Studying Language Properties and Divergences Via Automatically Classified Errors

Introduction:

Previous studies have examined the usefulness of a generalized unified framework analyzing cross-lingual syntactic divergences. However, there are few empirical studies that have examined the syntactic divergences across language pairs quantitatively.

In a recent study, Nikolaev and colleagues, proposed a framework for extracting divergence patterns across language pairs from a parallel corpus, building on Universal Dependencies. They defined Universal Dependencies (UD) as *'a framework for treebank annotation, whose objectives include satisfactory analyses of individual languages, providing a suitable basis for bringing out cross linguistic parallelism, suitability for rapid consistent annotation and accurate automatic parsing, ease of comprehension by non-linguists, and effective support for downstream tasks' (*Nikolaev et al.,2020).

They used '*The Parallel Universal Dependencies (PUD) corpus consists of 1000 sentences translated into various languages by professional translators'*

In this work we took a different approach, using a parallel corpus, consisting of sentences written in English by kids of various nationalities, and the corrected sentences.

Also worthwhile to mention "Classifying Syntactic Errors in Learner Language" (Choshen et al. 2020.) This work is mentioned in the footnotes as well, when we used their github repository. The abstract:

> '*We present a method for classifying syntactic errors in learner language, namely errors whose correction alters the morphosyntactic structure of a sentence. The methodology builds on the established Universal Dependencies syntactic representation scheme, and provides complementary information to other error-classification systems. Unlike existing error classification methods, our method is applicable across languages, which we showcase by producing a detailed picture of syntactic errors in learner English and learner Russian. We further demonstrate the utility of the methodology for analyzing the outputs of leading Grammatical Error Correction (GEC) systems.'*

The Current Study

The current study examines error patterns made in English by individuals of various nationalities. We used the EFCAMDAT dataset which specifically records nationalities and not languages. *Nationality is used as a proxy for language when comparing to other work on languages.* In this work we took a different approach to language comparisons, using a parallel corpus, consisting of sentences written in English by kids of various nationalities, and the corrected sentence, classifying the errors and analysing the error POS, both in the original and corrected sentence, and comparing to previous work on these Cross-Linguistic Syntactic Divergences.

This work is exploratory in nature.

We classified error-types in EFCAMDAT, and show visualisations of the distribution according to POS. We sliced the data according to nationality and proficiency level as well.
We measured distance between languages, computed from their error-profile matrix, and compared that to distances computed from their lan2vec vector. Measuring distances between languages with our error-profile compared to lang2vec, showed high correlation for most languages, with the highest being Spanish (mexico).

Jp (Japanese) and to a lesser extent de showed the least total similarity to other languages in both distance measurements.
We showed some prominent missed POS, (‘the’, ‘a’, ‘,’ , ‘.’) as well prominent missed POS in the DET POS, with ‘the’ and ‘a’ dominating this category.

For brevity sake we used the following names:

POS - part of speech.
DS - Dataset.
DP - data points (usually referring to a single error, and its metadata.)
DF - Dataframe. A python object that holds all the relevant processed data, in a convenient format for analysis.

Background:

Datasets:

The main Dataset we are working with is the EFCAMDAT dataset.
(https://philarion.mml.cam.ac.uk/)
From the site:

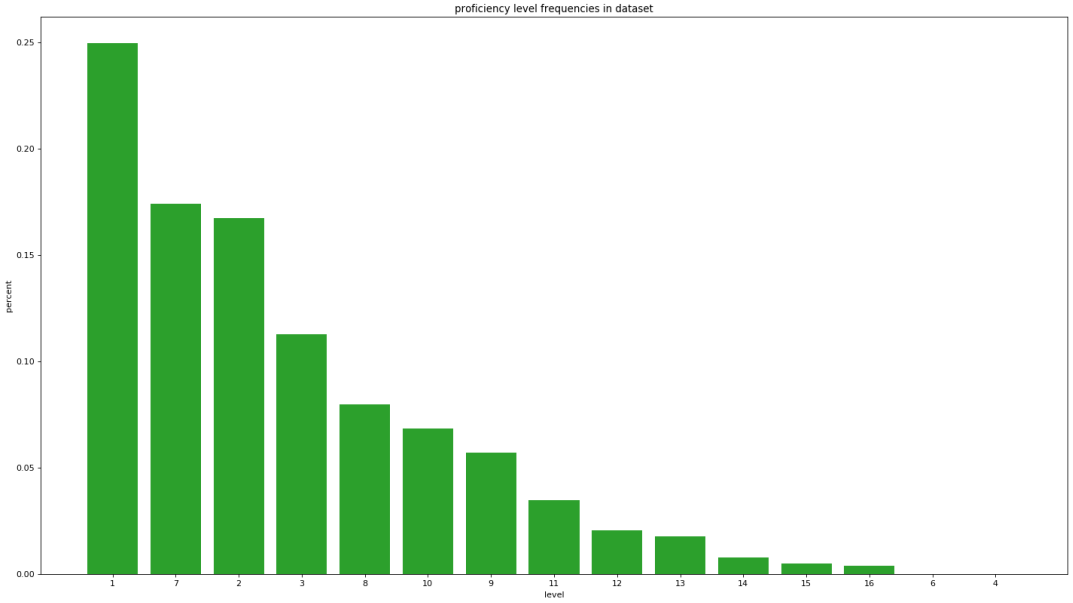
> *The EF-Cambridge Open Language Database (EFCAMDAT) is a publicly available resource to facilitate second language research and teaching.*
> *It contains written samples from thousands of adult learners of English as a second language, world wide.*
>
> *EFCAMDAT currently contains over 83 million words from 1 million assignments written by 174,000 learners, across a wide range of levels (CEFR stages A1-C2).*
> *This text corpus includes information on learner errors, part of speech, and grammatical relationships.*
> *Researchers can search for language patterns using a range of criteria, including learner nationality and level.*
>
> *The resource is actively developed by the Department of Theoretical and Applied Linguistics at the University of Cambridge in partnership with Education First.*

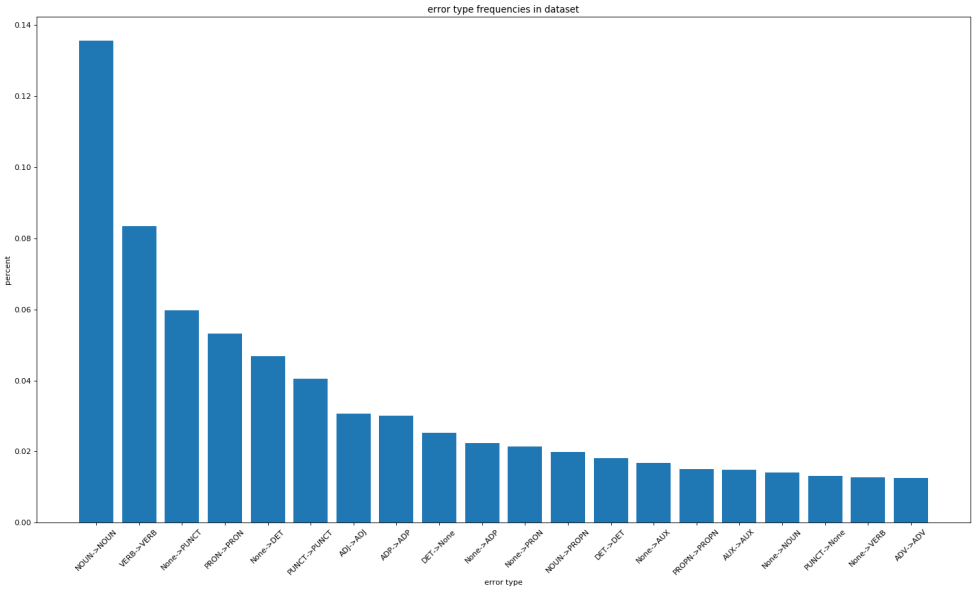
Dataset statistics:

Total nationalities: 186
Total proficiency levels: 15. (1-16, no level 5)
Total students: 114,554
(obtained from step 2 described in STEPS):
Total amount of classified errors: 4,564,523

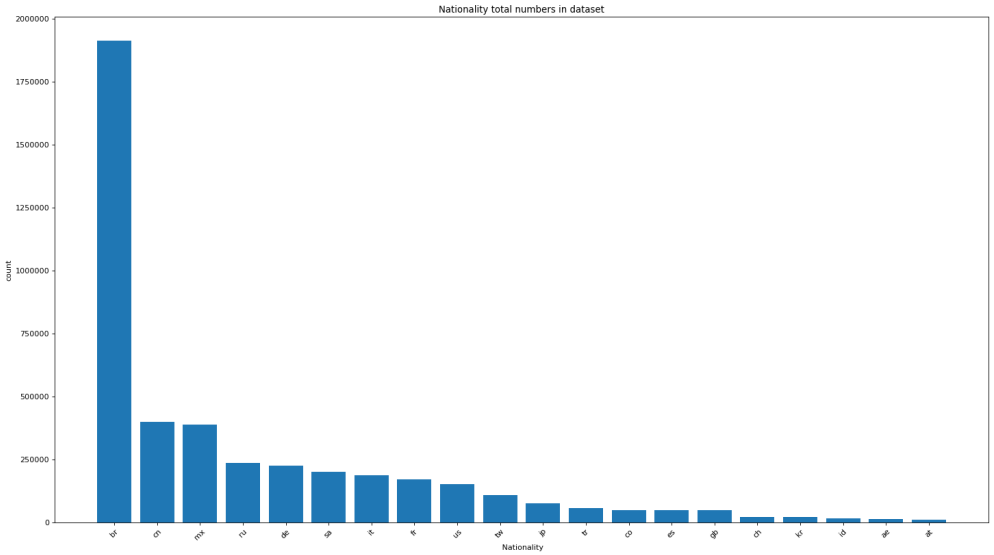
The EFCAMDAT dataset provided us with the original and corrected sentences.

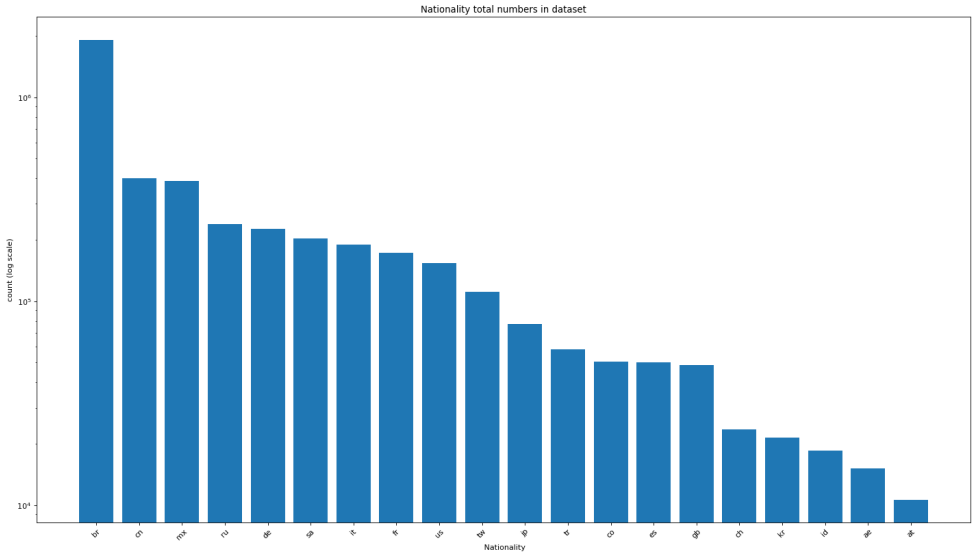
Also student uid, nationality, date, topic, grade, level.
We decided to provide our own error-type classification using the ErrAnt [1]model, for consistency and reproducibility.
Lets see the distribution of various parameters in the dataset, proficiency level, nationality and error-types:





[1] https://github.com/chrisjbryant/errant see references.


Nationality total numbers in dataset
count
2000000
1750000
1500000
1250000
1000000
750000
500000
250000
0
br
cn
mx
ru
de
sa
it
fr
us
tw
jp
tr
co
es
gb
ch
kr
id
ae
pk
Nationality


Nationality total numbers in dataset
count (log scale)
10^6
10^5
10^4
br
cn
mx
ru
de
sa
it
fr
us
tw
jp
tr
co
es
gb
ch
kr
id
ae
pk
Nationality

Main:

Methodology:

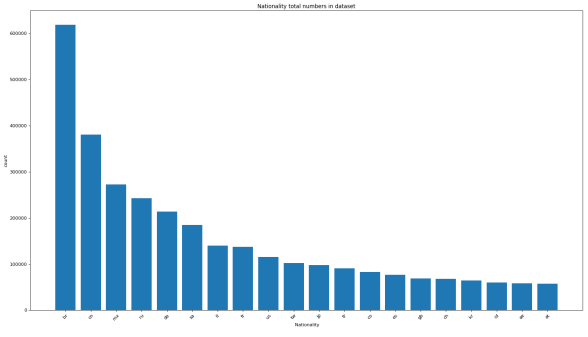
The first big effort was made to parse the DS into a DF

Steps:

1. Parsing the dataset XML, extracting the original sentence and the corrected sentence.
The DS included correction type tagging, but for consistency and reproducibility we decided to annotate the correction type ourselves.
For this we created parallel files the model can use to annotate.
2. Annotating the correction type with [2][ERRANT]
We used a trained model[3] to annotate the correction.
3. Creating an m2 file of the corrections
This is needed for step 5
4. Saving the metadata in a Dataframe.
5. creating CoNLL-U files for the original and corrected sentences using udpipe [4] library.
6. Convert the m2 and CoNLL-U files to a syntax based (p.o.s.) m2 file using [5]
GEC_UD_divergences
7. Add syntax based correction-types to the Dataframe (DF).
Now we have the DF, We can run our analysis. In our DF, a data-point is a single error, classified by the model.

Decisions:

Levels 6, 4 were not included in the analysis, as they had negligible data-points.
(15 and 8 respectively.)



