

# Intersecting Register and Genre: Understanding the Contents of Web-Crawled Corpora

Amanda Myntti, Veronika Laippala, Erik Henriksson,  
Elia Freyermuth \*

University of Turku, \*National Graduate School of Engineering of Caen



**TURKUNLP**  
**.ORG**



**UNIVERSITY  
OF TURKU**

*DHNB 2024 in Reykjavik*

# Introduction



Web-scale corpora present significant opportunities for diverse fields of research. However, lack of metadata often limits their use in humanities and social sciences.

Advances in text classification, specifically with *registers*, allow us to extend the metadata of documents. Register categories remain broad, so we investigate the benefits of using *genre* labelling scheme alongside register classification.

**RQ1:** What kind of intersection do register and genre labels have?

**RQ2:** How could the intersection be of use for other research?

First experiments in English due to data availability.

# Definitions for this presentation



We define both register and genre based on the corpora used in this study

## Register

- Situationally defined text varieties (Biber and Conrad, 2019; Biber and Egbert, 2018)
- focus on linguistic features that are functionally adapted to the communicative purpose at hand
- Interactive Discussion, Narrative, Opinion etc.
- E.g. Laippala et al. 2022, Kuzman et al. 2023\*

Main register	Sub register
How-to or Instruction (HI)	Recipe, Other HI
Interactive discussion (ID)	-
Informational description (IN)	Encyclopedia article, Research article, Description of a thing/person, FAQ, Legal, Other IN
Informational persuasion (IP)	Description with intent to sell, News & opinion blog / editorial, Other IP
Lyrical (LY)	-
Narrative (NA)	News report, Sports report, Narrative blog, Other NA
Opinion (OP)	Review, Opinion blog, Religious blog/sermon, Advice, Other OP
Spoken (SP)	Interview, Other SP

\*different terminology used

# Definitions for this presentation



We define both register and genre based on the corpora used in this study

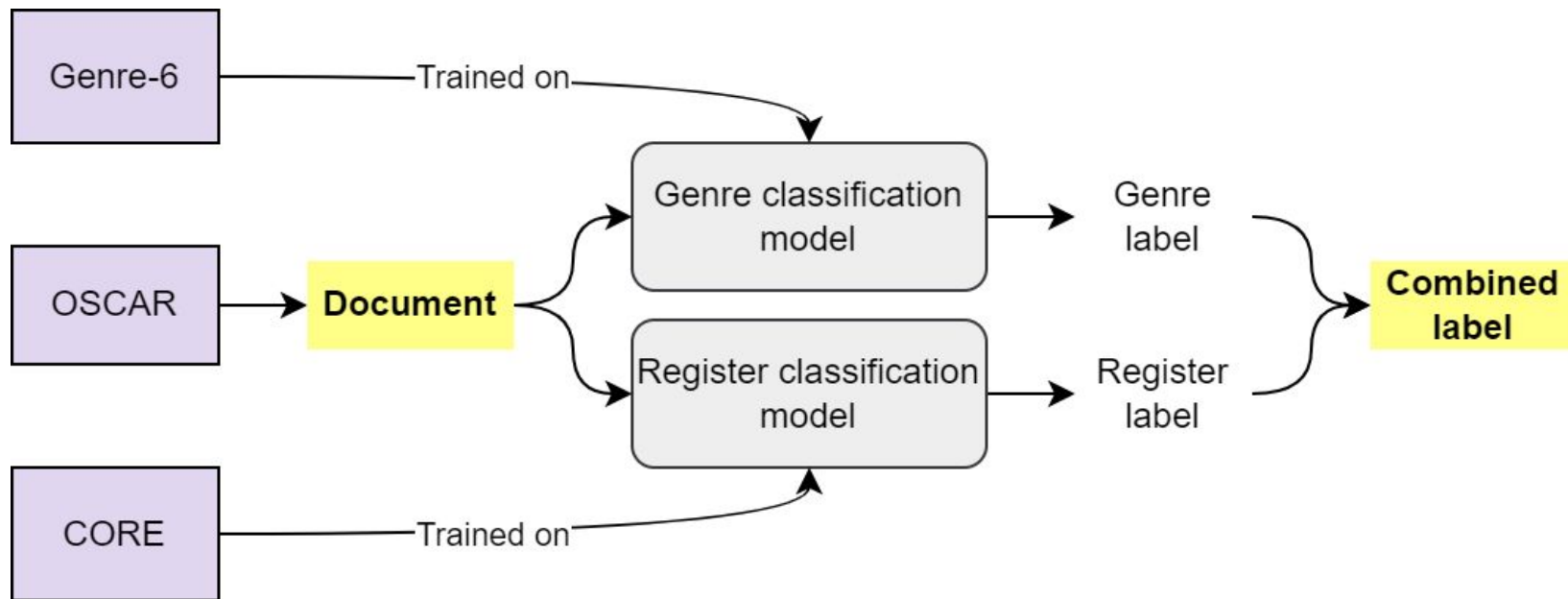
## Genre

- categories to which works of literature have been classified, akin to a library or bookshop categorisation system
- focus on content, context and narrative tools
- Science Fiction; Cookbooks, Food & Wine; etc.
- E.g. Goyal and Prakash 2022, Zhang et al. 2022

Genre*
Literature & Fiction
Engineering & Transportation
Medicine & Health Sciences
Romance
Arts & Photography
Religion
Science & Math
Politics & Social Sciences
Science Fiction

\*examples, not full scheme

# Our methodology



# Our corpora



*Corpus of Online Registers CORE* (Egbert et al. 2016; Laippala et al. 2022, 2023)

- Sample a English web (other languages also available)
- A hierarchical multilabel register annotation scheme
  - eight main classes, including narrative, informational description, and opinion
  - tens of subclasses, including news report, research article, and review
- Human annotated, combined label in event of annotator disagreement

## *Genre-6*

- Dataset from Hugging Face [marianna13/the-eye](https://huggingface.co/marianna13/the-eye)
  - books of different genres
  - internal cleaned version used in training, only excerpts of books used
- Multilabel genre annotation scheme from Kindle US&UK categories
  - labels assigned by authors

# Our corpora



*The OSCAR dataset (Suarez et al. 2019; Laippala et al. 2022)*

- Large scale dataset, consisting of CommonCrawl
- Cleaned and language classified, over 100 languages available
- Register classified (*Laippala et al. 2022*) version from Hugging Face

[TurkuNLP/register\\_oscar](https://turkunlp.github.io/register_oscar)

- 14 languages covered
- Additionally boilerplate text removed
- Chosen for size and quality

→ We do not use labels of this corpus for the study, since they only cover the main level labels.



# Models



*XLM-RoBERTA-Large (Conneau et al. 2020)*

- Two models fine tuned, one for register classification using the CORE corpus and one for genre classification using the Genre-6 corpus
- Chosen for ease, resource consumption and register classification prowess (Repo et al. 2021)
- Training done using the Hugging Face Transformers-library
- Classification threshold optimized for the F1 score



# Experiments



- Selection of chosen genre categories based on two criteria
  - achieved classifier performance
    - related to the large number of overlaps between the genre categories
    - performance suffers with more categories included
  - contents of the documents which are to be predicted

→ We chose to also include “None” category for uncertain classifications

- Experiments with Binary Cross-Entropy Loss and Focal Loss (Lin et al. 2020)
  - Better F1 with focal loss, but results visibly worse in manual evaluation

# Results



- Register classifier is able to reach an F1-score of 0.77\*
- Genre classifier performance at 0.70

Manual evaluation shows that genre label enhances the information given by the register label

