Exploring cross-linguistic representations of Web registers with a deep multilingual model

Aki-Juhani Kyröläinen, Filip Ginter and Veronika Laippala

TURKUNLP .ORG

UNIVERSITY OF TURKU

Introduction

Register (Biber, 1988)

- Text varieties such as news, how-to pages and encyclopedia articles
- Defined by their situational characteristics

Register variation and cross-linguistic similarities

- **Biber, 2014**
  - Oral vs. literate
  - Narrative vs. non-narrative discourse
- **Li, Dunn and Nini, 2022**
  - Register variation remains stable across 60 languages

Multilingual language models

Pretrained on very large amounts of massively multilingual data

- E.g., **XLM-Roberta** trained on 2.5TB of filtered CommonCrawl containing 100 languages

Trained using masked language modeling

- Predict a given masked token in the input
- *I have watched this [MASK] and it was awesome.*
- *What is [MASK] name?*

13 top layers can be fine-tuned to automatic register identification using manually register annotated data!

Multilingual language models

Multilingual register identification models work!

- *Zero-shot*: train on one language, evaluate on another
- *Zero shot* performance across three languages and eight register classes
  71% F1-score (Rönnqvist et al. 2021)

⇒ Registers have similarities across languages that allow for their identification

⇒ What can this multilingual register model tell us about the linguistic similarities of registers across languages?

Data

Register Oscar (Laippala et al. 2022)

- A sub-corpus of Oscar (Ortiz Suárez et al., 2020), a massive Web-crawled corpus
- 14 languages, 351M documents

Sample used in the study

- 72,000 documents
- Three languages: English, French, Finnish

Model

Model

- Fine-tuned XLM-R *base* to register classification using register annotations in Finnish, French and English
- Model performance: F1-score of 0.80

Narrative, Informational Description, Informational Persuasion, Opinion, How-to, Interactive Discussion

Document representations in the model

- Embeddings, i.e., vectors in a multilingual high-dimensional space (768 dims)
- Final layer of the fine-tuned XLM-R model

What has the model learnt from the registers?

Questions

Option 1:

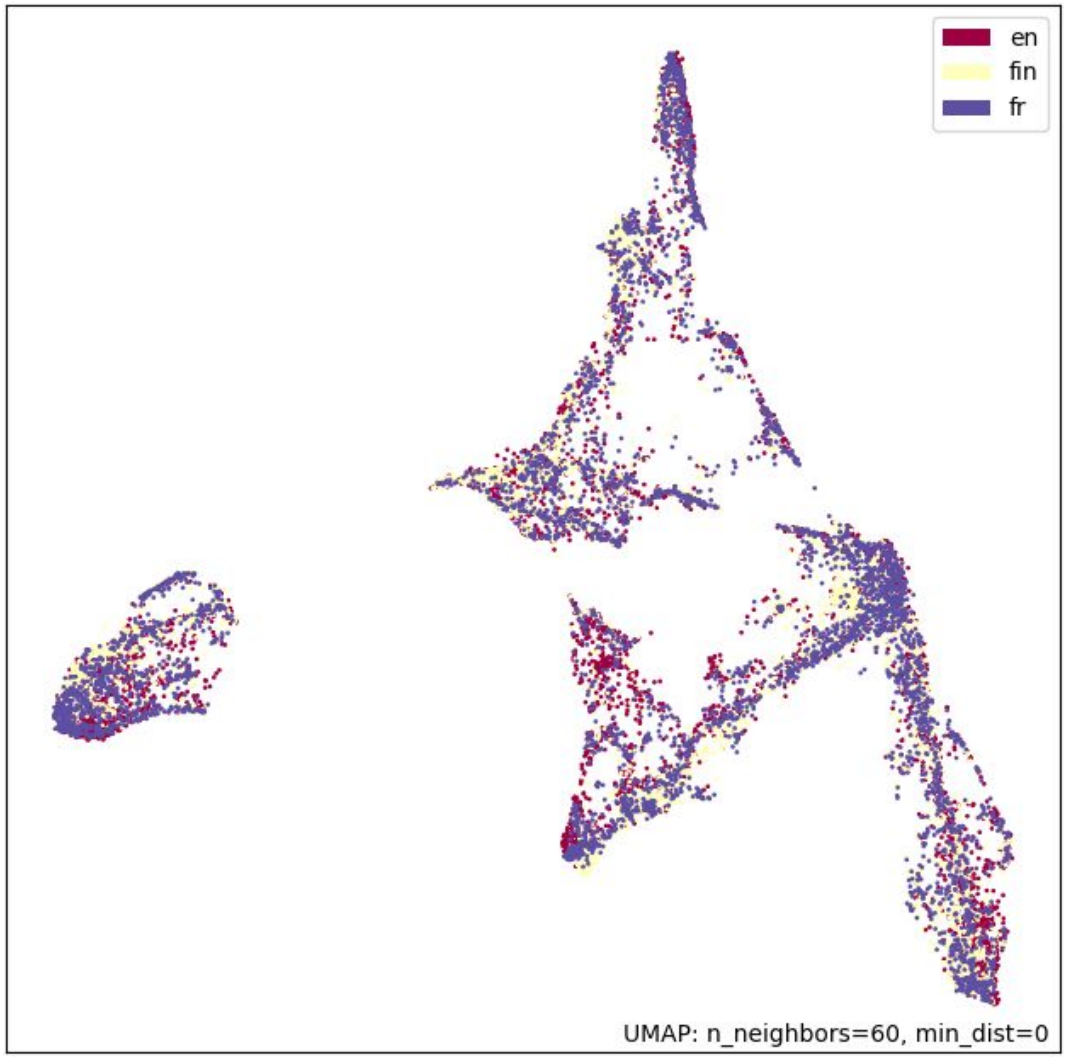
- Learned representations **do not display structuring** with respect to language/register/a combination of languages and registers