[2] ERRANT (https://github.com/chrisjbryant/errant)

[3] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3131/english-ewt-ud-2.5-191206.udpipe
[4] https://github.com/ufal/udpipe
[5] https://github.com/borgr/GEC_UD_divergences

Seeing there are186 nationalities, we decided to showcase mainly the largest 10, making the threshold 100,000 data points.
We decided to remove nationalities with less than 10,000 data-points. So the analysis refers to 20 nationalities.
Since the EFCAMDAT records nationalities and not languages, we use nationality as a proxy for language when comparing to other work on languages.
We only used for the comparison parts nationalities we can map confidently and without duplicates in our dataset.
The mapping is as follows:

Country	Name in EFCAMDAT	lang2vec	Language	Number DP in DS
Brazil	**br**	**pt**	**portuguese**	**476,817**
chine	cn	zh	chinese	165,162
mexico	mx	es	Spanish	87,259
russia	ru	ru	Russian	70K
Germany	de	de	German	54K
Saudi	sa	ar	Arabic	47,340
Italy	it	it	Italian	4K
France	fr	fr	French	41K
Japan	jp	jp	Japanese	21K

error profile

Create an error-profile for every language:

After creating the DF, I created an error-profile matrix, by using the syntax based (p.o.s.) error classification. Every cell in the matrix is the count of an error type.
Every line in the matrix is divided by its sum, to normalize it.
This was done for the whole dataset, as well as for single nationalities, student proficiency (level) and for a combination of both.
The same is shown after subtracting the average across all languages, per cell.

For a visual representation of the relationship between the proficiency and error profile, We created a matrix, where every cell shows a bar graph representing that cell's value, for all levels, ordered by proficiency.

lang to vec

Next we Downloaded the lang2vec[6] dataset and extracted the features vectors, for comparison. We Quantified the correlation between distance metrics of languages represented by their lang2vec vector, compared to the same languages represented by their error-profile matrix. Distances were computed with cosine similarity. For lang2vec this was done on the whole vector.
For our error-profile matrices, this was done per line, then averaged.
Quantifying the correlation between the 2 methods for measuring the distances was done by Spearman-Rank correlation.

[6] https://github.com/antonisa/lang2vec see references

Experiments and results:

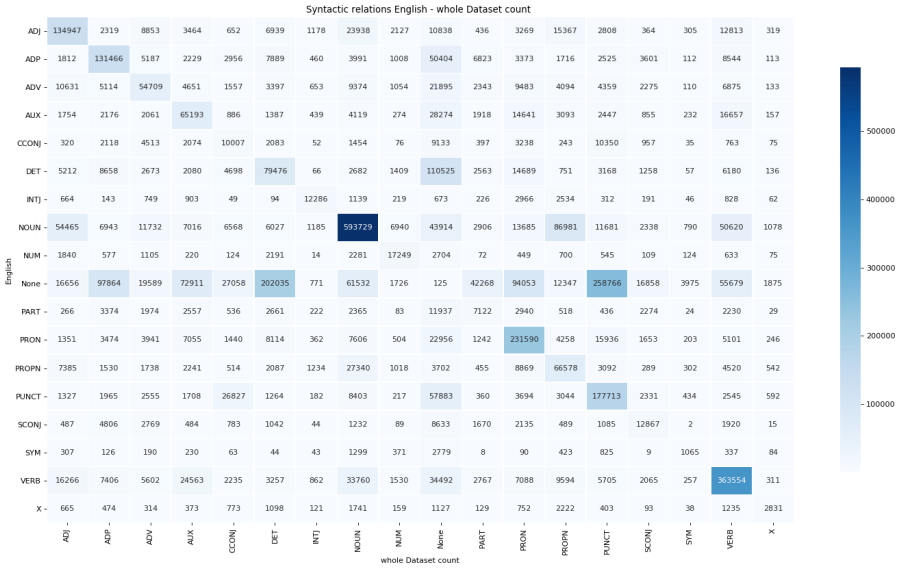
Questions

Are there specific features that distinguish specific languages?
Do they correlate with known Cross-Linguistic Syntactic Divergences?
Is there a correlation between proficiency level and some error type?

Graphs

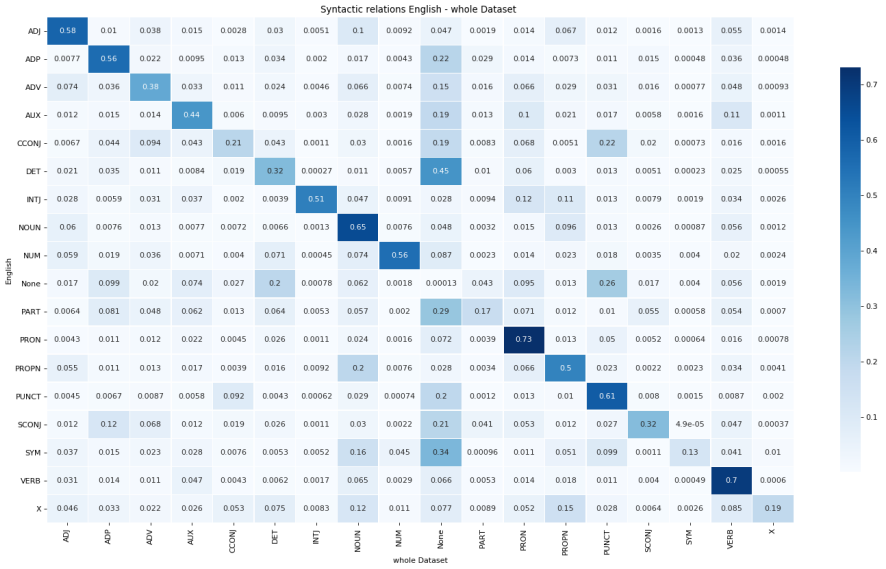
In the following graphs, the rows (marked ‘english’) are the corrected POS, and the columns (marked according to the current DS slice: the whole DS, nationality, Level etc.) are the original POS, as classified by ERRANT[7].

Classified Syntactic Errors for the Whole DS:

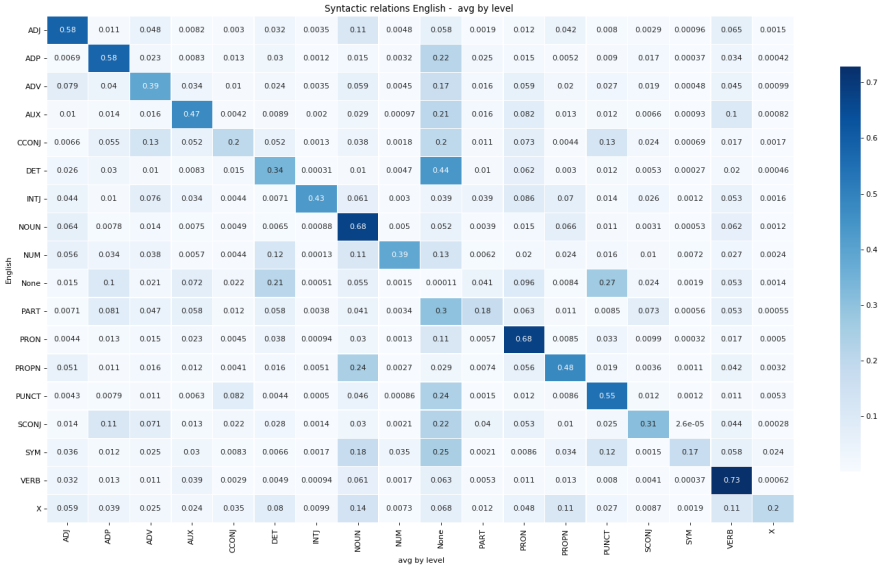


We can see the ‘NOUN->NOUN’ is prominent. Although it is attenuated when we normalize per row.
From here on all graphs are normalized per row. (Every cell is divided by the row total.)

[7] https://github.com/chrisjbryant/errant



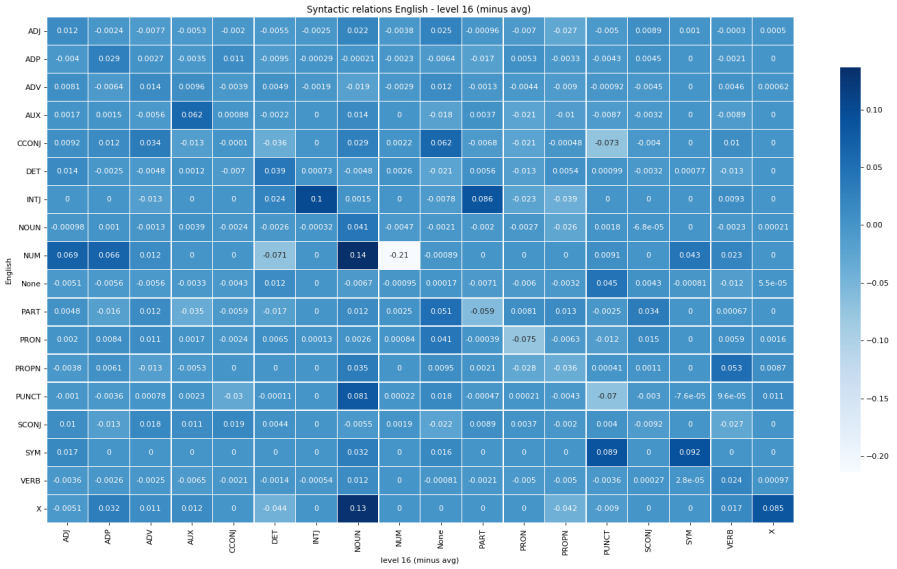
Classified Syntactic Errors for the Whole DS, averaged by level:
(each level gets the same weight, regardless of size.)

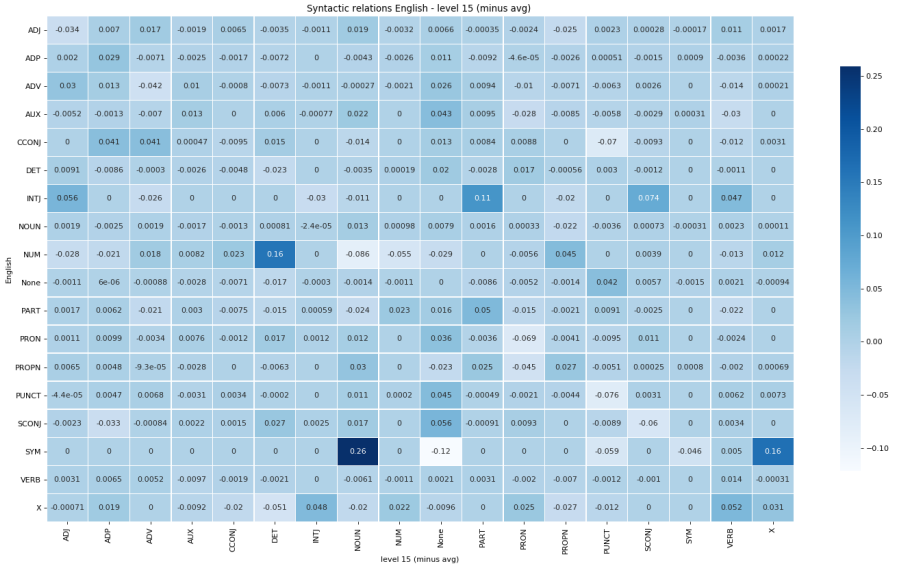


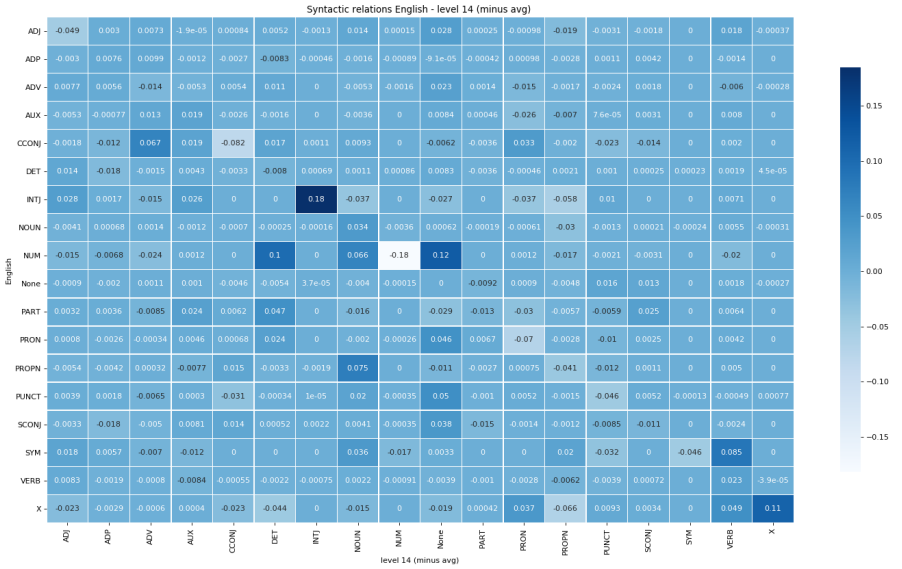
Classified Syntactic Errors by level, minus the average across levels:

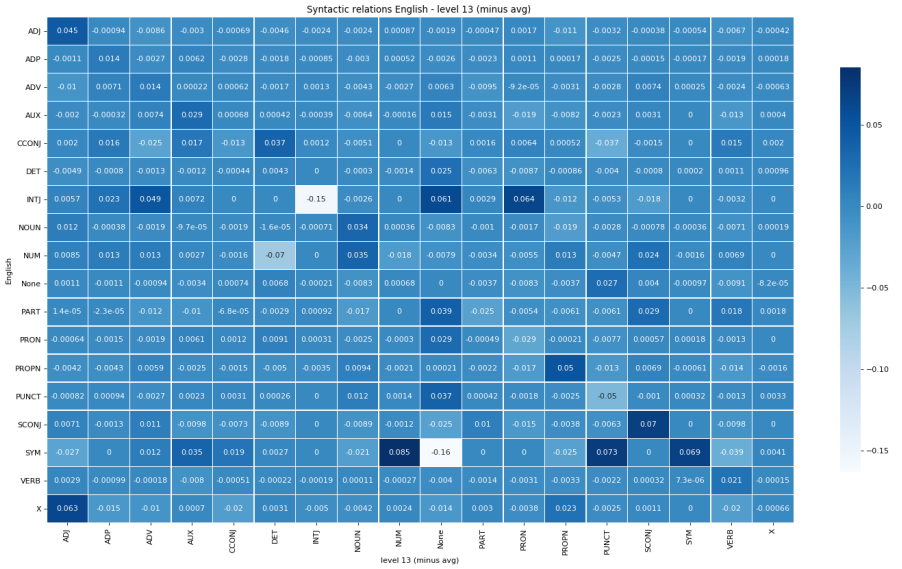
The error profile for the dataset minus the average value for each cell.