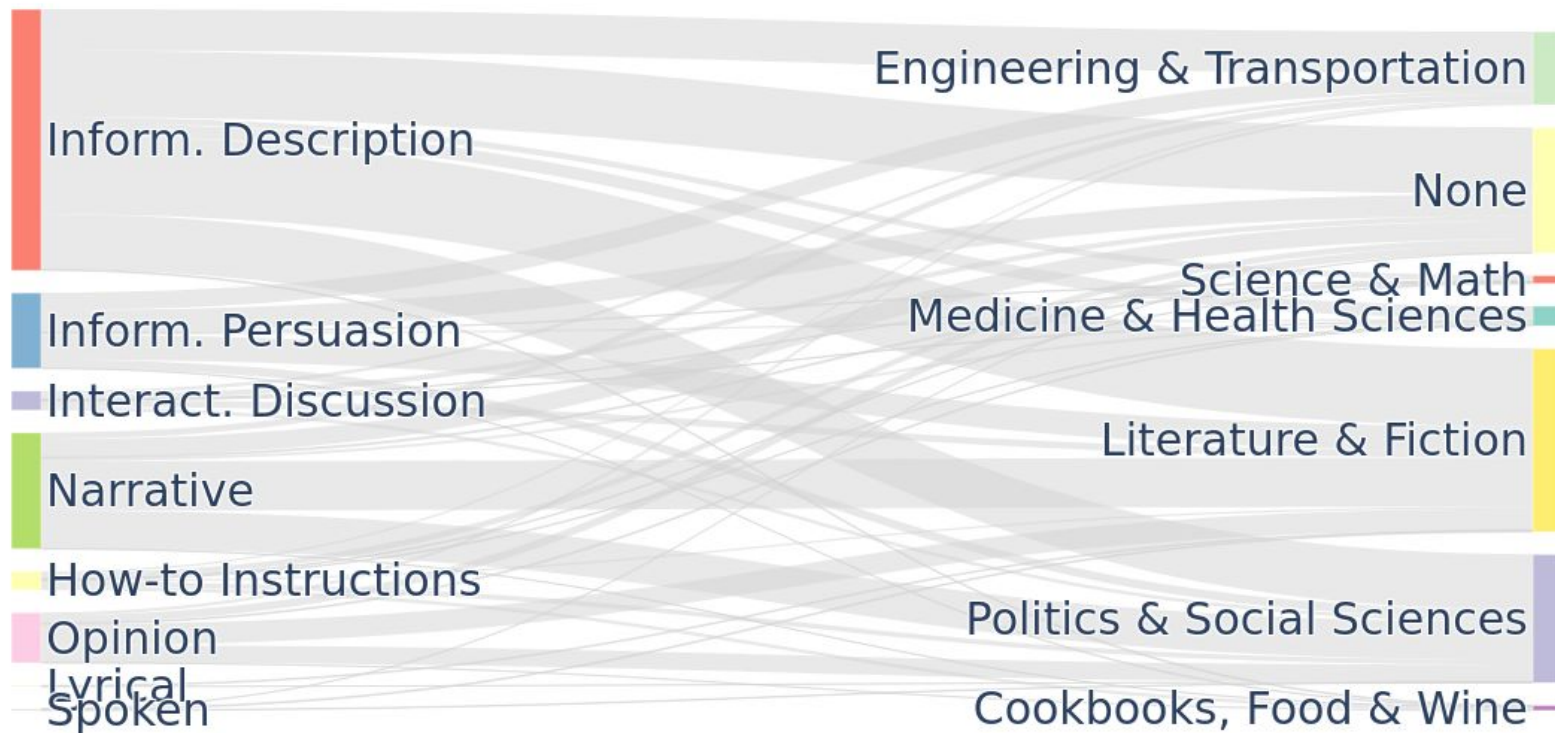
- E.g. Interactive Discussion + Engineering & Transportation = Discussion forum about technology
- Facilitates the use of the corpus in new ways in the study