Option 2:

- Learned representations display **language-specific and register-specific** structuring

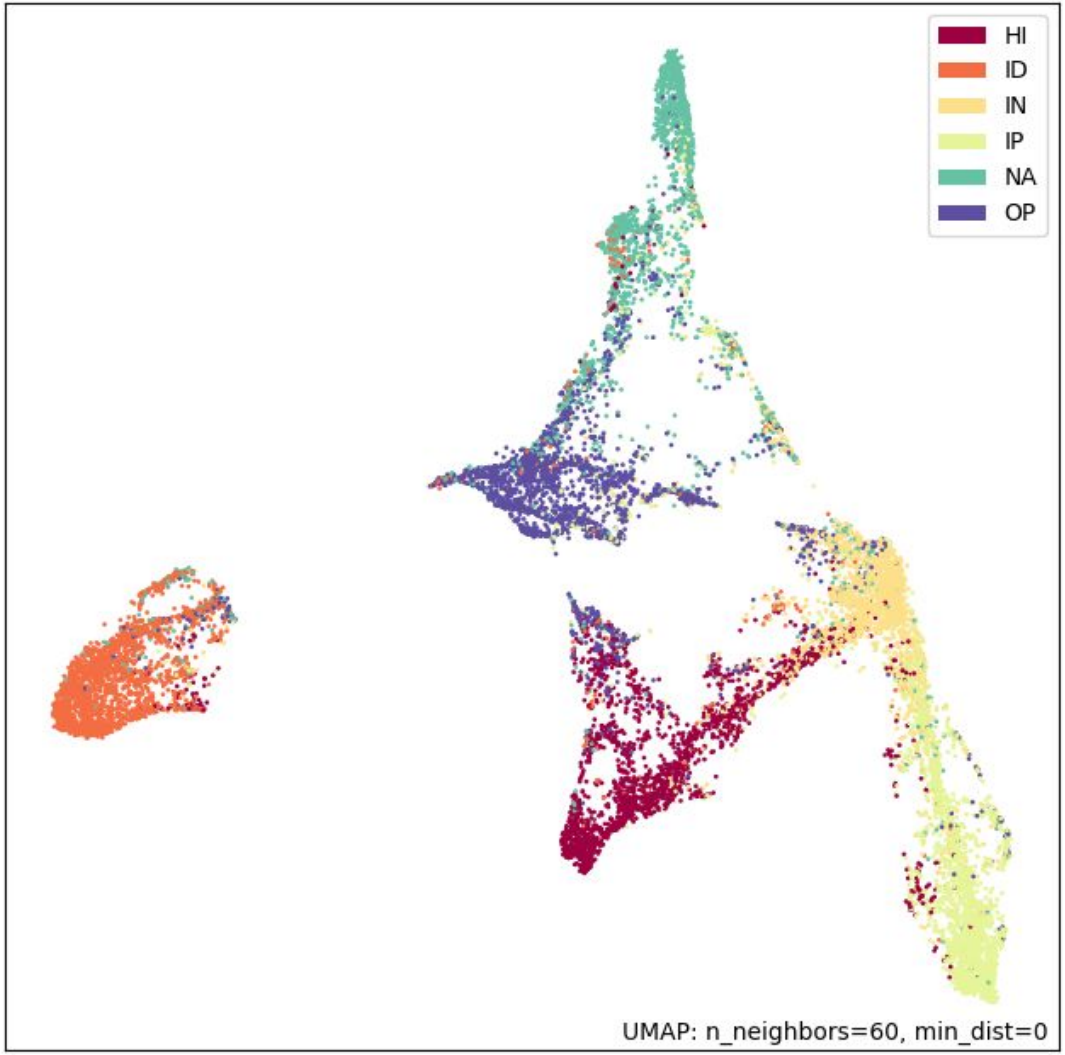
Option 3:

