Exploring register universals on the unrestricted web

Saara Hellström, Liina Repo, Valtteri Skantsi & Veronika Laippala

University of Turku & University of Oulu

TURKUNLP .ORG

UNIVERSITY OF TURKU

Our study

- Language use on the English, Finnish, French and Swedish web
- Focus on **registers**
  - situationally defined text varieties such as news, reviews, columns and advice (Biber, 1988)
- **Register universals**
  - similar patterns of variation across languages (Biber, 2014)

UNIVERSITY OF TURKU

Research in register universals

- Variation in language use across languages
  - different linguistic characteristics
  - different situations of language use
- Register studies *within* one language, e.g.,
  - English (e.g., Biber, 1988; Biber & Egbert, 2016, 2018)
  - Spanish (Parodi, 2007)
  - Czech (Cvrček et al., 2020)
- Register universals give a wider perspective on language use and its variation
  - Oral vs written language
  - Narrative vs non-narrative language

Material

Register classes

How-to/instructions
E.g., recipes

Interactive discussion
E.g., discussion forums

Informational description
E.g., encyclopedia articles

Informational persuasion
E.g., promoting texts

Narrative
E.g., news, personal blogs

Opinion
E.g., advice, opinion blogs

Lyrical
E.g., song lyrics, poems

Spoken
E.g., formal speeches

- **Four** corpora representing the **unrestricted** web
  - Corpus of Online Registers of English (**CORE**; Biber, Egbert & Davies, 2015) (40,944 documents)
  - **FinCORE** (8,373 documents) (Laippala et al., 2019)
  - **FreCORE** (1,446 documents) (Repo et al., 2021)
  - **SweCORE** (1,676 documents) (Repo et al., 2021)

  → manually annotated following the hierarchical annotation scheme for CORE

  → similar material and classification

  → direct comparison possible
- We focus on the **narrative** register

UNIVERSITY OF TURKU

Methods

- Supervised machine learning (text classification) & linguistic analysis of the most important features estimated by the classifier for the registers
- For each language, we

  1) **trained a linear classifier** to predict the register classes based on the grammatical features of the documents,

  2) extracted **the most important features** estimated by the classifier for the registers, and

  3) **compared** these features between the languages
- About the differences between the classification and MDA, see Egbert & Biber (2018)

UNIVERSITY OF TURKU

Classifier

Classification results			
EN	FI	FR	SW
52%	57%	55%	64%

- **Support Vector Machine** used as classifier *
- The classifier was run **100 times**
  - between every run the data was randomly divided into train and test
  - after every run the best 500 features were saved
  - **30 most frequent features** in the final analysis
  - represents the entire data
- The **syntactic features** of the documents used as predictors of the classifier
  - grammatical information was produced with state-of-the-art **dependency parsers** (Kanerva et al., 2018) which follow the language independent Universal Dependency scheme (Nivre et al. 2016)

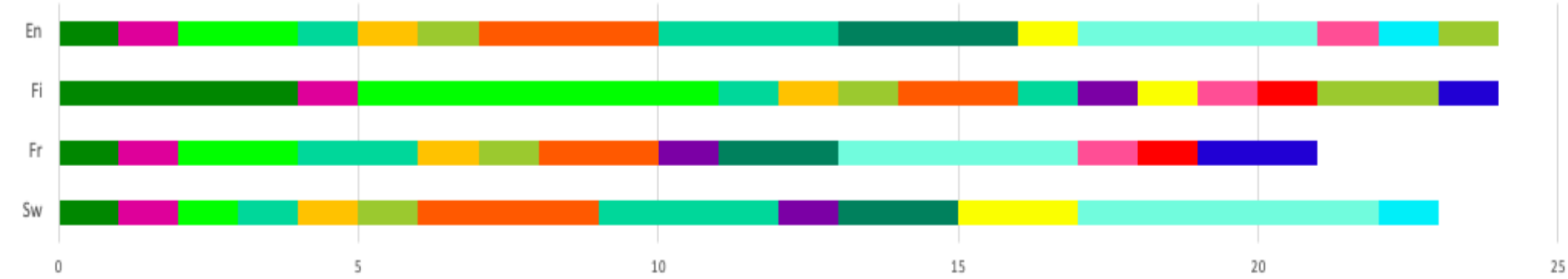
*Scikit-learn LinearSVC; l2 penalty; EN c=0.01; FR: c=0.1; FI c=0.01; SV: c=0.01, max_iter=10000, countvectorizer, min_df=0.05