Following examples use a register model with 0.74 wrt English

# Intersection of Registers and Genres



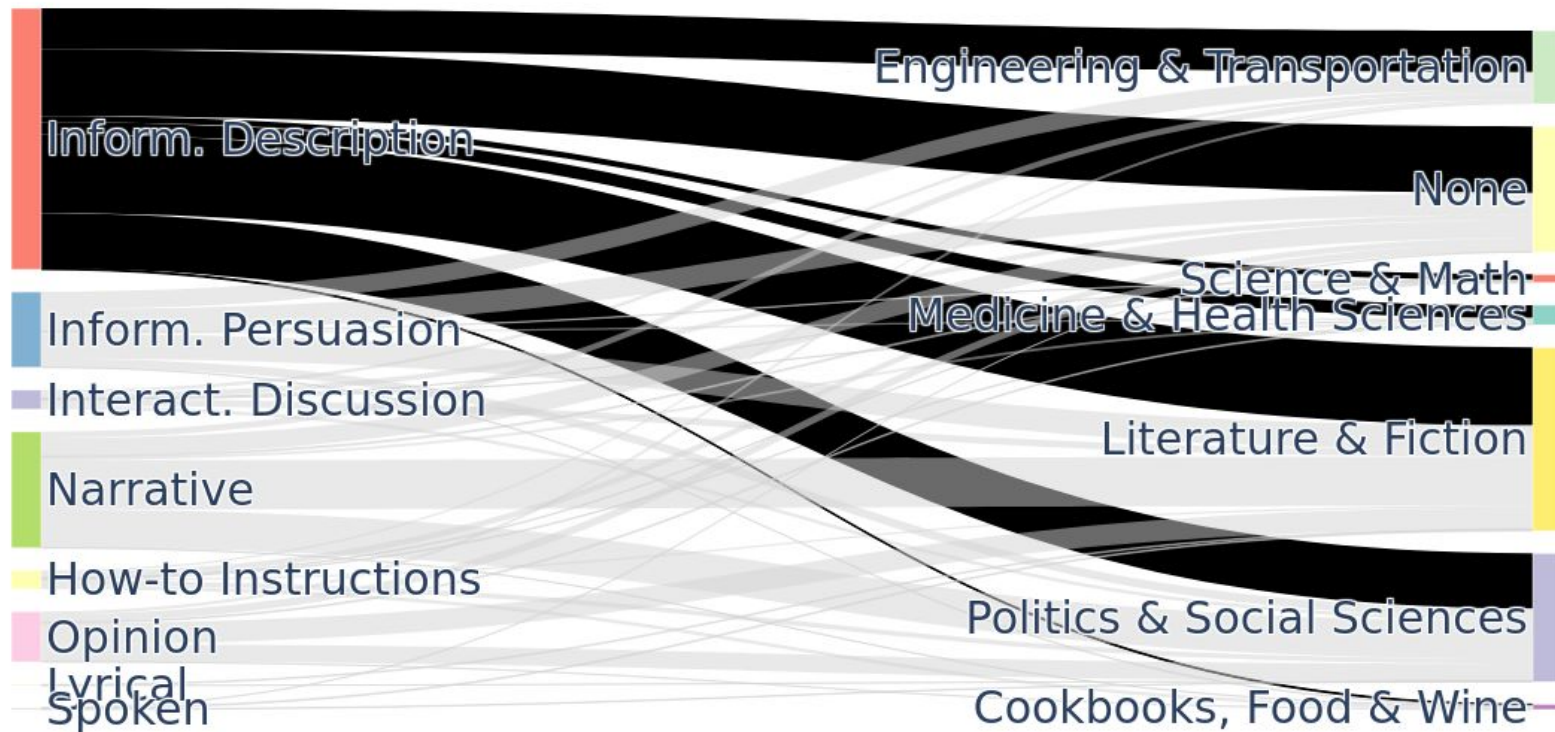
**TURKUNLP**  
.ORG



# Intersection of Registers and Genres



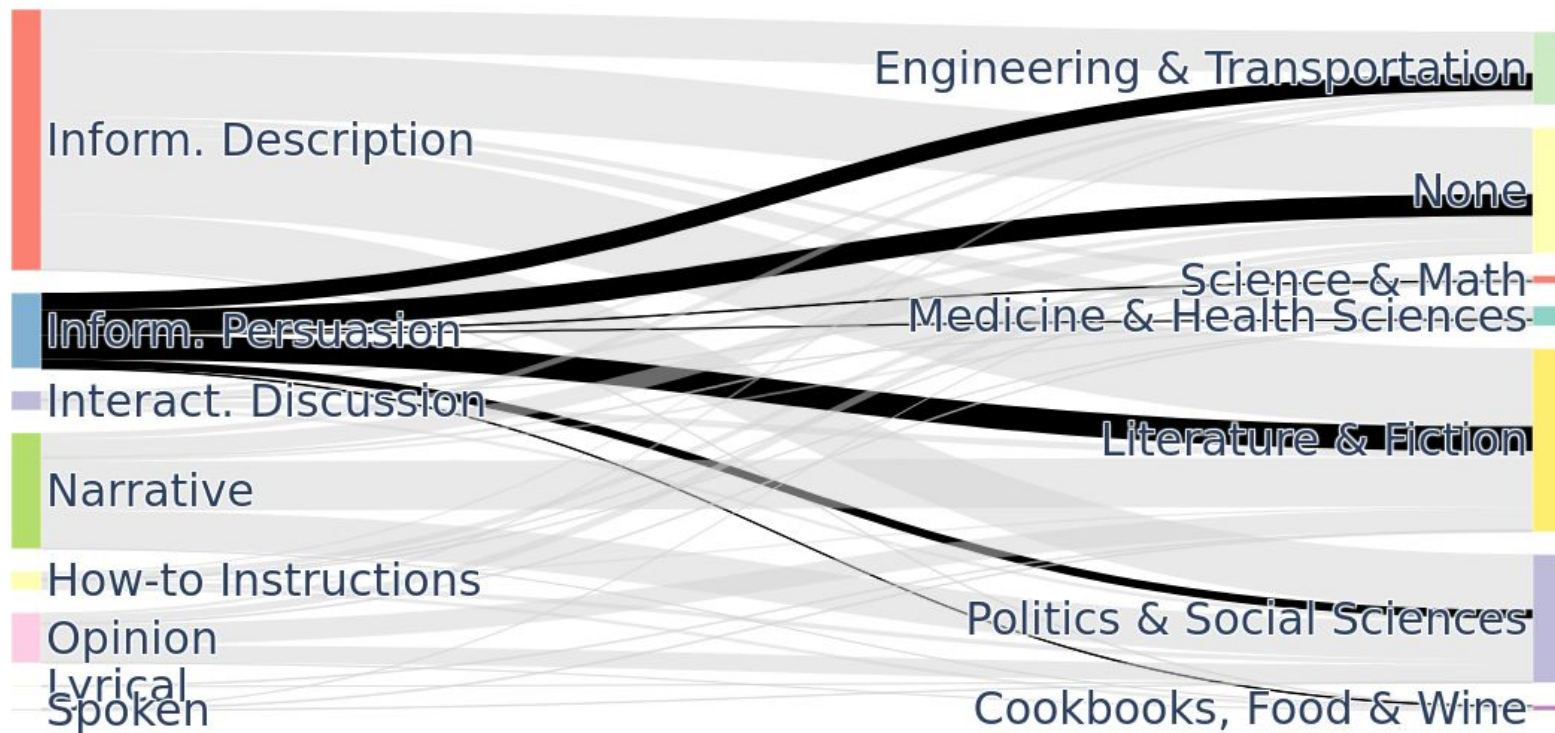
**TURKUNLP**  
.ORG



# Intersection of Registers and Genres



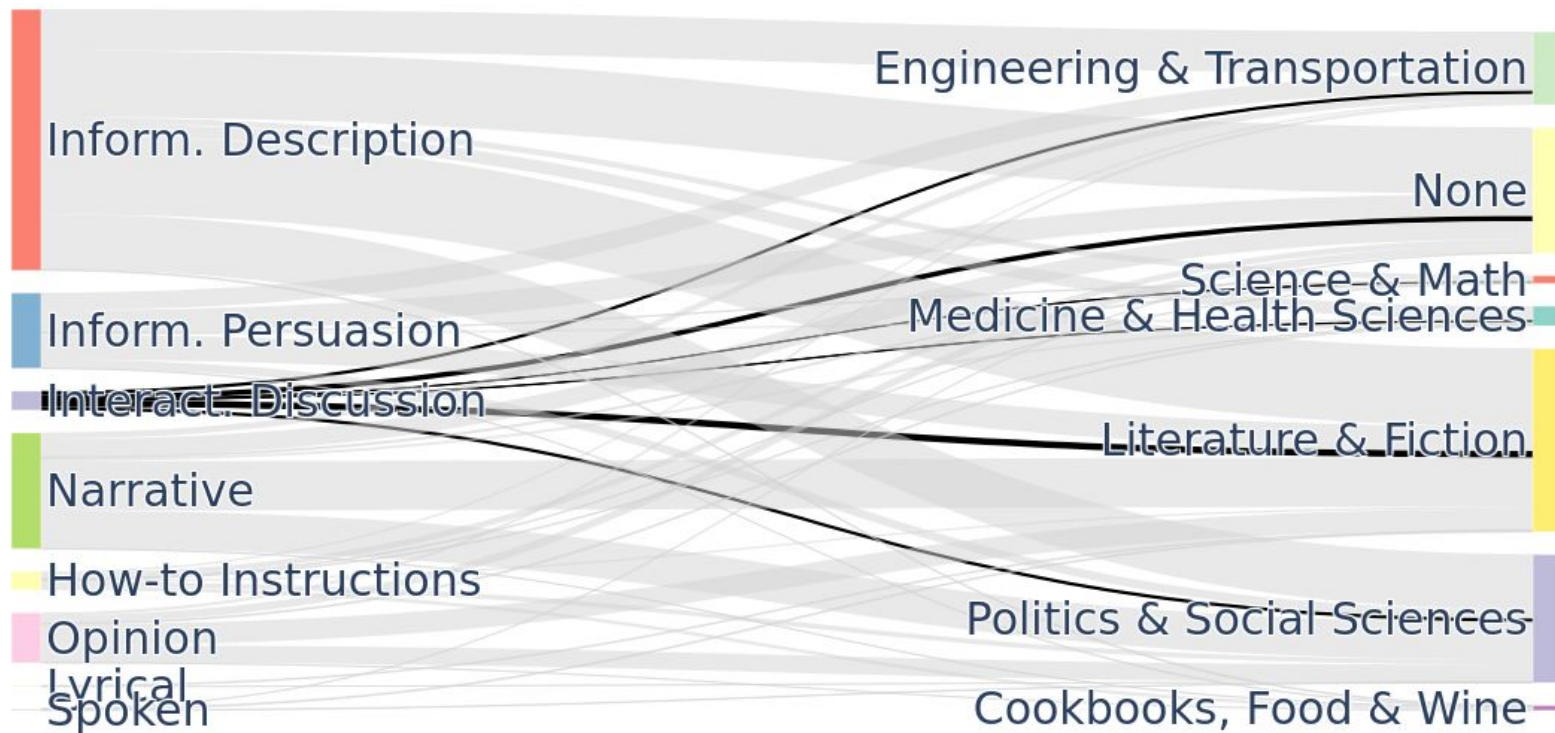
**TURKUNLP**  
.ORG



# Intersection of Registers and Genres



**TURKUNLP**  
.ORG

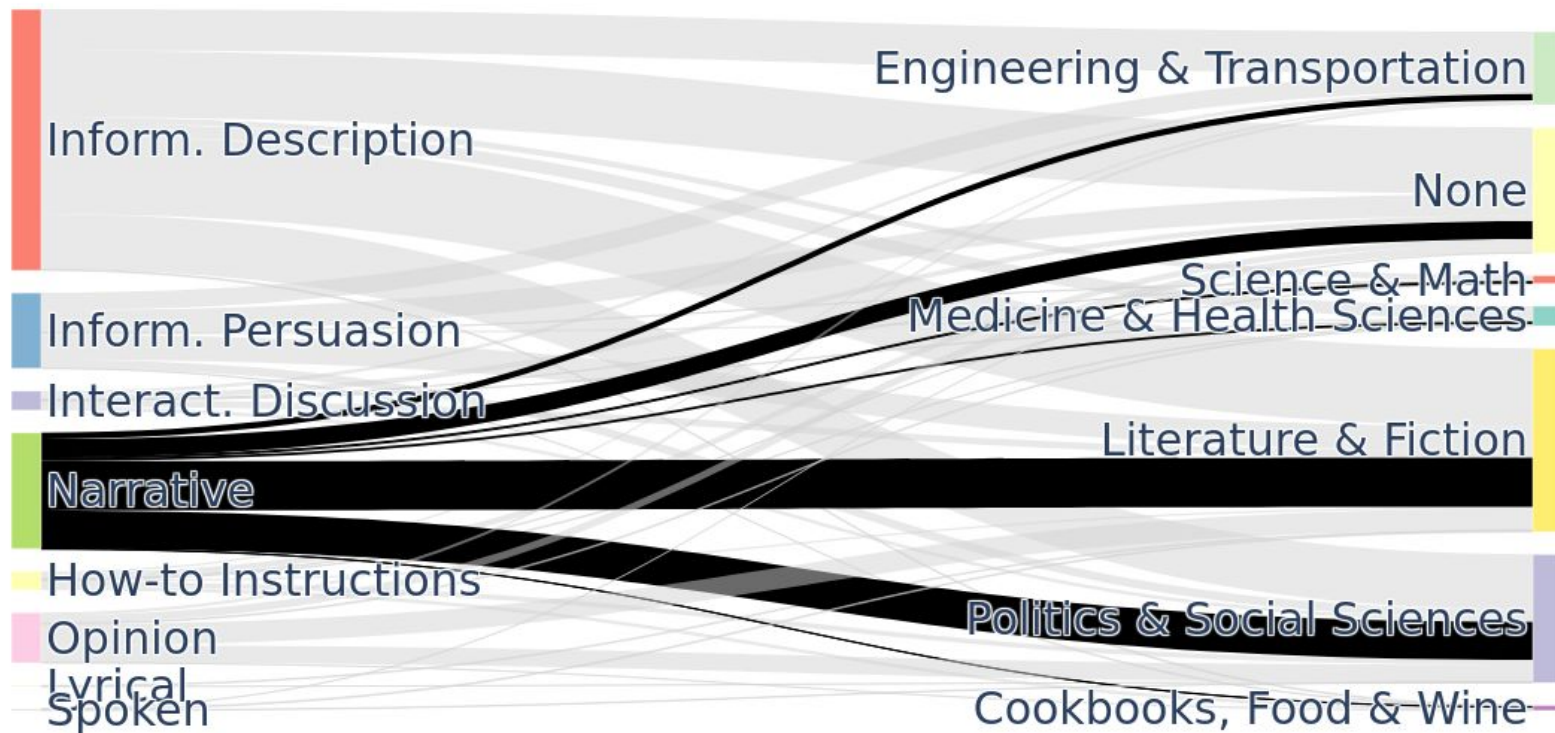




# Intersection of Registers and Genres



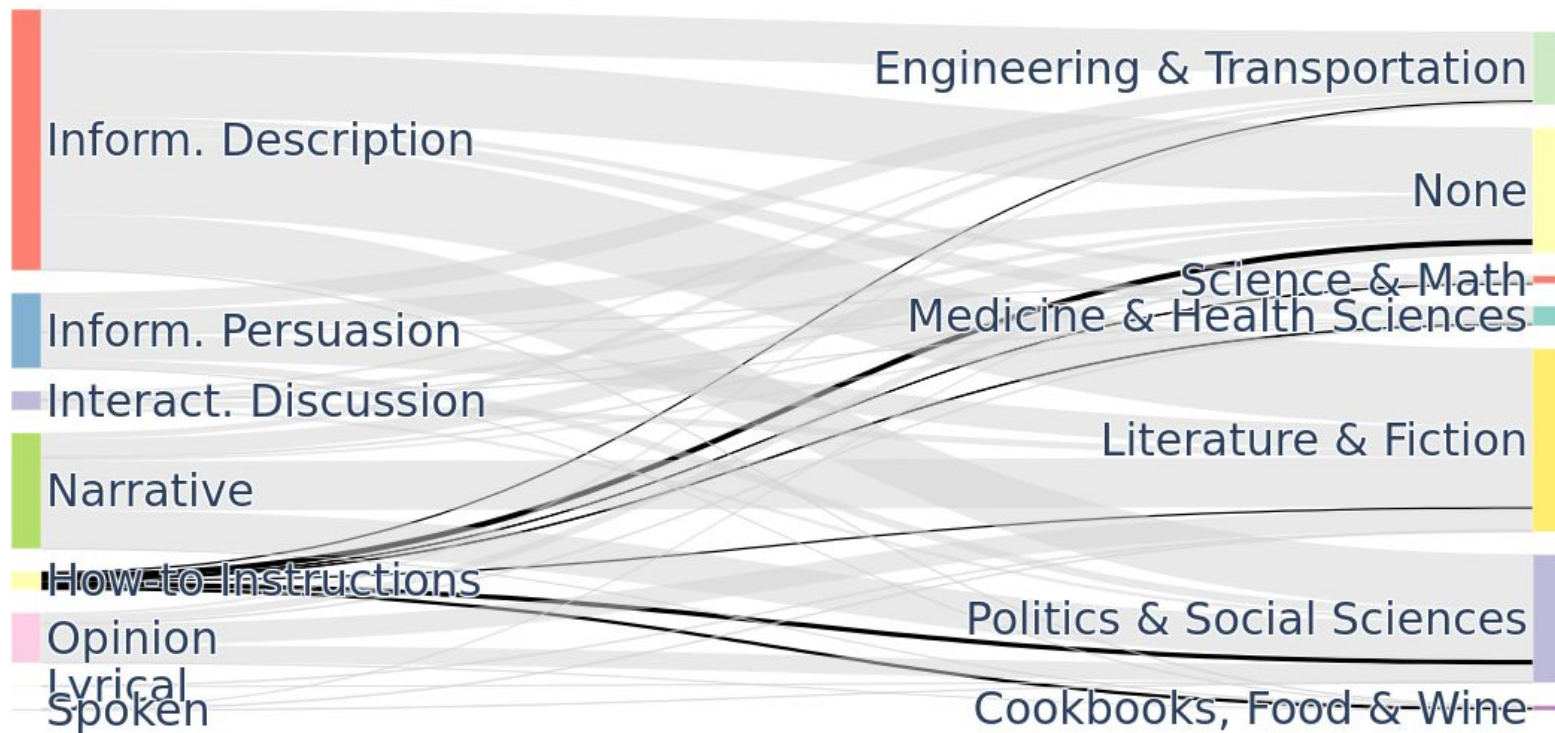
**TURKUNLP**  
.ORG



# Intersection of Registers and Genres



**TURKUNLP**  
.ORG

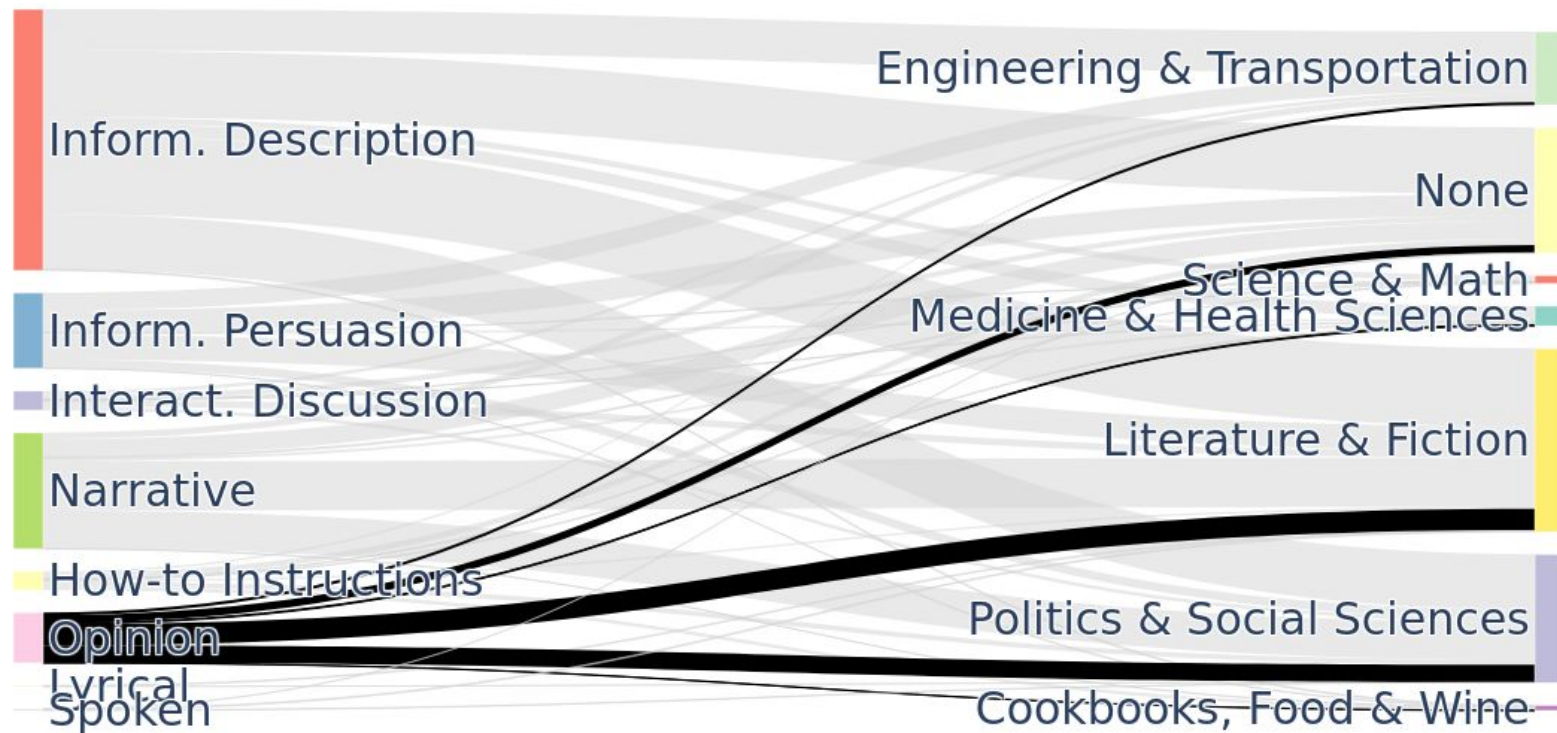




# Intersection of Registers and Genres



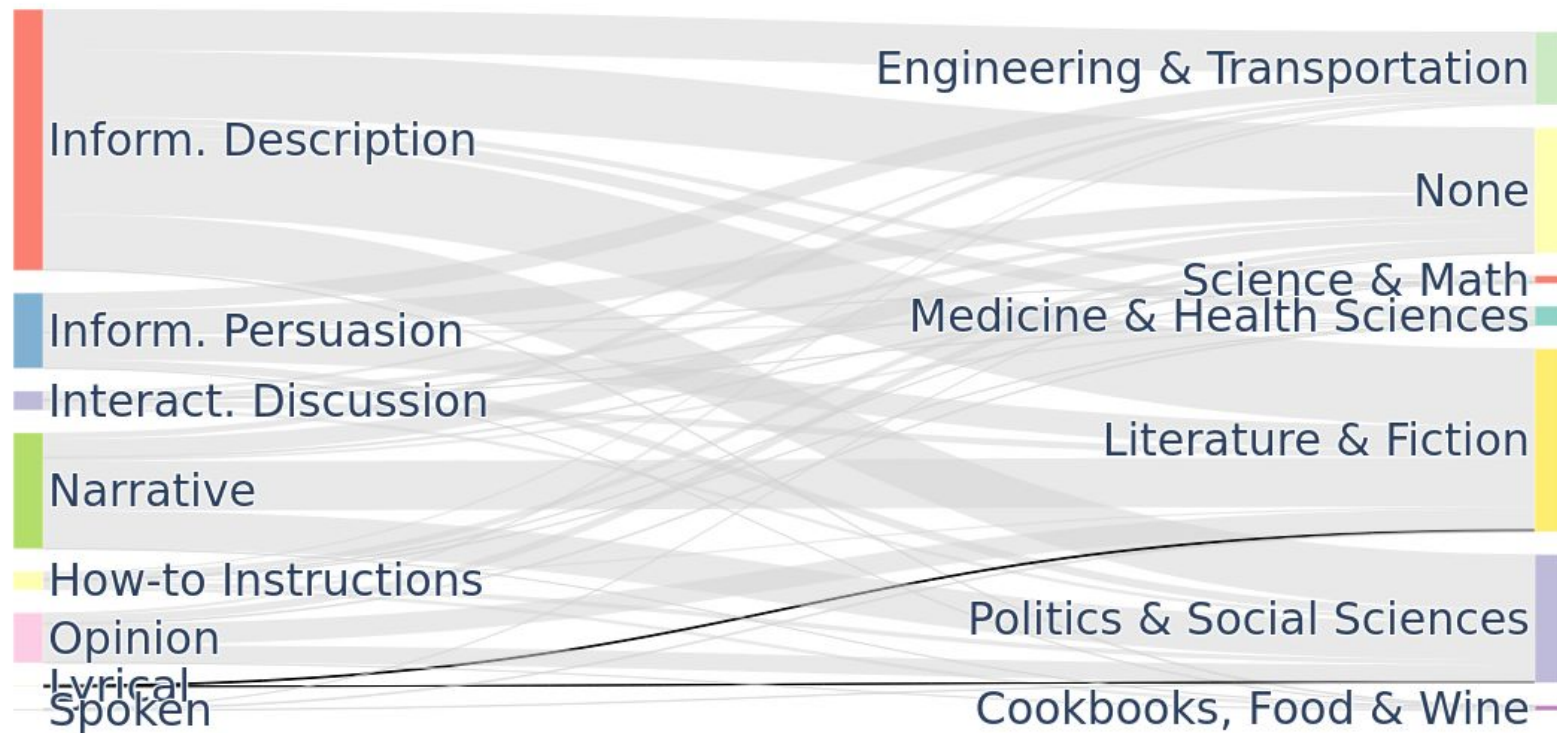
**TURKUNLP**  
.ORG



# Intersection of Registers and Genres



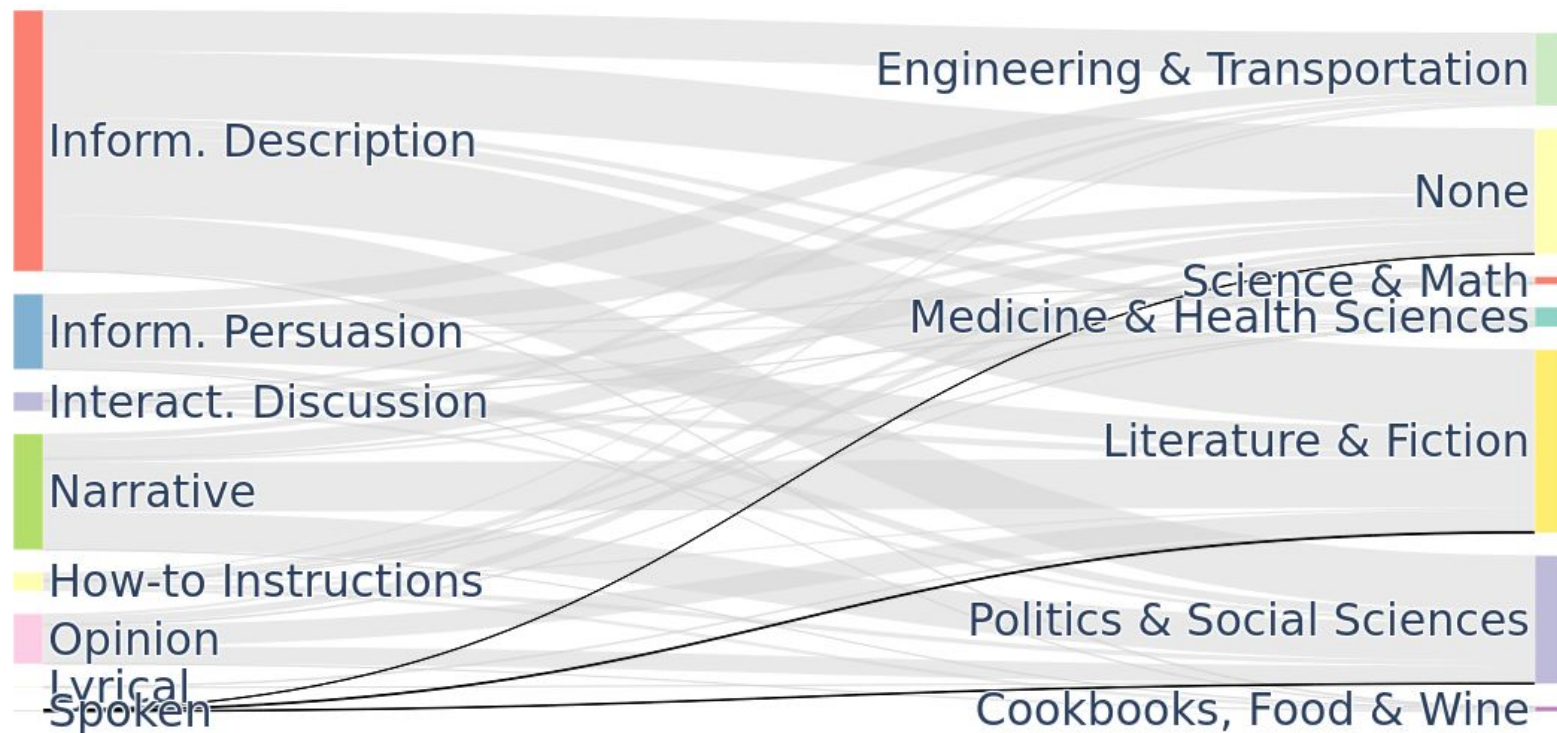
**TURKUNLP**  
.ORG



# Intersection of Registers and Genres



**TURKUNLP**  
.ORG



## Examples

Toddlers and nutrients do not constantly mix. Even though you started your little one out consuming a very high variety of healthy sound meals, sooner or later a young child will boycott all of your choices. It is their strategy to management. The easiest way to maintain diet can be your toddlers diet regime options is usually to hide well balanced meals in the food items that they can ingest, such as creating muffins, biscuits, and pancakes with invisible fresh fruit and veggies in them.



Opinion,  
Advice



Cookbooks  
Food &  
Wine

## Examples

There is a reason our hearing is a cherished sense, and many people fear the loss of it. Hearing loss can happen naturally over time, occur due to illness, or even due to an accident or injury. If you've suffered from hearing loss due to a medical accident, work-related accident, or even just a regular injury, you may be faced with unexpected challenges and expenses. So can you file a lawsuit for your hearing loss? The answer is—it depends.



Informat.  