- Learned representations display **register-specific** structuring across languages



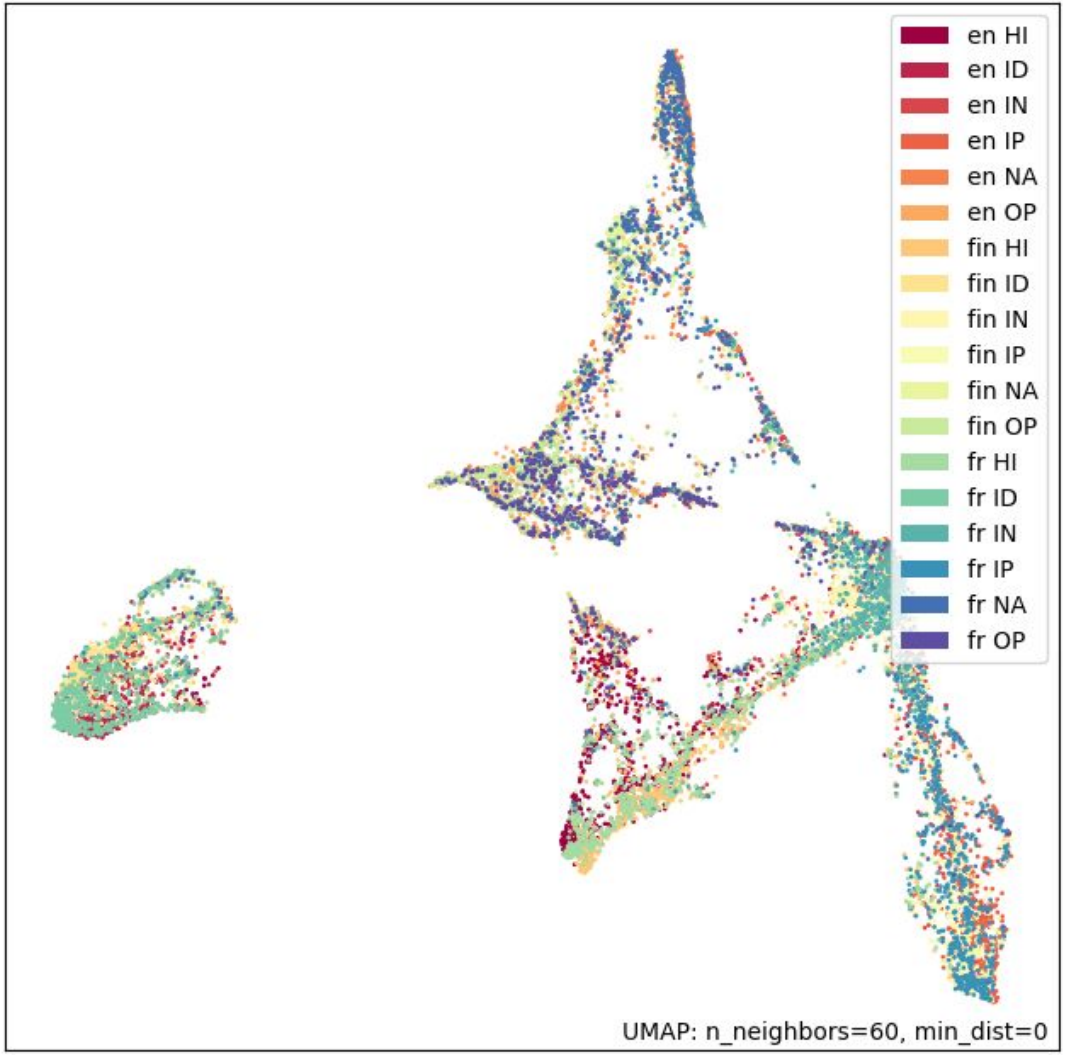
TURKUNLP .ORG

UNIVERSITY OF TURKU



TURKUNLP .ORG

UNIVERSITY OF TURKU

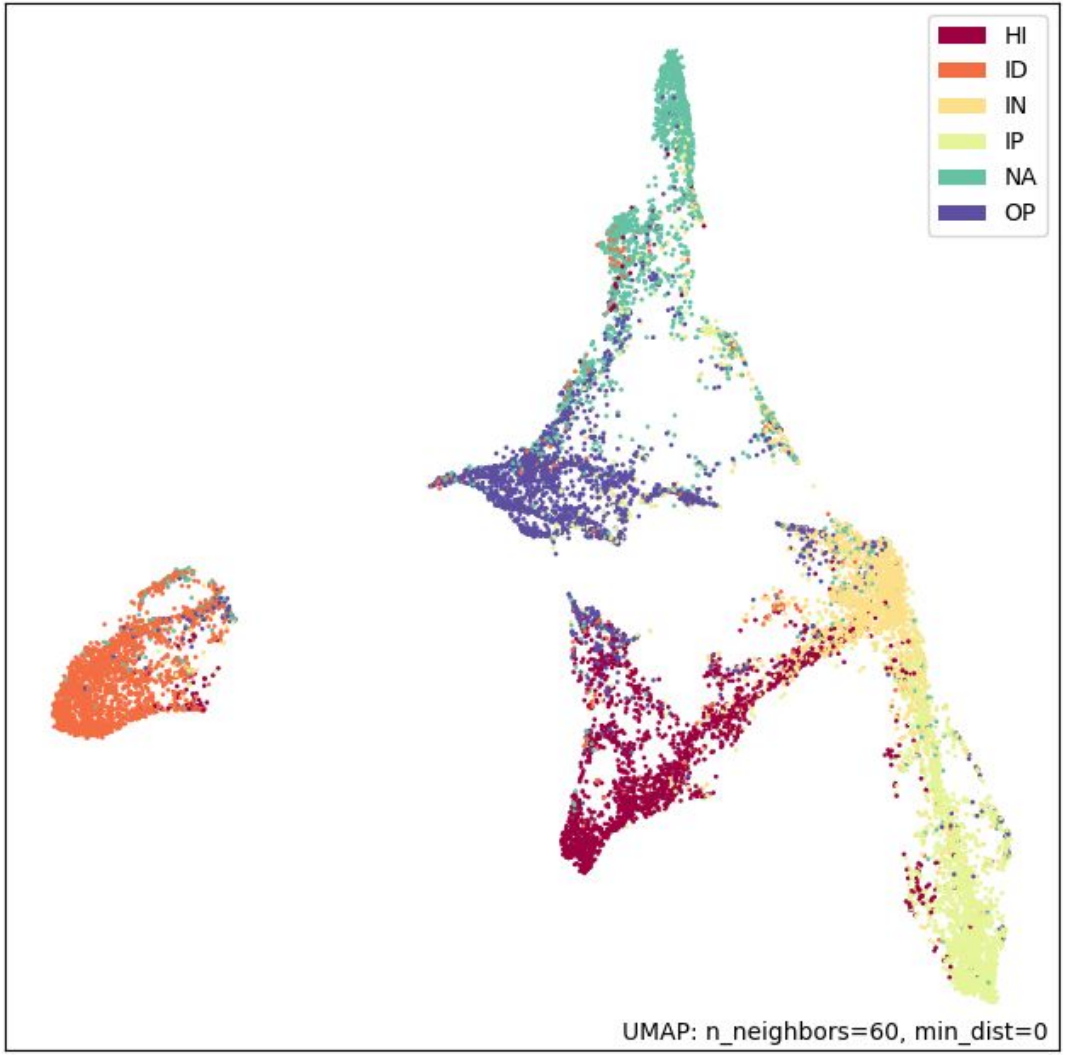


Evaluation

Cluster document embeddings using Kmeans

- 3 clusters corresponding to 3 languages
