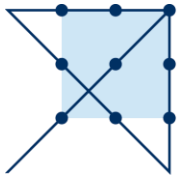


HITS
Heidelberg Institute for Theoretical Studies

Investigating Multilingual Coreference Resolution by Universal Annotations

Haixia Chai and **Michael Strube**

Findings of EMNLP 2023
Dec 6th, 2023

Multilingual Coreference Resolution (MCR)

Developing a general and robust system that can effectively handle multiple languages and a wide range of coreference phenomena (e.g., pronoun-drop).

1. **Most of work focus on a specific target language, especially English.**
    - Linguistic diversity and complexity of different languages.
    - Linguistic expertise in each individual language.

Multilingual Coreference Resolution (MCR)

2. **The two most studied types of languages.**
   - Pro-drop languages, e.g., Italian and Chinese.
   - Morphological-rich languages, e.g., German and Arabic.

3. **A few studies investigate coreference across multiple languages.**
   - Proposing MCR systems, e.g., latent structure learning.
   - Statistical analysis on multilingual coreference datasets.

In this work, we aim to fill the gap by studying **linguistic coreference analysis** in diverse languages?

Dataset – CorefUD (Nedoluzhko et al., 2022)

Universal Dependencies (de Marneffe et al., 2021)

- A framework for representing the syntactic structure in a consistent way.
- To provide a set of cross-linguistically consistent syntactic annotations
  part-of-speech tags, morphological features, dependency relations, and more.

Dataset: 17 datasets for 12 European languages.

Harmonization Scheme

- Coreference Annotations
- UD Annotations

CorefUD serves as a resource for the CRAC 2022–2023 shared task on MCR.

Questions

CorefUD

- Morphological features
- UPOS tags
- UD relations

1. **Are there any universal features/patterns that are common to all languages?**

2. **To what extent can the features contribute to a MCR system?**

Our Work

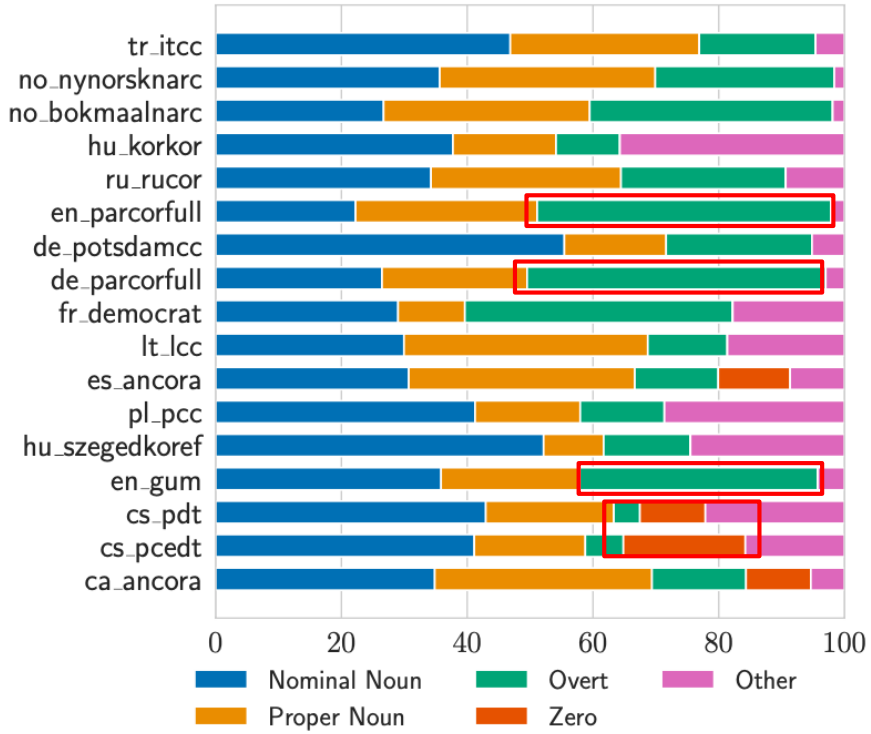
1. **Linguistic Analysis on CorefUD**
    - Mention
    - Entity
    - Document
2. **Error Analysis of MCR systems**

3. **Modeling with Universal Annotations**

Linguistic Analysis on CorefUD – Mention (1/2)

Mention Types

We categorize five types of mentions by the universal part-of-speech (UPOS) tags of the head words in gold mentions.

