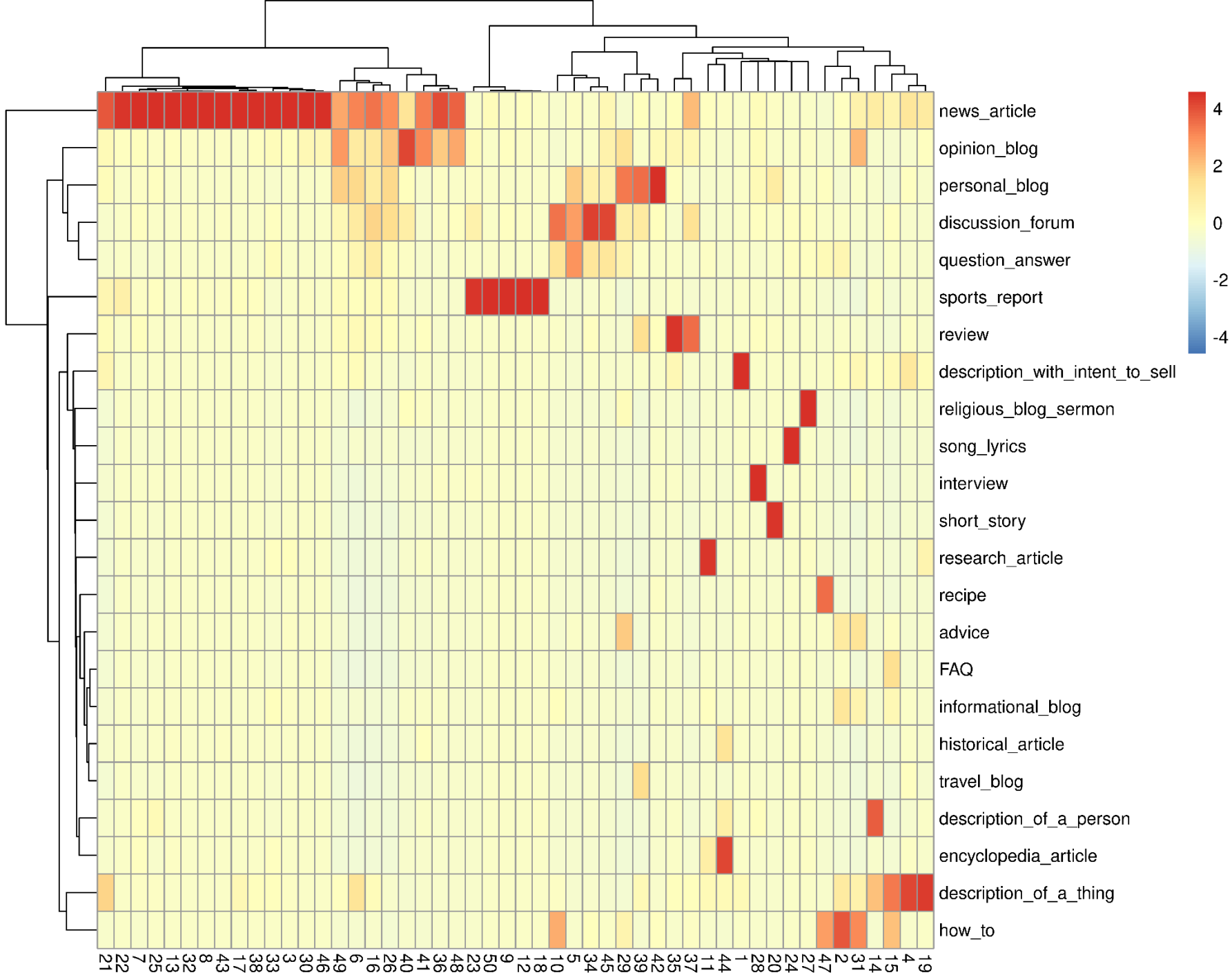
Document vectors predicting the registers