- 6 clusters corresponding to 6 registers
- 18 clusters corresponding to a combination of languages and registers

Comparison of the cluster solution to the ground truth with Adjusted Rand Index



3 clusters = ARand 0.0

6 clusters = ARand 0.58

18 clusters = ARand 0.23

TURKUNLP .ORG

UNIVERSITY OF TURKU

TURKUNLP .ORG

Keywords derived from the model!

UNIVERSITY OF TURKU

Keywords derived from the model

Common tool for analysing corpus characteristics (Scott 1997, Egbert & Biber 2019)

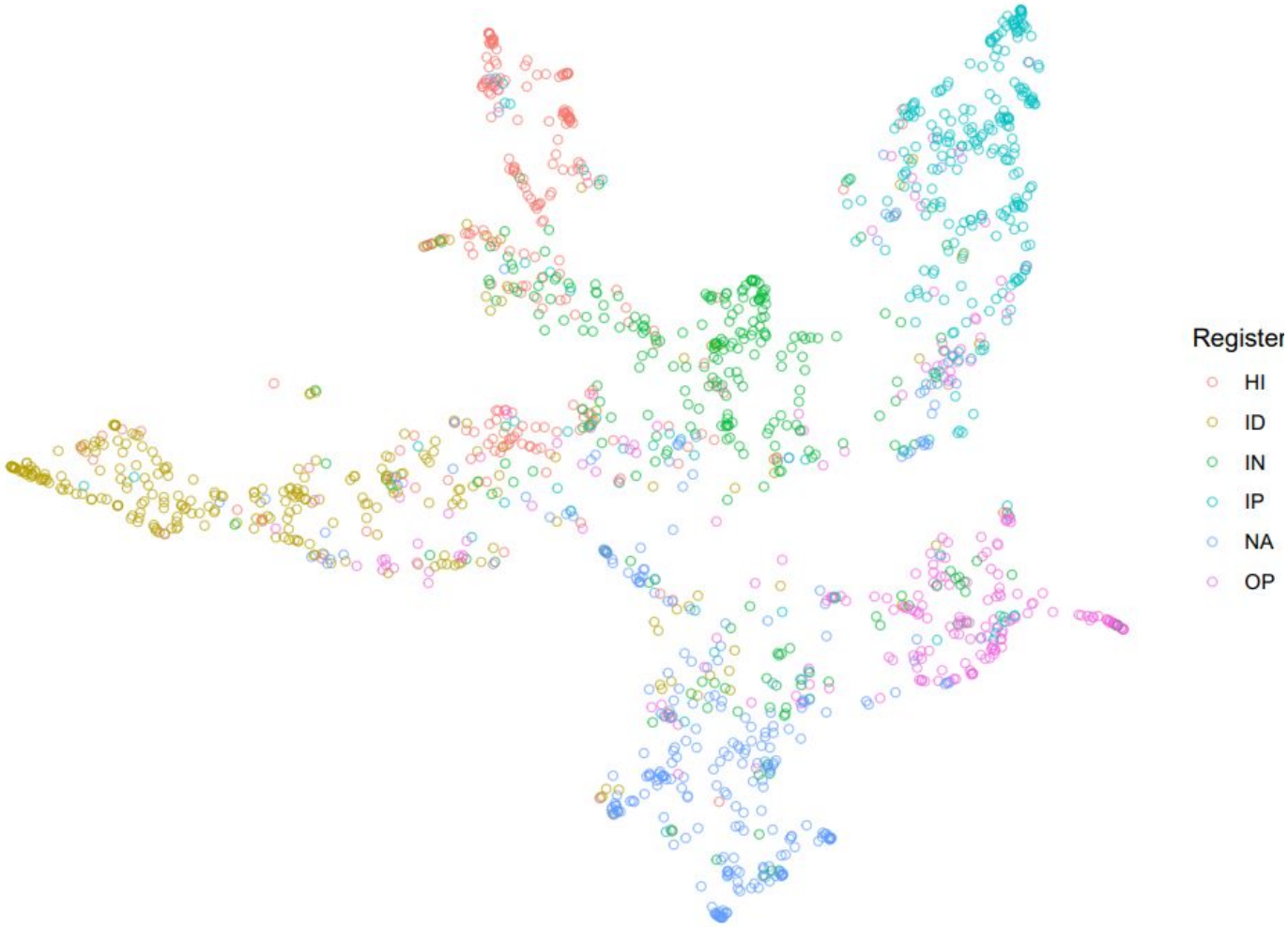
- Typically frequency-based, but can also be predicted (Kyröläinen & Laippala 2022)