UNIVERSITY OF TURKU

Analysis

UNIVERSITY OF TURKU

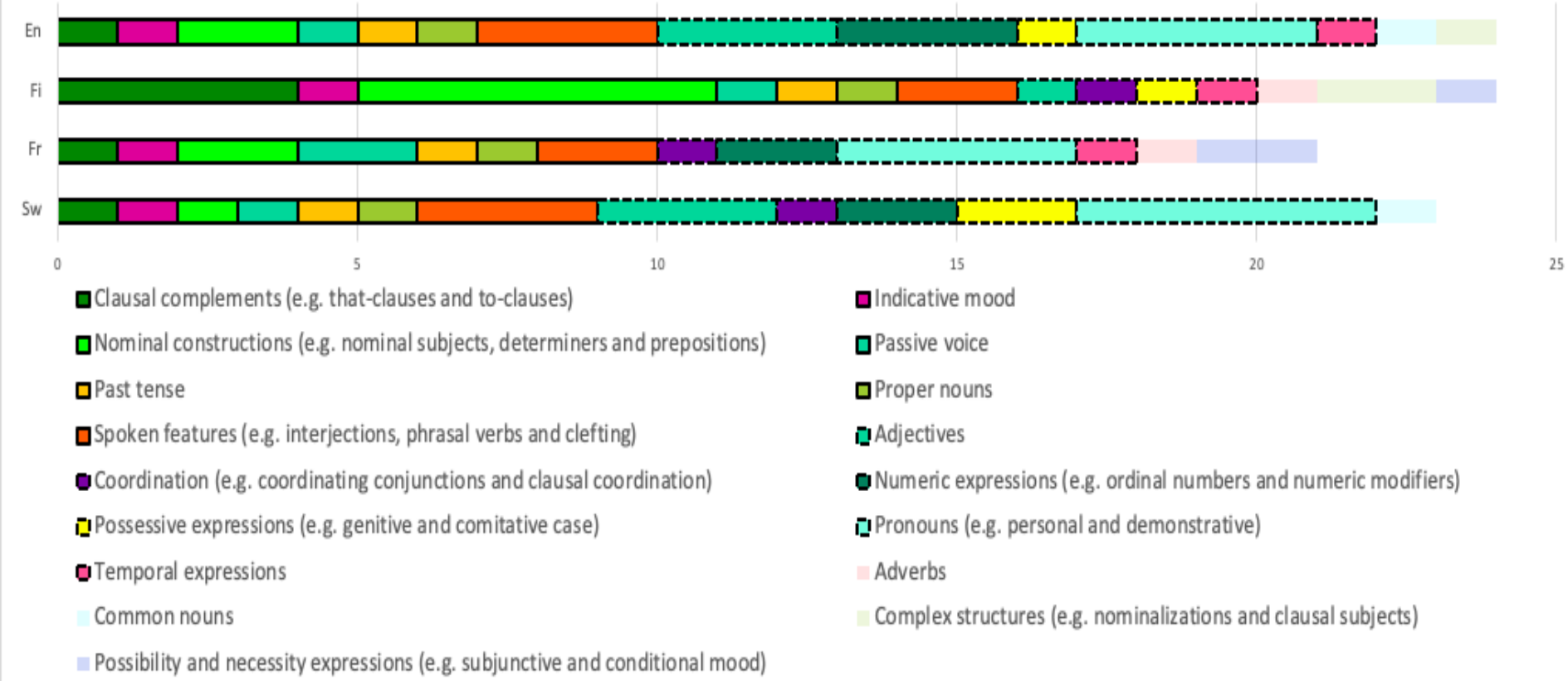
The most important syntactic features estimated for *Narrative*



- Clausal complements (e.g. that-clauses and to-clauses)
- Indicative mood
- Nominal constructions (e.g. nominal subjects, determiners and prepositions)
- Passive voice
- Past tense
- Proper nouns
- Spoken features (e.g. interjections, phrasal verbs and clefting)
- Adjectives
- Coordination (e.g. coordinating conjunctions and clausal coordination)
- Numeric expressions (e.g. ordinal numbers and numeric modifiers)
- Possessive expressions (e.g. genitive and comitative case)
- Pronouns (e.g. personal and demonstrative)
- Temporal expressions
- Adverbs
- Common nouns
- Complex structures (e.g. nominalizations and clausal subjects)
- Possibility and necessity expressions (e.g. subjunctive and conditional mood)