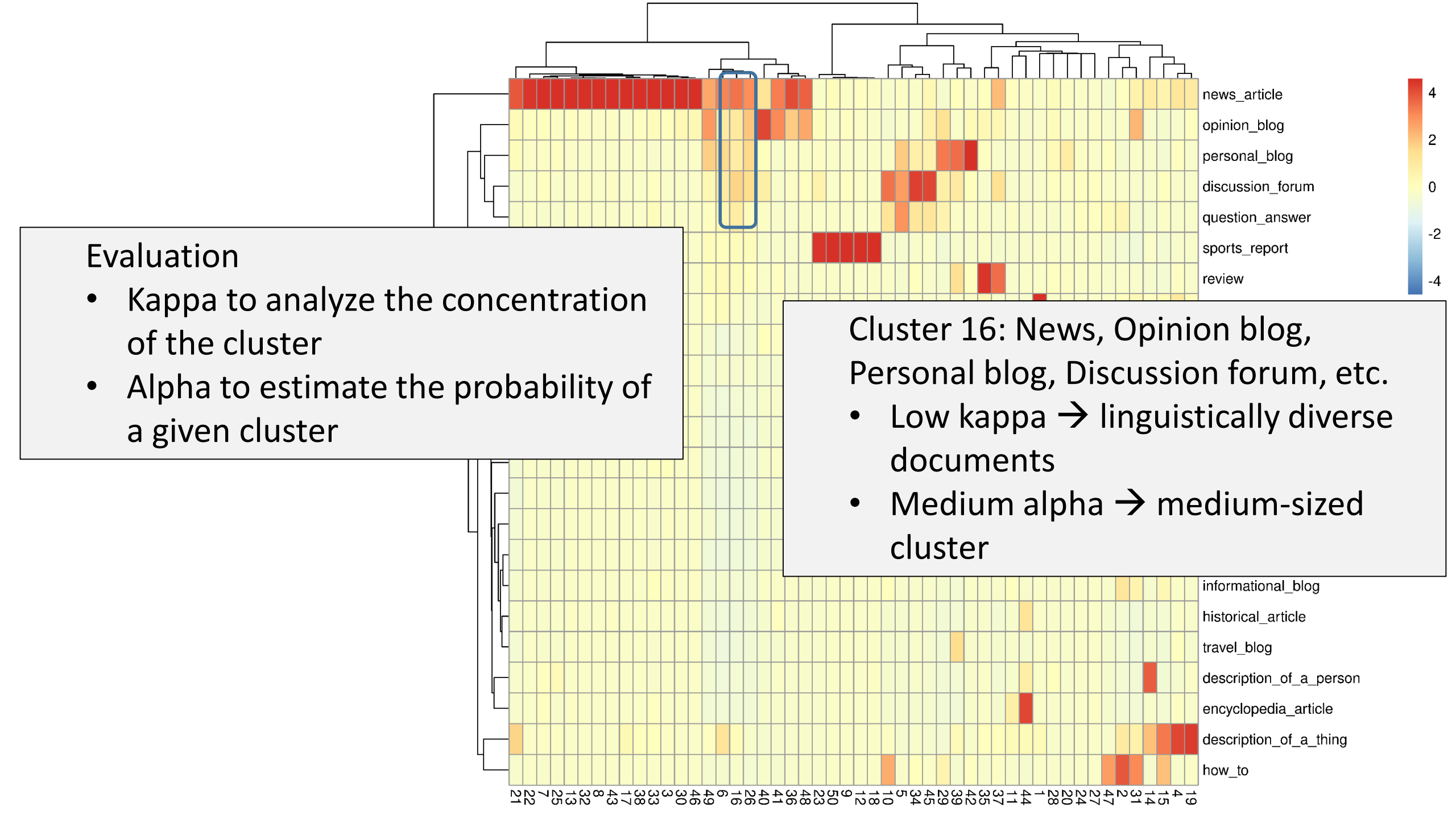
- Text classification with a polynomial support vector machine
  - svm.SVC with a polynomial kernel in Scikit learn, 80%/20% train / test sets
- Two experiments
  - Full dataset
  - Naïve dataset with four registers with equal class sizes (*Travel blog, news, discussion forum, sport report)*