description



Medicine  
& Health  
Sciences

## Examples

It would still be against Federal Law. And this is one law the feds prosecute as Poker Stars and that other bunch of thieves, Full Tilt Poker found out. I can't believe that any big name is going risk opening up a site in NJ. There are a lot of states where online poker is not against state law(almost all, in fact), but none of them have it. Yeah, we'll see how this plays out in six months plus.



Interactive  
Discussion



Politics &  
Social  
Sciences

## Future work



- Topic modelling and keyword analysis of the combined classes
- Improving performance
  - classes chosen for genre classifier training; Literature & Fiction encompasses too much
  - model architecture experiments

# References



Biber, Douglas, and Conrad, Susan. 2019. "Register, Genre, and Style". 2nd ed. of Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781108686136>

Biber, Douglas, and Egbert, Jesse. 2018. "Register variation online". Register Studies. 2. 166-171. 10.1075/rs.19018.smi.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. "Unsupervised Cross-lingual Representation Learning at Scale". In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Goyal, Anshaj, and Prakash, V. Prem. 2022. "Statistical and Deep Learning Approaches for Literary Genre Classification". In *Advances in Data and Information Sciences*, 297–305. Lecture Notes in Networks and Systems. Singapore: Springer Singapore. [https://doi.org/10.1007/978-981-16-5689-7\\_26](https://doi.org/10.1007/978-981-16-5689-7_26)

Kuzman, Taja, Rupnik, Peter, and Ljubešić, Nikol. 2023. "Get to Know Your Parallel Data: Performing English Variety and Genre Classification over MaCoCu Corpora". In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 91–103, Dubrovnik, Croatia. Association for Computational Linguistics

Laippala, Veronika, Rönqvist, Samuel, Oinonen, Miika, Kyröläinen, Aki-Juhani, Salmela, Anna, Biber, Douglas, Egbert, Jesse, Pyysalo, Sampo. 2023. "Register identification from the unrestricted open Web using the Corpus of Online Registers of English". *Lang Resources & Evaluation*, 57, 1045–1079.