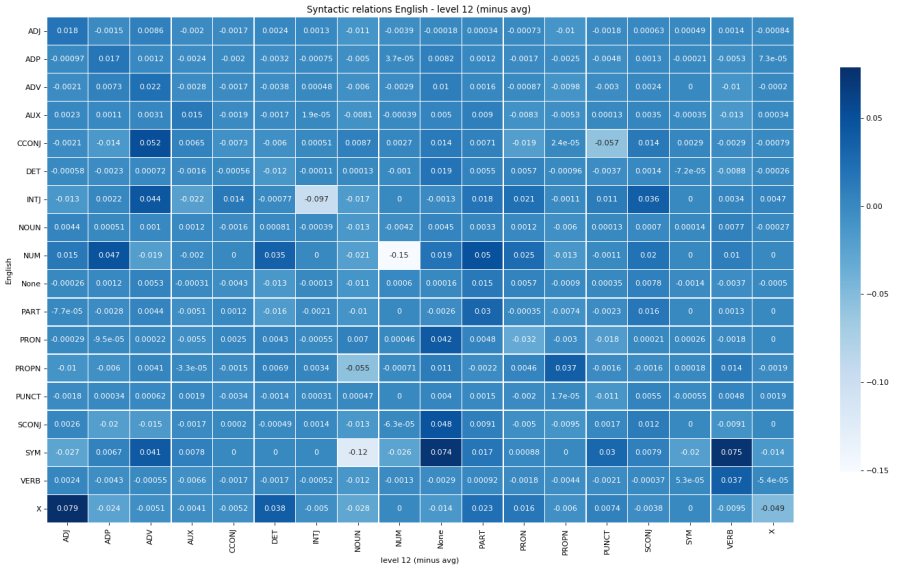
The average is computed across Levels.

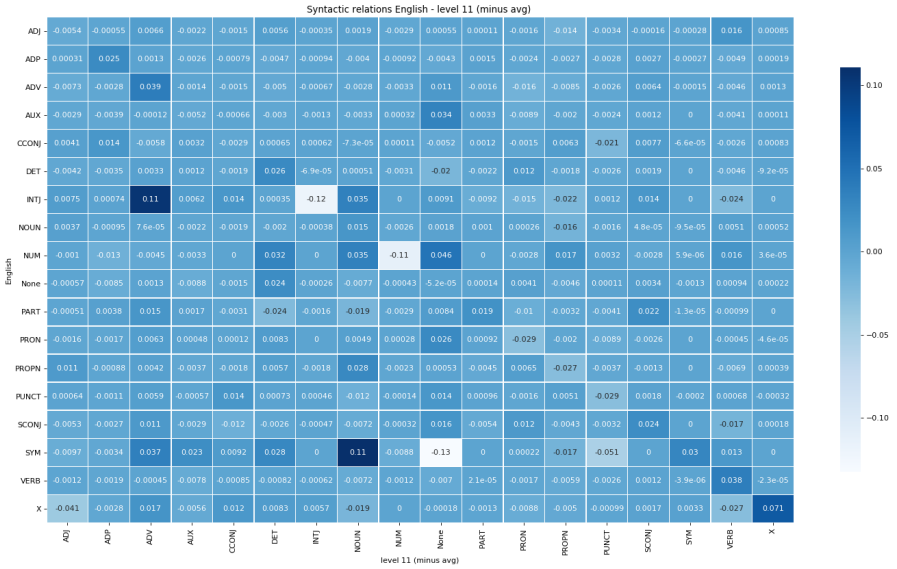


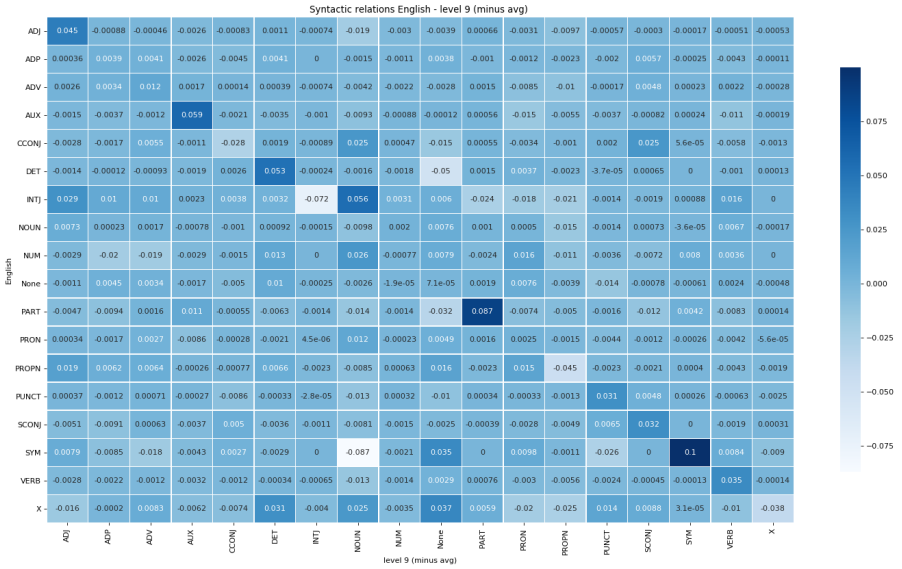


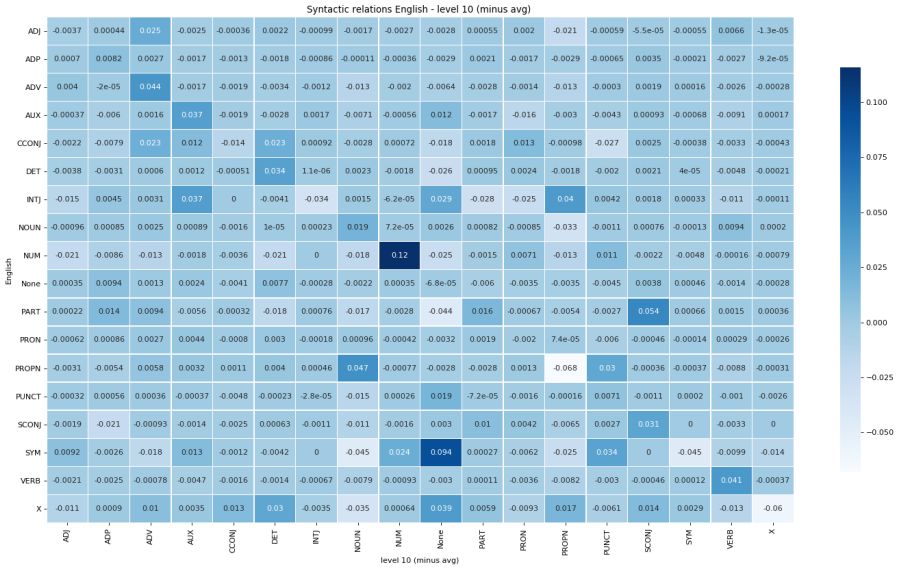

Syntactic relations English - level 14 (minus avg)
English
level 14 (minus avg)
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.15 0.10 0.05 0.00 −0.05 −0.10 −0.15

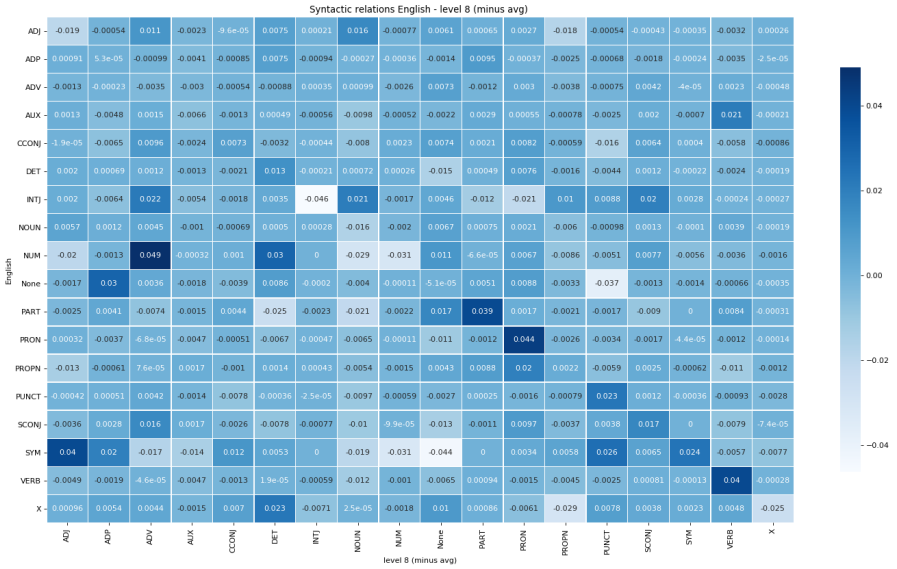

Syntactic relations English - level 13 (minus avg)
English
level 13 (minus avg)
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.05 0.00 −0.05 −0.10 −0.15

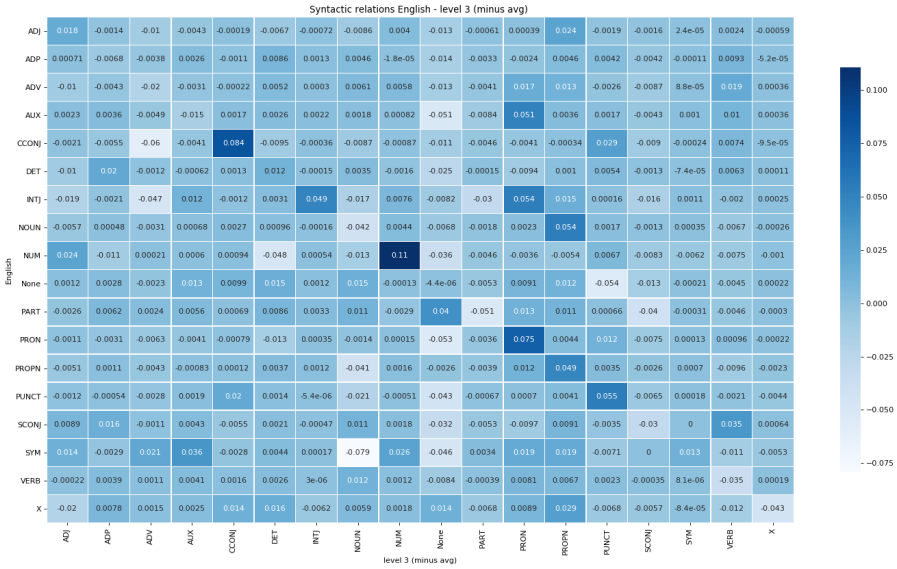

Syntactic relations English - level 12 (minus avg)
English
level 12 (minus avg)
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.05
0.00
−0.05
−0.10
−0.15

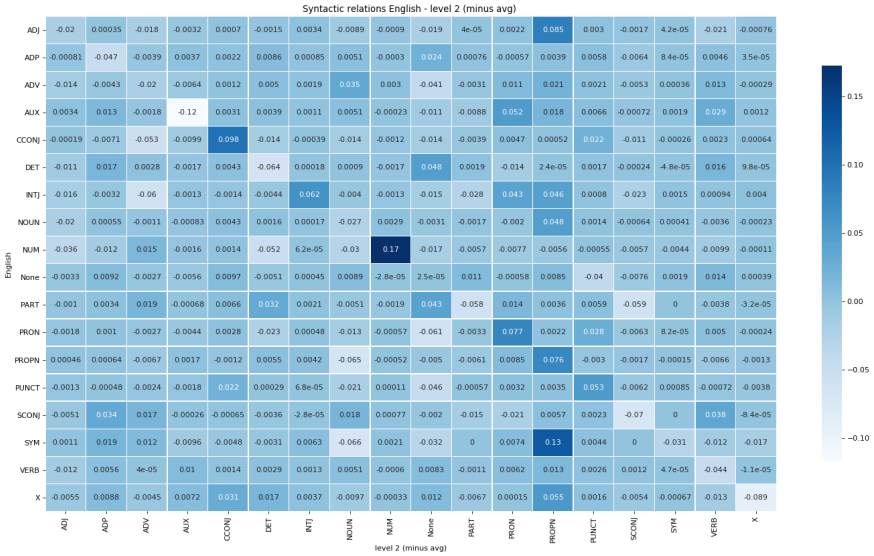

Syntactic relations English - level 11 (minus avg)
English
level 11 (minus avg)
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.10
0.05
0.00
−0.05
−0.10

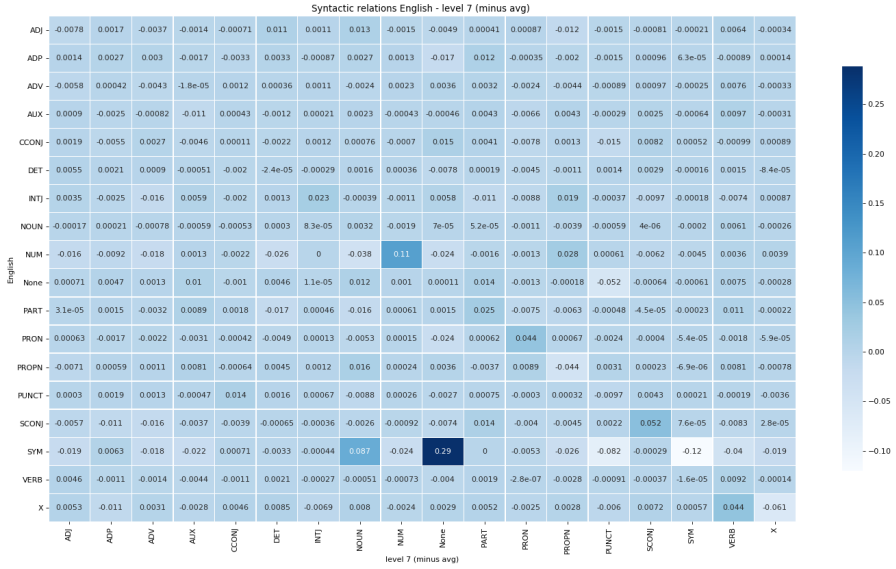

Syntactic relations English - level 9 (minus avg)
English
level 9 (minus avg)
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075

