To do liste:

1. Skriv “DANSK Descriptive stats” for DANSK
2. “DaCy model curation”
   1. Skriv outline
   2. Skriv det ud med tekst
3. “SOTA evaluation”
   1. Skriv outline
   2. Skriv det ud med tekst
4. Download DANSK, switch dev and test, and upload to HF
5. Fix DaCy download link til DANSK
6. Add information in Introduction about silver standard and gold standard datasets and how dataset curation usually takes place (to be able to frame the methods section)
7. Skriv mail til Kenneth
   1. Send “DANSK curation”, “Text sampling” underafsnittet til Kenneth og spørg om det er rigtigt.
   2. Spørg om han fik gjort så modellerne er implementeret i DaCy (og der derfor kan merges). Lige nu kan man nok ikke adde dem til en pipeline med en transformer pipe i sig.
      1. Hvis ikke: Henvis i thesis til branchen med min DaCy version.
8. Skriv mail til Rebekah:
   1. Du tog billede af mit spørgsmål til sidste møde. Har du svar?
   2. Ved ikke om DANSK kan udgives m. OntoNotes tingene, og med licenses fra alle de sources vi har det fra.
9. Skriv results
10. I results, sørg for at performance tabellerne opfylder følgende krav:
    1. Bedste score med **BOLD**
    2. Næstbedste med *ITALIC*
    3. Order: Large øverst, mine modeller øverst, sattrup næstøverste, så daCy, og så spaCy
    4. Ligner derfor følgende:

|  | | Domain | | | |
| --- | --- | --- | --- | --- | --- |
| Dannet | SoME |  |  |
| Model | da\_dacy\_small\_trf | **.9** | .5 |  |  |
| da\_dacy\_medium\_trf | *.8* | **.86** |  |  |
| da\_dacy\_medium\_trf | .7 | *.75* |  |  |

1. Ændr manuelt i de 4 plots med DANSK initial så der er flere LANGUAGE i FULL, TRAIN og DEV, (så der ikke er en så skæv fordeling -> ved at gøre dette så passer alle resten af plots’ne siden de er på test-settet som forbliver uforandret).
2. Ensure that all information from “DANSK\_eval/Thesis” is included
3. Switch dev and test on HF (Matches with the switch all other places (repos, tables, etc) except for on HF)
4. Add info til HF og GitHub for
   1. Datasæt
      1. Readme
   2. Modeller
      1. Readme’s
   3. DaCy
      1. meta\_json.py scriptet
   4. Add information on datasettet til DaCy repo'et inden træning i og evt. i readme's for Modeller og dataset
5. Automatisk dannet metadata
   1. Gem dataen fra øverste i model card, her: <https://huggingface.co/datasets/chcaa/DANSK> og indsæt det i “DANSK\_eval/Thesis/<some\_fold>”
   2. Indsæt det i Appendix
6. Lav DaCy review kommentarerne færdige
7. Request et nyt review af Kenneth. Spørg her også hvordan vi gør så man kan bruge modellerne gennem DaCy
8. Få merged min PR
9. Lav visualiseringer af brug til Release sektionen og gem dem under “DANSK\_eval/Thesis/Release”
10. Få styr på OntoNotes regler:
    1. Skriv til Rebekah:
       1. “*From reading the information on OntoNotes v5 from* [*https://catalog.ldc.upenn.edu/LDC2013T19*](https://catalog.ldc.upenn.edu/LDC2013T19)*, I am having trouble understanding whether I am allowed to disclose the annotation guidelines of the dataset.* 
          1. *Can I do that for the thesis?*
          2. *Can I do that for the dataset on HF, upon publishing it?*
          3. *Also; should I make the dataset public?”*
11. Skriv approx. sideantal for hver sektion og evt. undersektion. Skal passe med 30 sider, for til sidst skal der altid skrives ~5 sider ekstra til at dumb it down og give rød tråd.
12. Skriv in-depth outline for hele paperet. (Relevant points, citations, zotero)
    1. Brug litteraturen fra “Literature” i OneNote til introduktion og diskussionen
    2. Ikke add unødvendig teori i Introduktionen
13. Skriv Glossary her i GDocs (senere til OverLeaf Glossary)
14. Skriv hele opgaven ud fra sillebensstrukturen.
    1. **Brug chatGPT til inspiration - specielt til introduktionen.**
    2. **Brug \* hver gang der er noget der ikke er færdigt eller skal vendes tilbage til senere.**
15. Overvej at lave visualiseringer til “DANSK/METHODS” hvis det kan give clarity.
16. Sørg for at ALT fra “DANSK\_eval/Thesis” er med i opgaven, eller i appendix.
17. Sørg for at appendix bliver refereret til
18. Ensure that tables are ordered in the same fashion (large on top)
19. Sørg for at der er en meget klar og explicit deskription af tingene JEG har lavet og det jeg IKKE har. Det kan trække meget ned i vurderingen hvis jeg ikke har det.
20. Sørg for opgaven lever op til Formalia for Kursus beskrivelsen:
    1. Thesis statement:
       1. Using background theory, methodology, empirical data appropriately to develop thesis statement
       2. Specify and narrow down a question/task/problem, and explicitly state how this thesis will answer or solve it, using academic methods.
       3. CLEAR, delimited thesis statement
       4. Argue clearly for the thesis statement
       5. Coherent work, following the delimited thesis statement
    2. General writing has to be:
       1. Clear and comprehensible paper
       2. Clear connection(!) between analytical/theoretical/empirical work AND product
       3. Assume a naïve reader in the writing
    3. In the thesis, demonstrate:
       1. Critical thinking - investigating/reflecting on issues of the field
       2. Critical thinking by investigating potential shortcomings of own methods/work
       3. Demonstrate awareness of relevant literature and relate it to thesis topic
       4. Knowledge of how current research and practices within the field relate to thesis topic and paper
       5. Relevant theories and methods and how it relates to other similar areas
       6. Identify, interpret, integrate and build upon relevant existing academic literature
       7. In-depth knowledge of thesis topic
21. Strømlin sprogbrug
    1. Brug rater om personer og ikke annotated?
    2. Named-Entity Recognition? Eller uden -?
    3. Texts or documents
    4. “Other models”, “SOTA Models” Strømlin sprogbrug så man altid er klar over hvilke der refereres til
22. Lav sidste småændringer
    1. Sørg for at plots er kodet af mig selv (eneste regel for ikke at snyde) eller at det er rightly attributed.
23. Sæt alting ind i Overleaf
    1. Lav først backup(s)
    2. Sæt ind
    3. Lav flere backups
24. Få Malte, Stinne, Rebekah og Kenneth til at læse det igennem og få feedback
    1. Giv review pointer til dem hver især
    2. Rebekah + Kenneth:
       1. Bestemte sektioner
       2. Bestemte feedback punkter
          1. Er der noget jeg har rapporteret der kunne diskvalificere mit arbejde pga. snyd?
          2. Manglende kontekst til censor der ikke kender NLP og NER så godt?
          3. …
          4. …
    3. Stinne og Malte:
       1. Forståelse for udefrakommende
          1. Manglende information?
          2. Manglende kontekst?
          3. Andre ting der er iøjnefaldende.
25. Afvent feedback fra alle.
26. Omskriv på baggrund af alles feedback
27. Grammarly/ChatGPT gennemgang (tal med Johan om at splejse)
28. Print opgaven og læs den igennem.
29. Lav evt. ændringer ud fra den printede gennemgang
30. Gennemgå formalia og læs vejledningsplanen ved aflevering af specialet
    1. Vær opmærksom på, at du ud over at uploade vejledningsplanen via studieportalen også skal vedlægge vejledningsplanen som bilag til specialet, når du afleverer i Digital Eksamen. Hvis din vejledningsplan hverken er uploadet via Studieportalen eller vedlægges ved aflevering, kan dit speciale ikke blive bedømt."
31. Få alting eksporteret fra OneNote (da min account m. indhold bliver lukket ved endt studie)
32. Få indskrevet “Sommer to-do list” ind i min alm. to-do liste.
33. Lav de ting der skal laves inden jeg går ud af studiet.
34. **Aflevér. Men notér at man ikke kan genaflevere når den først er indsendt. Den skal derfor være HELT færdig.**

Resources

* Papers:
  + connectedpapers.com
  + elicit.org
  + reserachrabbit.ai
  + zotero
  + OneNote "Literature" (I wrote notes on most important papers)
* Models and datasets:
  + OnoNote: "Overview of models and datasets" (I wrote notes on all datasets)
* Making theoretical plots:
  + draw.io
  + LucidChart

DANSK and DaCy 3.0.0:

Text domain expansion within the NER-task in Danish NLP

[**Introduction 9**](#_715d7cyhqqor)

[Current state of Named-Entity Recognition in Danish 9](#_yksltvt1jxmq)

[Limitations of the current state of NER in Danish 10](#_zig5s4cebi5)

[Alleviating limitations 11](#_px1zqazdjc4b)

[Aim of paper 11](#_ndq5eqwztfc8)

[**Release 12**](#_dssp9f42bm9h)

[DANSK 12](#_9mwy4x9i3y3e)

[DaCy 18](#_d4e92vshluek)

[**Extra methods?? 18**](#_byifjasnad4x)

[Named-Entity match types 18](#_99qk88wkma8x)

[**Tools 18**](#_drvs78hsh1te)

[**DANSK curation 19**](#_tturle2sezqb)

[Text sampling 19](#_c2wye7o3hooz)

[Annotation guidelines 20](#_kemp2743em7d)

[Initial annotation 21](#_xnisjug6b0nc)

[Quality assessment of annotations 22](#_5mictcqavc7b)

[Annotation improvement of inter-annotated texts 25](#_5w73rtzd3fpf)

[Annotation improvement of single-annotated texts 26](#_l234ceele5ss)

[Resolving remaining inconsistencies 27](#_ybt9iofnb5p)

[Partitioning 27](#_qr5t661buif5)

[DANSK quality assessment 27](#_w7rey8ucr09l)

[**DaCy model curation 29**](#_8v3kfjxdcpdk)

[Methods 29](#_ocvhw7exrbmv)

[Model specifications 29](#_jpgmemg37td)

[Domain and entity-level performance 30](#_9vkkl97fgf2g)

[Metrics 30](#_5jchqjq9rcsk)

[Results 31](#_382youdpcqpq)

[**SOTA generalizability 37**](#_dhnzo7bs65y0)

[Methods 37](#_9kysmyf0fw82)

[Models 37](#_cobqg1gfg3as)

[Entity label transfer 37](#_ocvxy57ol8i1)

[Domain and entity-level performance 37](#_w07c633wlg3u)

[Metrics 37](#_onzhxwl8tab9)

[Results 37](#_ap101bjwd620)

[**Discussion 39**](#_2wnouxbqidal)

[DANSK dataset limitations 39](#_ezyv4cj04s70)

[DaCy models 43](#_a4cvkg540m6o)

[SOTA models and generalizability evaluation 43](#_wjs3mygx3qx8)

[Product usage and further research 43](#_mx3kv6v1rjfu)

[**Conclusion 44**](#_68s305g5l7w5)

[**Appendix 44**](#_6tjx5ahs351e)

[**Backup 44**](#_20iapsw1b78m)

# Introduction

## Current state of Named-Entity Recognition in Danish

1. NER (what is it?)
   1. NLP
   2. Tags/labels
   3. Machine learning and annotations
   4. Widely used
      1. NER in- and of itself not useful BUT
      2. NER is an important wheel in providing insights from data since most data is text data.
      3. Examples:
         1. Biomedicine
         2. Making historical newspaper on cultural heritage searchable (Europeana)
         3. \* More examples \*
2. Datasets (in Danish for NER)
   1. Good gold-standard dataset DaNE, in Universal Dependencies format
      1. DaNE (and its previous datasets)
         1. Age
         2. Domains
         3. “Classical labeling” (Coarse-grained)
   2. Universal Dependencies
3. Models (in Danish for NER)
   1. Best models for Danish NER currently
      1. Scandeval
         1. <https://scandeval.github.io/>
      2. DaCy
         1. What is DaCy
         2. DaCy models
      3. Sattrupp/model
         1. Best (scandeval)
         2. Cross-trained on languages
      4. spaCy models
         1. Most frequently used
         2. Most famous
   2. *High performance with transformer model architecture models*
      1. *Masked Model transformer embedding layer and different heads*
         1. *E.g. Danish Gigaword Corpus (huge, but does not include annotations and therefore cannot be used for training NER heads directly)*
      2. *Less data required*

## Limitations of the current state of NER in Danish

1. Three main issues to unpack:
   1. No measures of generalizability outside DaNE
   2. No models without this potential generalizability issue
   3. No models with fine-grained tagging
2. **No measures of generalizability outside DaNE**
   1. DaNE (standard for NER in Danish)
      1. Models are trained and evaluated on DaNE
      2. Good dataset with highly performing models
      3. DaNE lacks diversity in domains and time-periods
      4. As such, the high performance does not necessarily generalize
         1. Research indicates that domain shifts have been known to cause drops in performance. Time will necessarily also play a factor.
         2. Examples
            1. Example 1 of domain shift problems
            2. Example 2 of domain shift problems
3. **No models without this potential generalizability issue**
   1. Issue closely tied to the “no measures of generalizability” issue;
   2. All models are trained on DaNE
   3. If generalizability issue, then all models suffer from low generalizability
   4. Models are being used in new domains today
      1. Examples:
         1. Example 1
         2. Example 2
   5. These uses may be problematic and for Danish NLP and NER, we need alternative models
4. **No models have fine-grained tagging**
   1. As all models are trained on DaNE, which uses classic, coarse-grained tagging
   2. Opposed to high ressource languages (i.e. English, Spanish, Chinese)
   3. Higher levels of granularity means better use, if performance doesn’t suffer.
   4. Examples of use that is NOT possible in Danish
      1. Example 1 of fine-grained NER in English (that is not possible in Danish)
      2. Example 2 of fine-grained NER in English (that is not possible in Danish)
5. *Summary of limitations:*
   1. *No measures of generalizability outside DaNE*
   2. *No models without this potential generalizability issue*
   3. *No models with fine-grained tagging*

## Alleviating limitations

1. The limitations of …
   1. No measures of generalizability outside DaNE
   2. No models without this potential generalizability issue
   3. No models with fine-grained tagging
2. May be alleviated by curating a dataset with:
   1. More domains and newer texts
   2. More fine-grained annotation
3. This allows for:
   1. Estimating generalizability outside DaNE
   2. Training new models without the potential generalizability issue
   3. Training new models with fine-grained tagging
4. This would benefit:
   1. Generally
      1. Know which models to use for which domains (because the old models are likely still very useful
      2. Fine-grained tagging
   2. Research
      1. Example 1.
      2. Example 2.
   3. Industry
      1. Example 1.
      2. Example 2.