Our solution

- Stable Attribution Class Explanation method (SACX) (Rönnqvist et al., 2022)
- Based on Integrated Gradients (IG) (Sundararajan et al., 2017)
  - Identifies important features from each document
- SACX combines important words identified in individual documents to class-level keywords

Results

UNIVERSITY OF TURKU



Verbs of communication

Narrative				
English	Frech		Finnish	
announced	communiquées	‘communicated’	kommentoi	‘commented’
reported	annoncée	‘announced’	sanoo	‘says’
commented	explique	‘explains’	uutisoitiin	‘was reported in news’
replied	souligne	‘underline’	kertoi	‘told’
blogged	révèle	‘reveal’	julkisti	‘published’

Directions for making something

How to/Instructional (HI)				
English	French		Finnish	
recipe	recette	‘recipe’	sekoita	‘mix’
tutorial	tuto	‘tutorial’	käyttöohje	‘instructions for use’ (sng)
tutorials	tutoriel	‘tutorial’	reseptillä	‘with a recipe’
recipes	recettes	‘recettes’	viimeistään	‘at the latest’
tips	devez	‘devez’	käyttöohjeet	‘instructions’ (pl)
guide	conseils	‘conseils	asennusohje	‘installation instructions’
threaded	recommandé	‘recommended’	turvallisuusohjeet	‘security instructions’ (pl)
remove	préparer	‘prepare’	turvaohjeet	‘security instructions’ (pl)
steps	guide	‘guide’	ohjeet	‘instructions’ (pl)
accordance	faudra	‘have to’ (fut)	reseptiä	‘of a recipe’

References to writings

Opinion				
English	French		Finnish	
article	article	‘article’	kirjoituksessani	‘in-my-article’
blog	blog	‘blog’	blogissani	‘in-my-blog’
recommend	recommanderais	‘I-would-recommend’	suosittelen	‘I-recommend’
criticized	regrettons	‘we-regret’	ajattelin	‘I-thought’
impressed	satisfaite	‘pleased’	tyytyväinen	‘pleased’
pleasant	géniale	‘great’	loistavalla	‘with-a-great’
disappointed	inutile	‘useless’	absurdi	‘absurd’
complaint	apprécié	‘appreciated’	ironista	‘ironic’
greatful	suffisant	‘sufficient’	hienot	‘nice’