Data	Response	Precision	Recall
Full data (N=25,038)	Registers (N=23)	71%	64%
Naïve data (N=472)	Registers (N=4)	96%	96%

Clustering document embeddings

- Generative mixture-model approach based on the von Mises-Fisher (MVMF) distribution (Banerjee et al., 2005)
  - Distribution arises naturally for data distributed on the unit (hyper)sphere, such as our vectors with normalized lengths
  - Performs well with high-dimensional data
- Explain the numbers!!
  - Kappa to analyze the concentration of the cluster (similar to standard deviation in Gaussian distributions), how concentrated the documents vectors are around the mean vector
  - Alpha to estimate the probability of a given cluster


news_article
opinion_blog
personal_blog
discussion_forum
question_answer
sports_report
review
description_with_intent_to_sell
religious_blog_sermon
song_lyrics
interview
short_story
research_article
recipe
advice
FAQ
informational_blog
historical_article
travel_blog
description_of_a_person
encyclopedia_article
description_of_a_thing
how_to
4
2
0
-2
-4
21 22 7 25 13 32 8 43 17 38 33 3 30 46 49 6 16 26 40 41 36 48 23 50 9 12 18 10 5 34 45 29 39 42 35 37 11 44 1 28 20 24 27 47 2 31 14 15 4 19



Evaluation

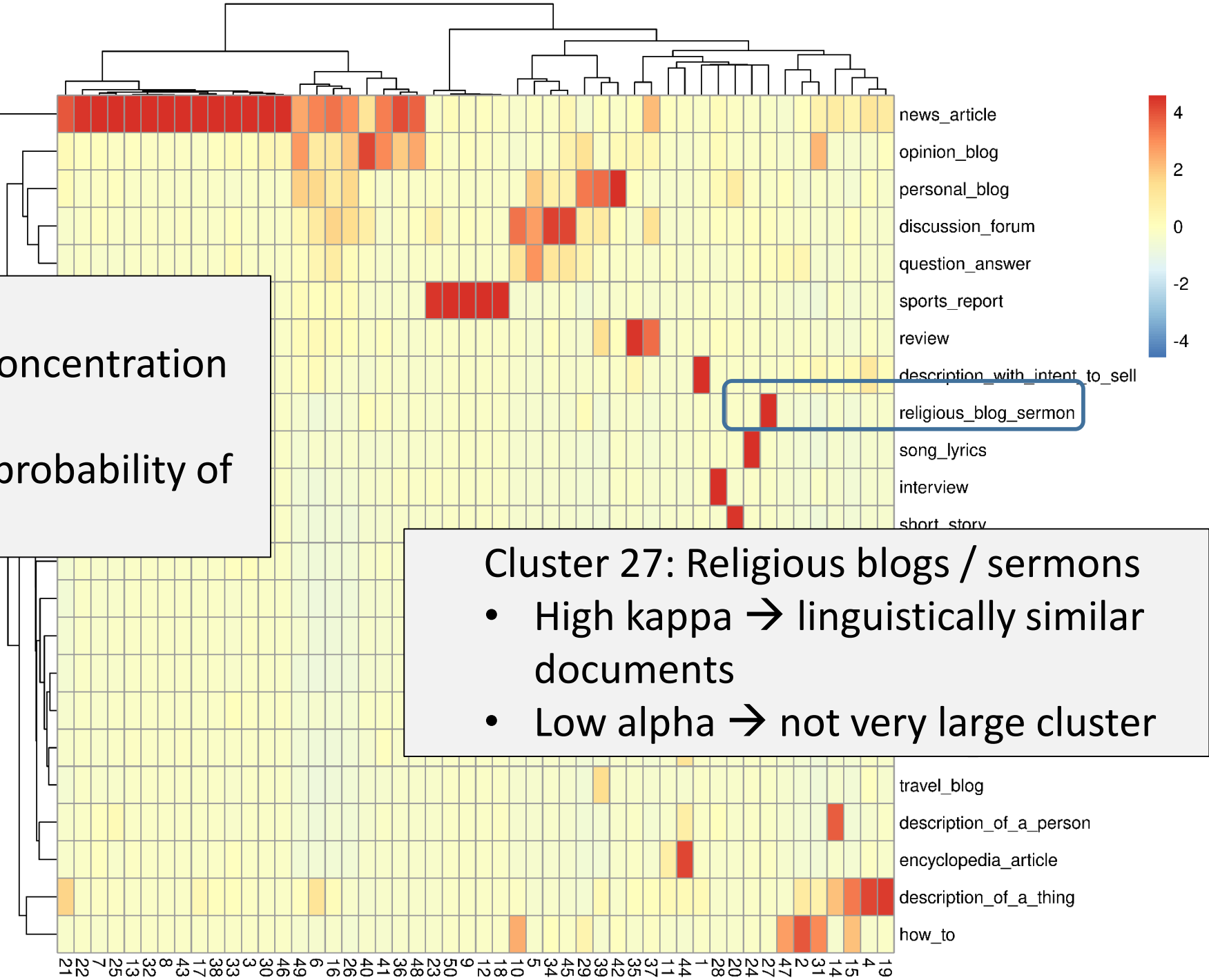
- Kappa to analyze the concentration of the cluster
- Alpha to estimate the probability of a given cluster