## Aim of paper

***Very condensed with aim and how I practically structure the paper and carry it out. Keep additional info in “alleviating domains barriers”***

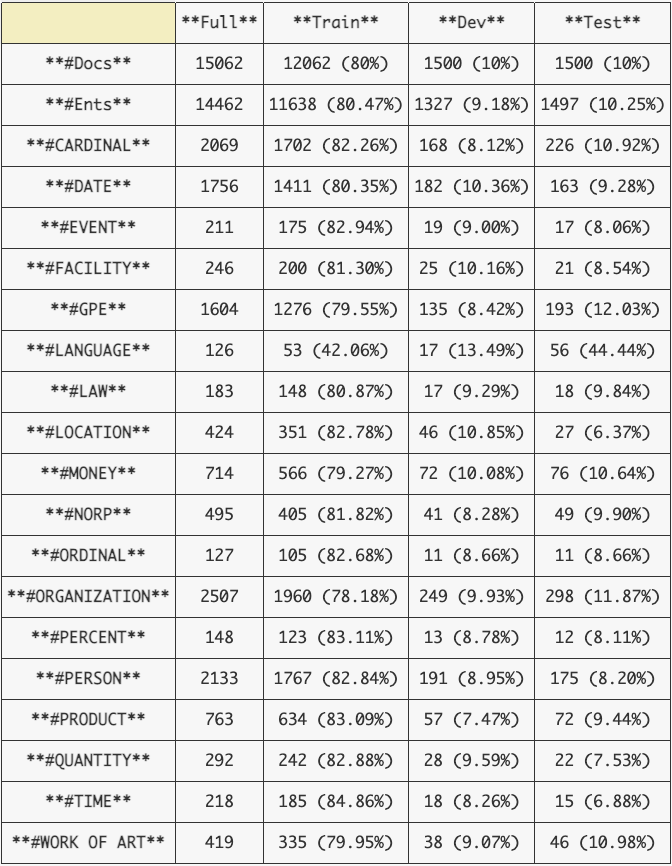
1. As such, this aim of this paper is thus threefold:
   1. **To introduce DANSK - Danish Annotations for NLP Specific tasKs**
      1. First version, only including NER annotations
   2. **To introduce three new models trained on DANSK in: DaCy 3.0.0**
      1. New version, including new models specifically for fine-grained NER on new domains
   3. **Evaluate current SOTA models to get estimate of generalizability**
2. This paper will therefore contain the following sections:
   1. Release
      1. Access and use of:
         1. DANSK
         2. DaCy models
   2. DANSK curation
      1. Brug dele af dette:
         1. “The procurement of the DANSK dataset would thus be an intricate process which involved a large number of customized processing steps. The outline of the previously mentioned steps is depicted in figure X\* in order to provide a general overview.”
         2. “While this description of the formation of the DANSK dataset curation may provide a more conspicuous understanding, an in-depth insight of the intricacies of the process may be assessed in the following sections: “Section X, Section X1,””
      2. Methods
         1. \* Short summary \*
      3. Results
         1. \*List type of results\*
         2. Interrater reliability
         3. Descriptive statistics
   3. Model curation
      1. Methods
         1. \* Short summary \*
      2. Results
         1. \*List type of results\*
         2. Within tag and domain
   4. Generalizability assessment
      1. Methods
         1. \* Short summary \*
      2. Results
         1. Within tag and domain
         2. Across models (also new)

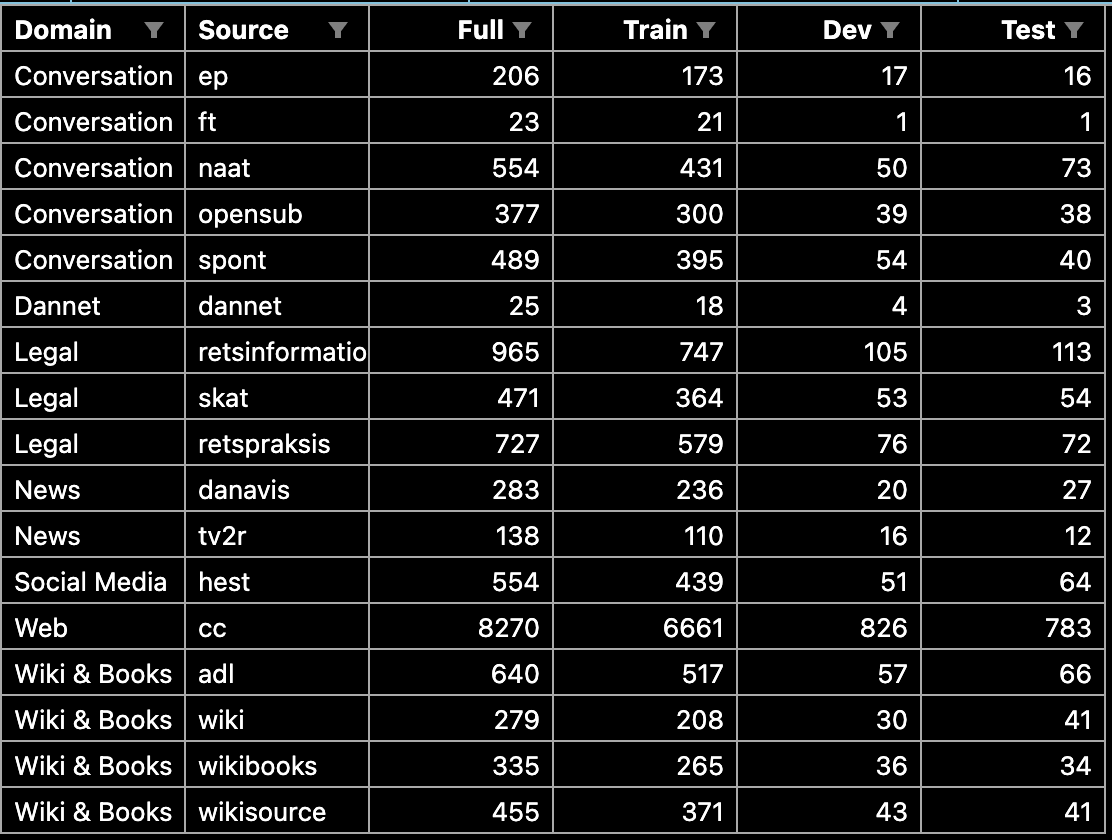
# Release

## DANSK

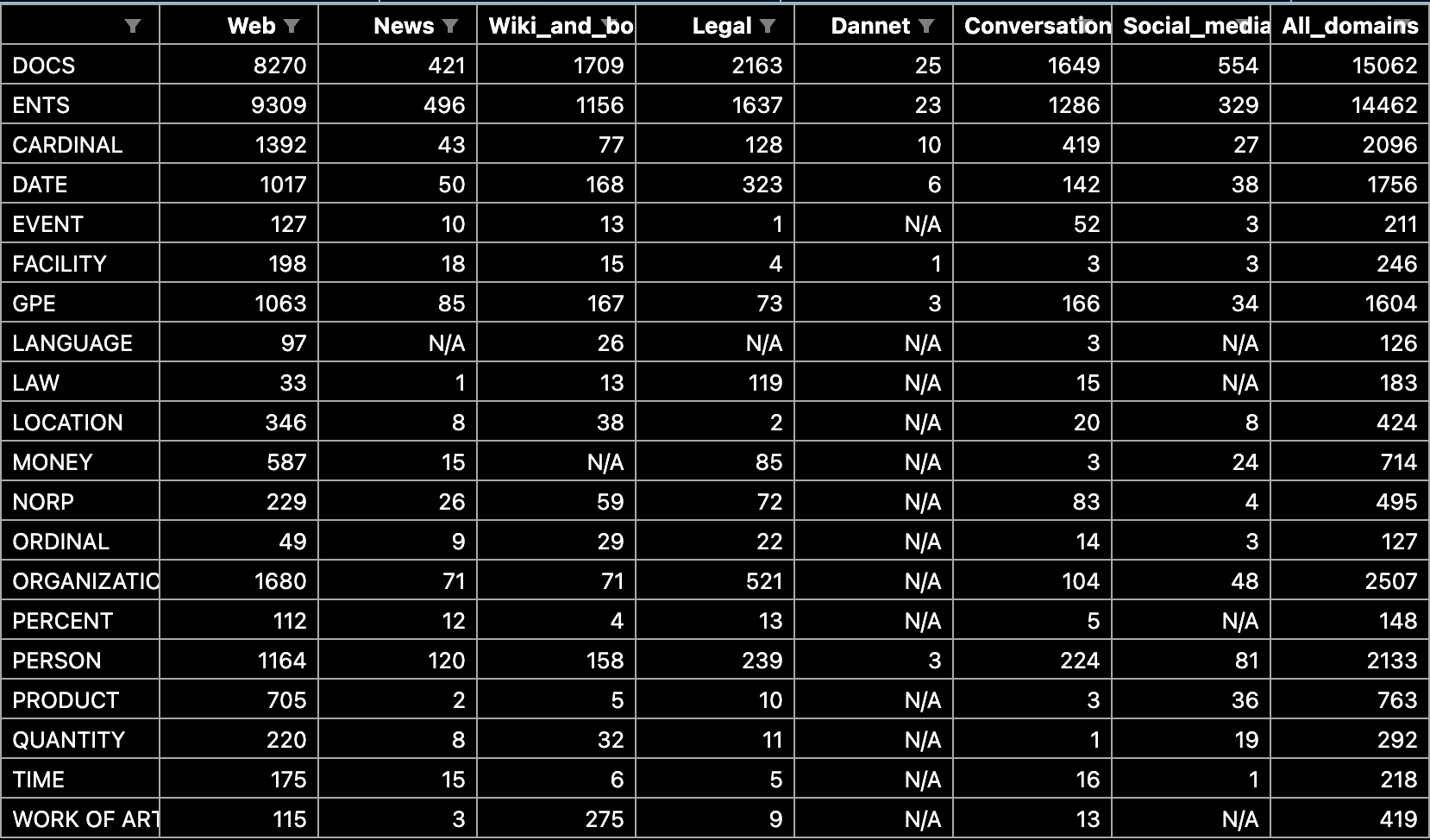
1. Generelt om datasættet
   1. First version
2. Hvad kan det bruges til?
3. Hvordan bruger man det?
4. Available on HuggingFace
5. Samplede tekster DAGW, fra forskellige domæner
6. Licensing:
   1. Creative Commons Attribution Share Alike 4.0 International
   2. <https://github.com/centre-for-humanities-computing/DANSK/blob/main/LICENSE>
   3. OR????
   4. <https://huggingface.co/datasets/DDSC/partial-danish-gigaword-no-twitter#source-data>
7. Descriptive stats

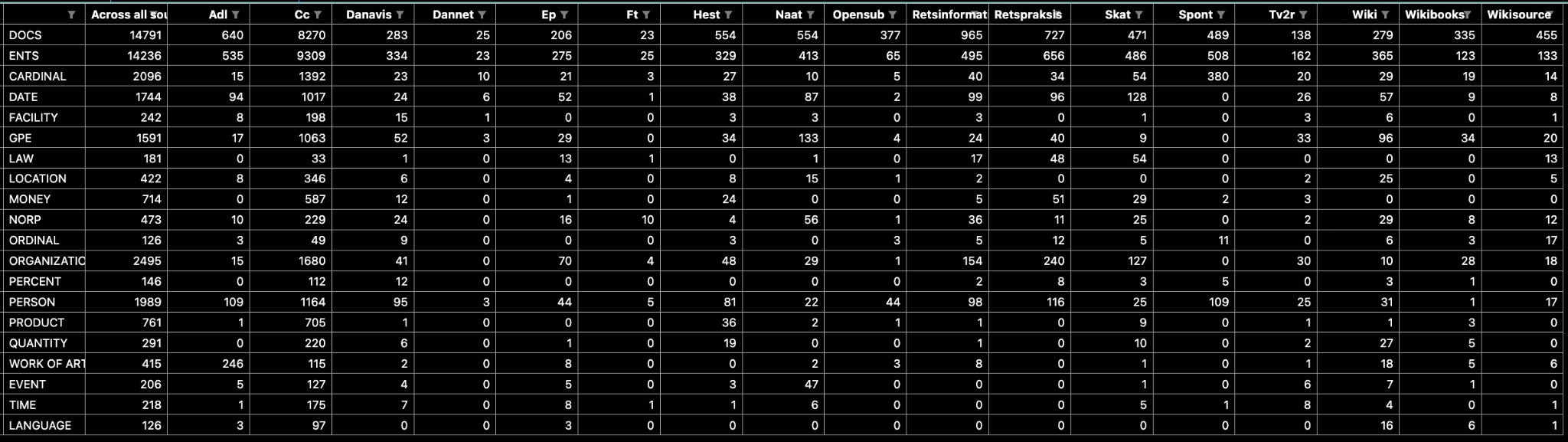
DANSK descriptive partitions (overvej at rykke til “Release” \*):





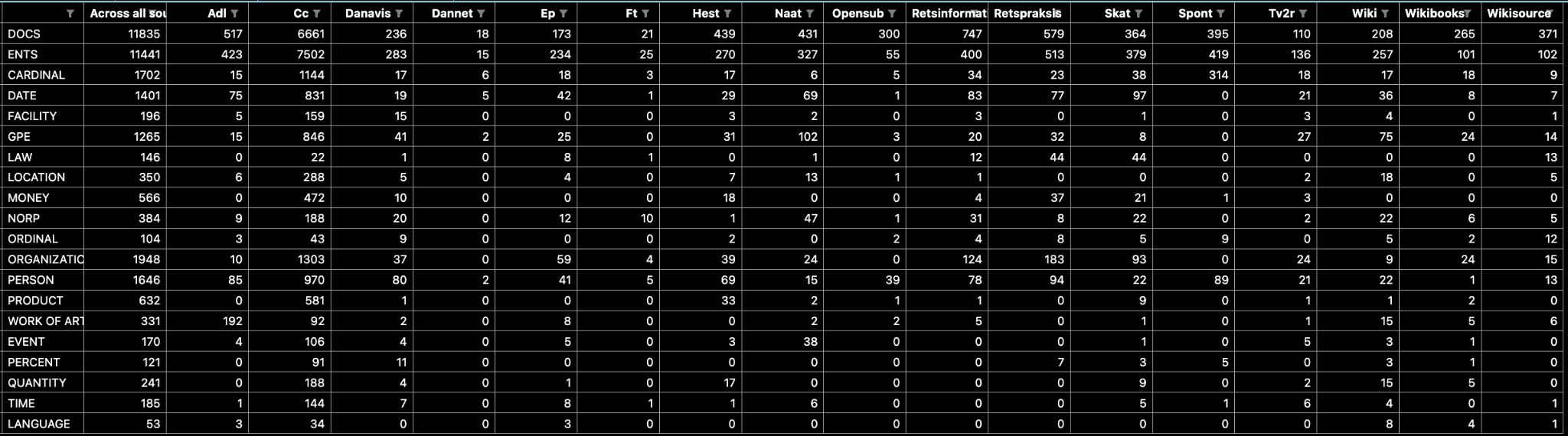
Full:



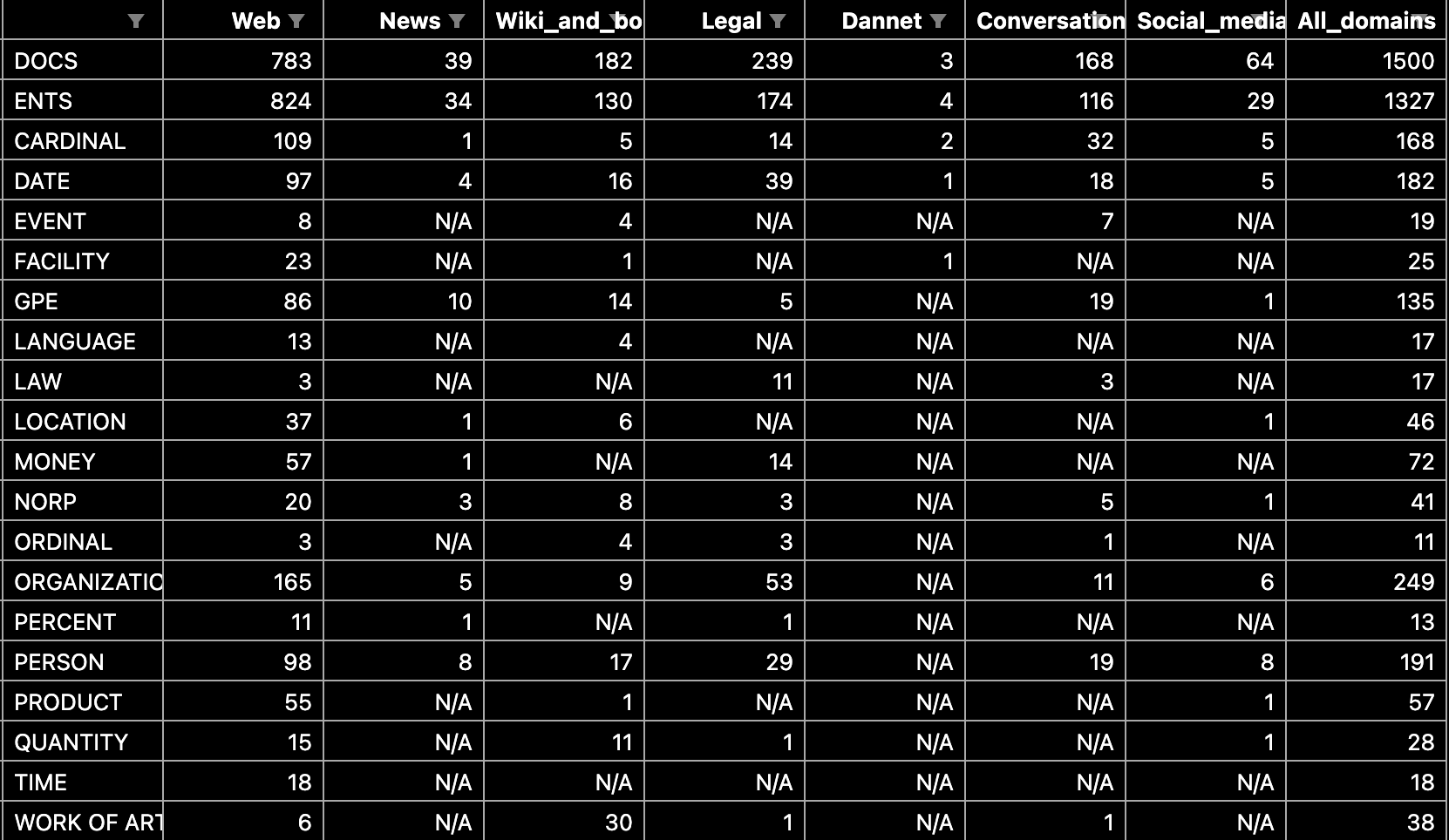


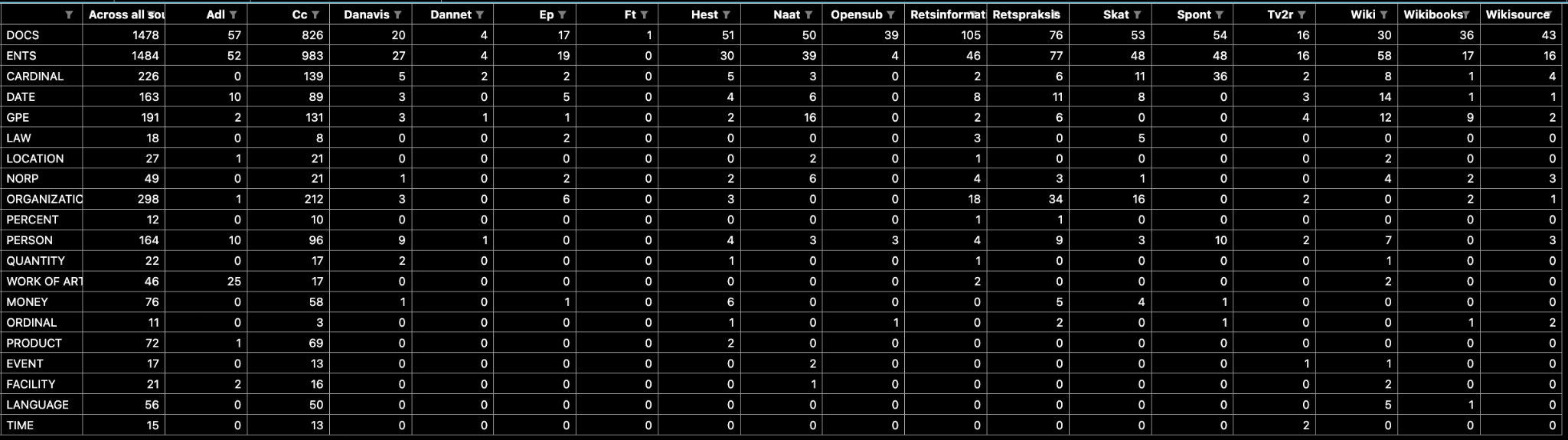
Train:



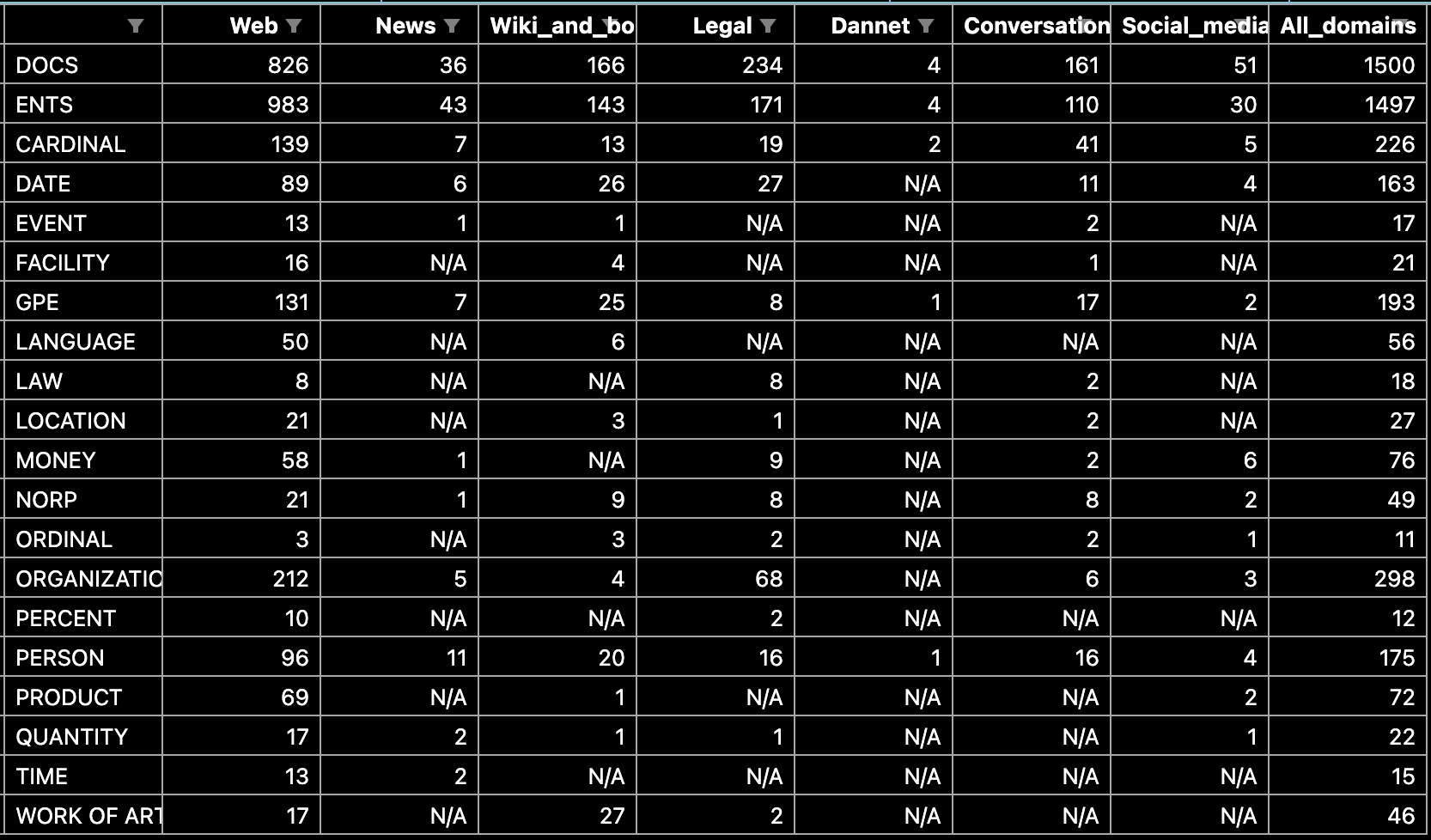


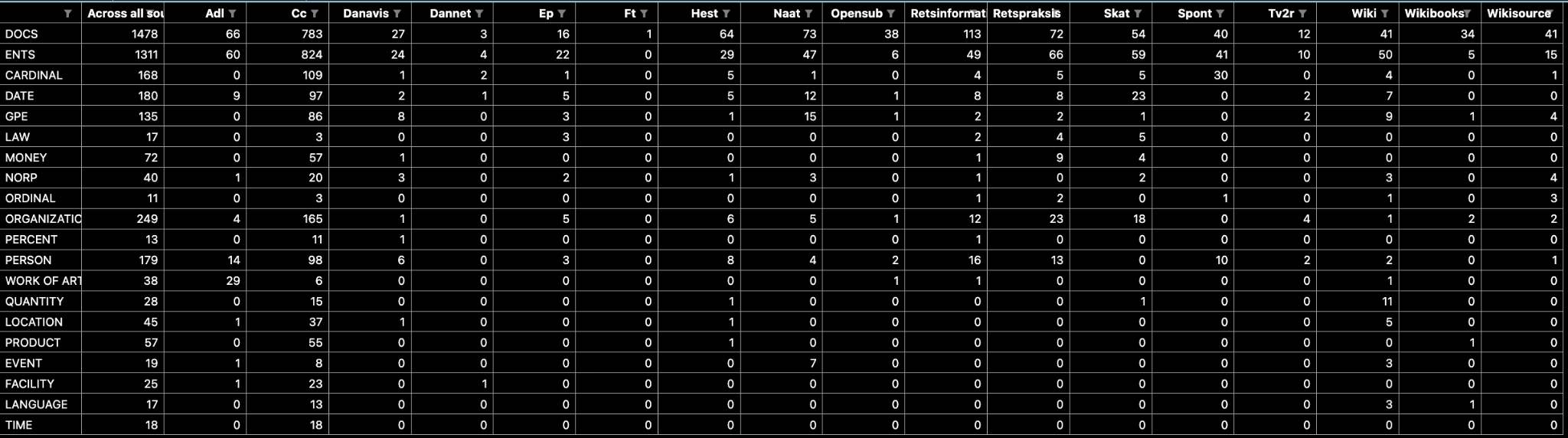
Dev:





Test:





## DaCy

1. Generelt om nyt content som følge af DaCy opdateringen
   1. Tutorial
   2. Nye modeller
   3. Screenshot af displaCy af det
2. Hvad kan det bruges til?
3. Hvordan man bruger det?
4. Generelt om de nye modeller
   1. Arkitektur
   2. Performance (referér til results)

# Extra methods??

## Named-Entity match types

1. Differentiation between these terms are both used in the Dataset preprocessing and in the evaluation of models and thus also results.
2. Description of this is also paramount for interpreting results, as these are from strict matches.
3. <https://pypi.org/project/nervaluate/>

## Tools

1. Anything annotation related that was manual:
   1. Prodigy - Explosion
2. Anything annotation related that was not manual:
   1. Python
   2. spaCy - Explosion
3. Anything model related:
   1. spaCy
   2. HuggingFace

# DANSK curation

## Text sampling

The texts in the DANSK dataset were sampled from the Danish Gigaword Corpus (DAGW); a publically available Danish text corpus which includes one billion words (Cite\* The Danish Gigaword Corpus <https://arxiv.org/pdf/2005.03521.pdf>). It covers a wide array of texts authored by people across socio-economic backgrounds and Danish dialects. Furthermore, the dataset has been curated with texts from different sources and time periods.

The texts were sampled randomly from a select list of sources, up until a select number of tokens was reached. Both the list of sources along with their sampling numbers can be seen in table X\*.

Subsequent to sampling, the texts were filtered so that no texts had origins from before the year 2000. Apart from the texts from Johannes V. Jensen that were completely removed, only NAAT, Danish Literature, Gutenberg, WikiSource, and Religious texts included older texts and were thus the only sources in which filtering took place.

While the source of the texts indicates the direct origin of a text, an additional abstractional layer of labeling was added to the texts; namely the domains of the texts. The term domain is for the duration of this paper used to refer to a more general origin of a text and - while more crude than source origin - should give a broader, more comprehensible understanding of the text. The mapping between domains and sources can be seen in table X\* below and is derived from (cite\* The Danish Gigaword Corpus <https://arxiv.org/pdf/2005.03521.pdf> ). Their paper clarifies the domain definitions as well as the reasoning behind the mapping.



## Annotation guidelines

It was paramount to use a set of annotation guidelines for the DANSK dataset that matched the high granularity that can be seen in high resource languages. By employing the set of annotation guidelines that had already been developed in the successful OntoNotes project, it would be feasible to have DANSK become tantamount to OntoNotes. These guidelines provided 18 different Named-Entity types. Please refer to table X\* for a shorthand description of each of the tags. This shorthand annotation scheme was thus utilized for the initial annotations by the employed raters.

Shorthand annotation scheme:



Later in the process (upon assessing the quality of the initial annotation, section X\*), however, it became evident that the shorthand descriptions seen in table X\* were too simplistic. The descriptions allowed for many ambiguous interpretations and did not provide enough information for properly annotating the texts.

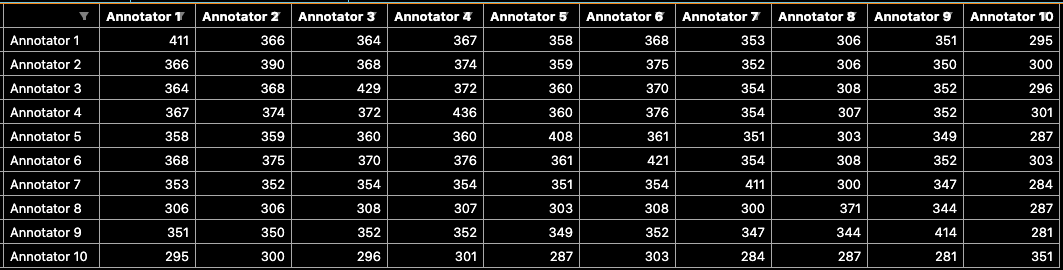
All manual resolvement of conflicts and further annotation in the later stages of the DANSK curation (section X and X and X\*), therefore utilized the extended OntoNotes NER tagging guidelines (cite\*). The 24\* pages of tagging guidelines and 5\* pages of language-specific supplemental material provided much more clarity. However, they did not provide guidelines for the edge-cases met in the curation of the DANSK dataset. For this reason, the author developed an additional supplement to the guidelines for DANSK, namely Supplement D: Danish Addendum. The addendum was developed by taking off-set in the logic in the OntoNotes extended guidelines and clarified edge-cases further. They were moreover developed to encapsulate the particularities of the Danish language, since the original OntoNotes extended guidelines were not developed with Danish in mind. Please refer to appendix X\* for an overview of the Danish Addendum.

## Initial annotation

Following the sampling of texts for DANSK, the texts underwent an annotation process.

For this, 10 English Linguistics Master’s programme students from Aarhus University were employed. They worked 10 hours/week for six weeks from October 11, 2021 to November 22, 2021. Their annotation tasks included Part-of-Speech tagging, Dependency Parsing and NER tagging. For POS, the annotators merely corrected the silver-standard POS predictions that had been created by a POS model that previously had been trained by the CHCAA. The Dependency Parsing was done from scratch.

For the NER annotation the texts were first divided up equally for the 10 annotators, with a 10% overlap between the texts to allow for calculating and assessing inter-rater reliability. The overlap was varying and thus not identical between raters meaning that overlapping documents could have anywhere from 2 to all 10 annotators. Refer to table X\* for an overview of the text overlap between raters.

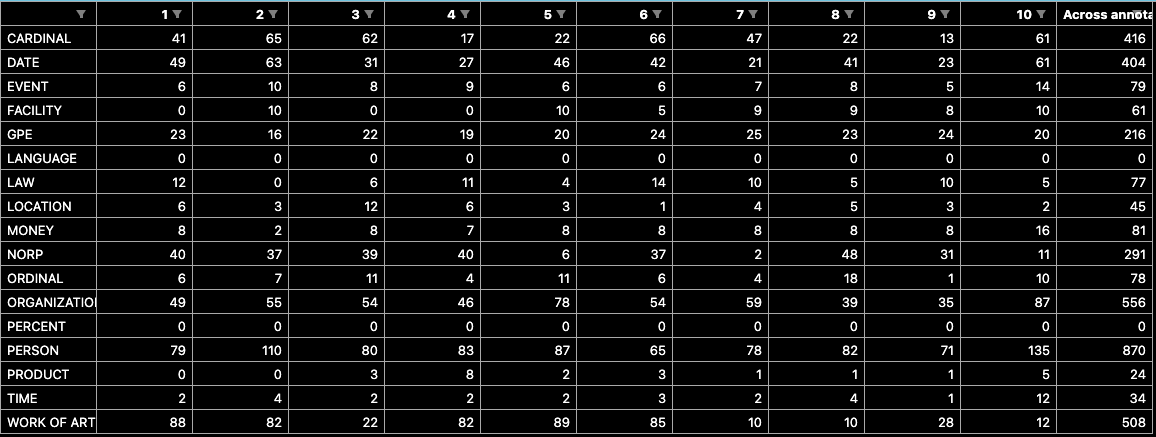


After assigning the appropriate texts to each of the annotators, the annotators were instructed in accordance with the 18 shorthand descriptions developed and utilized by OntoNotes V5 (cite\*). The descriptions of the different Named-Entity types can be seen in table X\* depicting the annotation scheme. The texts were annotated through the software Prodigy; a scriptable data annotation tool suitable for this type of work (Cite \* explosion).