Veronika Laippala, Anna Salmela, Samuel Rönqvist, Alham Fikri Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marc Pàmies, Deise Prina Dutra, Valtteri Skantsi, Lintang Sutawika, and Sampo Pyysalo. 2022. "Towards better structured and less noisy Web data: Oscar with Register annotations". In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 215–221, Gyeongju, Republic of Korea. Association for Computational Linguistics.

T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318-327, 1 Feb. 2020, doi: 10.1109/TPAMI.2018.2858826. keywords: {Detectors;Training;Object detection;Entropy;Proposals;Convolutional neural networks;Feature extraction;Computer vision;object detection;machine learning;convolutional neural networks},

Liina Repo, Valtteri Skantsi, Samuel Rönqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 2021. [Beyond the English Web: Zero-Shot Cross-Lingual and Lightweight Monolingual Classification of Registers](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 183–191, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Suarez, Benoit Sagot, and Laurent Romary. 2019. "Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures". *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019*. Cardiff, 22nd July 2019 (pp. 9 – 16). Leibniz-Institut für Deutsche Sprache.

Zhang, Jinbin, Yann Ciarán Ryan, Iiro Rastas, Filip Ginter, Mikko Tolonen and Rohit Babbar. 2022. "Detecting Sequential Genre Change in Eighteenth-Century Texts". In F. Karsdorp, A. Lassche, & K. Nielbo (Eds.), *Proceedings of the Computational Humanities Research Conference 2022* (pp. 243-255). (CEUR Workshop Proceedings; Vol. 3290). CEUR-WS.org.





**UNIVERSITY  
OF TURKU**

School of Languages and Translation Studies, 2024

# Genre

```
{"test_loss": 0.2669629454612732, "test_f1": 0.6944444444444444, "test_f1_th02": 0.6846846846846847,