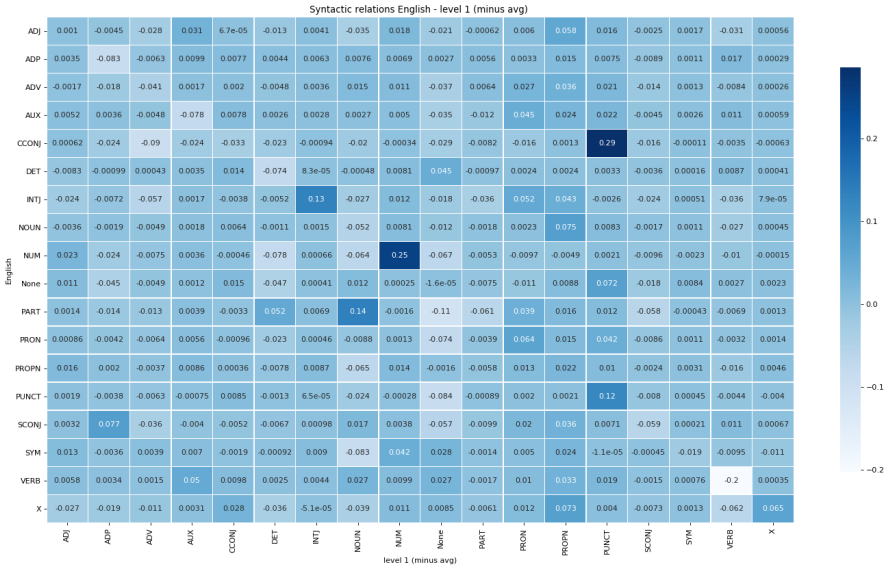

Syntactic relations English - level 10 (minus avg)
English
level 10 (minus avg)
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050


Syntactic relations English - level 8 (minus avg)
English
level 8 (minus avg)
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.04
0.02
0.00
−0.02
−0.04


Syntactic relations English - level 3 (minus avg)
English
level 3 (minus avg)
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075

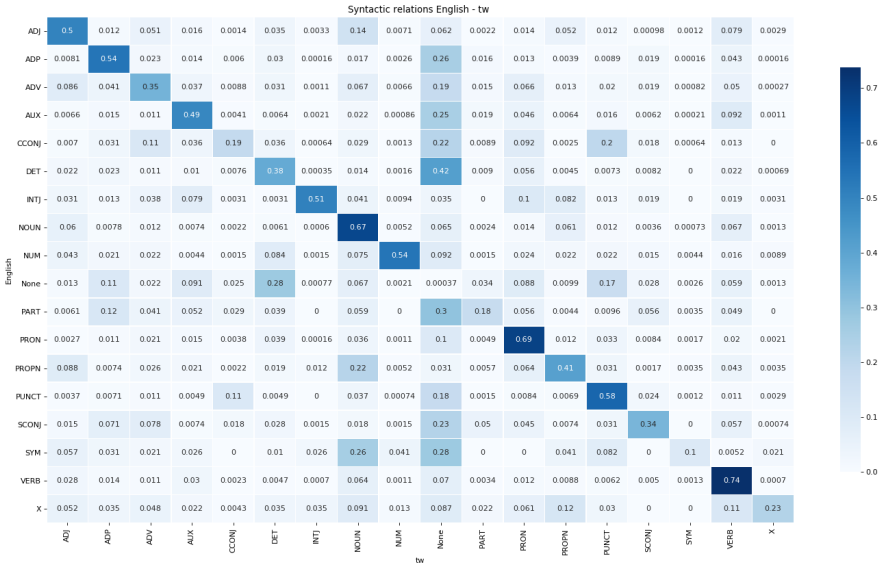

Syntactic relations English - level 2 (minus avg)
English
level 2 (minus avg)
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.15
0.10
0.05
0.00
−0.05
−0.10

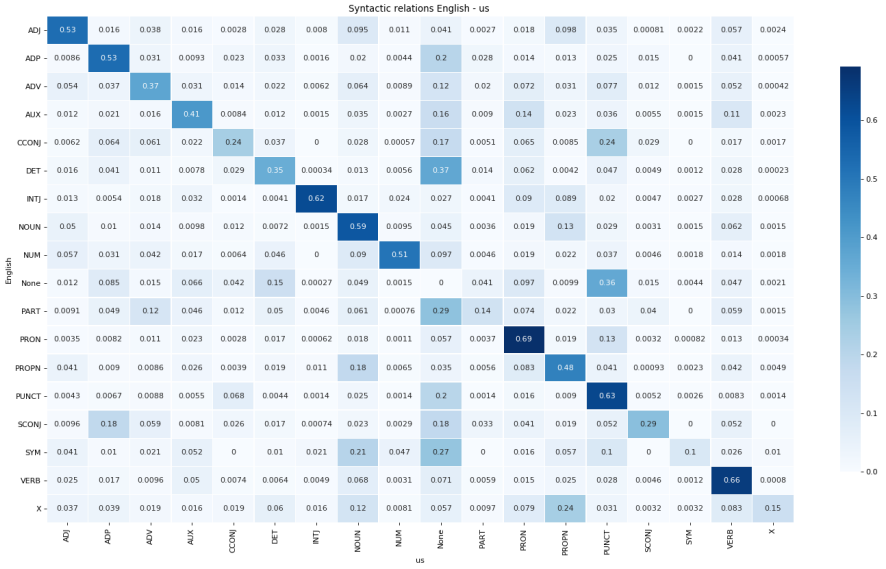

Syntactic relations English - level 7 (minus avg)
English
level 7 (minus avg)
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10

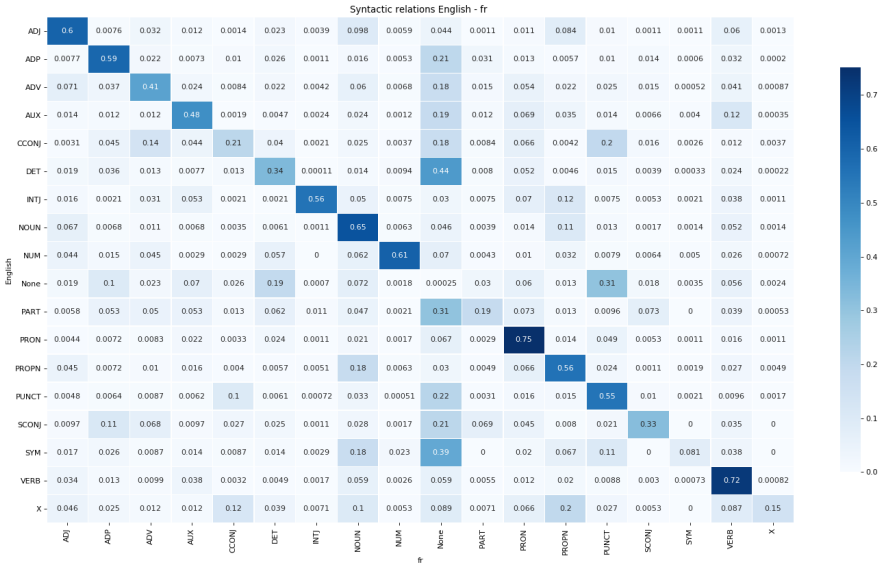

Syntactic relations English - level 1 (minus avg)
English
level 1 (minus avg)
ADJ
ADP
ADV
AUX
CCONJ
DET
INTJ
NOUN
NUM
None
PART
PRON
PROPN
PUNCT
SCONJ
SYM
VERB
X
0.2
0.1
0.0
−0.1
−0.2

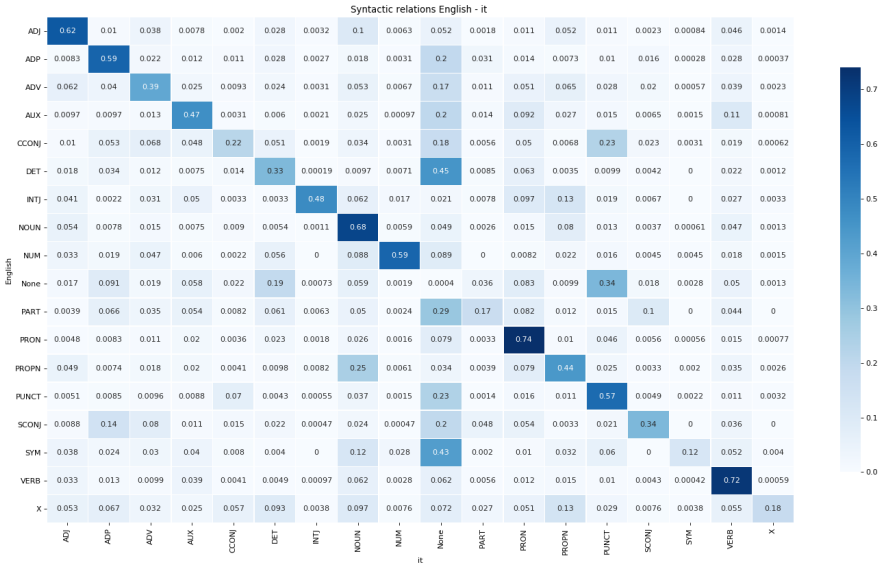
The error profile for each nationality:

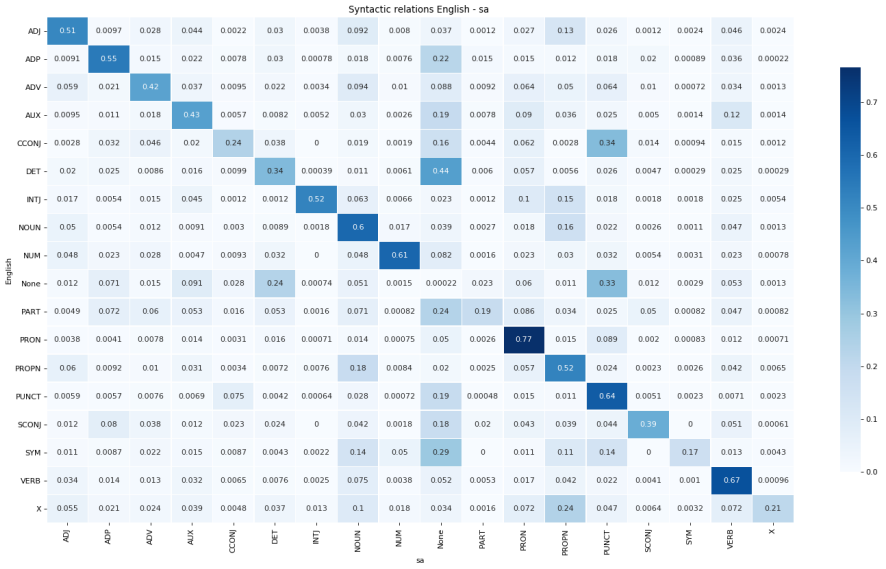
Classified Syntactic Errors by nationality:
The error profile for the dataset minus the average value for each cell.
The average is computed across languages.