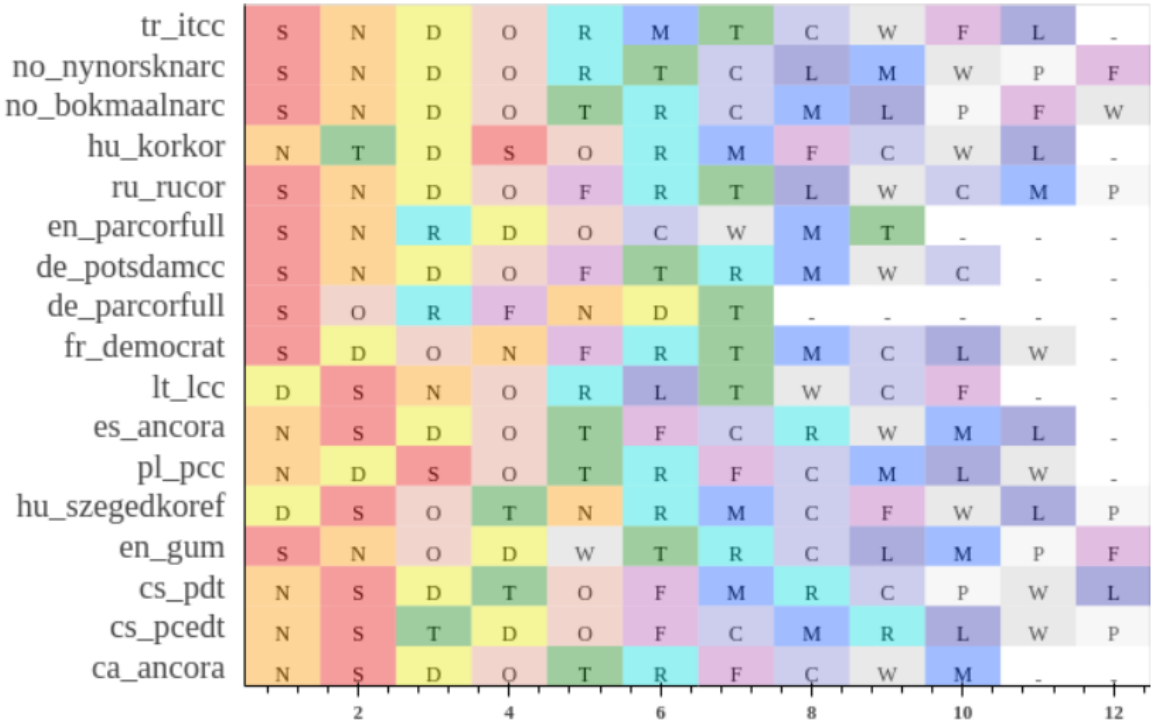


- Germanic languages (e.g., English and German) are the languages using most overt pronouns.
- Resolving zero pronouns is more crucial in the Czech datasets.

Linguistic Analysis on CorefUD – Mention (2/2)

Anaphor-Antecedent Relation

We analyze the UD category of the closest antecedent to an anaphor based on its mention types, e.g., **core arguments_subject** – <u>nominal noun</u>.