"test_f1_th03": 0.6955767562879445, "test_f1_th04": 0.6944444444444444, "test_f1_th05": 0.687099725526075,
"test_f1_th06": 0.6742209631728044, "test_accuracy": 0.6689834926151172, "test_runtime": 27.6341,
"test_samples_per_second": 41.651, "test_steps_per_second": 5.211}
```

Best th=0.3

	precision	recall	f1-score	support
Cookbooks, Food & Wine	0.69	0.51	0.59	35
Engineering & Transportation	0.60	0.72	0.65	172
Literature & Fiction	0.73	0.92	0.81	535
Medicine & Health Sciences	0.77	0.50	0.61	72
Politics & Social Sciences	0.66	0.44	0.53	194
Science & Math	0.71	0.33	0.45	144
micro avg	0.69	0.70	<b>0.70</b>	1152
macro avg	0.69	0.57	<b>0.61</b>	1152
weighted avg	0.69	0.70	0.68	1152
samples avg	0.68	0.70	0.69	1152

```
map_full_names = {  
    "MT": "Machine translated (MT)",  
    "LY": "Lyrical (LY)",  
    "SP": "Spoken (SP)",  
    "it": "Interview (it)",  
    "os": "Other SP",  
    "ID": "Interactive discussion (ID)",  
    "NA": "Narrative (NA)",  
    "ne": "News report (ne)",  
    "sr": "Sports report (sr)",  
    "nb": "Narrative blog (nb)",  
    "on": "Other NA",  
    "HI": "How-to or instructions (HI)",  
    "re": "Recipe (re)",  
    "oh": "Other HI",  
    "IN": "Informational description (IN)",  
    "en": "Encyclopedia article (en)",  
    "ra": "Research article (ra)",  
    "dtp": "Description: thing / person (dtp)",  
    "fi": "FAQ (fi)",  
    "lt": "Legal (lt)",  
    "oi": "Other IN",  
    "OP": "Opinion (OP)",  
    "rv": "Review (rv)",  
    "ob": "Opinion blog (ob)",  
    "rs": "Religious blog / sermon (rs)",  
    "av": "Advice (av)",  
    "oo": "Other OP",  
    "IP": "Informational persuasion (IP)",  
    "ds": "Description: intent to sell (ds)",  
    "ed": "News & opinion blog / editorial (ed)",  
    "oe": "Other IP",  
}
```

## Register...



- Situationally defined text varieties (Biber and Conrad, 2019)
  - focus on linguistic features that are functionally adapted to the communicative purpose at hand
- Interactive Discussion, Narrative, Opinion etc.