Cluster 16: News, Opinion blog, Personal blog, Discussion forum, etc.

- Low kappa → linguistically diverse documents
- Medium alpha → medium-sized cluster

Evaluation

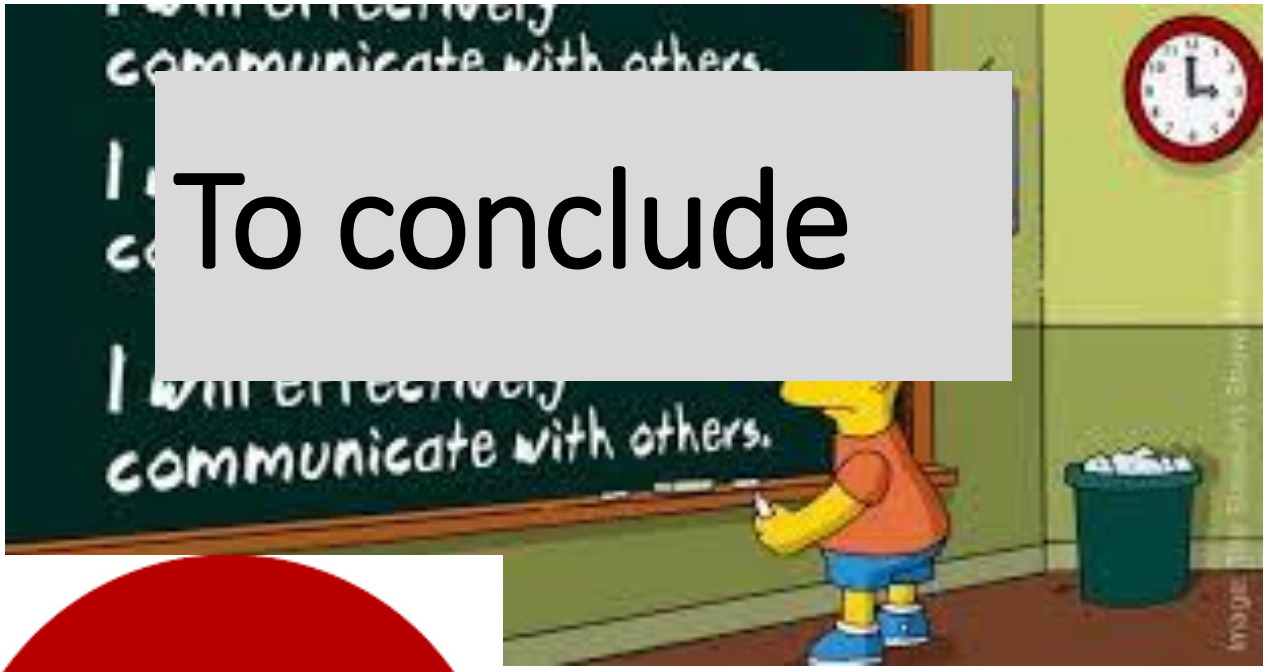
- Kappa to analyze the concentration of the cluster
- Alpha to estimate the probability of a given cluster