nominal noun

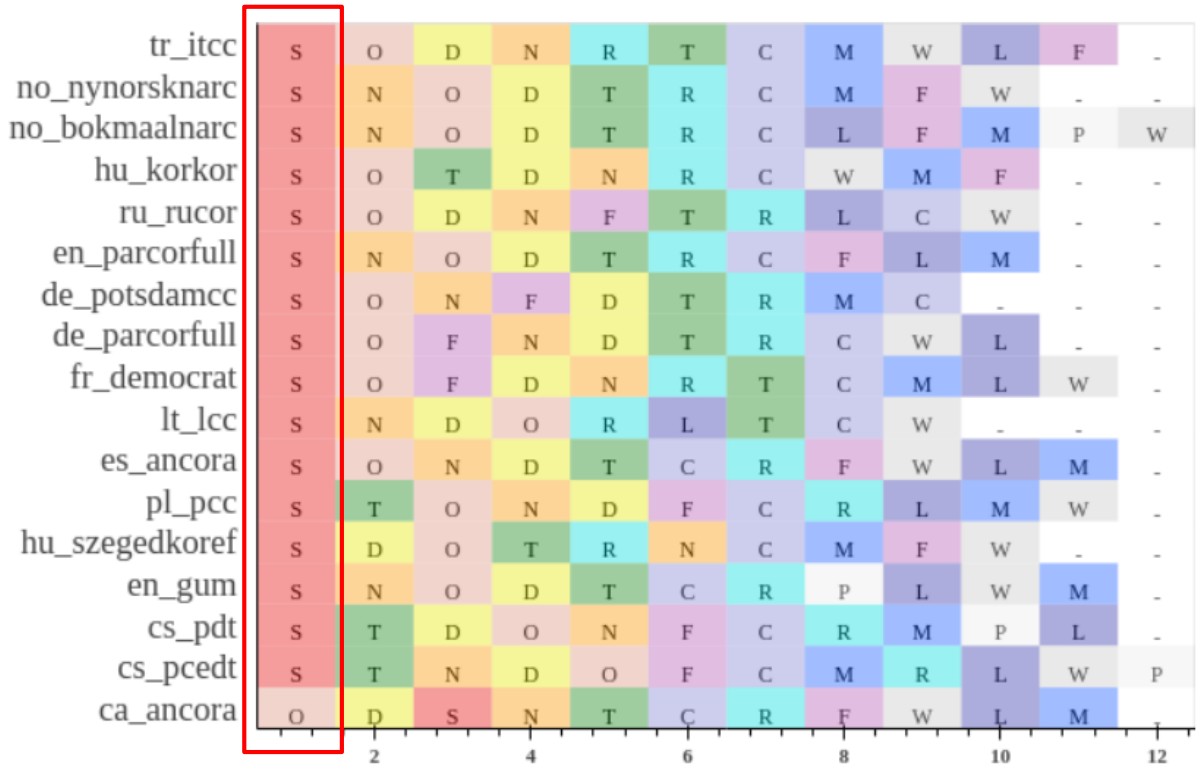
E.g., *Sam, my brother, John 's cousin, arrived.*

UD Categories	UD Relations
core arguments_ subject (S)	nsubj
core arguments_ object (O)	obj, iobj
non-core dependents_ nominals (D)	obl, vocative, expl, dislocated
nominal dependents_ nominals (N)	nmod, appos, nummod
clauses (C)	csubj, ccomp, xcomp, advcl, acl
modifier words (M)	advmod, discourse, amod
function words (F)	aux, cop mark det, clf, case
coordination (R)	conj, cc
MWE (W)	fixed, flat, compound
loose (L)	list, parataxis
special (P)	orphan, goeswith, reparandum
other (T)	punct, root, dep

Linguistic Analysis on CorefUD – Mention (2/2)

Anaphor-Antecedent Relation

We analyze the UD category of the closest antecedent to an anaphor based on its mention types, e.g., **core arguments_subject** – nominal noun.



overt pronoun

UD Categories	UD Relations
core arguments_ subject (S)	nsubj
core arguments_ object (O)	obj, iobj
non-core dependents_ nominals (D)	obl, vocative, expl, dislocated
nominal dependents_ nominals (N)	nmod, appos, nummod
clauses (C)	csubj, ccomp, xcomp, advcl, acl
modifier words (M)	advmod, discourse, amod
function words (F)	aux, cop mark det, clf, case
coordination (R)	conj, cc
MWE (W)	fixed, flat, compound
loose (L)	list, parataxis
special (P)	orphan, goeswith, reparandum
other (T)	punct, root, dep

Linguistic Analysis on CorefUD – Mention (2/2)

Anaphor-Antecedent Relation

We analyze the UD relation of the closest antecedent to an anaphor based on its mention types, e.g., **core arguments_subject** – nominal noun.

- Nominal noun
  non-core dependents, nominal dependents, core arguments_subject and core arguments_object.
- Overt pronoun
  core arguments_subject and core arguments_object.

These findings/patterns are applicable across all languages and are independent of any specific language.

Linguistic Analysis on CorefUD – Entity

First Mention

- The first mention within a mention chain serves to introduce the entity into a context.

- In Catalan, for example, 97% of first mentions belong to mention types of **nominal noun** and **proper noun**.

Consistent trend across all languages:

- The ratio of entities with the first mention being the longest mention in the entity ranges from 70% to 90%.
- *E.g., A person* vs. *A person that works at Penn*.