(Stance) verbs

Opinion				
English	French		Finnish	
article	article	‘article’	kirjoituksessani	‘in-my-article’
blog	blog	‘blog’	blogissani	‘in-my-blog’
recommend	recommanderais	‘I-would-recommend’	suosittelen	‘I-recommend’
criticized	regrettons	‘we-regret’	ajattelin	‘I-thought’
impressed	satisfaite	‘pleased’	tyytyväinen	‘pleased’
pleasant	géniale	‘great’	loistavalla	‘with-a-great’
disappointed	inutile	‘useless’	absurdi	‘absurd’
complaint	apprécié	‘appreciated’	ironista	‘ironic’
greatful	suffisant	‘sufficient’	hienot	‘nice’

Evaluation

Opinion				
English	French		Finnish	
article	article	‘article’	kirjoituksessani	‘in-my-article’
blog	blog	‘blog’	blogissani	‘in-my-blog’
recommend	recommanderais	‘I-would-recommend’	suosittelen	‘I-recommend’
criticized	regrettons	‘we-regret’	ajattelin	‘I-thought’
impressed	satisfaite	‘pleased’	tyytyväinen	‘pleased’
pleasant	géniale	‘great’	loistavalla	‘with-a-great’
disappointed	inutile	‘useless’	absurdi	‘absurd’
complaint	apprécié	‘appreciated’	ironista	‘ironic’
greatful	suffisant	‘sufficient’	hienot	‘nice-pl’

Part-of-Speech?
Tense?

UNIVERSITY
OF TURKU

Part-Of-Speech

Register	Adj	Noun	Verb	Other
HI	10	150	121	19
ID	11	172	60	44
IN	40	198	46	13
IP	68	172	44	13
NA	3	221	68	7
OP	48	209	34	8
V = .22 95% CI[.20, .26]				

Part-Of-Speech

Register	Adj	Noun	Other	Verb
HI	-3.69	-2.85	0.35	7.33
ID	-3.35	-0.67	6.65	-0.02
IN	1.81	0.78	-1.05	-2.06
IP	6.92	-1.12	-1.05	-2.32
NA	-4.95	2.36	-2.51	0.67
OP	3.23	1.49	-2.27	-3.62

Tense

Register	Other	Past	Present
HI	61	11	49
ID	17	10	33
IN	14	17	15
IP	18	15	11
NA	9	33	26
OP	9	12	13
V = .28			

Tense

Register	Other	Past	Present
HI	3.02	-3.69	0.19
ID	-0.79	-1.45	1.92
IN	-0.45	1.41	-0.73
IP	0.75	1.01	-1.52
NA	-2.97	3.58	-0.15
OP	-0.78	1.03	-0.11

TURKUNLP .ORG

How-to

Set whether or not students have access to the calculator. When checked (default), students will be able to access the calculator. To restrict access to the calculator, uncheck the box. This setting governs access to the calculator in pathway lessons only.

Avant de commencer votre paie pour la première fois, vous devez ajouter vos employés. Pour démarrer, suivez les étapes ci-dessous : Se connecter à votre compte. Si c'est la première fois vous avez accédé à votre compte, [...]

infinitive

UNIVERSITY OF TURKU

Register always matters.

Doug Biber 2013

UNIVERSITY OF TURKU

Register always matters, even across languages!

UNIVERSITY OF TURKU

TURKUNLP .ORG

Join SIGWAC!

ACL Special Interest Group in Web-as-Corpus

Officers

- Nikola Ljubešić (co-president)
- Benoît Sagot (co-president)
- Veronika Laippala (co-secretary)
- Pedro Ortiz Suarez (co-secretary)

SIGWAC

https://www.sigwac.org.uk/

UNIVERSITY OF TURKU