Cluster 27: Religious blogs / sermons

- High kappa → linguistically similar documents
- Low alpha → not very large cluster

To conclude



PEDIA
cyclopedia

BBC NEWS

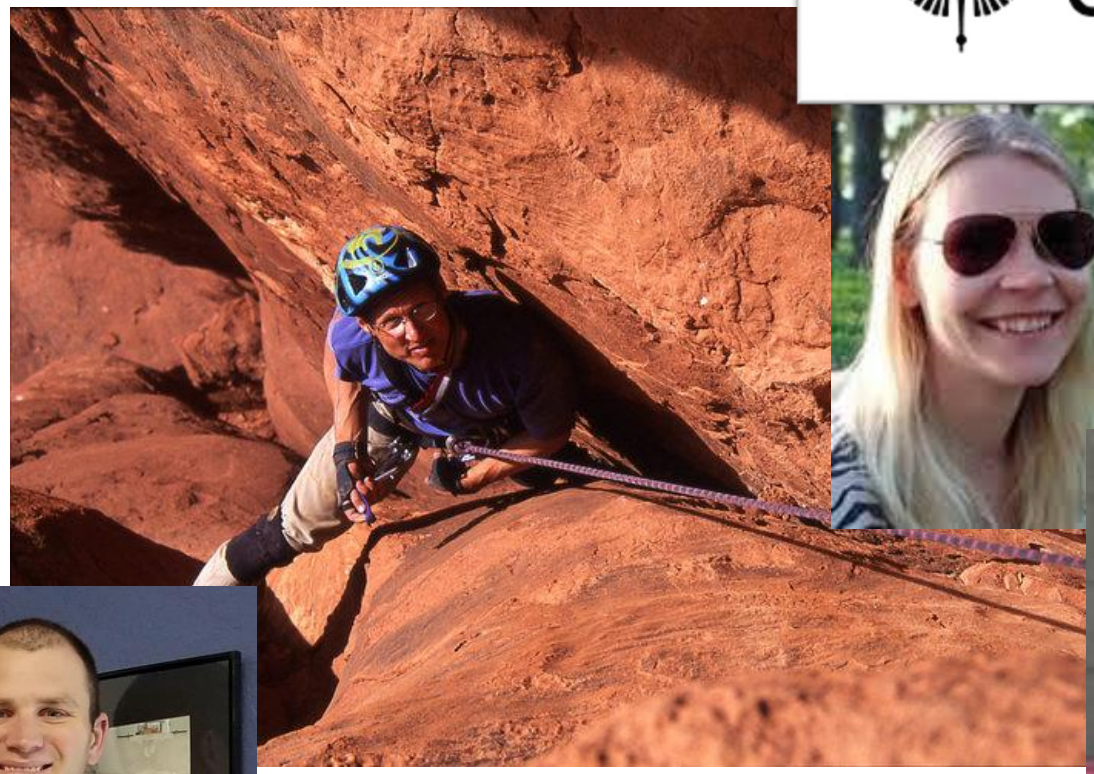
SUOM

Douglas Biber
Jesse Egbert
Filip Ginter
Jenna Kanerva
Roosa Kyllönen
Aki Kyröläinen
Sampo Pyysalo

Douglas Biber
Jesse Egbert
Filip Ginter
Jenna Kanerva
Roosa Kyllönen
Aki Kyröläinen
Sampo Pyysalo

McMaster University

UNIVERSITY OF TURKU



Brock University

Psst: we are hiring…
https://www.utu.fi/fi/yliopisto/tule-meille-toihin/avoimet-tehtavat

Fulbright Finland