	Entities (%)
CA_ANCORA	76.92
CS_PCEDT	87.10
CS_PDT	75.22
EN_GUM	69.55
HU_SZEGEDKOREF	80.11
PL_PCC	83.13
ES_ANCORA	77.25
LT_LCC	81.80
FR_DEMOCRAT	82.46
DE_PARCORFULL	88.02
DE_POTSDAMCC	77.48
EN_PARCORFULL	89.24
RU_RUCOR	83.41
HU_KORKOR	82.27
NO_BOKMAALNARC	81.62
NO_NYNORSKNARC	78.11
TR_ITCC	70.36

Linguistic Analysis on CorefUD – Document

Competing Antecedents of Pronominal Anaphors

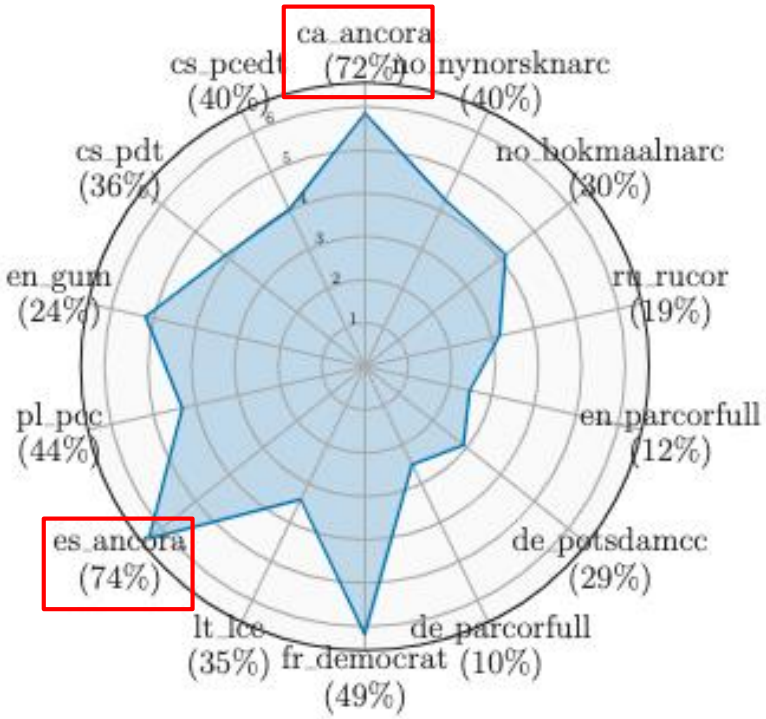
- The resolution of pronouns can become difficult due to their ambiguity caused by the presence of multiple potential antecedents from distinct entities or singletons.

The study of how **[people]**$_8$, as **[fans]**$_8$, access and manage information within a transmedia system provides valuable insight that contributes not only to **[practitioners]**$_7$ and **[scholars of the media industry]**$_6$, but to the wider context of cultural studies, by offering findings on this new model of **[the fan]**$_5$ as **[consumer]**$_4$ and **[information-user]**$_3$. For **[us]**$_1$, as **[digital humanists]**$_1$, defining **[the "transmedia fan"]**$_2$ is of particular relevance as **[we]**$_1$ seek to understand contemporary social and cultural transformations engendered by digital technologies.

Linguistic Analysis on CorefUD – Document

Competing Antecedents of Pronominal Anaphors – Overt Pronoun

- In ca_ancora and es_ancora, it is more difficult in distinguishing the true antecedent(s) of the pronoun among a pool of antecedents.