## Quality assessment of annotations

Upon having the annotators finishing the annotations, the resulting quality of the annotations were assessed.

Tag counts, per rater in the overlapping data:

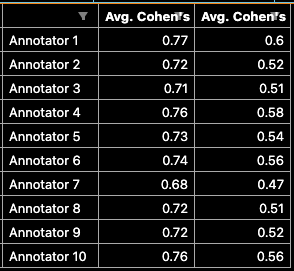


Moreover, to assess inter-rater reliability, Cohen’s Kappa scores were calculated for all annotators, as seen in table X\*. Cohen’s Kappa can be interpreted as a quantitative measure of reliability across raters, that is corrected for rater agreement by mere chance (cite\*).

\*\* INSERT COHEN’S KAPPA FORMULA \*\*

An interpretation of some magnitude in Cohen’s Kappa may be considered arbitrary as it only truly can be contextualized through comparisons to other values for similar tasks. However, as a rule of thumb, values between .40 and .75 have generally been considered fair to good, while values lower or larger have largely been considered poor and excellent, respectively \*cite (JL Fleiss et al (1981, Statistical methods for rates and proportions).

Average Cohen’s kappas per rater table:



To assess the annotation consensus between annotators on a NER-tag level, additional F1-mean scores were calculated for all annotators (see section X\* for a thorough description of the nature of F1-scores). The mean F1-scores for each tag can be seen in table X\*.

F1-tag strict per tag:



The assessment of the combination of Cohen’s Kappa scores, F1-scores as well as a manual inspection of the annotated texts yielded insights that shaped the course of action for the procurement of the DANSK dataset. Generally, the consensus between the annotators was - across the board - quite low. It was therefore deemed necessary for the annotated texts to undergo additional processing before they could be unified into a coherent, high quality dataset. Moreover, the annotation quality of rater 2, 8 and 10 was especially poor. Rater 2 and 10 seemed to have misunderstood the task to some extent and were therefore excluded. Rater 8 had incorrectly annotated pronouns as Named-Entities (such as “[man](PERSON) burde gøre” and “[sig selv](PERSON)”), despite the non-named nature of these pronouns. However, as pronouns often are excluded in the preprocessing for NLP due the limited semantic content, they are included in lists of so-called stop words. All spans that had been annotated from rater 8 were therefore excluded if the span included a word in the stopword list curated by Jens Dahl Møllerhøj that is utilized in spaCy (cite \*). This stopword list can be inspected in appendix X\*..

A final issue was evident in table X\* in which it can be seen that some Named-Entity type annotations were severely underrepresented in the overlapping data. PRODUCT, PERCENT and LANGUAGE had only 24, 0 and 0 annotations, respectively.

To accommodate these limitations, a number of steps would be taken to enhance the quality of the annotations. The annotated texts could be divided up into two distinct groups. The two groups consisted of:

1. Inter-annotated texts, that had overlapping annotations from multiple raters
2. Single-annotated texts, that had only been annotated by a single rater

As the two groups of texts differ on the number of annotations per text, the processing and improvement of the annotations would be done separately.

## Annotation improvement of inter-annotated texts

The inter-annotated texts had an annotation version for each rater that had rated it. However, as the DANSK dataset was to include only unique texts, each text had to have its versions be merged into a single annotated text. As such, any texts with conflicting annotations had to be resolved. The curation of other successful datasets have utilized a single annotator for manual resolvement of conflicts between raters (cite\* OntoNotes). However, relying on a single individual for resolving conflicts manually naturally skews the annotations towards the opinion of a single annotator, rather than the general consensus across raters. In order to resolve conflicts while diminishing this skew, an automated procedure would be employed. This automated procedure resolved conflicts in texts with predominant consensus between annotators from a set of rules.

The rules for automated resolvement of conflicts with predominant consensus were as follows:

1. Named-Entity annotations that matched across span and label in >= 50% of raters were accepted, if no other annotations in the same span lived up to the same criteria.
2. Named-Entity annotations that had zero matches on span (regardless of label) were rejected.
3. Only texts annotated by more than 4 annotators underwent the automated procedure of conflict resolvement. This circumvented the issues that any annotation in a text annotated by e.g. 2 annotators would adhere to the 50% of raters and thus be accepted.
4. In cases of two annotations existing within the same document with 50% annotation agreement between annotators, none of them would be accepted. Rather these cases would stay conflicted. E.g. “[President Trump] (PER) is the president”, “President [Trump] (PER) is the president”

\*\*Make graphical depiction of the accept + decline + do nothing rules and what it would result in, name it: “Automated resolvement of conflicts with predominant consensus”\*\*

After employing the automated procedure, the 886 inter-annotated texts went from having 513 (58%) texts with complete rater agreement to 789 (89%). As such, 97 (21%) of the inter-annotated texts had remaining annotation conflicts. The remaining texts with conflicting annotations across annotators were resolved manually. For the manual resolvement, 3 of these texts were of such bad quality that they were rejected and excluded.

Finally, to ensure that any Named Entities of the type “Language”, “Percent” and “Product” had not been missed by the annotators (as conjectured in section X\*), an extra measure was taken. The model TNER/Roberta-Large-OntoNotes5 was used to add these types of annotations to the inter-annotated texts (cite\* <https://aclanthology.org/2021.eacl-demos.7/> ). The model predictions were then manually assessed, but none of the predictions matched the annotation guidelines and were thus not added to the texts.

This step concluded the processing of the inter-annotated texts, which resulted in a total of 883 texts that were added to the DANSK dataset.

## Annotation improvement of single-annotated texts

The single-annotated texts had only a single set of annotations per text. Since the annotation quality assessment of the inter-annotated texts revealed poor quality as well as low consensus between raters (section X\*, table X\*, etc.), it was assumed that the single-annotated texts suffered the same limitations. Silver annotations (i.e. predictions of a model) have previously been used in the literature to validate gold-standard annotations (cite\*). The logic here being that if a gold annotation matches that of a highly performing model, then the gold annotation is more likely to be correct. However, no Danish datasets nor models employ the fine-grained Named Entity annotations that are utilized here. However, a model matching the annotation schema would be trained by utilizing the inter-annotated texts that had been already been to the preliminary version of the DANSK dataset (see section X\*). A such model would thus add the needed silver annotations. Texts with agreement between model and rater would have a higher chance of being correct, and would on this basis be accepted and integrated into the DANSK dataset. The remaining texts with conflicts between model and rater would be manually resolved and subsequently also integrated into the DANSK dataset.

This method of improving the single-annotated texts thus consisted of:

a) model training on preliminary DANSK dataset

b) model predictions on non-processed texts

c) manual resolvement of conflicts

The preliminary DANSK dataset only included relatively few texts; inadequate for training a well-performing model. To ensure that the model had acceptable performance, 3 models were trained and evaluated, in order to find the best model. All three models were versions of the multilingual model “xlm-roberta-large” that has been found to perform well across languages (cite\* <https://arxiv.org/pdf/1911.02116.pdf> ). Their training differed on the account of the data that they were fine-tuned on, however. Model 1 was fine-tuned on 80% of the preliminary DANSK dataset. Model 2 was fine-tuned on the same 80% + the English OntoNotes v5 dataset. As the English OntoNotes v5 dataset is considerably larger than the preliminary version of the DANSK dataset, it was deemed possible that the second model would perform poorly due to the large representation of the English texts. As such, model 3 was trained on 80% of the preliminary DANSK dataset, that had been upscaled (by duplicating the annotated texts) to reach the size of the English OntoNotes v5, while also being fine-tuned on the English OntoNotes v5. The latter model had the best performance on the remaining 20% of the DANSK gold overlapping texts. This model was then used to add silver annotations for all the texts that had been annotated by rater 1. To account for the low numbers of annotations of the types “Language”, “Percent” and “Product” (see table X\*), TNER/Roberta-Large-OntoNotes5 was used to add these types of annotations to the texts as well (same manner as section X\* which further unravels the use). This annotation process added 38 additional annotations.

Out of the 1412 texts from rater 1, the silver annotations by the models were in agreement in 759 (54%) of the texts. The remaining 653 documents had conflicting annotations. These conflicts were manually resolved and added to the DANSK dataset, except for 43 cases where the texts were of such bad quality that they were rejected and excluded. Ultimately 1370 processed texts from rater 1 were added to the DANSK dataset.

A second iteration of the silver annotation process was carried out. An instance of xlm-roberta-large was trained on the preliminary DANSK dataset that this time not only included the inter-annotated texts, but also the processed single-annotated texts from rater 1. This more than doubled the size of the preliminary DANSK dataset compared to the first iteration. This likely contributed to better model predictions for this second iteration, and is the motivation for only predicting and resolving the texts from a single rater in the first iteration. Rater 1, in particular, was chosen since his annotation standard was highest, and would thus have the best data to include for the model training (see table X\*), in this second iteration. Similarly to before, the annotated texts were then duplicated until reaching the size of the OntoNotes v5 and a model was trained on a concatenation of the two. This model was then used to add silver annotations for all the remaining inter-annotated texts, i.e. those for rater 3, 4, 5, 6, 7, 8 and 9.

Out of the 13373 single-annotated texts from rater 3, 4, 5, 6, 7, 8, 9, the silver annotations by the model were in agreement in 5502 (41%) of the texts. The remaining 7871 documents had conflicting annotations. These conflicts were manually resolved, apart from 564 texts of poor quality that were rejected and excluded. The 12809 texts that had been processed into an acceptable standard were added to the preliminary DANSK dataset.

## Resolving remaining inconsistencies

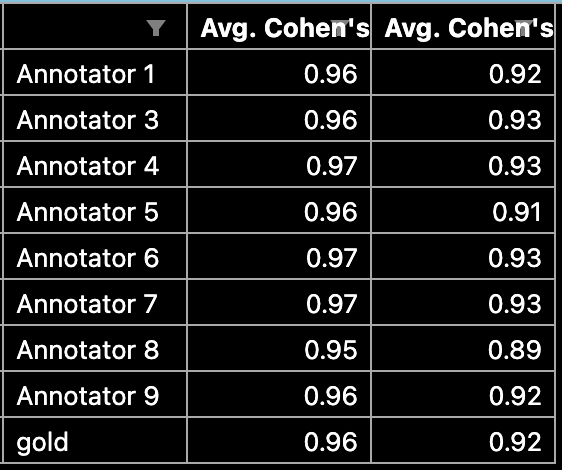
Finally, as the great number of reviews yielded insights in common mistakes for the annotations, all documents were screened using a number of regex patterns (see appendix X\*). If the pattern found a match in a document, the given document was manually reviewed and annotated in accordance with the OntoNotes extended tagging guidelines (cite \* ) and the newly developed additional set of rules that clarify any remaining ambiguousness for Danish (appendix X\*). The regex pattern search yielded matches in 449 texts, all of which were then manually assessed and any potential inconsistencies with the annotation schemes were resolved.

## Partitioning

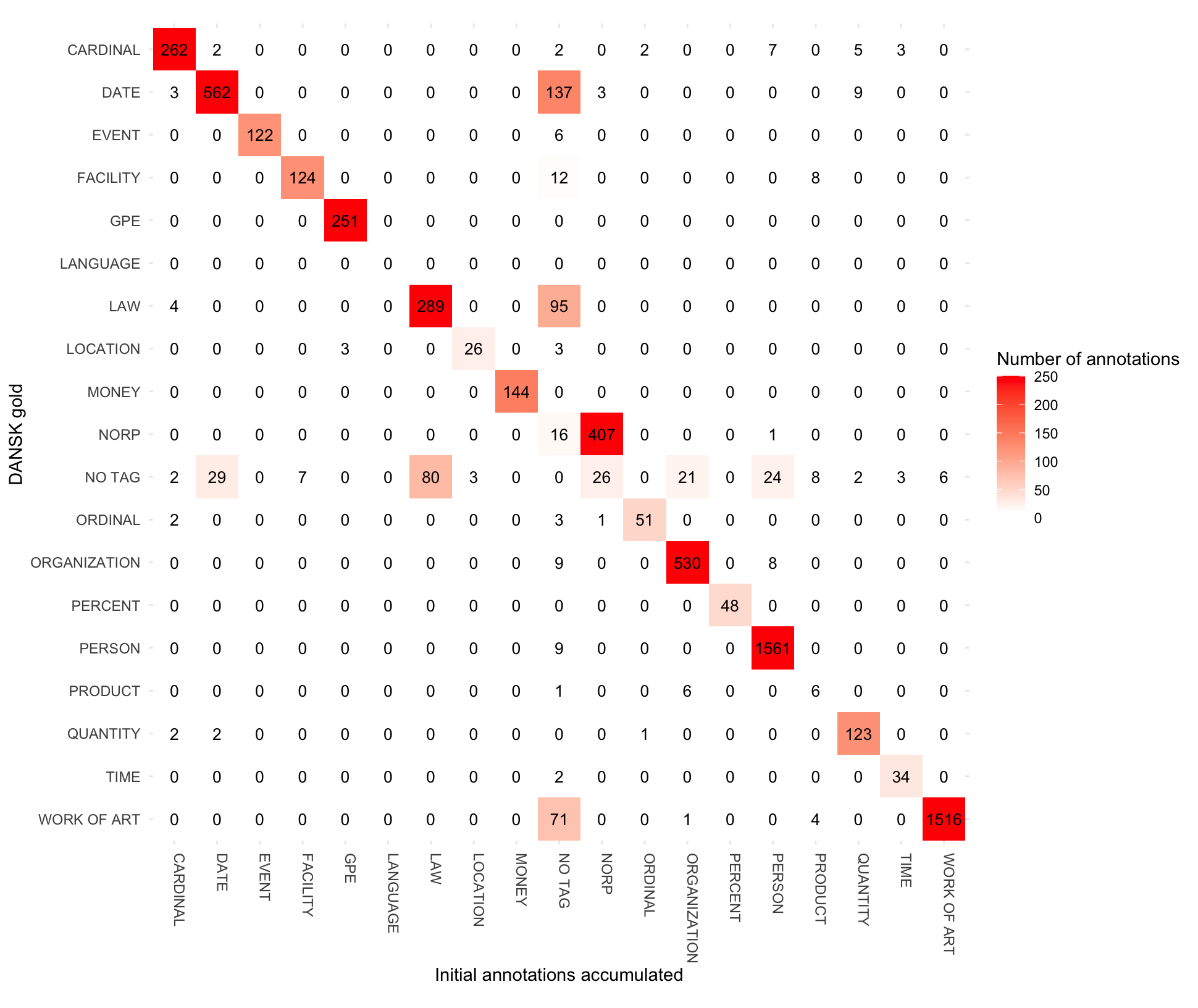
After resolving the remaining inconsistencies, the full DANSK dataset was randomly split up into 3 subsets; a training set consisting of 80% of the texts, as well as a validation and a test set that each held 10% of the remaining texts.