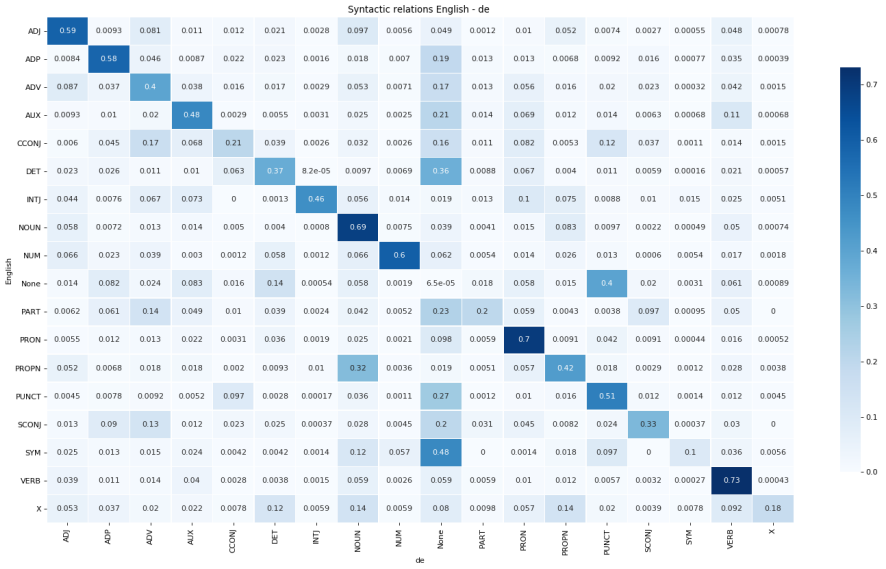


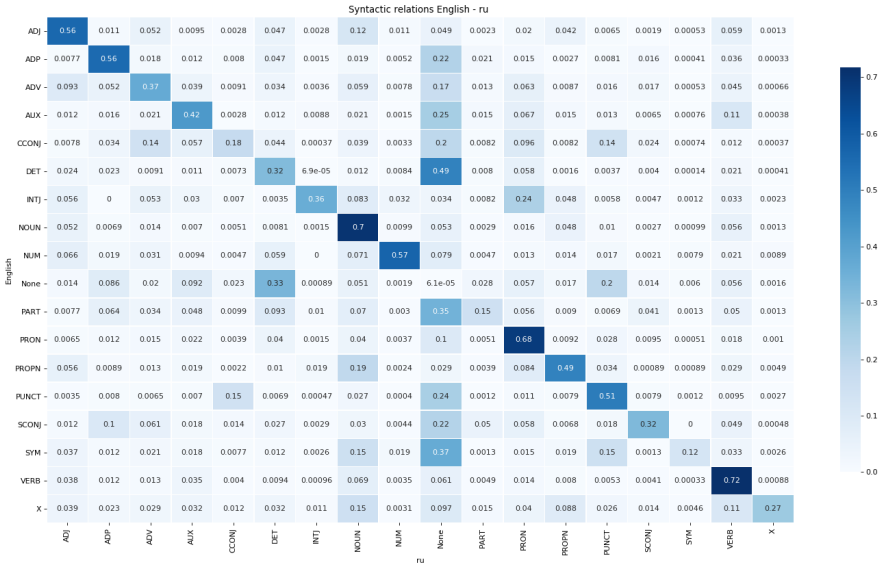


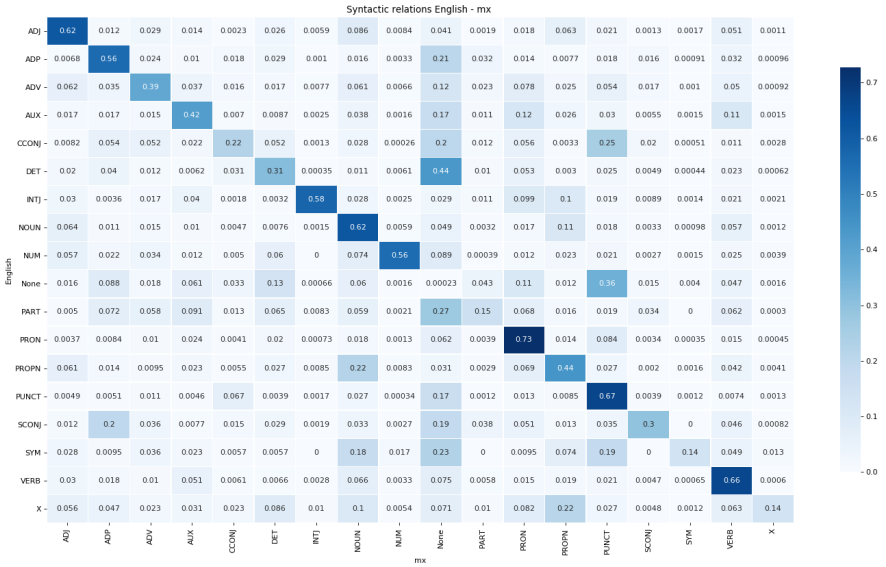

Syntactic relations English - fr
English
fr
ADJ 0.6 0.0076 0.032 0.012 0.0014 0.023 0.0039 0.098 0.0059 0.044 0.0011 0.011 0.084 0.01 0.0011 0.0011 0.06 0.0013
ADP 0.0077 0.59 0.022 0.0073 0.01 0.026 0.0011 0.016 0.0053 0.21 0.031 0.013 0.0057 0.01 0.014 0.0006 0.032 0.0002
ADV 0.071 0.037 0.41 0.024 0.0084 0.022 0.0042 0.06 0.0068 0.18 0.015 0.054 0.022 0.025 0.015 0.00052 0.041 0.00087
AUX 0.014 0.012 0.012 0.48 0.0019 0.0047 0.0024 0.024 0.0012 0.19 0.012 0.069 0.035 0.014 0.0066 0.004 0.12 0.00035
CCONJ 0.0031 0.045 0.14 0.044 0.21 0.04 0.0021 0.025 0.0037 0.18 0.0084 0.066 0.0042 0.2 0.016 0.0026 0.012 0.0037
DET 0.019 0.036 0.013 0.0077 0.013 0.34 0.00011 0.014 0.0094 0.44 0.008 0.052 0.0046 0.015 0.0039 0.00033 0.024 0.00022
INTJ 0.016 0.0021 0.031 0.053 0.0021 0.0021 0.56 0.05 0.0075 0.03 0.0075 0.07 0.12 0.0075 0.0053 0.0021 0.038 0.0011
NOUN 0.067 0.0068 0.011 0.0068 0.0035 0.0061 0.0011 0.65 0.0063 0.046 0.0039 0.014 0.11 0.013 0.0017 0.0014 0.052 0.0014
NUM 0.044 0.015 0.045 0.0029 0.0029 0.057 0 0.062 0.61 0.07 0.0043 0.01 0.032 0.0079 0.0064 0.005 0.026 0.00072
None 0.019 0.1 0.023 0.07 0.026 0.19 0.0007 0.072 0.0018 0.00025 0.03 0.06 0.013 0.31 0.018 0.0035 0.056 0.0024
PART 0.0058 0.053 0.05 0.053 0.013 0.062 0.011 0.047 0.0021 0.31 0.19 0.073 0.013 0.0096 0.073 0 0.039 0.00053
PRON 0.0044 0.0072 0.0083 0.022 0.0033 0.024 0.0011 0.021 0.0017 0.067 0.0029 0.75 0.014 0.049 0.0053 0.0011 0.016 0.0011
PROPN 0.045 0.0072 0.01 0.016 0.004 0.0057 0.0051 0.18 0.0063 0.03 0.0049 0.066 0.56 0.024 0.0011 0.0019 0.027 0.0049
PUNCT 0.0048 0.0064 0.0087 0.0062 0.1 0.0061 0.00072 0.033 0.00051 0.22 0.0031 0.016 0.015 0.55 0.01 0.0021 0.0096 0.0017
SCONJ 0.0097 0.11 0.068 0.0097 0.027 0.025 0.0011 0.028 0.0017 0.21 0.069 0.045 0.008 0.021 0.33 0 0.035 0
SYM 0.017 0.026 0.0087 0.014 0.0087 0.014 0.0029 0.18 0.023 0.39 0 0.02 0.067 0.11 0 0.081 0.038 0
VERB 0.034 0.013 0.0099 0.038 0.0032 0.0049 0.0017 0.059 0.0026 0.059 0.0055 0.012 0.02 0.0088 0.003 0.00073 0.72 0.00082
X 0.046 0.025 0.012 0.012 0.12 0.039 0.0071 0.1 0.0053 0.089 0.0071 0.066 0.2 0.027 0.0053 0 0.087 0.15
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0

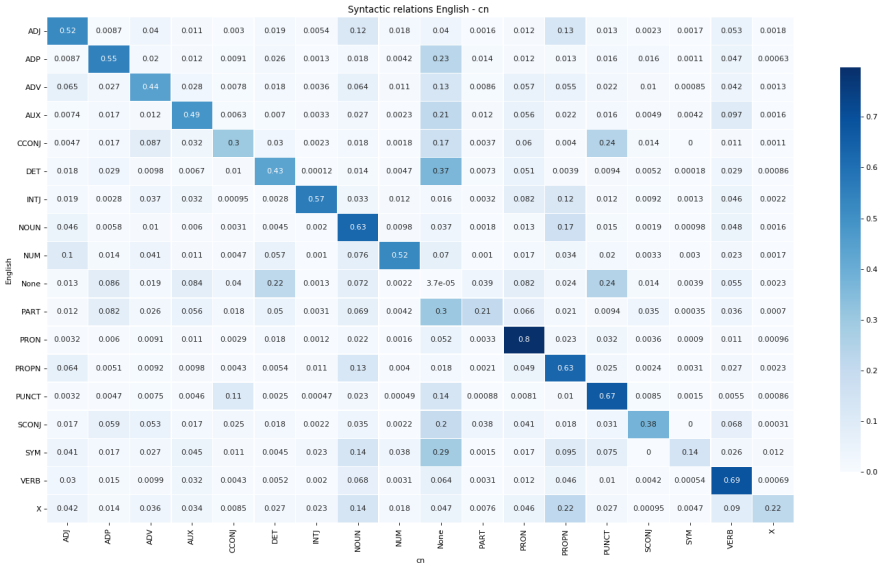

Syntactic relations English - it
English
it
ADJ 0.62 0.01 0.038 0.0078 0.002 0.028 0.0032 0.1 0.0063 0.052 0.0018 0.011 0.052 0.011 0.0023 0.00084 0.046 0.0014
ADP 0.0083 0.59 0.022 0.012 0.011 0.028 0.0027 0.018 0.0031 0.2 0.031 0.014 0.0073 0.01 0.016 0.00028 0.028 0.00037
ADV 0.062 0.04 0.39 0.025 0.0093 0.024 0.0031 0.053 0.0067 0.17 0.011 0.051 0.065 0.028 0.02 0.00057 0.039 0.0023
AUX 0.0097 0.0097 0.013 0.47 0.0031 0.006 0.0021 0.025 0.00097 0.2 0.014 0.092 0.027 0.015 0.0065 0.0015 0.11 0.00081
CCONJ 0.01 0.053 0.068 0.048 0.22 0.051 0.0019 0.034 0.0031 0.18 0.0056 0.05 0.0068 0.23 0.023 0.0031 0.019 0.00062
DET 0.018 0.034 0.012 0.0075 0.014 0.33 0.00019 0.0097 0.0071 0.45 0.0085 0.063 0.0035 0.0099 0.0042 0 0.022 0.0012
INTJ 0.041 0.0022 0.031 0.05 0.0033 0.0033 0.48 0.062 0.017 0.021 0.0078 0.097 0.13 0.019 0.0067 0 0.027 0.0033
NOUN 0.054 0.0078 0.015 0.0075 0.009 0.0054 0.0011 0.68 0.0059 0.049 0.0026 0.015 0.08 0.013 0.0037 0.00061 0.047 0.0013
NUM 0.033 0.019 0.047 0.006 0.0022 0.056 0 0.088 0.59 0.089 0 0.0082 0.022 0.016 0.0045 0.0045 0.018 0.0015
None 0.017 0.091 0.019 0.058 0.022 0.19 0.00073 0.059 0.0019 0.0004 0.036 0.083 0.0099 0.34 0.018 0.0028 0.05 0.0013
PART 0.0039 0.066 0.035 0.054 0.0082 0.061 0.0063 0.05 0.0024 0.29 0.17 0.082 0.012 0.015 0.1 0 0.044 0
PRON 0.0048 0.0083 0.011 0.02 0.0036 0.023 0.0018 0.026 0.0016 0.079 0.0033 0.74 0.01 0.046 0.0056 0.00056 0.015 0.00077
PROPN 0.049 0.0074 0.018 0.02 0.0041 0.0098 0.0082 0.25 0.0061 0.034 0.0039 0.079 0.44 0.025 0.0033 0.002 0.035 0.0026
PUNCT 0.0051 0.0085 0.0096 0.0088 0.07 0.0043 0.00055 0.037 0.0015 0.23 0.0014 0.016 0.011 0.57 0.0049 0.0022 0.011 0.0032
SCONJ 0.0088 0.14 0.08 0.011 0.015 0.022 0.00047 0.024 0.00047 0.2 0.048 0.054 0.0033 0.021 0.34 0 0.036 0
SYM 0.038 0.024 0.03 0.04 0.008 0.004 0 0.12 0.028 0.43 0.002 0.01 0.032 0.06 0 0.12 0.052 0.004
VERB 0.033 0.013 0.0099 0.039 0.0041 0.0049 0.00097 0.062 0.0028 0.062 0.0056 0.012 0.015 0.01 0.0043 0.00042 0.72 0.00059
X 0.053 0.067 0.032 0.025 0.057 0.093 0.0038 0.097 0.0076 0.072 0.027 0.051 0.13 0.029 0.0076 0.0038 0.055 0.18
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0

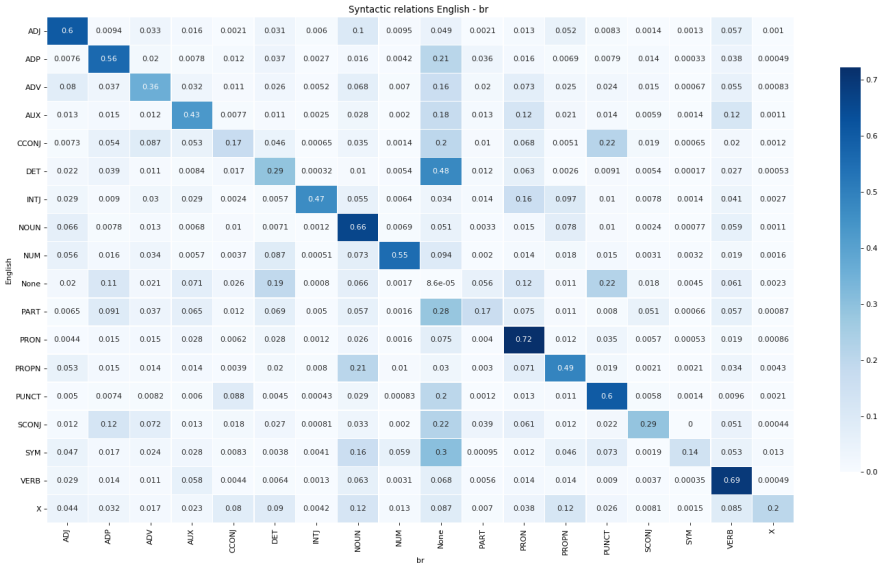

Syntactic relations English - sa
English
sa
ADJ 0.51 0.0097 0.028 0.044 0.0022 0.03 0.0038 0.092 0.008 0.037 0.0012 0.027 0.13 0.026 0.0012 0.0024 0.046 0.0024
ADP 0.0091 0.55 0.015 0.022 0.0078 0.03 0.00078 0.018 0.0076 0.22 0.015 0.015 0.012 0.018 0.02 0.00089 0.036 0.00022
ADV 0.059 0.021 0.42 0.037 0.0095 0.022 0.0034 0.094 0.01 0.088 0.0092 0.064 0.05 0.064 0.01 0.00072 0.034 0.0013
AUX 0.0095 0.011 0.018 0.43 0.0057 0.0082 0.0052 0.03 0.0026 0.19 0.0078 0.09 0.036 0.025 0.005 0.0014 0.12 0.0014
CCONJ 0.0028 0.032 0.046 0.02 0.24 0.038 0 0.019 0.0019 0.16 0.0044 0.062 0.0028 0.34 0.014 0.00094 0.015 0.0012
DET 0.02 0.025 0.0086 0.016 0.0099 0.34 0.00039 0.011 0.0061 0.44 0.006 0.057 0.0056 0.026 0.0047 0.00029 0.025 0.00029
INTJ 0.017 0.0054 0.015 0.045 0.0012 0.0012 0.52 0.063 0.0066 0.023 0.0012 0.1 0.15 0.018 0.0018 0.0018 0.025 0.0054
NOUN 0.05 0.0054 0.012 0.0091 0.003 0.0089 0.0018 0.6 0.017 0.039 0.0027 0.018 0.16 0.022 0.0026 0.0011 0.047 0.0013
NUM 0.048 0.023 0.028 0.0047 0.0093 0.032 0 0.048 0.61 0.082 0.0016 0.023 0.03 0.032 0.0054 0.0031 0.023 0.00078
None 0.012 0.071 0.015 0.091 0.028 0.24 0.00074 0.051 0.0015 0.00022 0.023 0.06 0.011 0.33 0.012 0.0029 0.053 0.0013
PART 0.0049 0.072 0.06 0.053 0.016 0.053 0.0016 0.071 0.00082 0.24 0.19 0.086 0.034 0.025 0.05 0.00082 0.047 0.00082
PRON 0.0038 0.0041 0.0078 0.014 0.0031 0.016 0.00071 0.014 0.00075 0.05 0.0026 0.77 0.015 0.089 0.002 0.00083 0.012 0.00071
PROPN 0.06 0.0092 0.01 0.031 0.0034 0.0072 0.0076 0.18 0.0084 0.02 0.0025 0.057 0.52 0.024 0.0023 0.0026 0.042 0.0065
PUNCT 0.0059 0.0057 0.0076 0.0069 0.075 0.0042 0.00064 0.028 0.00072 0.19 0.00048 0.015 0.011 0.64 0.0051 0.0023 0.0071 0.0023
SCONJ 0.012 0.08 0.038 0.012 0.023 0.024 0 0.042 0.0018 0.18 0.02 0.043 0.039 0.044 0.39 0 0.051 0.00061
SYM 0.011 0.0087 0.022 0.015 0.0087 0.0043 0.0022 0.14 0.05 0.29 0 0.011 0.11 0.14 0 0.17 0.013 0.0043
VERB 0.034 0.014 0.013 0.032 0.0065 0.0076 0.0025 0.075 0.0038 0.052 0.0053 0.017 0.042 0.022 0.0041 0.001 0.67 0.00096
X 0.055 0.021 0.024 0.039 0.0048 0.037 0.013 0.1 0.018 0.034 0.0016 0.072 0.24 0.047 0.0064 0.0032 0.072 0.21
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0