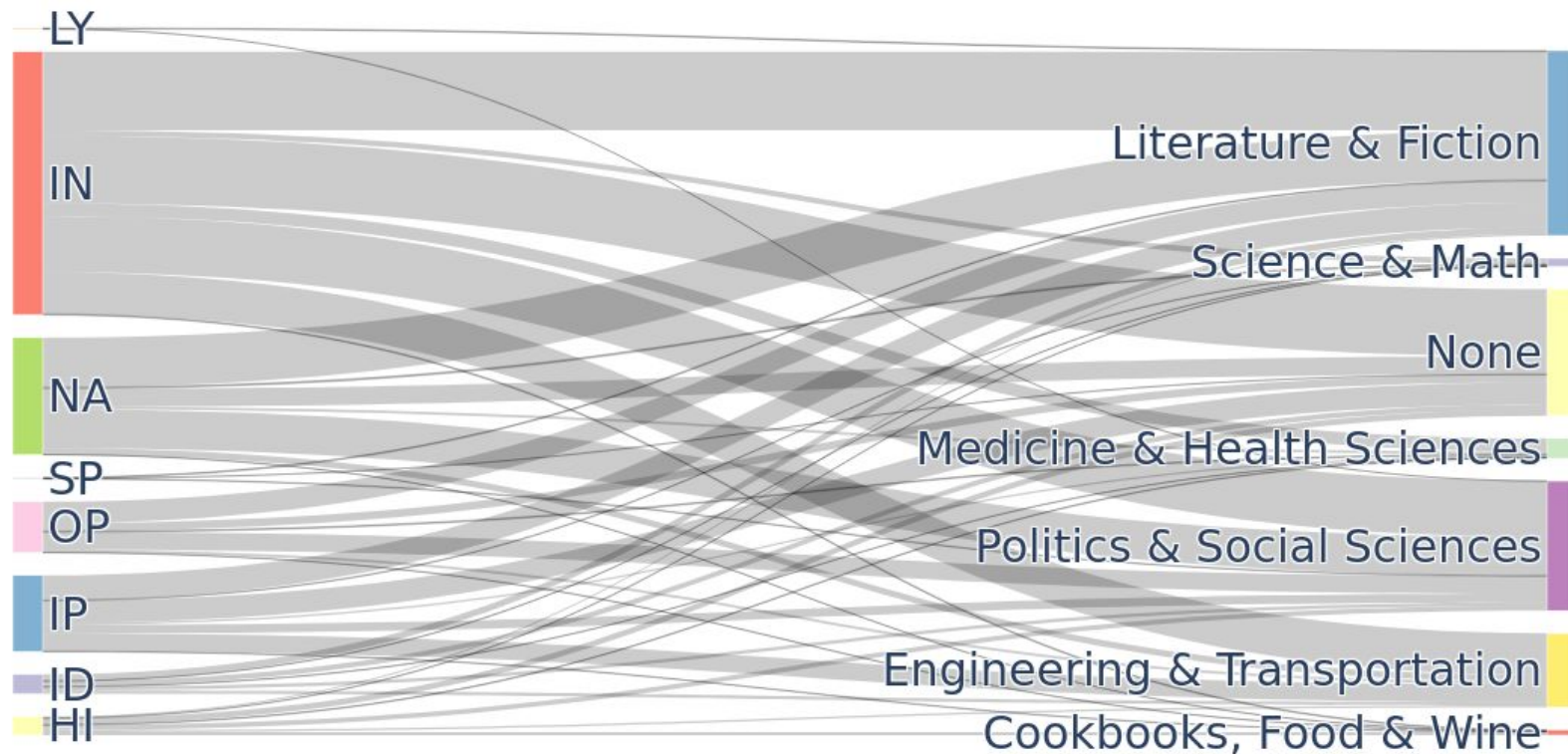
## ...and Genre

- categories to which works of literature can be classified (Goyal and Prakash 2022)
  - focus on similarities between the style of writing, surrounding temporal and geographical context, and narrative techniques
- Science Fiction; Cookbooks, Food & Wine; etc.