## DANSK quality assessment

Kappa with streamlined multi:



Gold vs. rest confusion matrix:



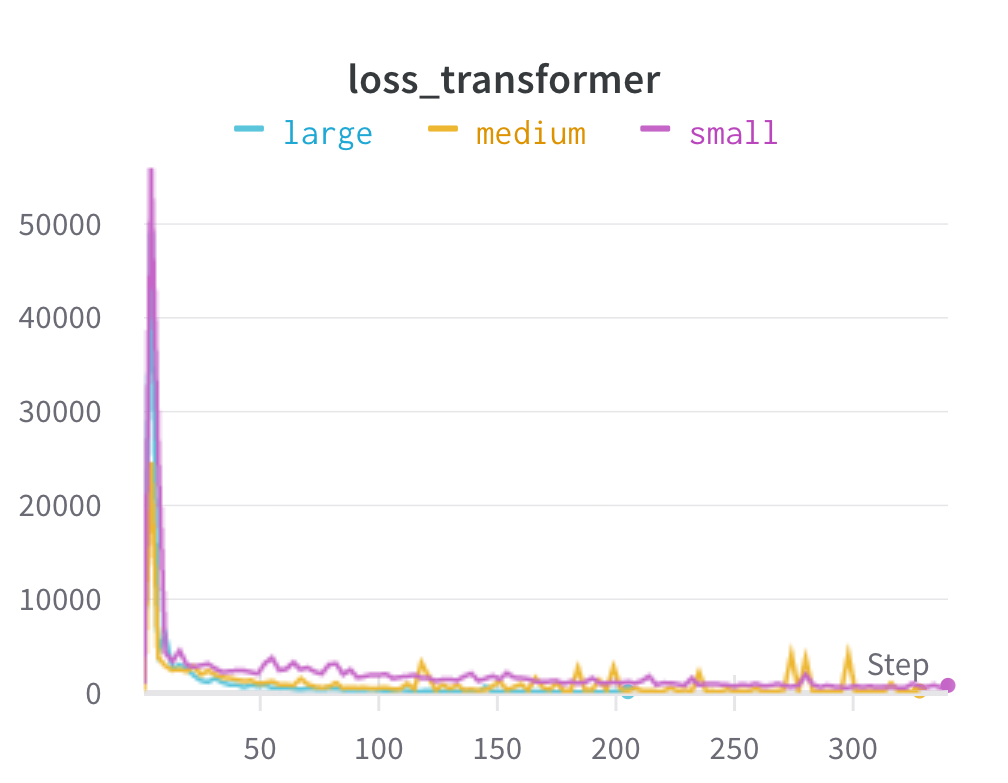
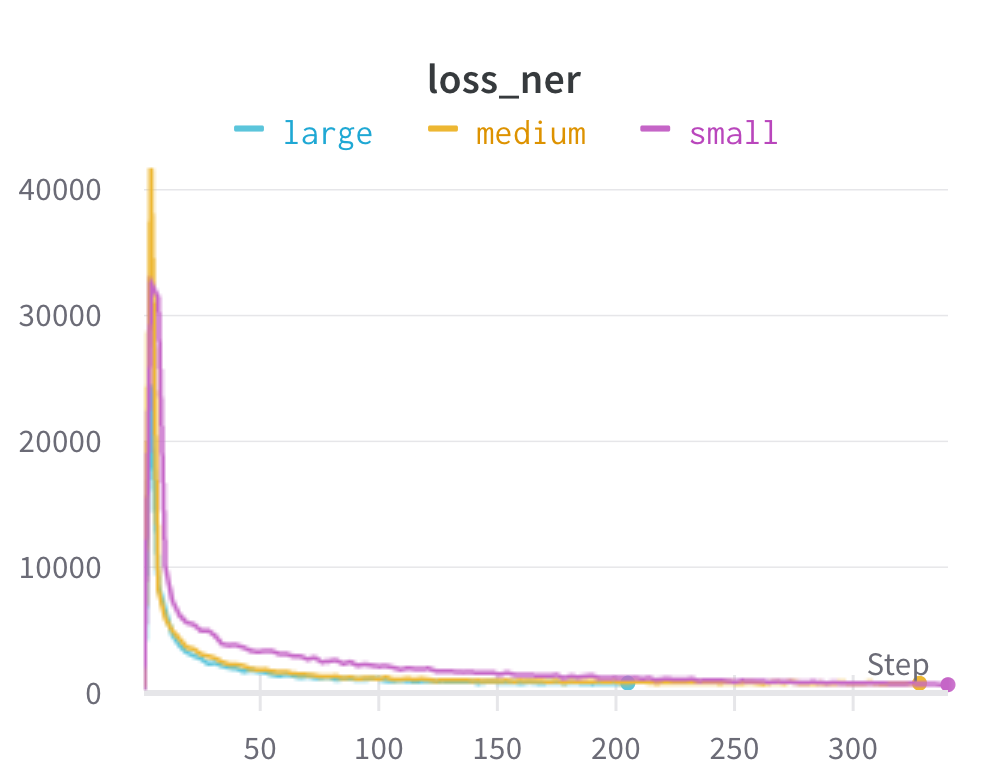
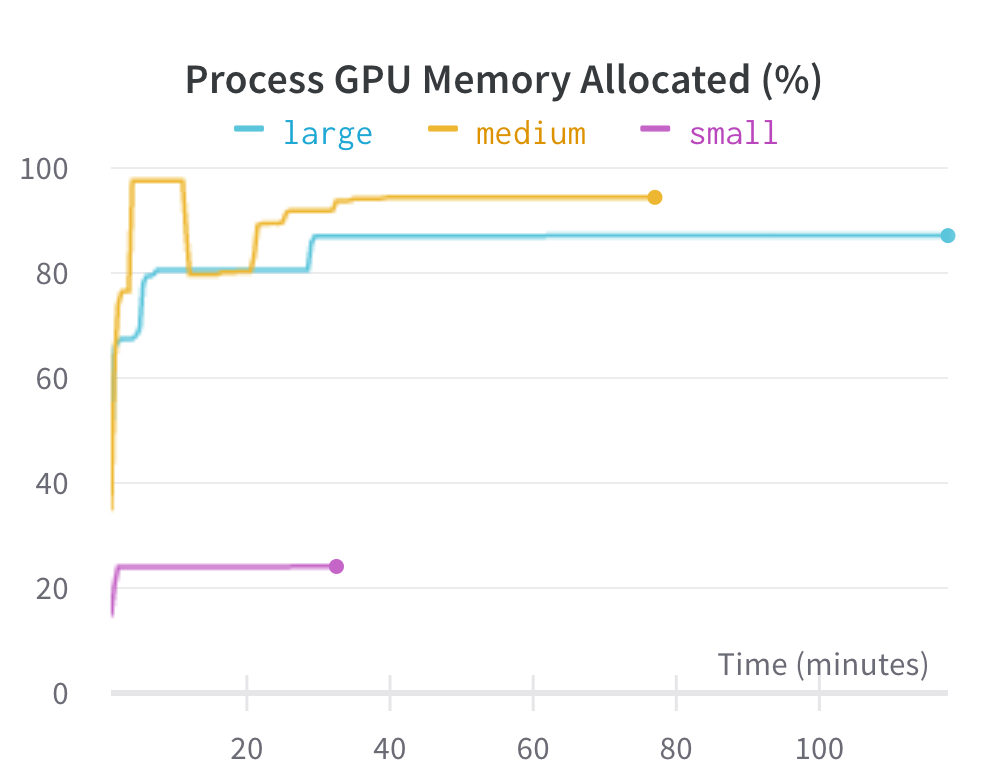
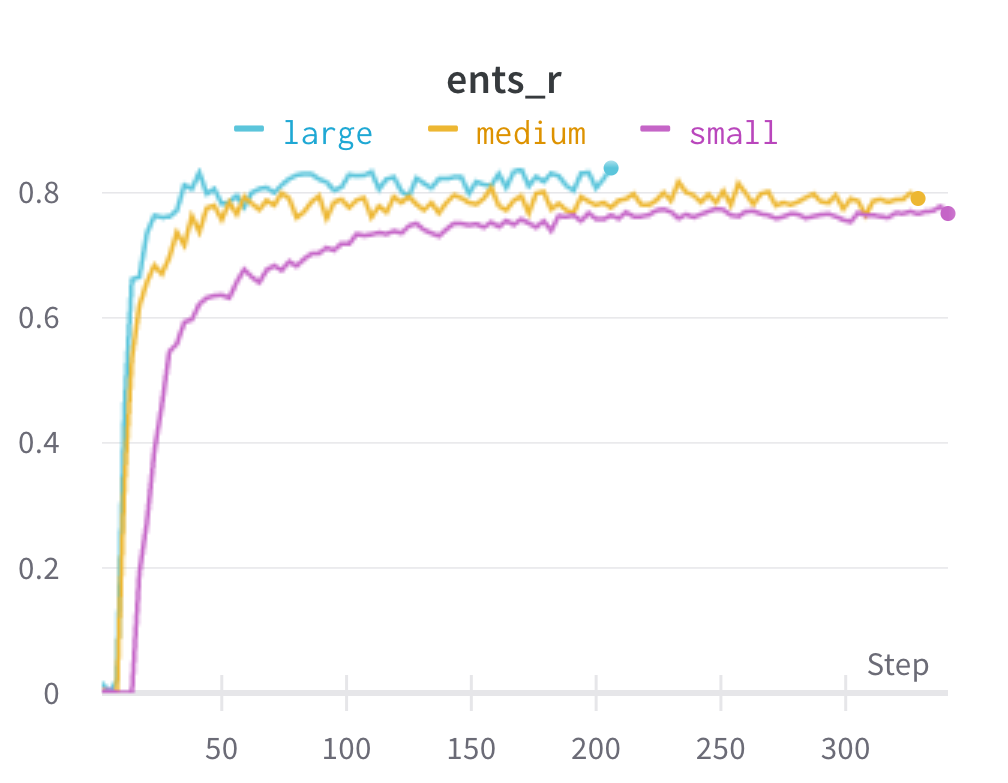
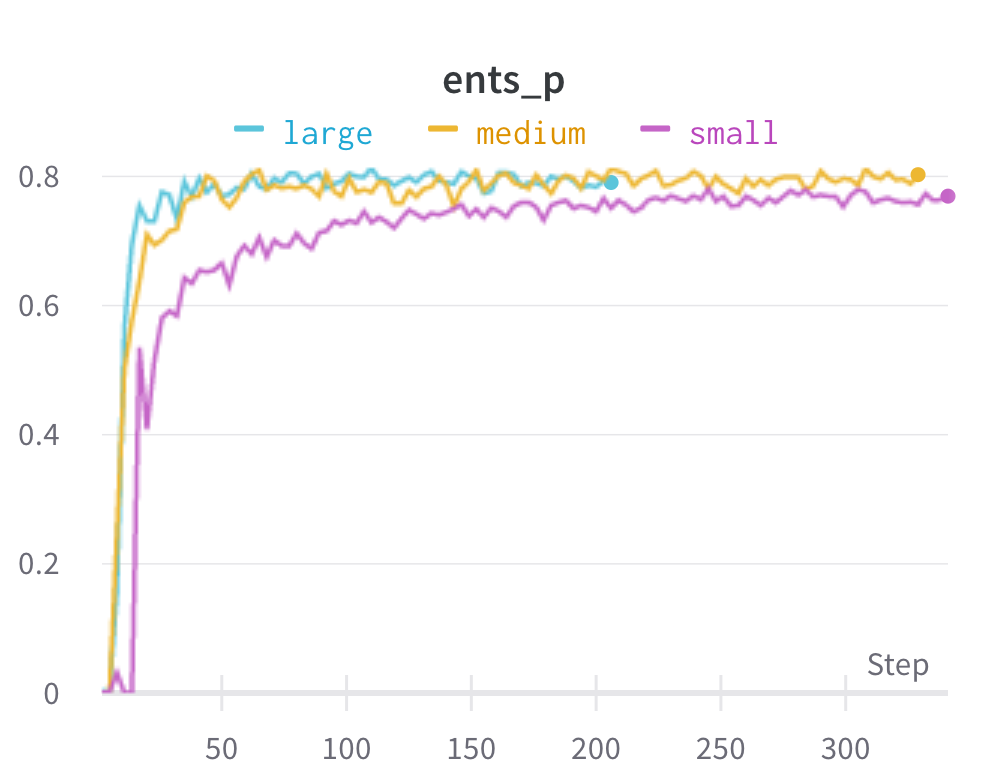
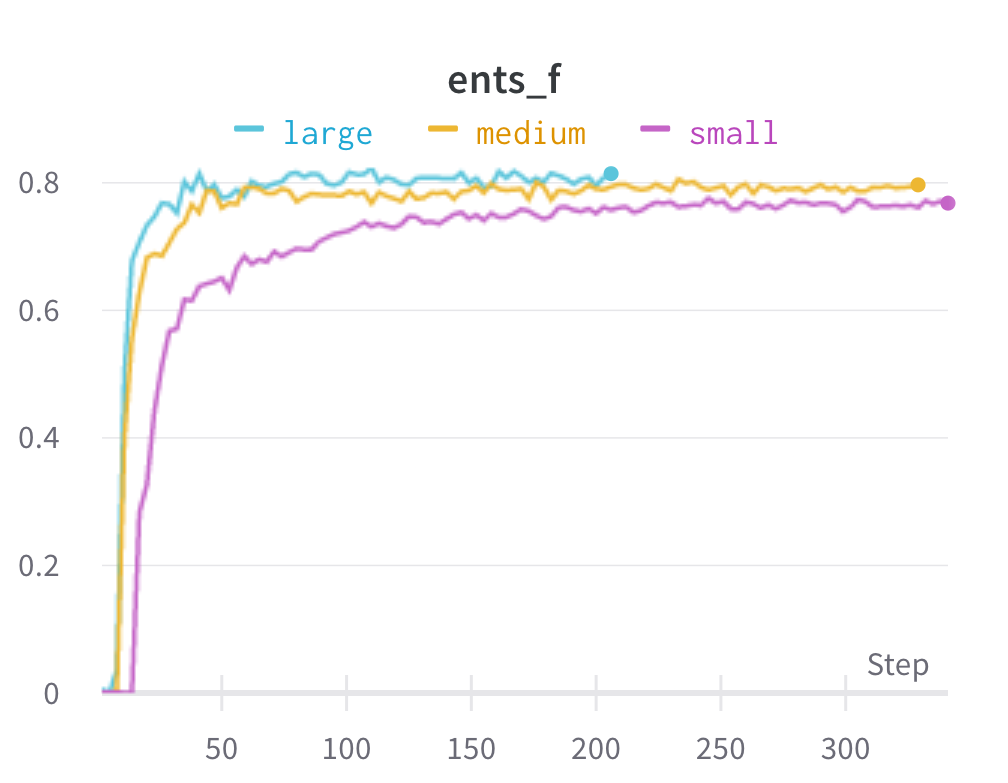
# DaCy model curation

## Methods

### Model specifications

1. Full model specs:
   1. daCy/ner\_extended/configs/
2. Transformer model:
   1. jonfd/electra-small-nordic
   2. NbAiLab/nb-roberta-base-scandi
   3. chcaa/dfm-encoder-large-v1
3. All had the hyperparameters:
   1. except large that had accumulate gradient=3
      1. # Avoids CUDA being out of memory. torch.cuda.OutOfMemoryError: CUDA out of memory. Tried to allocate 78.00 MiB (GPU 0; 14.76 GiB total capacity; 13.48 GiB already allocated; 52.75 MiB free; 13.73 GiB reserved in total by PyTorch) If reserved memory is >> allocated memory try setting max\_split\_size\_mb to avoid fragmentation. See documentation for Memory Management and PYTORCH\_CUDA\_ALLOC\_CONF
   2. Optimizers:
      1. @optimizers = "Adam.v1"
      2. beta1 = 0.9
      3. beta2 = 0.999
      4. L2\_is\_weight\_decay = true
      5. L2 = 0.01
      6. grad\_clip = 1.0
      7. use\_averages = false
      8. eps = 0.00000001
   3. NER head:
      1. @architectures = "spacy.TransitionBasedParser.v2"
      2. state\_type = "ner"
      3. extra\_state\_tokens = false
      4. hidden\_width = 64
      5. maxout\_pieces = 2
   4. Gradient descent:
      1. Dropout = 0.1
      2. Max\_steps = 20000
      3. Warm\_up\_steps = 250
      4. Initial LR = 0.0005
4. Trained on GPU
   1. # Setup for initial training of models for Virtual Machine
   2. # System: image based on Cuda + Jupyter on Ubuntu v. 20.04
   3. # GPU: uc-t4-1 / uc-t4gpu # Access: UCloud, Aalborg University
5. For the progression of loss of the NER head, loss of the transformer, NER performance measured in recall, precision, f1-score, add WANDB plots here. Potentially later move to appendix.