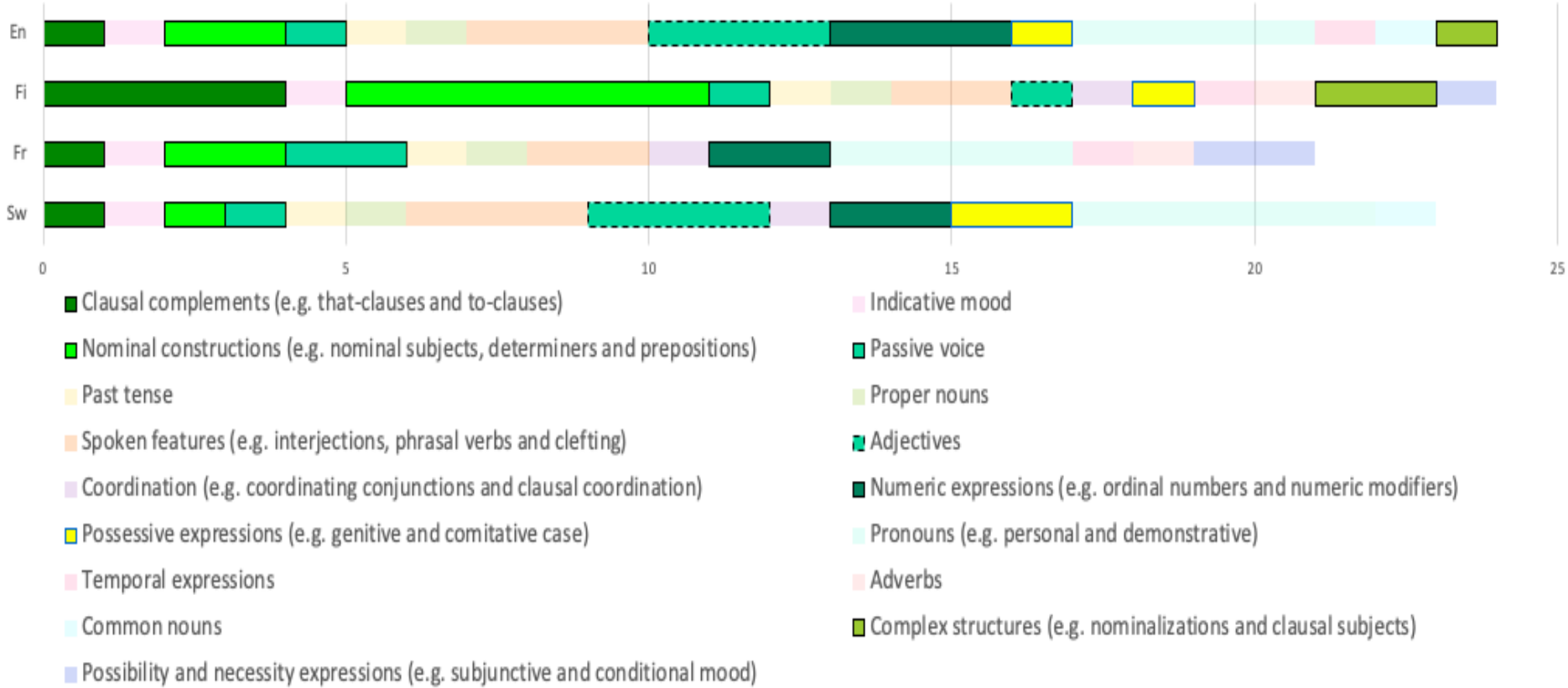
UNIVERSITY OF TURKU

The most important syntactic features estimated for *Narrative*



UNIVERSITY OF TURKU

The most important syntactic features estimated for *Narrative* - Informational discourse



Besnier (1988)
Biber & Egbert (2016)
Biber & Egbert (2018)
Kanoksilapatham (2007)

UNIVERSITY OF TURKU

Examples of informational features

English:
Gatland, who **was unveiled** as the **Lions'** boss earlier this week, has hinted that **outstanding** club form could be enough to book a place on the plane Down Under with the likes of **former** England internationals Jonny Wilkinson and Mike Tindall on **his** radar.

Finnish:
Oulussa **on opittu**, että **paras** hetki neuvontaan on heti **rakennushankkeen** alussa, kun vasta **mietitään** valintoja.

[It has been learned in Oulu that the best time for counseling is right at the beginning of a construction project, when choices are just being considered.]