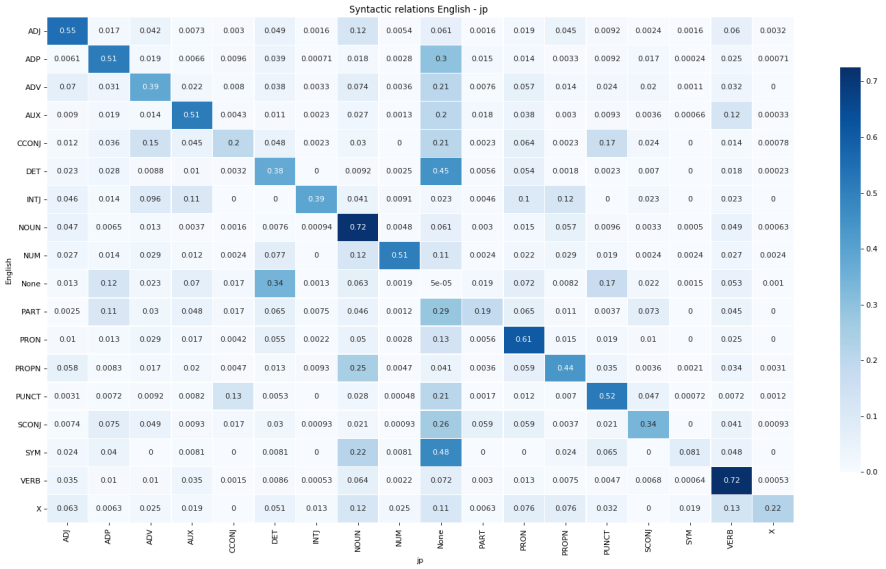

Syntactic relations English - de
English
de
ADJ 0.59 0.0093 0.081 0.011 0.012 0.021 0.0028 0.097 0.0056 0.049 0.0012 0.01 0.052 0.0074 0.0027 0.00055 0.048 0.00078
ADP 0.0084 0.58 0.046 0.0087 0.022 0.023 0.0016 0.018 0.007 0.19 0.013 0.013 0.0068 0.0092 0.016 0.00077 0.035 0.00039
ADV 0.087 0.037 0.4 0.038 0.016 0.017 0.0029 0.053 0.0071 0.17 0.013 0.056 0.016 0.02 0.023 0.00032 0.042 0.0015
AUX 0.0093 0.01 0.02 0.48 0.0029 0.0055 0.0031 0.025 0.0025 0.21 0.014 0.069 0.012 0.014 0.0063 0.00068 0.11 0.00068
CCONJ 0.006 0.045 0.17 0.068 0.21 0.039 0.0026 0.032 0.0026 0.16 0.011 0.082 0.0053 0.12 0.037 0.0011 0.014 0.0015
DET 0.023 0.026 0.011 0.01 0.063 0.37 8.2e-05 0.0097 0.0069 0.36 0.0088 0.067 0.004 0.011 0.0059 0.00016 0.021 0.00057
INTJ 0.044 0.0076 0.067 0.073 0 0.0013 0.46 0.056 0.014 0.019 0.013 0.1 0.075 0.0088 0.01 0.015 0.025 0.0051
NOUN 0.058 0.0072 0.013 0.014 0.005 0.004 0.0008 0.69 0.0075 0.039 0.0041 0.015 0.083 0.0097 0.0022 0.00049 0.05 0.00074
NUM 0.066 0.023 0.039 0.003 0.0012 0.058 0.0012 0.066 0.6 0.062 0.0054 0.014 0.026 0.013 0.0006 0.0054 0.017 0.0018
None 0.014 0.082 0.024 0.083 0.016 0.14 0.00054 0.058 0.0019 6.5e-05 0.018 0.058 0.015 0.4 0.02 0.0031 0.061 0.00089
PART 0.0062 0.061 0.14 0.049 0.01 0.039 0.0024 0.042 0.0052 0.23 0.2 0.059 0.0043 0.0038 0.097 0.00095 0.05 0
PRON 0.0055 0.012 0.013 0.022 0.0031 0.036 0.0019 0.025 0.0021 0.098 0.0059 0.7 0.0091 0.042 0.0091 0.00044 0.016 0.00052
PROPN 0.052 0.0068 0.018 0.018 0.002 0.0093 0.01 0.32 0.0036 0.019 0.0051 0.057 0.42 0.018 0.0029 0.0012 0.028 0.0038
PUNCT 0.0045 0.0078 0.0092 0.0052 0.097 0.0028 0.00017 0.036 0.0011 0.27 0.0012 0.01 0.016 0.51 0.012 0.0014 0.012 0.0045
SCONJ 0.013 0.09 0.13 0.012 0.023 0.025 0.00037 0.028 0.0045 0.2 0.031 0.045 0.0082 0.024 0.33 0.00037 0.03 0
SYM 0.025 0.013 0.015 0.024 0.0042 0.0042 0.0014 0.12 0.057 0.48 0 0.0014 0.018 0.097 0 0.1 0.036 0.0056
VERB 0.039 0.011 0.014 0.04 0.0028 0.0038 0.0015 0.059 0.0026 0.059 0.0059 0.01 0.012 0.0057 0.0032 0.00027 0.73 0.00043
X 0.053 0.037 0.02 0.022 0.0078 0.12 0.0059 0.14 0.0059 0.08 0.0098 0.057 0.14 0.02 0.0039 0.0078 0.092 0.18
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0


Syntactic relations English - ru
English
ru
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7


Syntactic relations English - mx
English
mx
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7

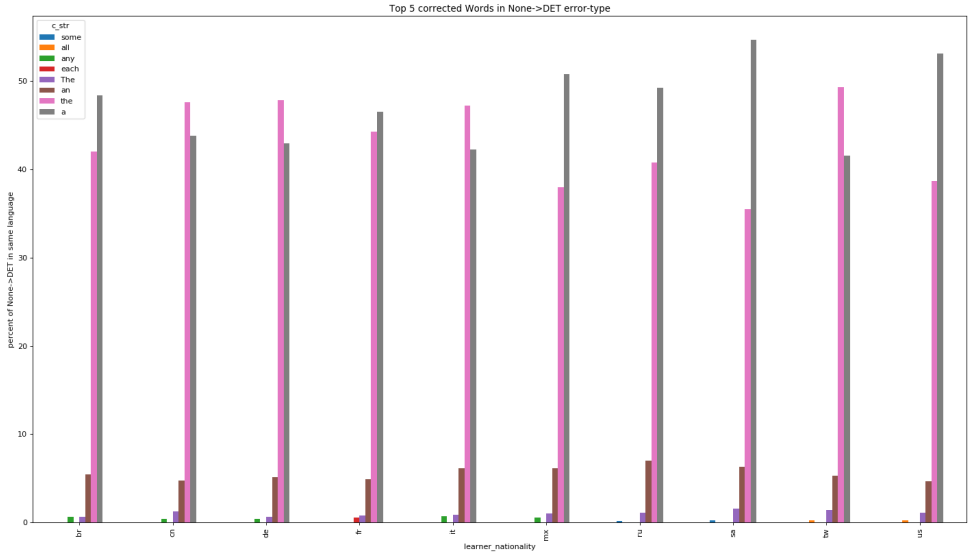

Syntactic relations English - cn
English
cn
ADJ 0.52 0.0087 0.04 0.011 0.003 0.019 0.0054 0.12 0.018 0.04 0.0016 0.012 0.13 0.013 0.0023 0.0017 0.053 0.0018
ADP 0.0087 0.55 0.02 0.012 0.0091 0.026 0.0013 0.018 0.0042 0.23 0.014 0.012 0.013 0.016 0.016 0.0011 0.047 0.00063
ADV 0.065 0.027 0.44 0.028 0.0078 0.018 0.0036 0.064 0.011 0.13 0.0086 0.057 0.055 0.022 0.01 0.00085 0.042 0.0013
AUX 0.0074 0.017 0.012 0.49 0.0063 0.007 0.0033 0.027 0.0023 0.21 0.012 0.056 0.022 0.016 0.0049 0.0042 0.097 0.0016
CCONJ 0.0047 0.017 0.087 0.032 0.3 0.03 0.0023 0.018 0.0018 0.17 0.0037 0.06 0.004 0.24 0.014 0 0.011 0.0011
DET 0.018 0.029 0.0098 0.0067 0.01 0.43 0.00012 0.014 0.0047 0.37 0.0073 0.051 0.0039 0.0094 0.0052 0.00018 0.029 0.00086
INTJ 0.019 0.0028 0.037 0.032 0.00095 0.0028 0.57 0.033 0.012 0.016 0.0032 0.082 0.12 0.012 0.0092 0.0013 0.046 0.0022
NOUN 0.046 0.0058 0.01 0.006 0.0031 0.0045 0.002 0.63 0.0098 0.037 0.0018 0.013 0.17 0.015 0.0019 0.00098 0.048 0.0016
NUM 0.1 0.014 0.041 0.011 0.0047 0.057 0.001 0.076 0.52 0.07 0.001 0.017 0.034 0.02 0.0033 0.003 0.023 0.0017
None 0.013 0.086 0.019 0.084 0.04 0.22 0.0013 0.072 0.0022 3.7e-05 0.039 0.082 0.024 0.24 0.014 0.0039 0.055 0.0023
PART 0.012 0.082 0.026 0.056 0.018 0.05 0.0031 0.069 0.0042 0.3 0.21 0.066 0.021 0.0094 0.035 0.00035 0.036 0.0007
PRON 0.0032 0.006 0.0091 0.011 0.0029 0.018 0.0012 0.022 0.0016 0.052 0.0033 0.8 0.023 0.032 0.0036 0.0009 0.011 0.00096
PROPN 0.064 0.0051 0.0092 0.0098 0.0043 0.0054 0.011 0.13 0.004 0.018 0.0021 0.049 0.63 0.025 0.0024 0.0031 0.027 0.0023
PUNCT 0.0032 0.0047 0.0075 0.0046 0.11 0.0025 0.00047 0.023 0.00049 0.14 0.00088 0.0081 0.01 0.67 0.0085 0.0015 0.0055 0.00086
SCONJ 0.017 0.059 0.053 0.017 0.025 0.018 0.0022 0.035 0.0022 0.2 0.038 0.041 0.018 0.031 0.38 0 0.068 0.00031
SYM 0.041 0.017 0.027 0.045 0.011 0.0045 0.023 0.14 0.038 0.29 0.0015 0.017 0.095 0.075 0 0.14 0.026 0.012
VERB 0.03 0.015 0.0099 0.032 0.0043 0.0052 0.002 0.068 0.0031 0.064 0.0031 0.012 0.046 0.01 0.0042 0.00054 0.69 0.00069
X 0.042 0.014 0.036 0.034 0.0085 0.027 0.023 0.14 0.018 0.047 0.0076 0.046 0.22 0.027 0.00095 0.0047 0.09 0.22
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0


Syntactic relations English - br
English
br
ADJ 0.6 0.0094 0.033 0.016 0.0021 0.031 0.006 0.1 0.0095 0.049 0.0021 0.013 0.052 0.0083 0.0014 0.0013 0.057 0.001
ADP 0.0076 0.56 0.02 0.0078 0.012 0.037 0.0027 0.016 0.0042 0.21 0.036 0.016 0.0069 0.0079 0.014 0.00033 0.038 0.00049
ADV 0.08 0.037 0.36 0.032 0.011 0.026 0.0052 0.068 0.007 0.16 0.02 0.073 0.025 0.024 0.015 0.00067 0.055 0.00083
AUX 0.013 0.015 0.012 0.43 0.0077 0.011 0.0025 0.028 0.002 0.18 0.013 0.12 0.021 0.014 0.0059 0.0014 0.12 0.0011
CCONJ 0.0073 0.054 0.087 0.053 0.17 0.046 0.00065 0.035 0.0014 0.2 0.01 0.068 0.0051 0.22 0.019 0.00065 0.02 0.0012
DET 0.022 0.039 0.011 0.0084 0.017 0.29 0.00032 0.01 0.0054 0.48 0.012 0.063 0.0026 0.0091 0.0054 0.00017 0.027 0.00053
INTJ 0.029 0.009 0.03 0.029 0.0024 0.0057 0.47 0.055 0.0064 0.034 0.014 0.16 0.097 0.01 0.0078 0.0014 0.041 0.0027
NOUN 0.066 0.0078 0.013 0.0068 0.01 0.0071 0.0012 0.66 0.0069 0.051 0.0033 0.015 0.078 0.01 0.0024 0.00077 0.059 0.0011
NUM 0.056 0.016 0.034 0.0057 0.0037 0.087 0.00051 0.073 0.55 0.094 0.002 0.014 0.018 0.015 0.0031 0.0032 0.019 0.0016
None 0.02 0.11 0.021 0.071 0.026 0.19 0.0008 0.066 0.0017 8.6e-05 0.056 0.12 0.011 0.22 0.018 0.0045 0.061 0.0023
PART 0.0065 0.091 0.037 0.065 0.012 0.069 0.005 0.057 0.0016 0.28 0.17 0.075 0.011 0.008 0.051 0.00066 0.057 0.00087
PRON 0.0044 0.015 0.015 0.028 0.0062 0.028 0.0012 0.026 0.0016 0.075 0.004 0.72 0.012 0.035 0.0057 0.00053 0.019 0.00086
PROPN 0.053 0.015 0.014 0.014 0.0039 0.02 0.008 0.21 0.01 0.03 0.003 0.071 0.49 0.019 0.0021 0.0021 0.034 0.0043
PUNCT 0.005 0.0074 0.0082 0.006 0.088 0.0045 0.00043 0.029 0.00083 0.2 0.0012 0.013 0.011 0.6 0.0058 0.0014 0.0096 0.0021
SCONJ 0.012 0.12 0.072 0.013 0.018 0.027 0.00081 0.033 0.002 0.22 0.039 0.061 0.012 0.022 0.29 0 0.051 0.00044
SYM 0.047 0.017 0.024 0.028 0.0083 0.0038 0.0041 0.16 0.059 0.3 0.00095 0.012 0.046 0.073 0.0019 0.14 0.053 0.013
VERB 0.029 0.014 0.011 0.058 0.0044 0.0064 0.0013 0.063 0.0031 0.068 0.0056 0.014 0.014 0.009 0.0037 0.00035 0.69 0.00049
X 0.044 0.032 0.017 0.023 0.08 0.09 0.0042 0.12 0.013 0.087 0.007 0.038 0.12 0.026 0.0081 0.0015 0.085 0.2
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0