Training progression:



### Domain and entity-level performance

1. On which levels were the models evaluated on?

### Metrics

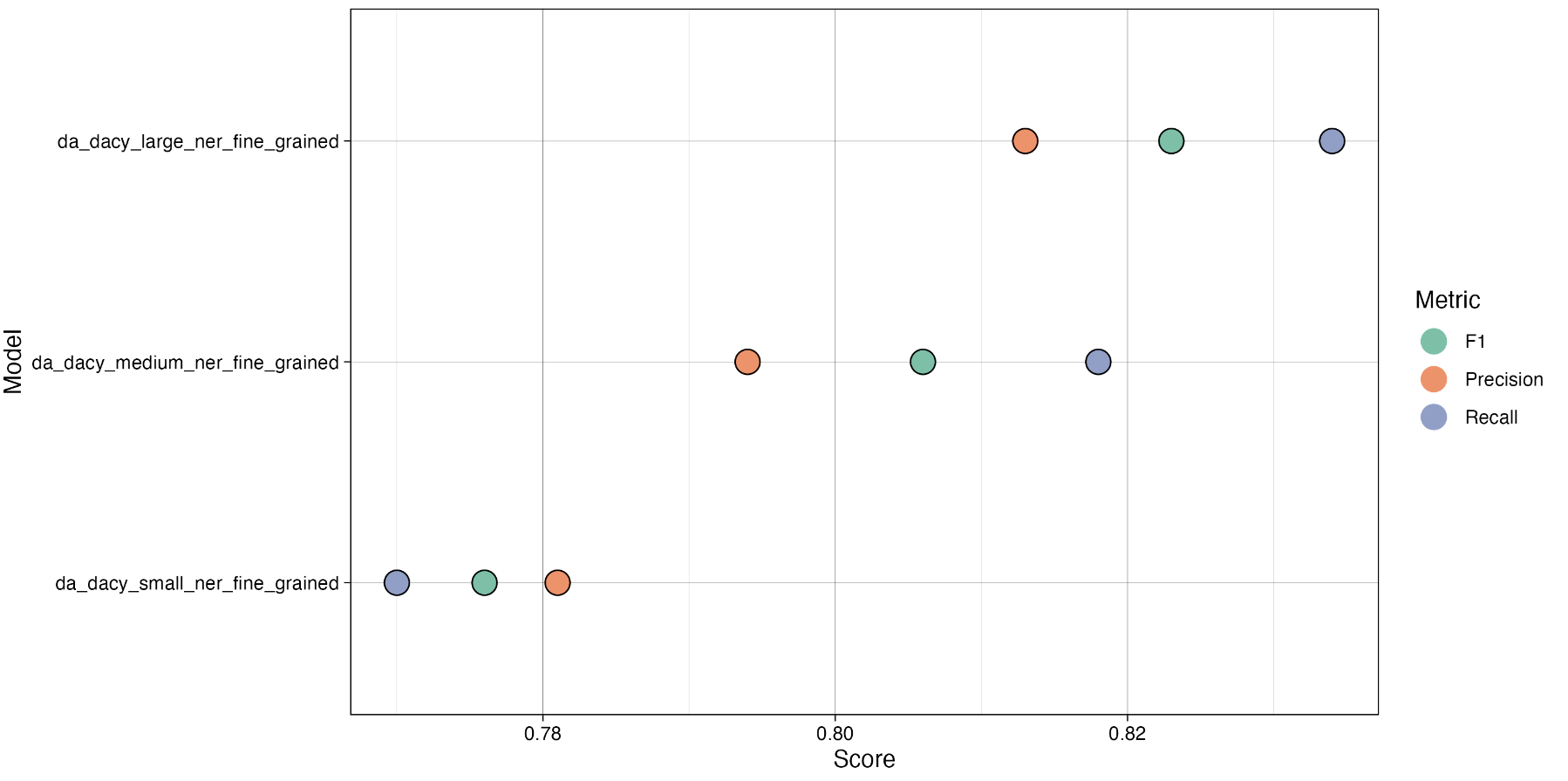
1. Recall,
2. Precision
3. F1-score
4. Steal a bit from previous exams, e.g. bachelors.

## Results

Overall F1-scores (and recall and precision):



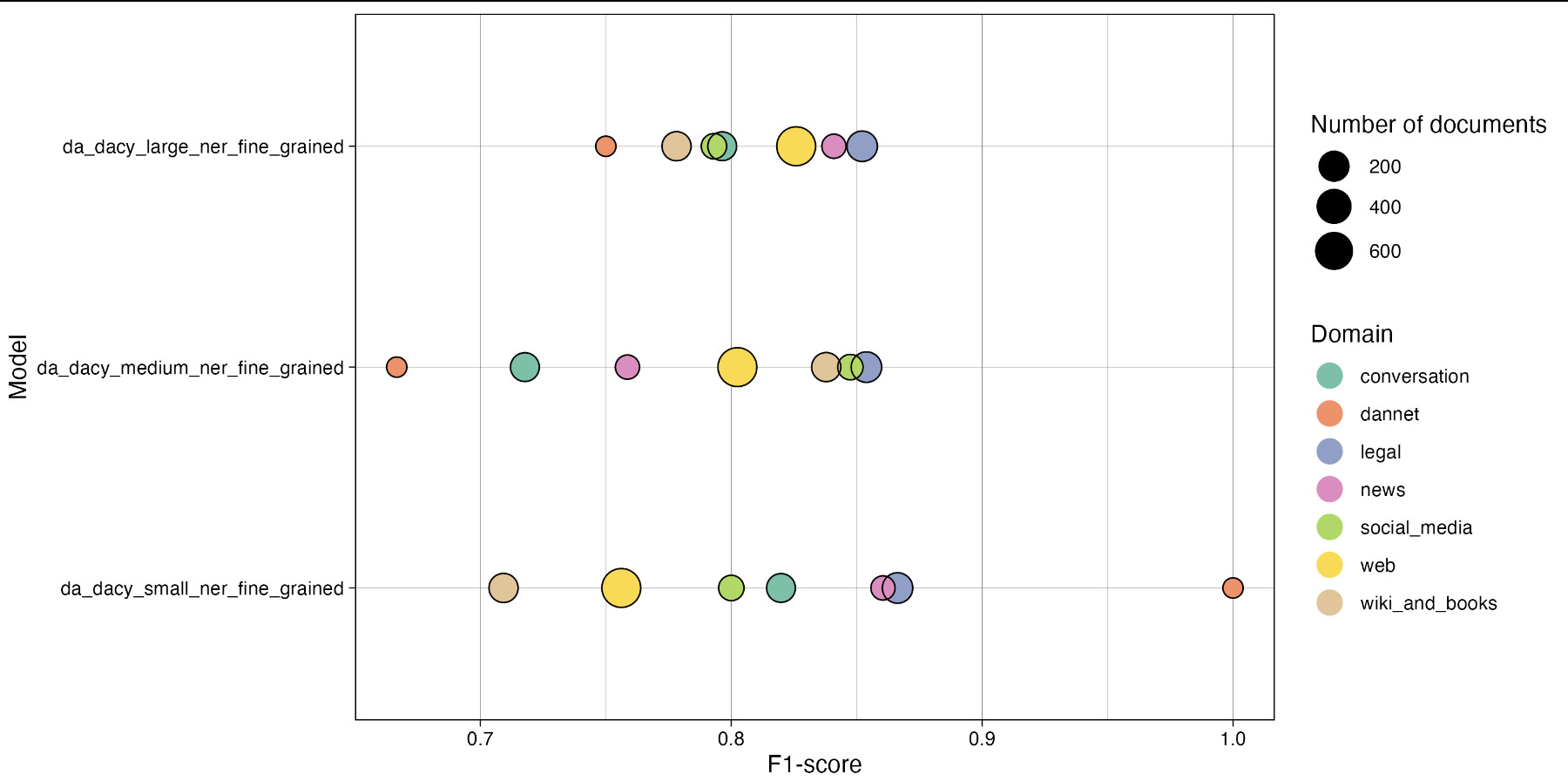
Overall F1-scores (and recall and precision) plot:



Domain F1-scores (wide format) table:



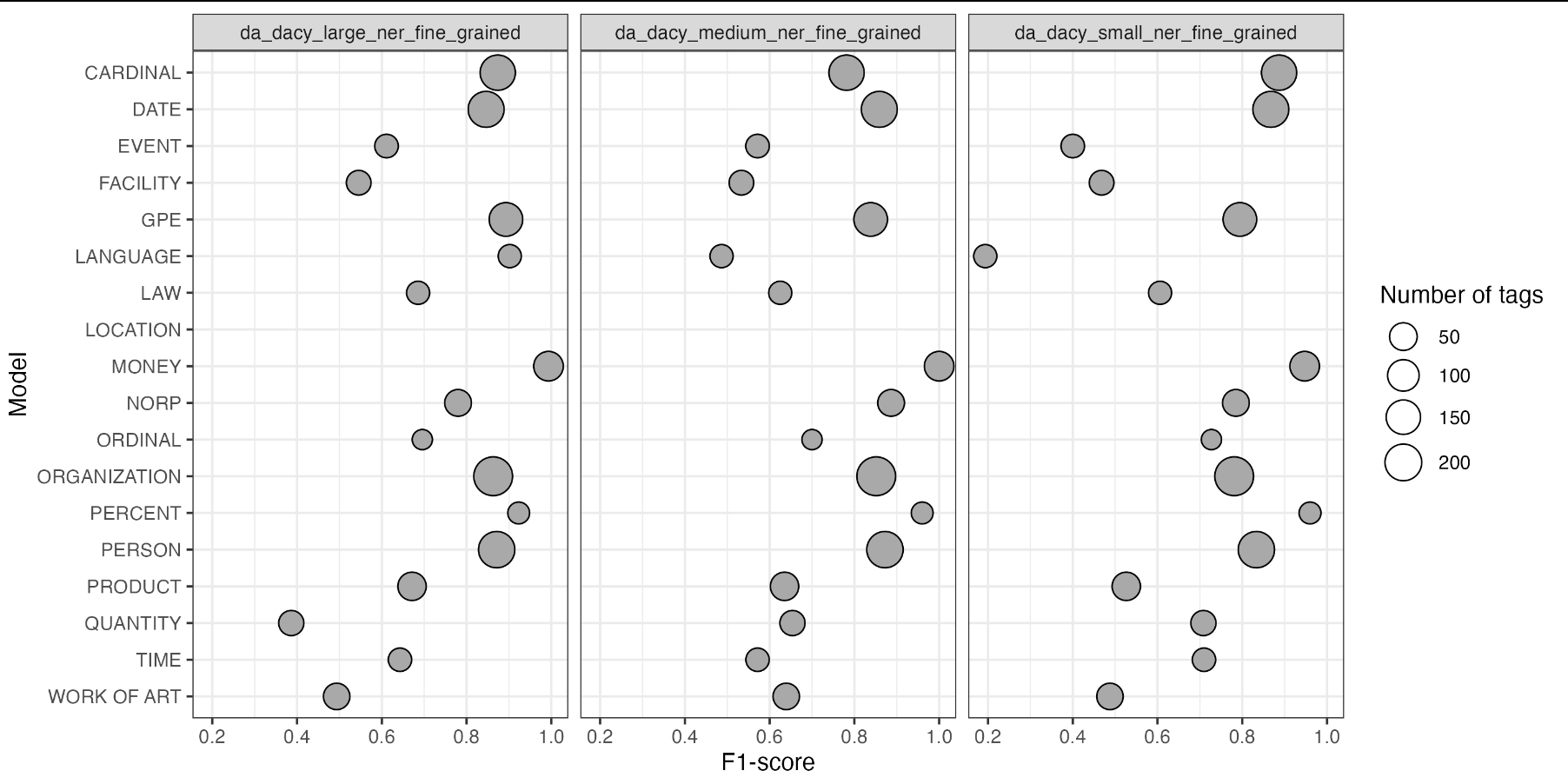
Domain F1-scores plot:



Tag F1-scores



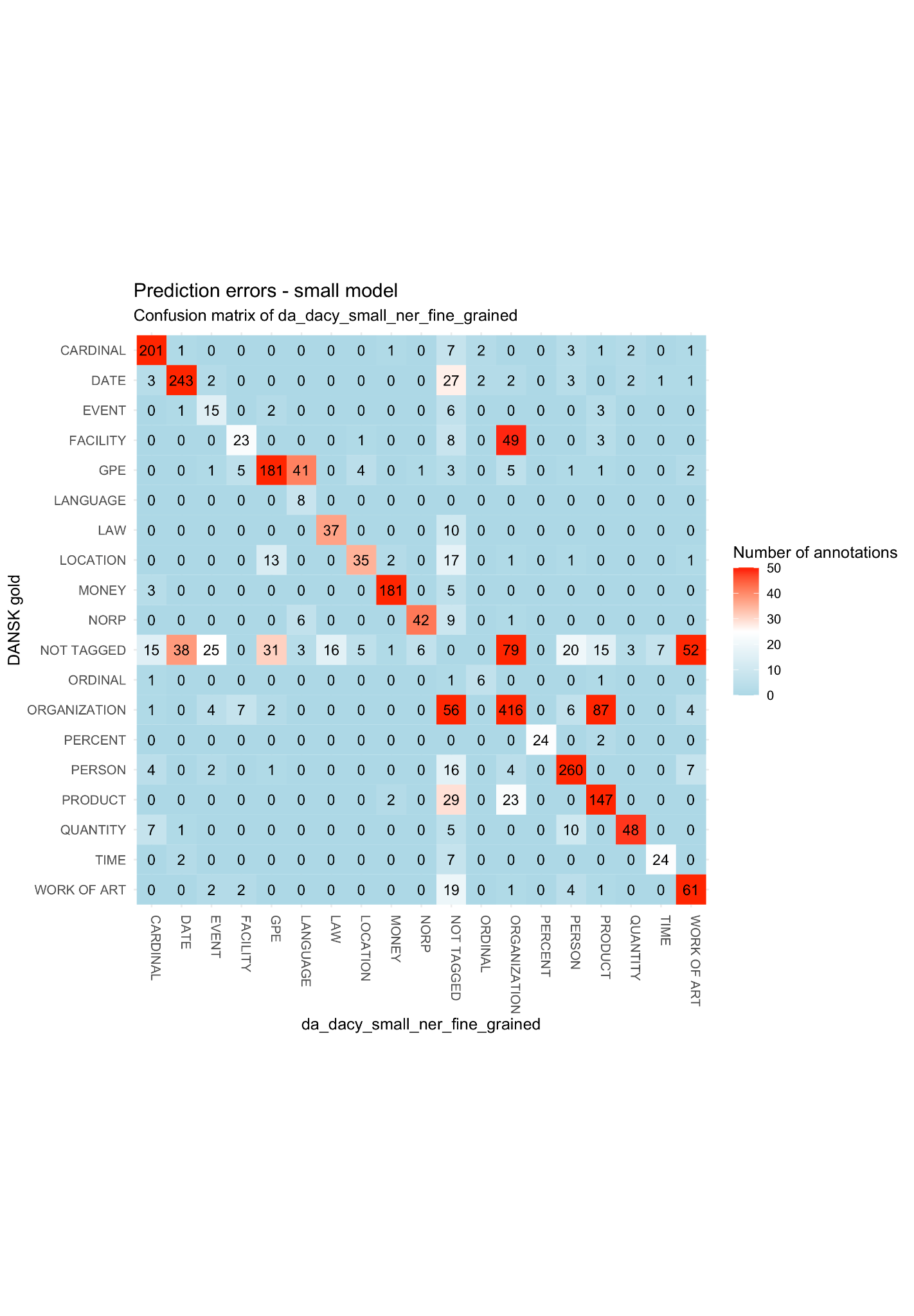
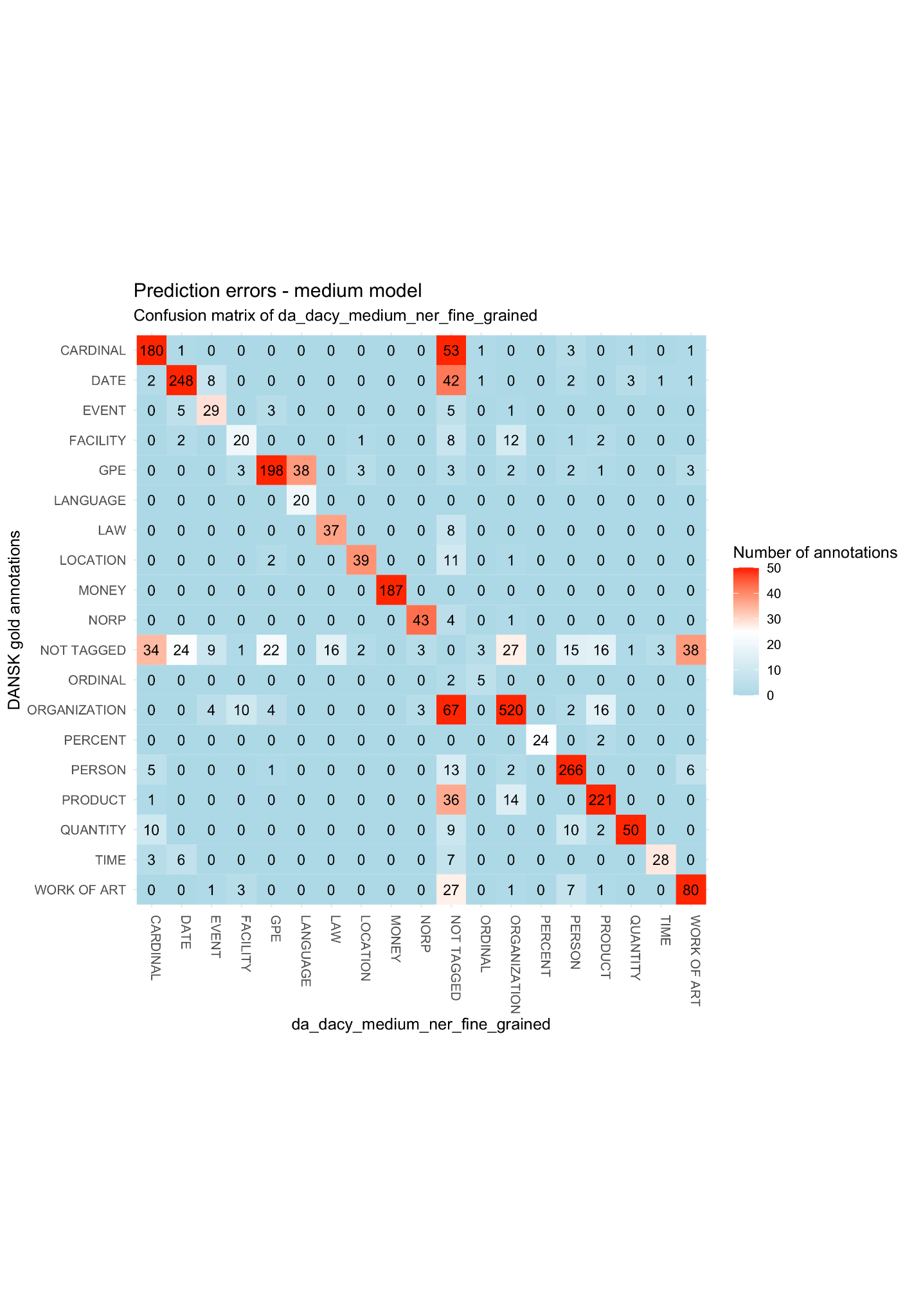
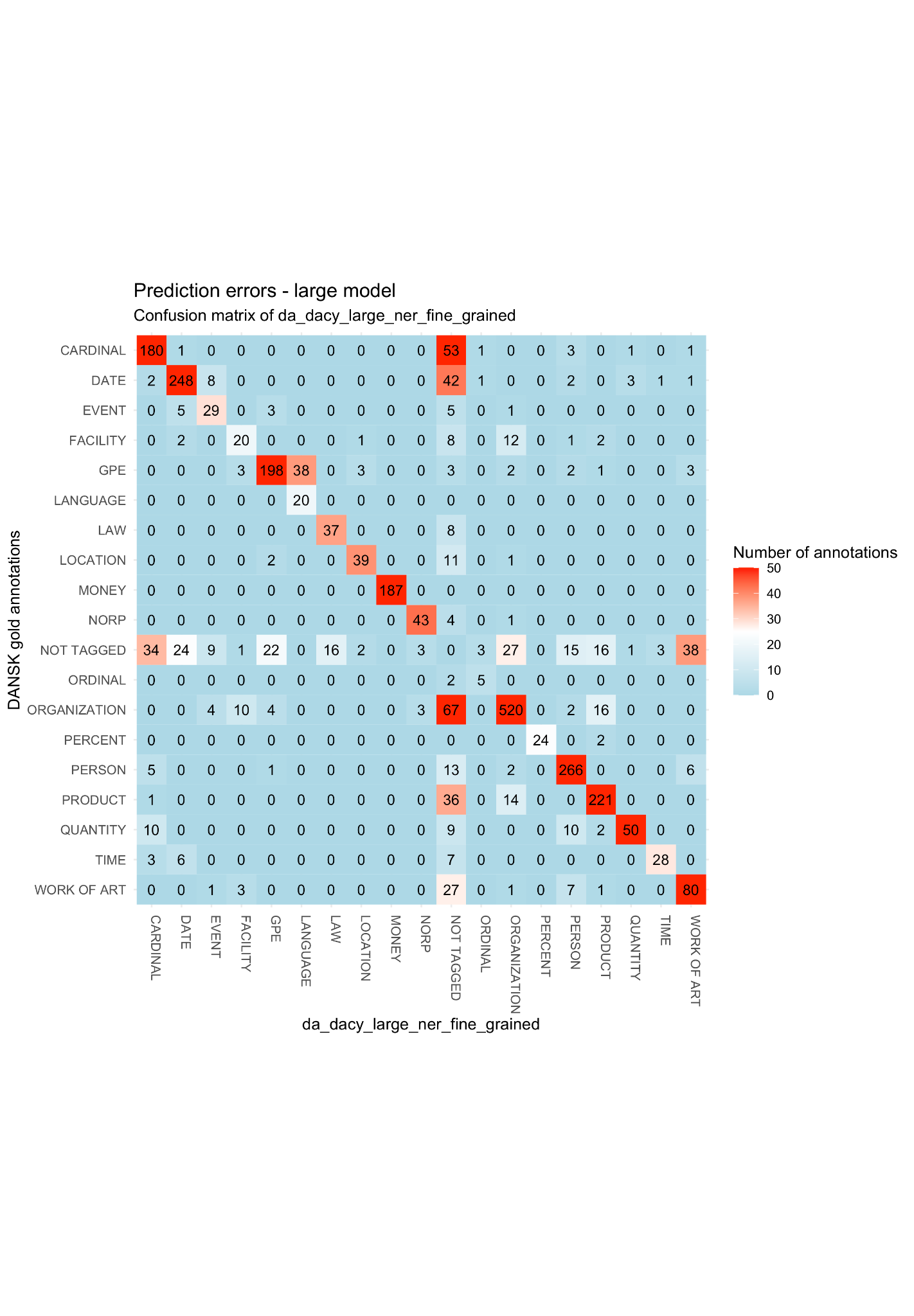
Tag f1-scores plot



Type F1-scores



Confusion matrices:



# SOTA generalizability

## Methods

### Models

1. Which models are SOTA

### Entity label transfer

1. Translation of condensed tag set to fine-grained tag set for DANSK use in SOTA models

### Domain and entity-level performance

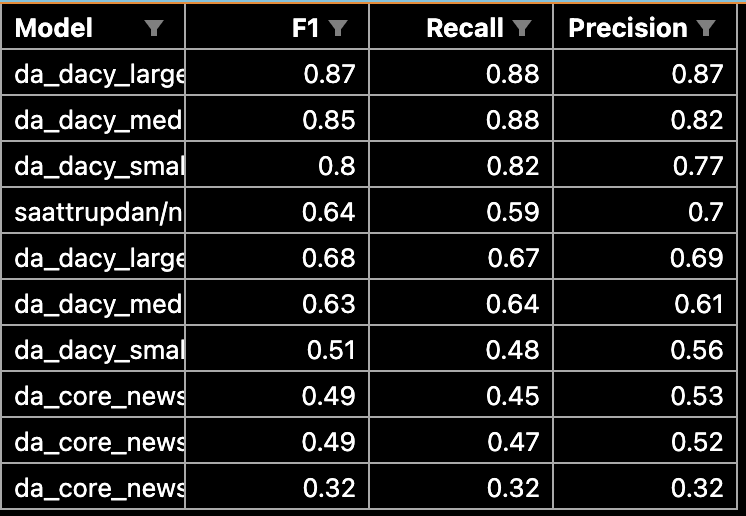
1. On which levels were the models evaluated on?

### Metrics

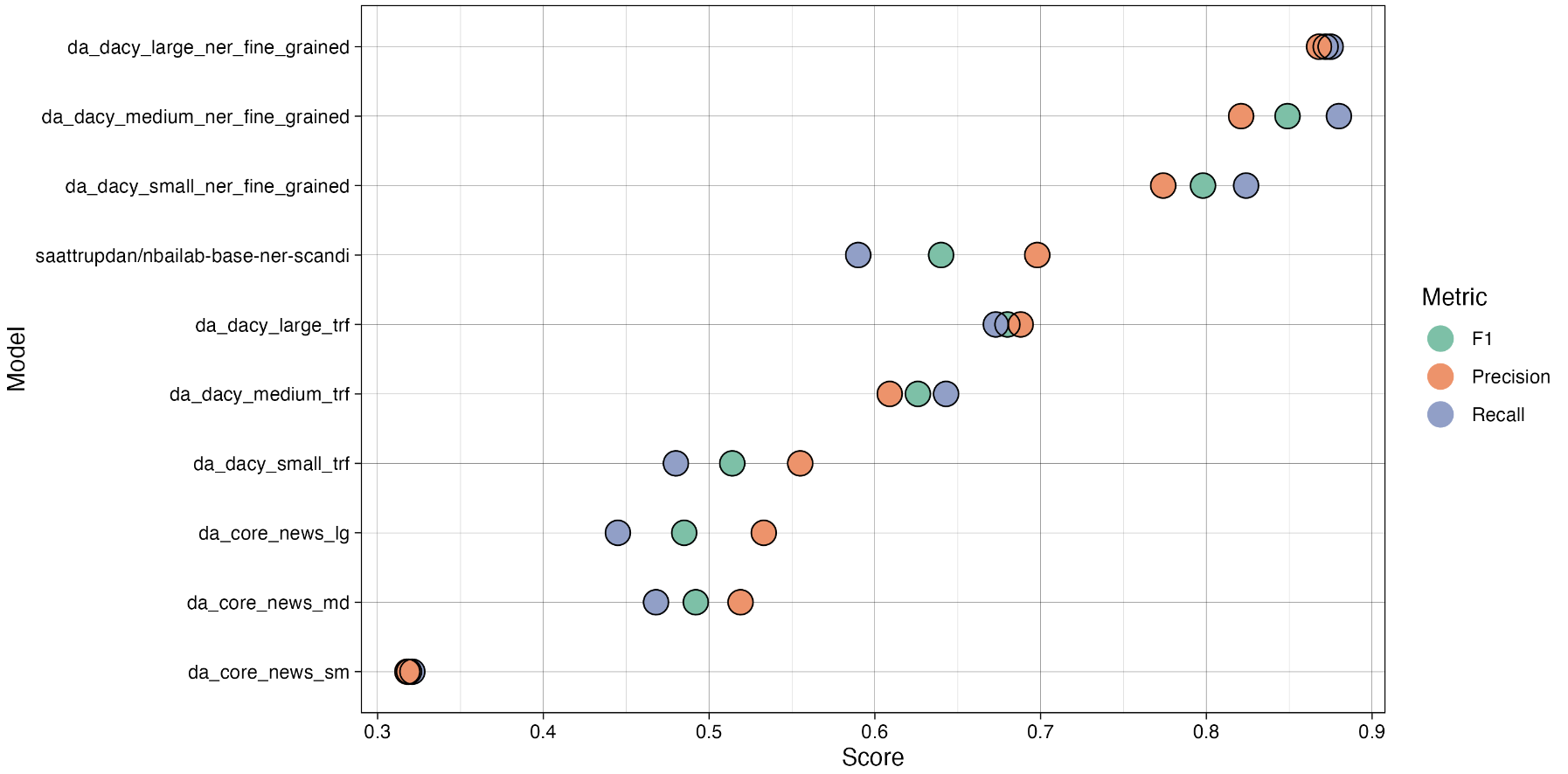
1. Recall,
2. Precision
3. F1-score
4. Steal a bit from previous exams, e.g. bachelors.

## Results

Overall performance:



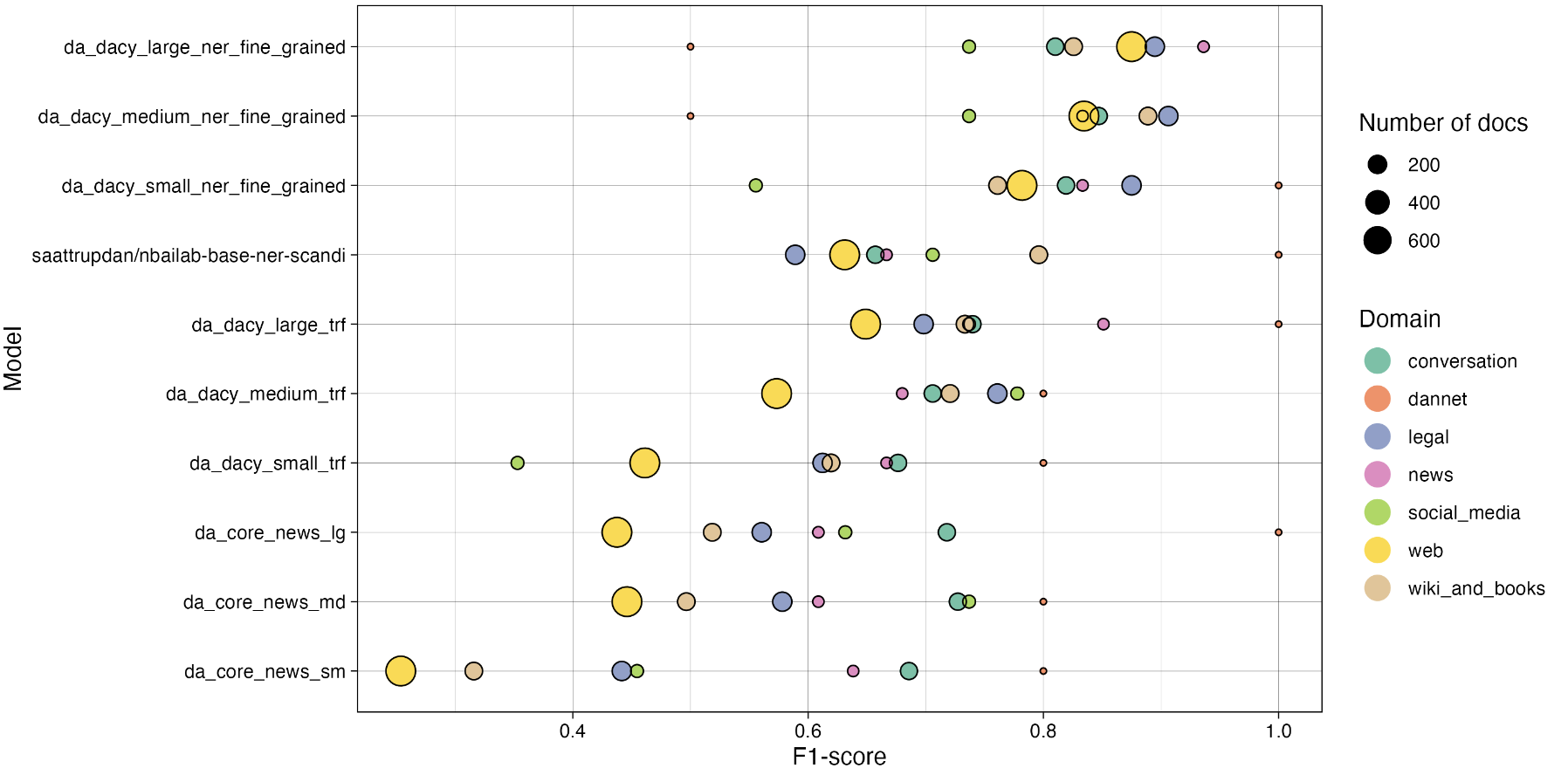
Overall performance plot:



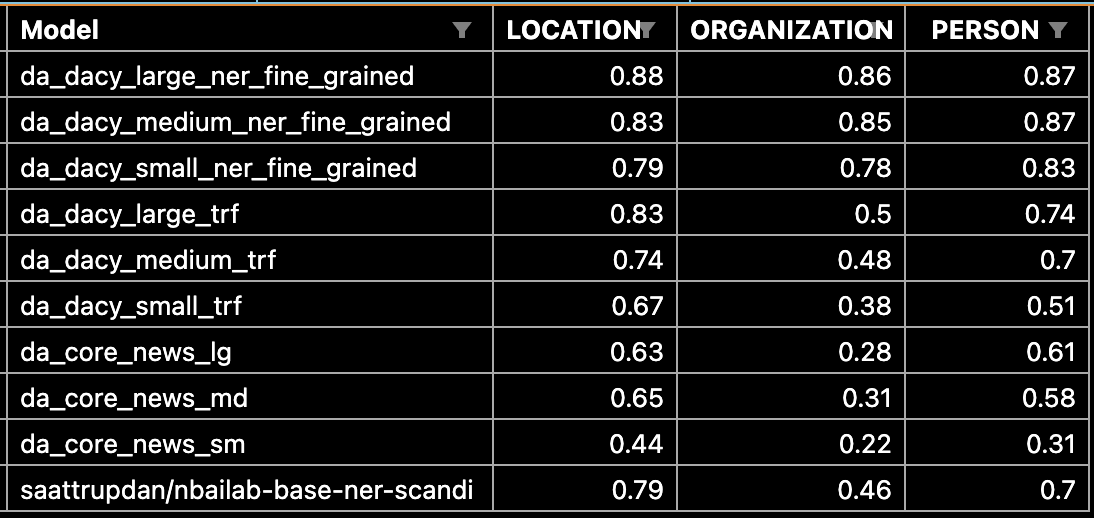
Domain F1-score



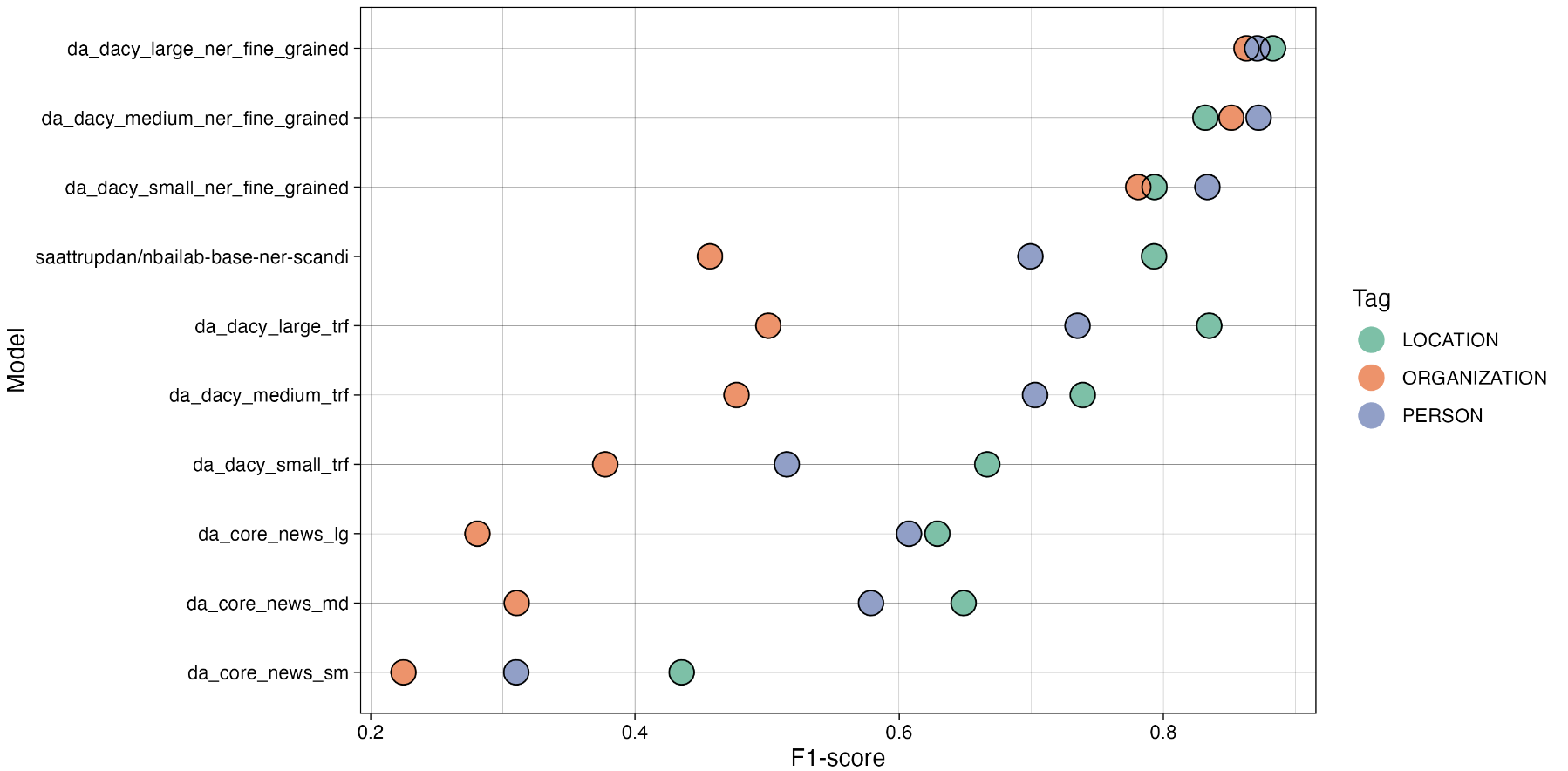
Domain F1-score plot



Tag F1-score

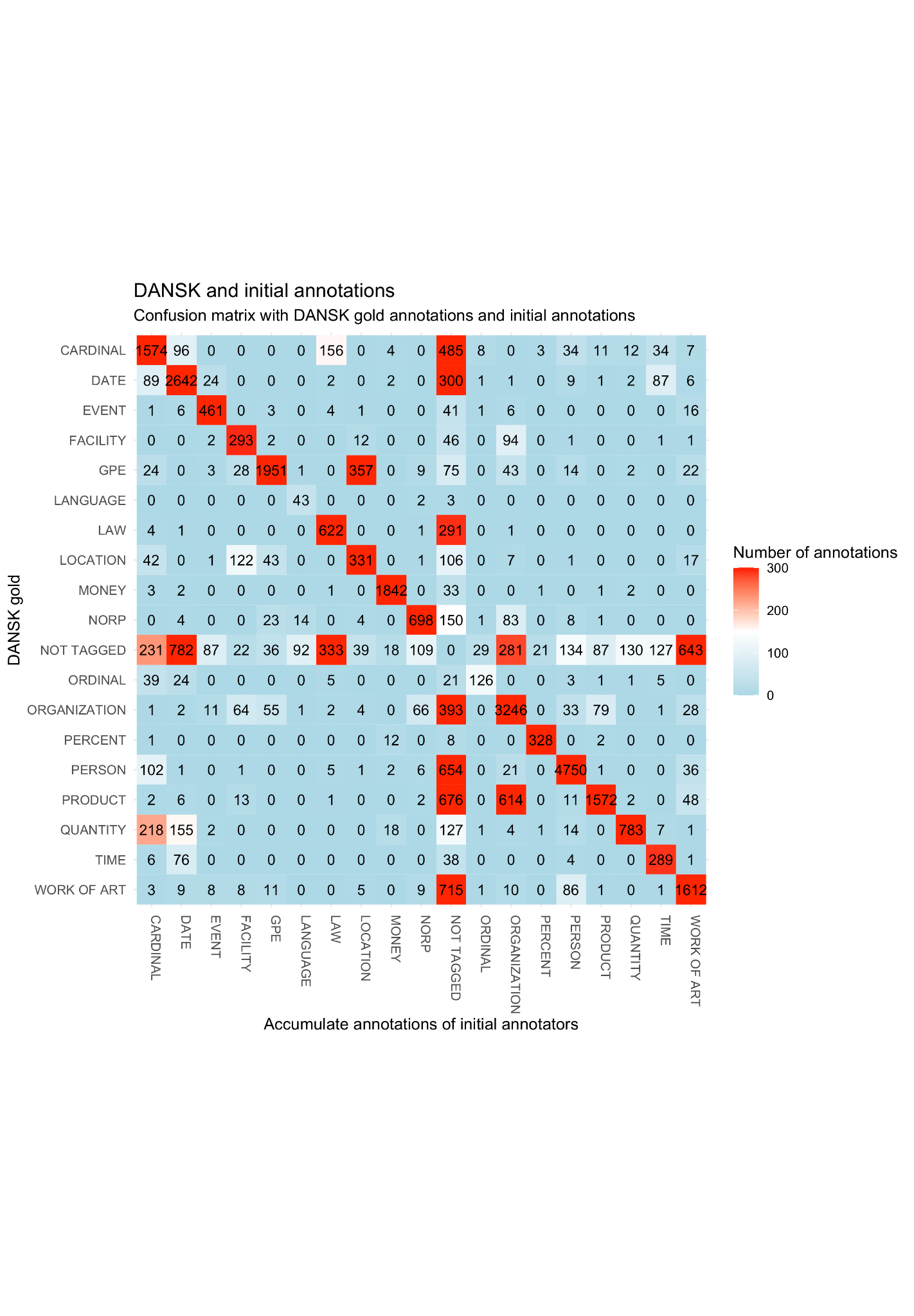


Tag F1-score plot:



# Discussion

## DANSK dataset limitations

1. Initial annotation issues:
   1. Bad sampling
      1. Many texts randomly sampled from Gigaword Corpus were of suboptimal quality or proper texts but in which it is very hard to properly tag. Amongst it, it included even:
         1. Dutch texts "\textit{Behoud van de waddenzee og nederlandse vereniging tot bescherming van vogels.}"
         2. Texts with information on hardware products (find eksempel)
      2. Not evenly sampled across domains
   2. The annotators were not properly trained:
      1. Only given access to the shortened description of the classes (not given access to the full ontonotes guidelines)
      2. No training period
   3. As such, annotations were poor.
      1. Refer to table: “Average Cohen’s kappas per rater table:”\*
      2. Show/refer to examples of bad annotations from annotators
      3. 
   4. Measures that should have been taken in the initial stages of the sampling and annotation process:
      1. Sampling:
         1. The sampling should have been carried out evenly across domains
         2. Low quality texts should have been discarded from the sampling
      2. Annotators:
         1. Should have been instructed to discard bad texts
         2. Should have had a trial/training period
         3. Should have been given access to the full description of the entity classes
2. Preprocessing issues
   1. Any issues with the dataset stem from poor initial quality prior to preprocessing. As a result, the preprocessing did not successfully build a dataset of the highest quality, despite its elaborateness.
   2. Regardless, it would be possible to mend and patch up the dataset in new versions using the following process to further enhance the quality of DANSK
      1. Further preprocessing should consist of:
         1. Train a model
         2. Use model to predict
         3. Manually go through errors
         4. Correcting cases in which the dataset was annotated incorrectly
      2. Repeating steps 1 through 4
      3. Adding POS, Dependency Parsing, etc.
3. Other less urgent issues included:
   1. The Ontonotes guidelines are not adequate and moreover not directly suitable to Danish
      1. "The" may never be included (but are included in some dates)
      2. Contact information may never be included (but full addresses are included)
      3. Cases where in Danish words are combined into single tokens, as opposed to English. Which makes some tokens impossible to tag in the same manner as OntoNotes (or does it?) see if there are examples in "new-resolved-cases"
      4. Mention all the cases where I made my own rules (use regex\\_filter.py and new-resolved-edge-cases.txt)
   2. Would have been better to use a method such as the one described in the paper "A Joint Named-Entity Recognizer for Heterogeneous Tag-sets Using a Tag Hierarchy"
   3. Choice of partitioning;
      1. Could have used certain packages that allow for stratifying by domain and ents, rather than just doing a random split
      2. Could have partitioning so that the test had an equal number of texts from each domain. This way the uncertainty of performance on e.g. Dannet would have been better.
4. Regardless of limitations, DANSK is still very useful:
   1. Now
   2. With new versions of the dataset

## DaCy models

1. Generally
   1. Good performance, given the nature of the DANSK dataset
2. Error analysis (include heatmaps/plots, other)
   1. Within domains
   2. Within sources
   3. Within ents
3. Performance issue more likely to stem from data inconsistencies, rather than poor model training
   1. Even the case for perfect datasets (see paper: <https://nlp.stanford.edu/pubs/CICLing2011-manning-tagging.pdf>)
   2. Refer to table with interrater reliability in the end
   3. Refer to conf matrix plots for models
4. Include error analysis of model (using the heatmap from interrater reliability script

## SOTA models and generalizability evaluation

1. Poor performance generally
   1. Some models did better than others
      1. Cross-linguistic best?
   2. Explanation of why
2. Poor performance on certain domains
   1. Which
   2. Explanation of why
3. Not possible to get an accurate estimation of generalizability, as:
   1. The dataset is not perfect
   2. The old tagging system and the new do not naturally transfer
   3. The findings can therefore only be used as an indication of a generalizability problem

## Product usage and further research

1. New model performance on “old domains”?
   1. Other tagging system generates problems
   2. Perhaps this generalizability comes at a cost in performance on more conventional texts?
2. Crazy performance if also trained on English Ontonotes -> Like sattrupp/ model.
3. Error analysis would have been very interesting to look at
4. Barbara Planck version with nested is nice, and could have been nice to implement
5. E.g. the scandinavian models that are cross-trained on 3 languages, but seem to perform better on each individual language as well.
6. How will DANSK and DaCy fine-grained models help society

*Maybe include somewhere in the rest of the Discussion in the end if need more things to write about. These are relevant notes, don’t just delete!!!:*

1. *Error analysis (interrater reliability script, within each domain and across)*
2. *Still room for improvement in performance. Worth making an error analysis of the type used in https://nlp.stanford.edu/pubs/CICLing2011-manning-tagging.pdf.*
3. *Although not evaluated, the NER component will likely underperform on longer texts as the model has mostly been trained on 1 sentence at a time. They implicitly learn the entities only occur in the first sentence of a text (since they never occur later, as there are no later ents). This tendency is especially apparent in transformer models such as DaNLP, BERT and NERDA (see table in DaCy paper on performance under 5 and 10 sentences). Most likely this will not be as big an issue for the other components, as the annotated documents have been merged to chunks of 10 sentences. This was, however, not possible for DANSK, as the extensive review process meant that only parts of longer texts were annotated in isolation. Thus the out-of-sentence context was not used in the annotation.*
4. [*https://nlp.stanford.edu/pubs/CICLing2011-manning-tagging.pdf*](https://nlp.stanford.edu/pubs/CICLing2011-manning-tagging.pdf)

# Conclusion

# Appendix

1. Cohen’s kappa interrater reliability FULL table(s) X 4
   1. Only annotated/also-non annotated,
   2. Initial/End
2. Any plots or tables that do not fit in the text
3. Extra annotation rules
   1. /Users/emiltrencknerjessen/Desktop/priv/DANSK\_eval/output/Thesis/Annotation\_rules
4. Regex Patterns
   1. /Users/emiltrencknerjessen/Desktop/priv/DANSK\_eval/output/Thesis/Regex\_patterns
5. Stop words

# Backup

1. Model training
   1. 20% dev set
   2. 3 models:
      1. Overlapping\_resolved + Annotator\_1\_resolved
      2. Overlapping\_resolved + Annotator\_1\_resolved + OntoNotes
      3. **Upscaled\_overlapping\_resolved + upscaled\_annotator\_1\_resolved + OntoNotes**
2. Model prediction on annotator 3-9
   1. Descriptive stats
      1. Total (rater 3 - 9, excluding rater 1 and overlapping): # 13373
      2. Agreement: # 5502 (excluding rater 1 and overlapping)
      3. Disagreement: # 7871 (excluding rater 1 and overlapping)
      4. Rater 1 (only included in predictions from previous model, see above, though):
         1. Agreement: # 759
         2. Disagreement: # 653
      5. Rater 3:
         1. Agreement: # 638
         2. Disagreement: # 888
      6. Rater 4:
         1. Agreement: # 1114
         2. Disagreement: # 1363
      7. Rater 5:
         1. Agreement: # 422
         2. Disagreement: # 980
      8. Rater 6:
         1. Agreement: # 1046
         2. Disagreement: # 1213
      9. Rater 7:
         1. Agreement: # 754
         2. Disagreement: # 1148
      10. Rater 8:
          1. Agreement: # 622
          2. Disagreement: # 1076
      11. Rater 9:
          1. Agreement: # 906
          2. Disagreement: # 1203
   2. Manual resolvement of conflicts
      1. Accept: # 12809
      2. Ignore: # 0
      3. Reject: # 564
3. *Manual review of docs matching certain regex patterns*
   1. *Finally, as the great number of reviews yielded insights in common mistakes for the annotations, the entire DANSK was screened using a number of regex patterns. If the pattern found a match in document, the given document was manually reviewed.*
   2. *Patterns:*
      1. *Either list here, or include in appendix \**
   3. *The search yielded 449 cases that were manually reviewed and resolved.*