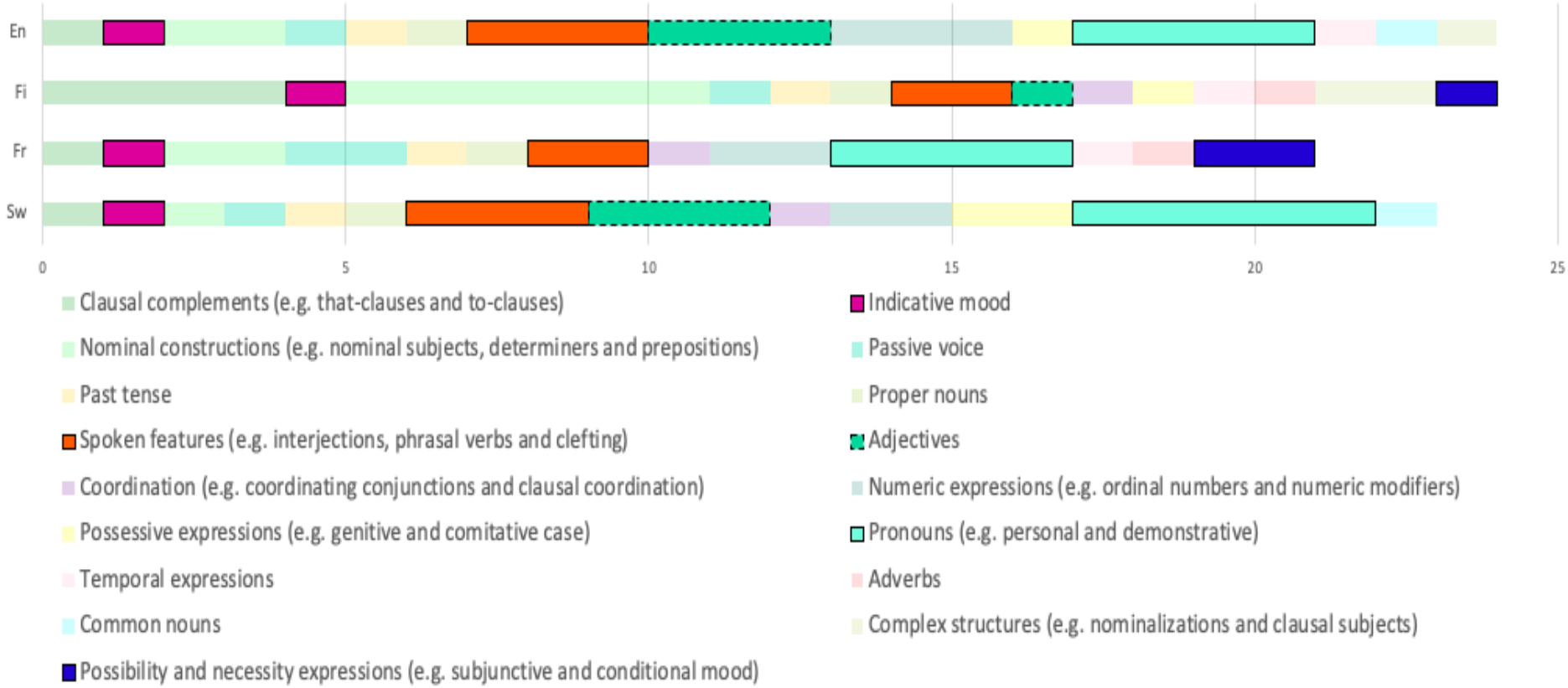
French:
De même, aucune date d'interdiction de ce métal dans les amalgames **dentaires** n'**a été précisée**, sous la pression du lobby des dentistes qui ont cependant admis que **son** utilisation doit diminuer.

[Similarly, no date for the ban of this metal in dental amalgams has been specified under the pressure from the lobby of dentists who have admitted, however, that its use must decrease.]

clausal complement, nominal constructions, **passive voice**, **adjective**, **possessive expression**

UNIVERSITY OF TURKU

The most important syntactic features estimated for *Narrative* - Oral-irrealis discourse



Biber & Egbert (2016)
Parodi (2007)

UNIVERSITY OF TURKU

Examples of oral + irrealis features

Finnish:

Olen sitä mieltä, että ihmiset **voisivat** peremmin [sic], jos **tekisivät** enemmän **meditatiivisia** asioita ja ylipäätään asioita käsillään: golfissa **on molemmat** aspektit mukana!