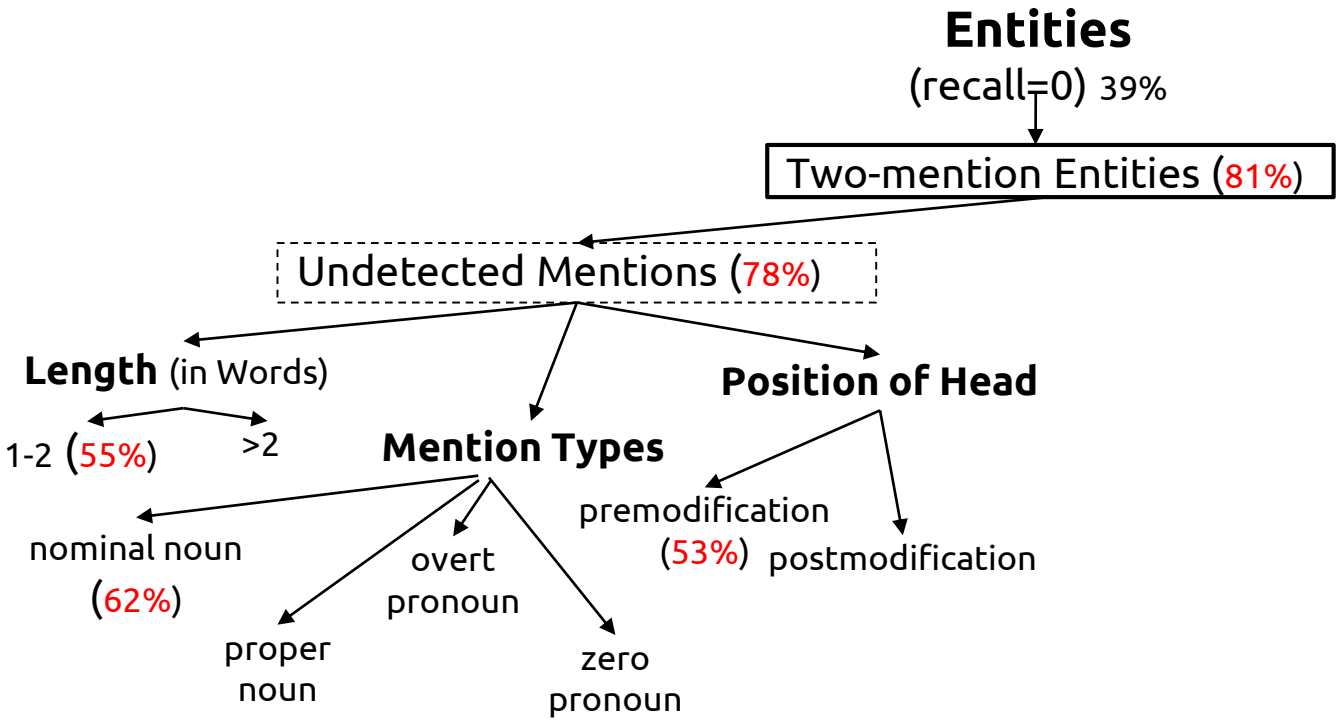


Centering theory:

- It suggests that pronoun tends to refer to the center or the most prominent entity in the preceding context. (Chai and Strube, 2022)
- It is applicable across all languages, as it is not dependent on any specific language.

Error Analysis of MCR Systems (1/2)