Syntactic relations English - jp
English
jp
ADJ 0.55 0.017 0.042 0.0073 0.003 0.049 0.0016 0.12 0.0054 0.061 0.0016 0.019 0.045 0.0092 0.0024 0.0016 0.06 0.0032
ADP 0.0061 0.51 0.019 0.0066 0.0096 0.039 0.00071 0.018 0.0028 0.3 0.015 0.014 0.0033 0.0092 0.017 0.00024 0.025 0.00071
ADV 0.07 0.031 0.39 0.022 0.008 0.038 0.0033 0.074 0.0036 0.21 0.0076 0.057 0.014 0.024 0.02 0.0011 0.032 0
AUX 0.009 0.019 0.014 0.51 0.0043 0.011 0.0023 0.027 0.0013 0.2 0.018 0.038 0.003 0.0093 0.0036 0.00066 0.12 0.00033
CCONJ 0.012 0.036 0.15 0.045 0.2 0.048 0.0023 0.03 0 0.21 0.0023 0.064 0.0023 0.17 0.024 0 0.014 0.00078
DET 0.023 0.028 0.0088 0.01 0.0032 0.38 0 0.0092 0.0025 0.45 0.0056 0.054 0.0018 0.0023 0.007 0 0.018 0.00023
INTJ 0.046 0.014 0.096 0.11 0 0 0.39 0.041 0.0091 0.023 0.0046 0.1 0.12 0 0.023 0 0.023 0
NOUN 0.047 0.0065 0.013 0.0037 0.0016 0.0076 0.00094 0.72 0.0048 0.061 0.003 0.015 0.057 0.0096 0.0033 0.0005 0.049 0.00063
NUM 0.027 0.014 0.029 0.012 0.0024 0.077 0 0.12 0.51 0.11 0.0024 0.022 0.029 0.019 0.0024 0.0024 0.027 0.0024
None 0.013 0.12 0.023 0.07 0.017 0.34 0.0013 0.063 0.0019 5e-05 0.019 0.072 0.0082 0.17 0.022 0.0015 0.053 0.001
PART 0.0025 0.11 0.03 0.048 0.017 0.065 0.0075 0.046 0.0012 0.29 0.19 0.065 0.011 0.0037 0.073 0 0.045 0
PRON 0.01 0.013 0.029 0.017 0.0042 0.055 0.0022 0.05 0.0028 0.13 0.0056 0.61 0.015 0.019 0.01 0 0.025 0
PROPN 0.058 0.0083 0.017 0.02 0.0047 0.013 0.0093 0.25 0.0047 0.041 0.0036 0.059 0.44 0.035 0.0036 0.0021 0.034 0.0031
PUNCT 0.0031 0.0072 0.0092 0.0082 0.13 0.0053 0 0.028 0.00048 0.21 0.0017 0.012 0.007 0.52 0.047 0.00072 0.0072 0.0012
SCONJ 0.0074 0.075 0.049 0.0093 0.017 0.03 0.00093 0.021 0.00093 0.26 0.059 0.059 0.0037 0.021 0.34 0 0.041 0.00093
SYM 0.024 0.04 0 0.0081 0 0.0081 0 0.22 0.0081 0.48 0 0 0.024 0.065 0 0.081 0.048 0
VERB 0.035 0.01 0.01 0.035 0.0015 0.0086 0.00053 0.064 0.0022 0.072 0.003 0.013 0.0075 0.0047 0.0068 0.00064 0.72 0.00053
X 0.063 0.0063 0.025 0.019 0 0.051 0.013 0.12 0.025 0.11 0.0063 0.076 0.076 0.032 0 0.019 0.13 0.22
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0

Prominent Words

We can see **None->DET** is standing out, So let's see what words are prominent in this error-type.

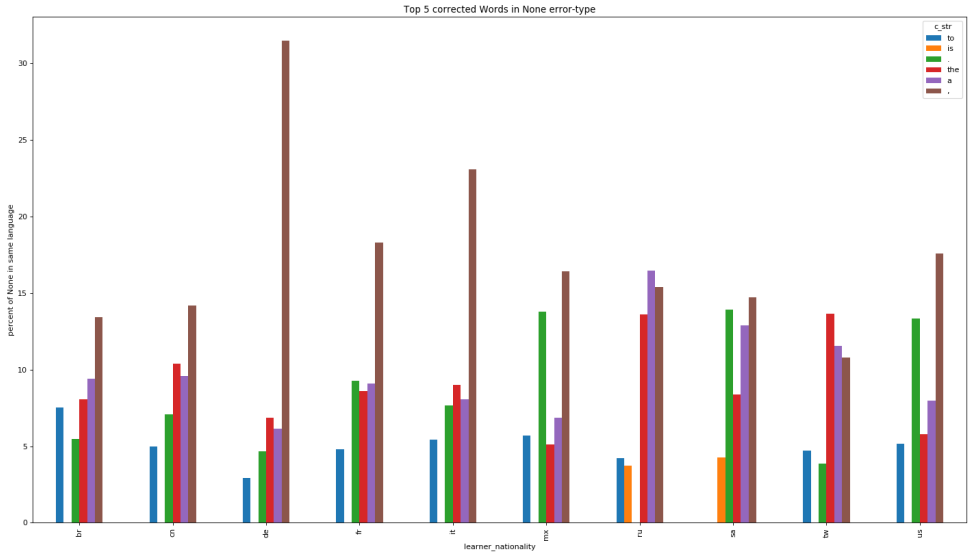


We can see there are 2 prominent words in every language. All the rest are less frequent.

'The' and 'a'. For cn, de, it and tw(taiwan, also chinese) 'the' is the most missed word.

For br, fr, mx, ru, and us 'a' is the most missed word.

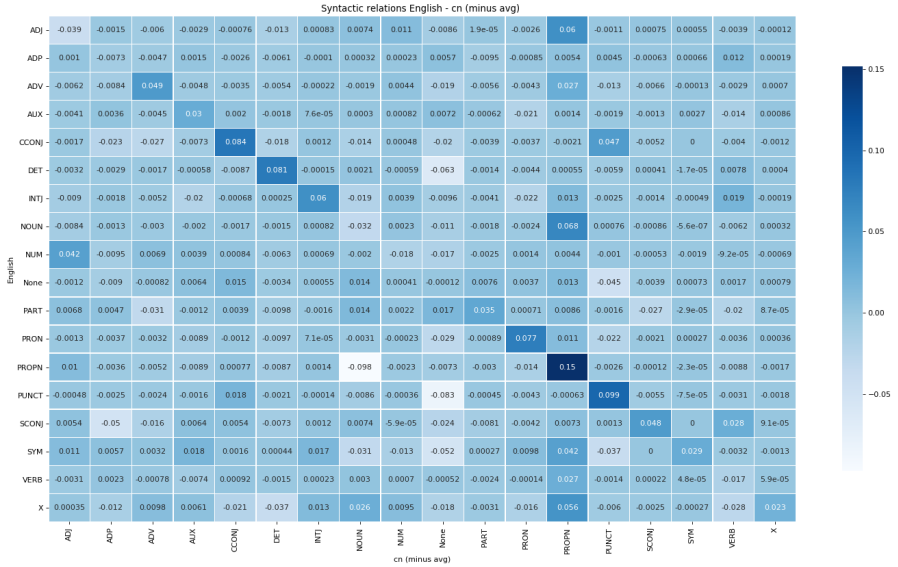
If we look at all missed words(**None->***):

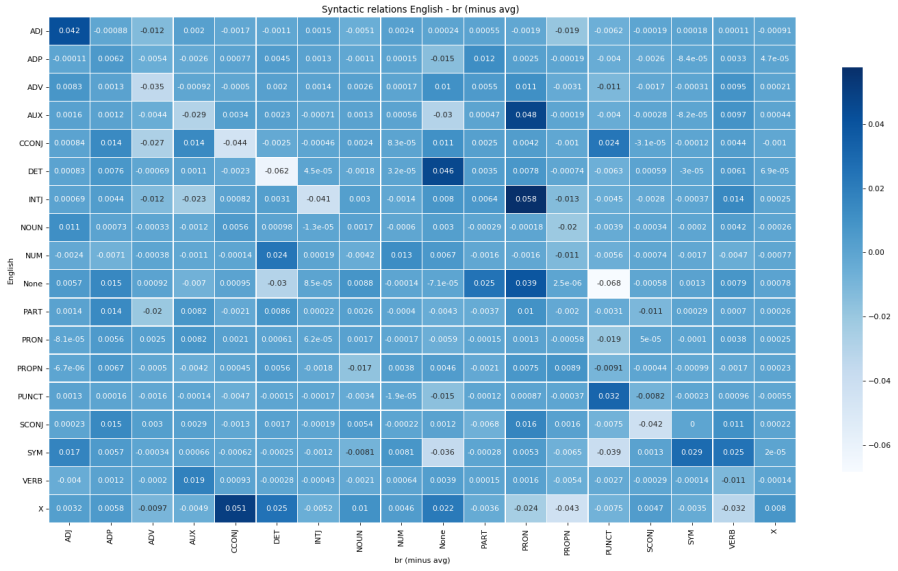


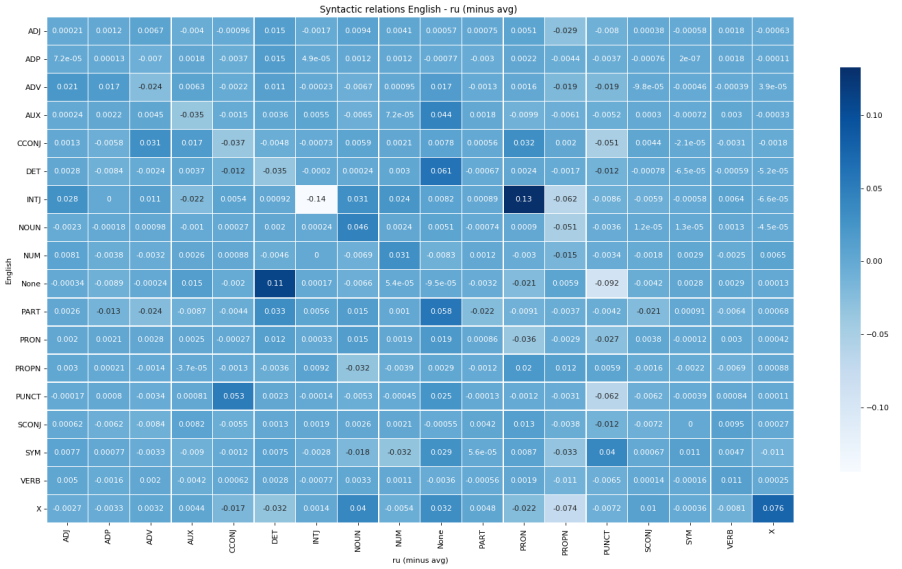
We can see that comma(,) is prominent, except for ru with 'a' and tw with 'the'.

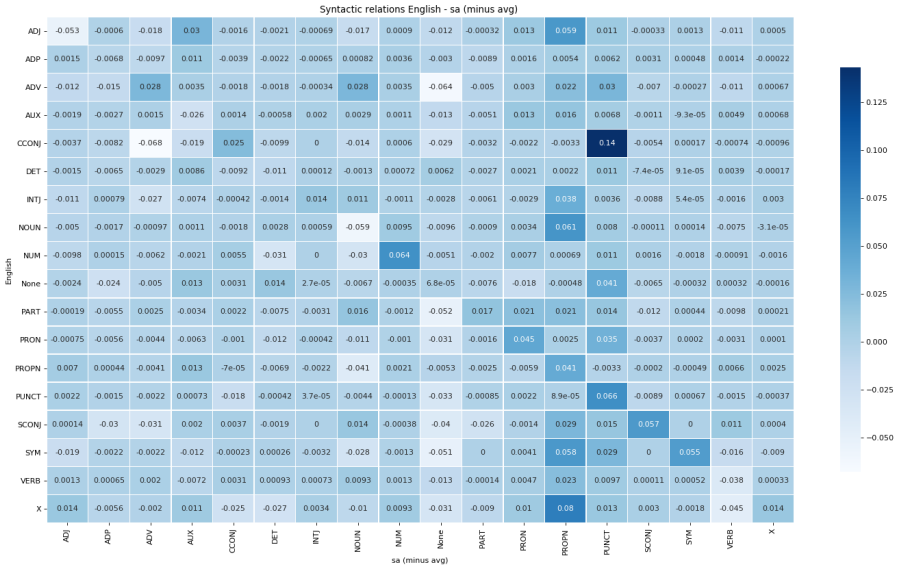
Interestingly, ‘to’ joined the top 5 missed words, except for sa(saudi - arabic) where ‘is’ joined instead. ‘Is’ joined for ru as well.
Interestingly the comma is over-represented in de. (Is comma less prevalent in German?)