[In my opinion, people would feel better if they did more meditative things and generally things with their hands: golf includes both aspects!]

French:

Et ce qui **me surprend** le plus... **c'est** la vitesse a [sic] **laquelle on consomme**!!

[What surprises me the most… is the speed at which it is consumed!!]

Swedish:

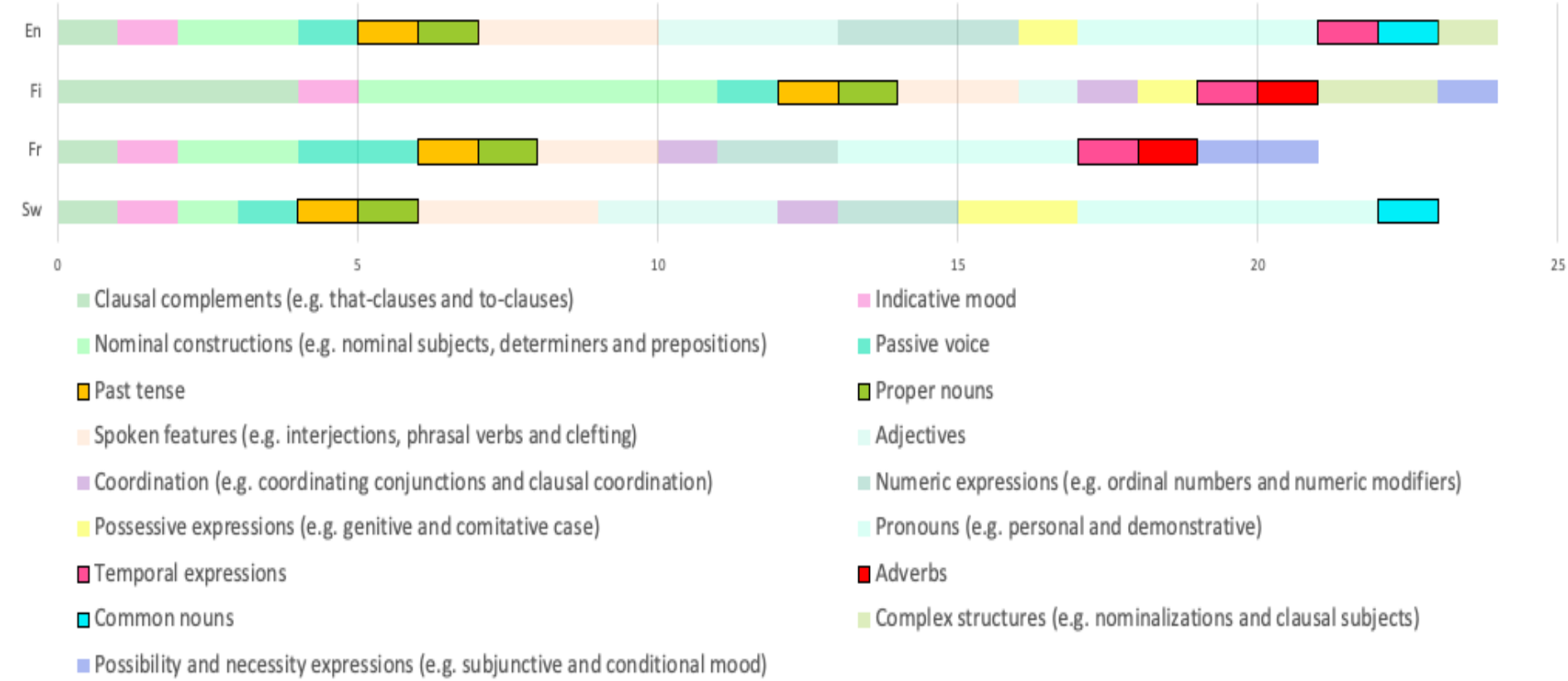
Ja grattis till utmärkelsen! Det **var** så **jag hittade** hit för någon vecka sedan.

[Yes congratulations for the award! That's how I found this site for a couple of weeks ago.]

pronoun, spoken features (**interjection**, dislocation, cleft, parataxis), **adjective**, possibility and necessity expressions (**conditional mood**), **indicative mood**

UNIVERSITY OF TURKU

The most important syntactic features estimated for *Narrative* - Narrative discourse



Biber & Egbert (2016)

UNIVERSITY OF TURKU

Examples of narrative features

Finnish:

Australiassa ollessa pyysin **pojille myös Suomen rokotusohjelmaan** kuuluvat **suojarokotteet**.