Undetected Mentions

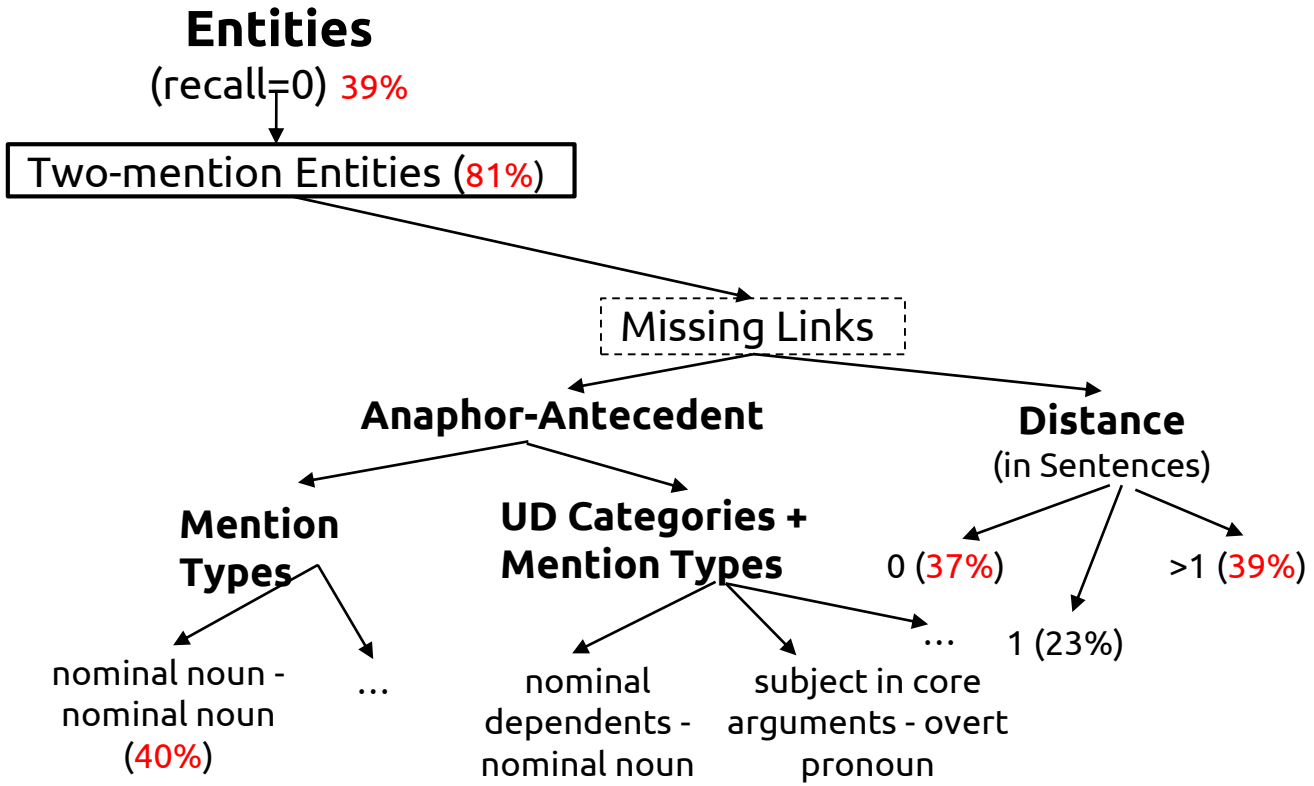


- Two-mention Entity: An entity includes only two mentions.
- More than 60% of the undetected mentions on average are nominal nouns.
- The highly variable nature of definiteness across languages.

Error Analysis of MCR Systems (2/2)

Missing Links

- Distance = 0: syntax information
- Distance > 1: knowledge extracted from the discourse structure of the text.
- Mention Types: nominal noun – nominal noun
- Anaphor-Antecedent Relation
  - nominal dependents – nominal noun
  - subject – overt pronoun



Discussion

- While multilingual datasets are partially harmonized to some extent, there are still cases where certain information, such as entity types, is only provided for a limited number of languages.

- We primarily focus on identity coreference. There exist various other anaphoric relations, such as bridging and discourse deixis, that remain unexplored.

- The languages examined in our study predominantly belong to the European languages.

Conclusions

- We analyze coreference across multiple languages by using the harmonized universal morphosyntactic and coreference annotations in CorefUD.
- We conduct error analysis of two MCR systems.
- We demonstrate the potential benefit of incorporating linguistic features in enhancing MCR system performance.

We will be in Findings 6 (poster) session **at 9:00 AM on December 10th**!

Modeling with Universal Annotations

- We adopt BASELINE (Prazak et al., 2021) in the CRAC 2022 shared task on MCR.

Incorporating Linguistic Information

$$\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$$

$$\mathbf{e}_c = [\mathbf{x}_{c_{start}}, \mathbf{x}_{c_{end}}, \hat{\mathbf{x}}_c, \phi(s_c)]$$

$$f_m(c) = \mathbf{FFNN}_m([\mathbf{e}_c, \phi(u_c)])$$

$$f(c, q) = \mathbf{FFNN}_s([\mathbf{e}_c, \mathbf{e}_q, \mathbf{e}_c \circ \mathbf{e}_q, \phi(c, q)])$$

$$\hat{P}(q) = \frac{\exp(f(c, q))}{\sum_{k \in Y(c)} \exp(f(c, k))}$$

UPOS tags, UD relations, mention types and UD categories of the span.