# Intersection of Registers and Genres



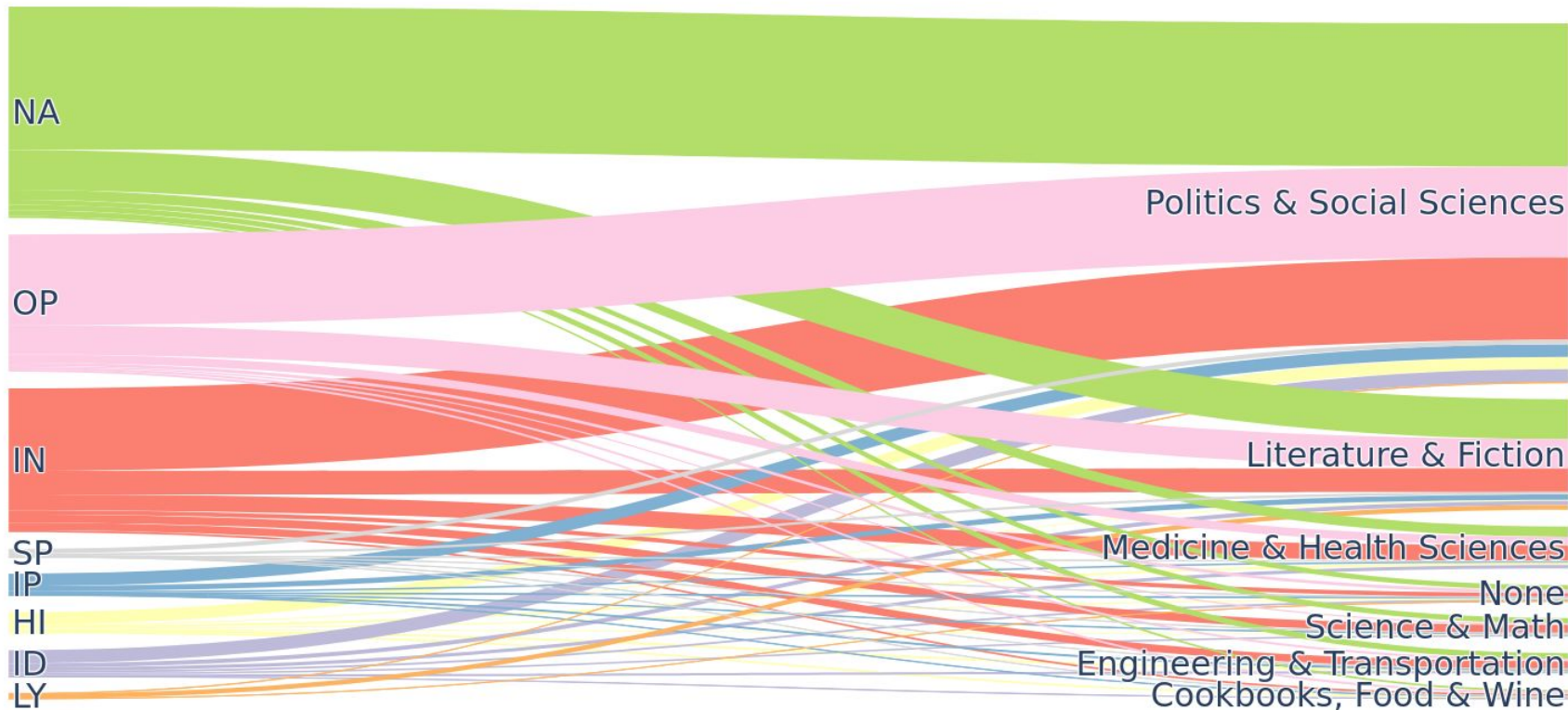
**TURKUNLP**  
.ORG



# Intersection of Registers and Genres



**TURKUNLP**  
**.ORG**



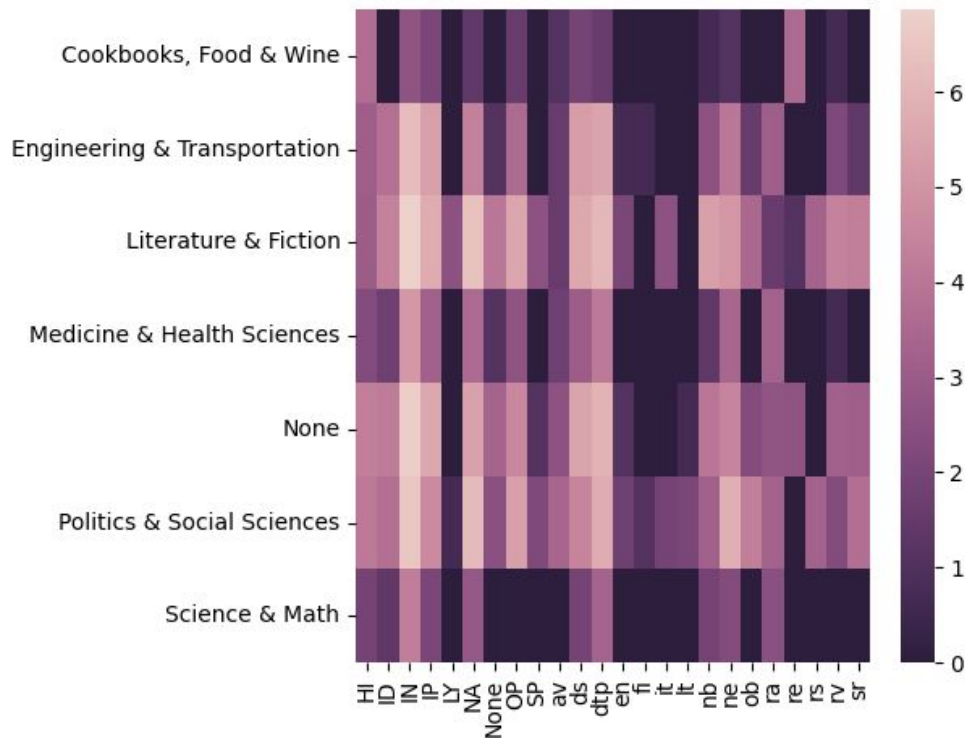
## Intersection of Registers and Genres

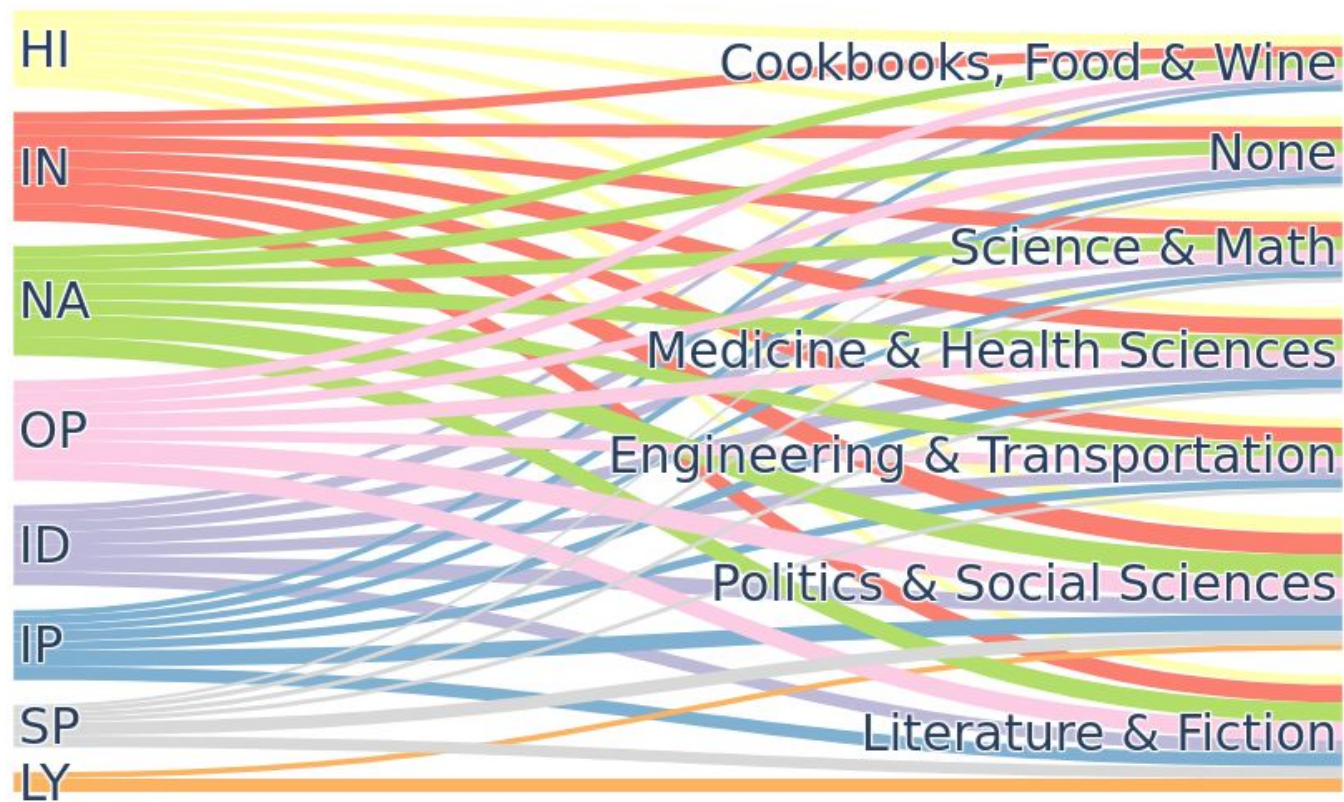


**TURKUNLP**  
**.ORG**

## Most common combinations

1. IN + Literature & Fiction
2. IN + Politics & Social Sciences
3. NA + Literature & Fiction
4. IN + Engineering & Transportation






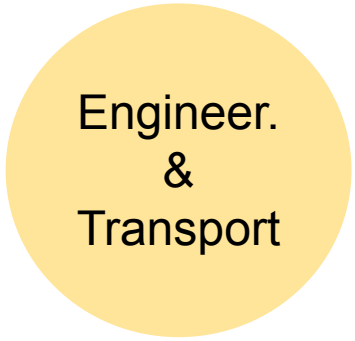


## Examples (CORE)

Join if you have a car, truck, bike, atv, anything with wheels or if you live in the Thunder Bay area. Our automotive community participates in cruises, meets, racing events (AutoX, Drag), 4x4ing, as well as online discussion. Click the register link. It's quick, easy and free! You will then be able to post and gain access to many additional features. To start viewing messages, select the forum that you want to visit from the selection below.




Interactive  
Discussion



Engineer.  
&  
Transport

## Examples (CORE)

Australian inflation plummets as the fiscal vandals undermine the economy The Australian Bureau of Statistics released the Consumer Price Index, Australia data for the March 2012 quarter today and the inflation rate has plummeted in the face of a slowing economy. The trend over the second half of 2011 was for inflation to ease. But the plunge in the first three months of 2012 that today's data reveals is pointing to a very sick economy.




Informational  
Narrative  
News report



Politics &  
Social  
Sciences

## Examples (CORE)

A cholesterol warning in the palm of your hand New device won't tell you how high your level is, but it can tell you if you should get it checked by a doctor Annie Wang, of London Drugs in Vancouver, demonstrates the use of a new method for measuring cholesterol on the skin. Photograph by: Stuart Davis, PNG , Vancouver Sun A hand-held device that can tell people whether they should head to their doctors for further cholesterol testing is being launched in Greater Vancouver this week.



Informational  
Narrative  
News report



Medicine  
& Health  
Sciences