[While in Australia I asked for the boys also the vaccines that are part of Finland's vaccine program.]

Swedish:

Men ofta saknar de både **pass** och **id-kort** och sedan några **år** har inga **avvisningar** gjorts eftersom **Marocko** har valt att inte ta **emot papperslösa**.

[But they often lack both passports and ID cards and during the past couple of years no one has been turned back since Morocco has decided not to receive any paperless migrants.]

proper nouns, temporal expressions, past tense, **common nouns**, **adverbs**

UNIVERSITY OF TURKU

The most frequent adjectives

English	Finnish	Swedish
other	toinen	annan
new	uusi	ny
good	hyvä	bra
old	vanha	gammal
own	oma	egen
little	pieni	liten
first	ensimmäinen	
	suuri, iso	stor
many		många
	koko	hel
last	viime	
long	pitkä	
great		fin
well	sellainen	själv
same	eri	en
best	tärkeä	svensk
more		olik
		sen

UNIVERSITY OF TURKU

Conclusions

- Informational, oral-irrealis and narrative features are prominent in narrative web registers in the studied languages
- Pointers to new register universals?
  - Irrealis and informational
  - Adjectives are very similar between languages

What next?

- Register universals in sub-registers

  → more homogenous classes
- Register universals in multilingual neural net models
  - What kind of linguistic features the models pay attention to?
  - How multilingual models differ from monolingual models qualitatively? What does this tell about registers?

UNIVERSITY OF TURKU

References

Besnier, N. (1988). The Linguistic Relationships of Spoken and Written Nukulaelae Registers. *Language*, 64:707–736.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Biber, D. (2014). Using Multi-dimensional Analysis to Explore Cross-linguistic Universals of Register Variation. *Languages in Contrast, 14*(1), 7–34.

Biber, D., & Egbert, J. (2016). Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics, 44*(2), 95–137.

Biber, D., & Egbert J. (2018). *Register variation online*. Cambridge University Press.

Biber, D., Egbert, J., & Davies, M. (2015). Exploring the composition of the searchable web: A corpus-based taxonomy of web registers. *Corpora,, 10*(1), 11–45.

Cvrček, V., Komrskova, Z., Lukes, D., Poukarová, P., Řehořková, A, Zasina, A., & Benko, V. (2020). Comparing web-crawled and traditional corpora. *Language Resources and Evaluation*, *54*(10), 713–745. https://doi.org/10.1007/s10579-020-09487-4.

Egbert, J., & Biber, D. (2018). Do all roads lead to Rome? Modeling register variation with factor analysis and discriminant analysis. *Corpus Linguistics and Linguistic Theory, 14*(2), 233–273.

Kanerva, J., Ginter, F., Miekka, N., Leino, A., & Salakoski.T. (2018). Turku Neural Parser Pipeline: An end-to-end system for the CoNLL 2018 Shared Task. In D. Zeman & J. Hajič (Eds.), *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies* (pp. 133–142). Association for Computational Linguistics. https://doi.org/10.18653/v1/K18-2013.

Kanoksilapatham, B. (2007). Rhetorical Moves in Biochemistry Research Articles. In D. Biber, U. Connor and T. A. Upton (Eds.). *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure* (pp. 73–120). John Benjamins.

Laippala, V., Kyllönen, R., Egbert, J., Biber, D., & Pyysalo, S. (2019). Toward multilingual identification of online registers. In M. Hartmann & B. Plank (Eds.), *Proceedings of the 22nd Nordic Conference on Computational Linguistics* (pp. 292–297). Linköping University Electronic Press.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C .D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R. & Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. https://www.aclweb.org/anthology/L16-1262.pdf

Parodi, G. (2007). Variation across Registers in Spanish. In G. Parodi (Ed), *Working with Spanish Corpora* (pp. 11–53). Continuum.

Repo, L., Skantsi, V., Rönnqvist, S., Hellström, S., Oinonen, M., Salmela, A., Biber, D., Egbert, J., Pyysalo S., & Laippala, V. (2021). Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. *Proceedings of the EACL 2021 Student Research Workshop*, (pp. 183–191). Association for Computational Linguistics.

UNIVERSITY OF TURKU

Thank you!

UNIVERSITY OF TURKU

ACADEMY OF FINLAND

EMIL AALTOSEN SÄÄTIÖ

TURKUNLP .ORG