General features: language and word order of the language

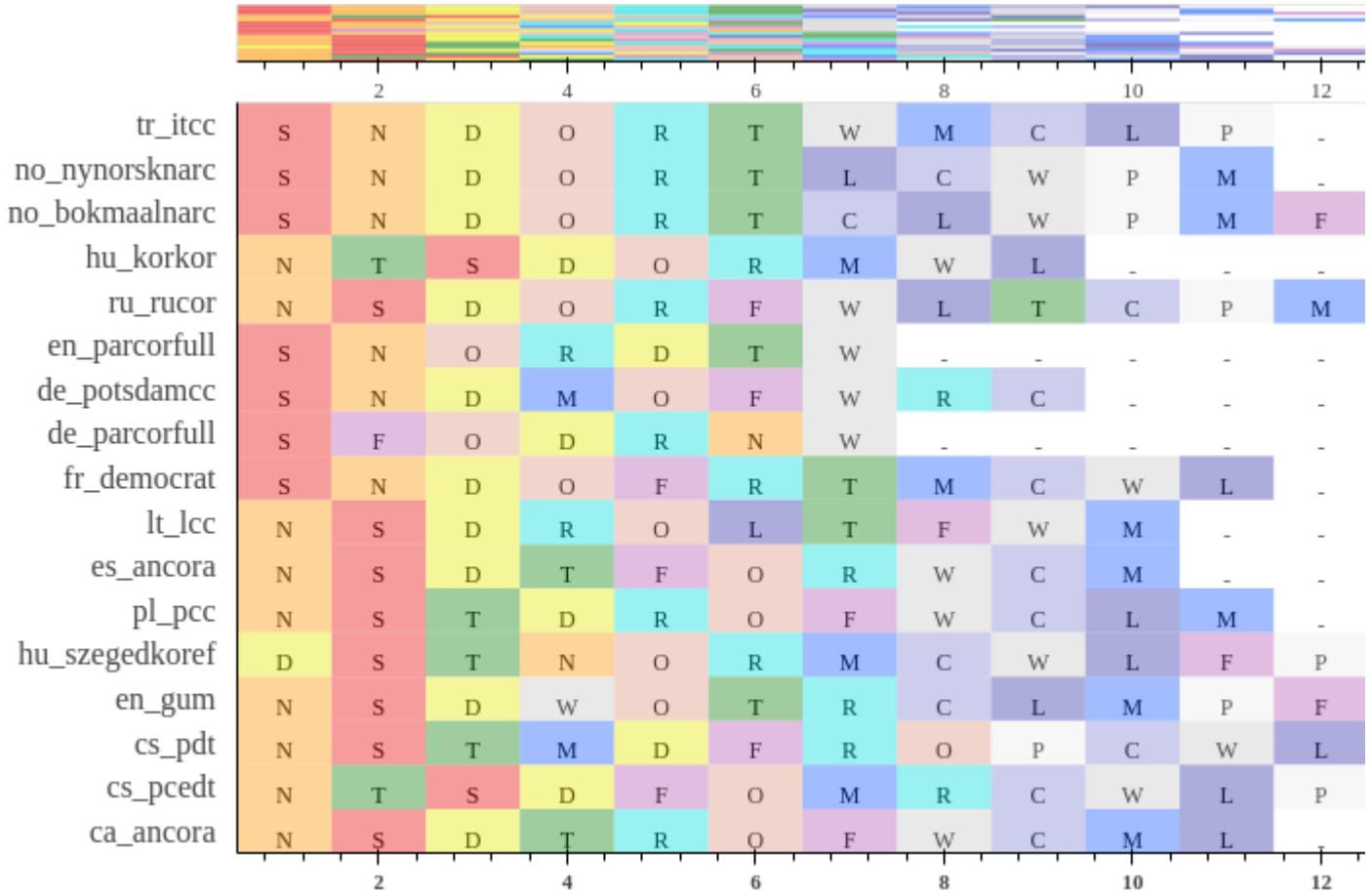
Modeling with Universal Annotations

MODELS	AVG	CA	CS PCED	CS PDT	EN GUM	HU	PL	ES	LT	FR	DE PARC	DE POTS	EN PARC	RU
BASELINE	53.7	55.2	68.4	64.3	48.8	46.4	50.2	57.6	**64.2**	57.0	33.7	43.3	43.0	**66.9**
Ours	**54.6**	**55.7**	**68.5**	**64.9**	**50.1**	**47.1**	**50.4**	**57.7**	62.1	**58.6**	**35.1**	**44.9**	**48.5**	66.5
⊖ ua	-0.51	-0.03	-0.23	0.00	-0.75	-0.21	-0.72	+0.34	+1.57	-1.78	+0.81	-2.90	-4.04	+1.36
⊖ lang	-0.87	-0.45	-0.11	-0.65	-1.29	-0.71	-0.19	-0.14	+2.09	-1.62	-1.44	-1.61	-5.58	+0.40

- A modest improvement over the baseline with a margin of 0.9% F1 score.
- In the ablation study, general features like language and word order also yield positive effects on performance.

Appendix – Anaphor-Antecedent Relation

- Proper noun



Appendix – Anaphor-Antecedent Relation

- Zero pronoun