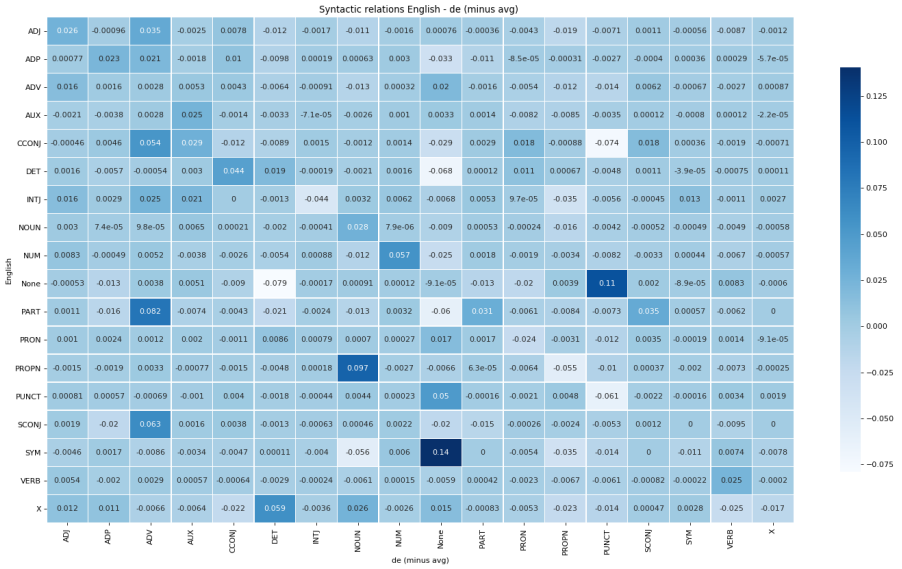
Classified Syntactic Errors by nationality, minus the average across nationalities:

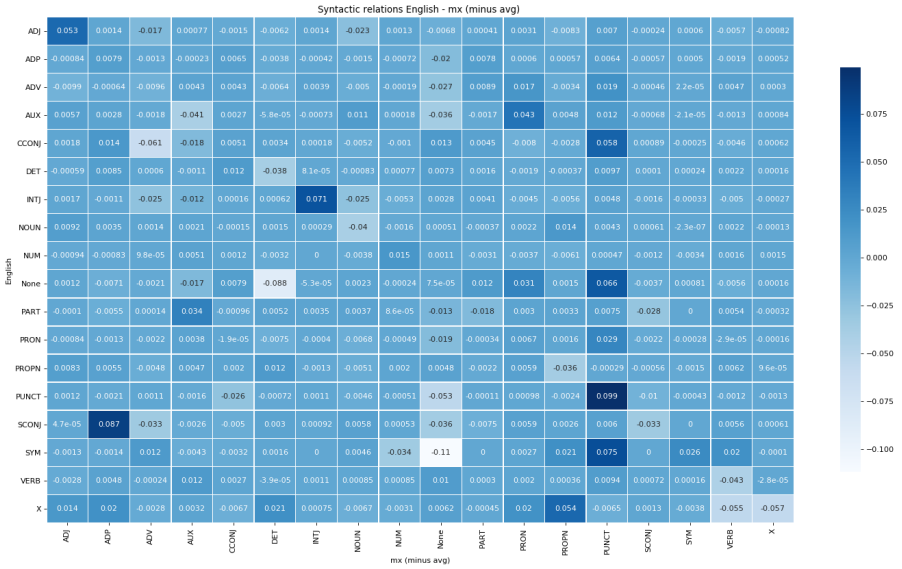




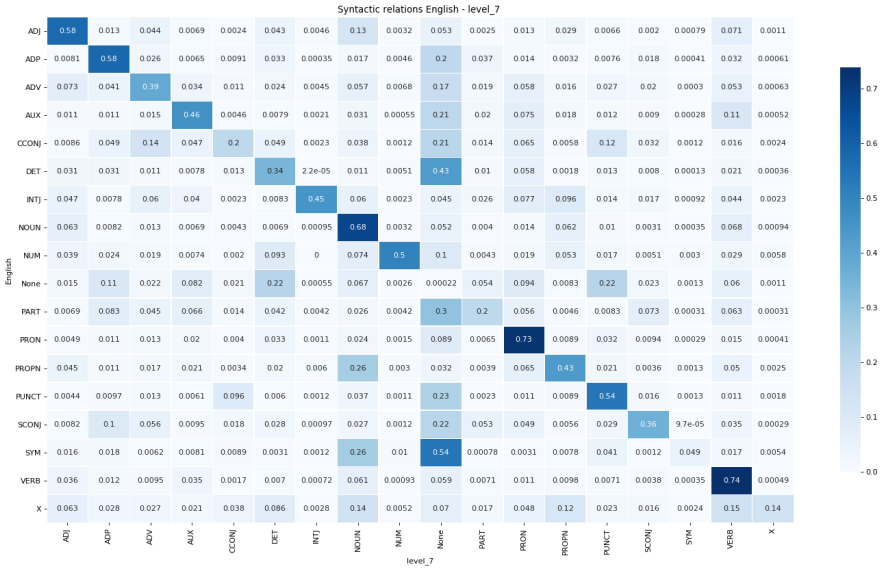

Syntactic relations English - ru (minus avg)
English
ADJ 0.00021 0.0012 0.0067 -0.004 -0.00096 0.015 -0.0017 0.0094 0.0041 0.00057 0.00075 0.0051 -0.029 -0.008 0.00038 -0.00058 0.0018 -0.00063
ADP 7.2e-05 0.00013 -0.007 0.0018 -0.0037 0.015 4.9e-05 0.0012 0.0012 -0.00077 -0.003 0.0022 -0.0044 -0.0037 -0.00076 2e-07 0.0018 -0.00011
ADV 0.021 0.017 -0.024 0.0063 -0.0022 0.011 -0.00023 -0.0067 0.00095 0.017 -0.0013 0.0016 -0.019 -0.019 -9.8e-05 -0.00046 -0.00039 3.9e-05
AUX 0.00024 0.0022 0.0045 -0.035 -0.0015 0.0036 0.0055 -0.0065 7.2e-05 0.044 0.0018 -0.0099 -0.0061 -0.0052 0.0003 -0.00072 0.003 -0.00033
CCONJ 0.0013 -0.0058 0.031 0.017 -0.037 -0.0048 -0.00073 0.0059 0.0021 0.0078 0.00056 0.032 0.002 -0.051 0.0044 -2.1e-05 -0.0031 -0.0018
DET 0.0028 -0.0084 -0.0024 0.0037 -0.012 -0.035 -0.0002 0.00024 0.003 0.061 -0.00067 0.0024 -0.0017 -0.012 -0.00078 -6.5e-05 -0.00059 -5.2e-05
INTJ 0.028 0 0.011 -0.022 0.0054 0.00092 -0.14 0.031 0.024 0.0082 0.00089 0.13 -0.062 -0.0086 -0.0059 -0.00058 0.0064 -6.6e-05
NOUN -0.0023 -0.00018 0.00098 -0.001 0.00027 0.002 0.00024 0.046 0.0024 0.0051 -0.00074 0.0009 -0.051 -0.0036 1.2e-05 1.3e-05 0.0013 -4.5e-05
NUM 0.0081 -0.0038 -0.0032 0.0026 0.00088 -0.0046 0 -0.0069 0.031 -0.0083 0.0012 -0.003 -0.015 -0.0034 -0.0018 0.0029 -0.0025 0.0065
None -0.00034 -0.0089 -0.00024 0.015 -0.002 0.11 0.00017 -0.0066 5.4e-05 9.5e-05 -0.0032 -0.021 0.0059 -0.092 -0.0042 0.0028 0.0029 0.00013
PART 0.0026 -0.013 -0.024 -0.0087 -0.0044 0.033 0.0056 0.015 0.001 0.058 -0.022 -0.0091 -0.0037 0.0042 -0.021 0.00091 -0.0064 0.00068
PRON 0.002 0.0021 0.0028 0.0025 -0.00027 0.012 0.00033 0.015 0.0019 0.019 0.00086 -0.036 -0.0029 -0.027 0.0038 -0.00012 0.003 0.00042
PROPN 0.003 0.00021 -0.0014 -3.7e-05 -0.0013 -0.0036 0.0092 -0.032 -0.0039 0.0029 -0.0012 0.02 0.012 0.0059 -0.0016 -0.0022 -0.0069 0.00088
PUNCT -0.00017 0.0008 -0.0034 0.00081 0.053 0.0023 -0.00014 -0.0053 -0.00045 0.025 -0.00013 -0.0012 -0.0031 -0.062 -0.0062 -0.00039 0.00084 0.00011
SCONJ 0.00062 -0.0062 -0.0084 0.0082 -0.0055 0.0013 0.0019 0.0026 0.0021 -0.00055 0.0042 0.013 -0.0038 -0.012 -0.0072 0 0.0095 0.00027
SYM 0.0077 0.00077 -0.0033 -0.009 -0.0012 0.0075 -0.0028 -0.018 -0.032 0.029 5.6e-05 0.0087 -0.033 0.04 0.00067 0.011 0.0047 -0.011
VERB 0.005 -0.0016 0.002 -0.0042 0.00062 0.0028 -0.00077 0.0033 0.0011 -0.0036 -0.00056 0.0019 -0.011 -0.0065 0.00014 -0.00016 0.011 0.00025
X -0.0027 -0.0033 0.0032 0.0044 -0.017 -0.032 0.0014 0.04 -0.0054 0.032 0.0048 -0.022 -0.074 -0.0072 0.01 -0.00036 -0.0081 0.076
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
ru (minus avg)
0.10
0.05
0.00
−0.05
−0.10

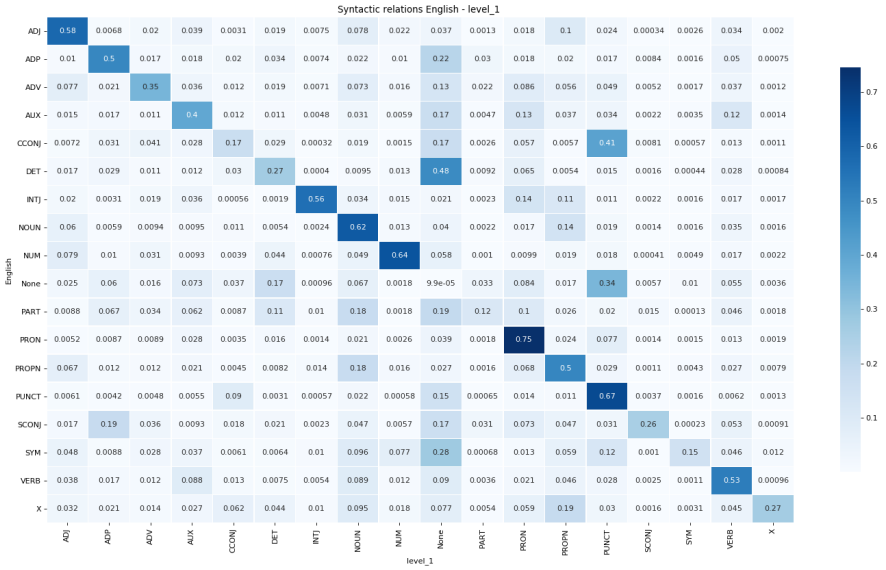

Syntactic relations English - sa (minus avg)
English
ADJ -0.053 -0.0006 -0.018 0.03 -0.0016 -0.0021 -0.00069 -0.017 0.0009 -0.012 -0.00032 0.013 0.059 0.011 -0.00033 0.0013 -0.011 0.0005
ADP 0.0015 -0.0068 -0.0097 0.011 -0.0039 -0.0022 -0.00065 0.00082 0.0036 -0.003 -0.0089 0.0016 0.0054 0.0062 0.0031 0.00048 0.0014 -0.00022
ADV -0.012 -0.015 0.028 0.0035 -0.0018 -0.0018 -0.00034 0.028 0.0035 -0.064 -0.005 0.003 0.022 0.03 -0.007 -0.00027 -0.011 0.00067
AUX -0.0019 -0.0027 0.0015 -0.026 0.0014 -0.00058 0.002 0.0029 0.0011 -0.013 -0.0051 0.013 0.016 0.0068 -0.0011 -9.3e-05 0.0049 0.00068
CCONJ -0.0037 -0.0082 -0.068 -0.019 0.025 -0.0099 0 -0.014 0.0006 -0.029 -0.0032 -0.0022 -0.0033 0.14 -0.0054 0.00017 -0.00074 -0.00096
DET -0.0015 -0.0065 -0.0029 0.0086 -0.0092 -0.011 0.00012 -0.0013 0.00072 0.0062 -0.0027 0.0021 0.0022 0.011 -7.4e-05 9.1e-05 0.0039 -0.00017
INTJ -0.011 0.00079 -0.027 -0.0074 -0.00042 -0.0014 0.014 0.011 -0.0011 -0.0028 -0.0061 -0.0029 0.038 0.0036 -0.0088 5.4e-05 -0.0016 0.003
NOUN -0.005 -0.0017 -0.00097 0.0011 -0.0018 0.0028 0.00059 -0.059 0.0095 -0.0096 -0.0009 0.0034 0.061 0.008 -0.00011 0.00014 -0.0075 -3.1e-05
NUM -0.0098 0.00015 -0.0062 -0.0021 0.0055 -0.031 0 -0.03 0.064 -0.0051 -0.002 0.0077 0.00069 0.011 0.0016 -0.0018 -0.00091 -0.0016
None -0.0024 -0.024 -0.005 0.013 0.0031 0.014 2.7e-05 -0.0067 -0.00035 6.8e-05 -0.0076 -0.018 -0.00048 0.041 -0.0065 -0.00032 0.00032 -0.00016
PART -0.00019 -0.0055 0.0025 -0.0034 0.0022 -0.0075 -0.0031 0.016 -0.0012 -0.052 0.017 0.021 0.021 0.014 -0.012 0.00044 -0.0098 0.00021
PRON -0.00075 -0.0056 -0.0044 -0.0063 -0.001 -0.012 -0.00042 -0.011 -0.001 -0.031 -0.0016 0.045 0.0025 0.035 -0.0037 0.0002 -0.0031 0.0001
PROPN 0.007 0.00044 -0.0041 0.013 -7e-05 -0.0069 -0.0022 -0.041 0.0021 -0.0053 -0.0025 -0.0059 0.041 -0.0033 -0.0002 -0.00049 0.0066 0.0025
PUNCT 0.0022 -0.0015 -0.0022 0.00073 -0.018 -0.00042 3.7e-05 -0.0044 -0.00013 -0.033 -0.00085 0.0022 8.9e-05 0.066 -0.0089 0.00067 -0.0015 -0.00037
SCONJ 0.00014 -0.03 -0.031 0.002 0.0037 -0.0019 0 0.014 -0.00038 -0.04 -0.026 -0.0014 0.029 0.015 0.057 0 0.011 0.0004
SYM -0.019 -0.0022 -0.0022 -0.012 -0.00023 0.00026 -0.0032 -0.028 -0.0013 -0.051 0 0.0041 0.058 0.029 0 0.055 -0.016 -0.009
VERB 0.0013 0.00065 0.002 -0.0072 0.0031 0.00093 0.00073 0.0093 0.0013 -0.013 -0.00014 0.0047 0.023 0.0097 0.00011 0.00052 -0.038 0.00033
X 0.014 -0.0056 -0.002 0.011 -0.025 -0.027 0.0034 -0.01 0.0093 -0.031 -0.009 0.01 0.08 0.013 0.003 -0.0018 -0.045 0.014
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
sa (minus avg)
0.125
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050

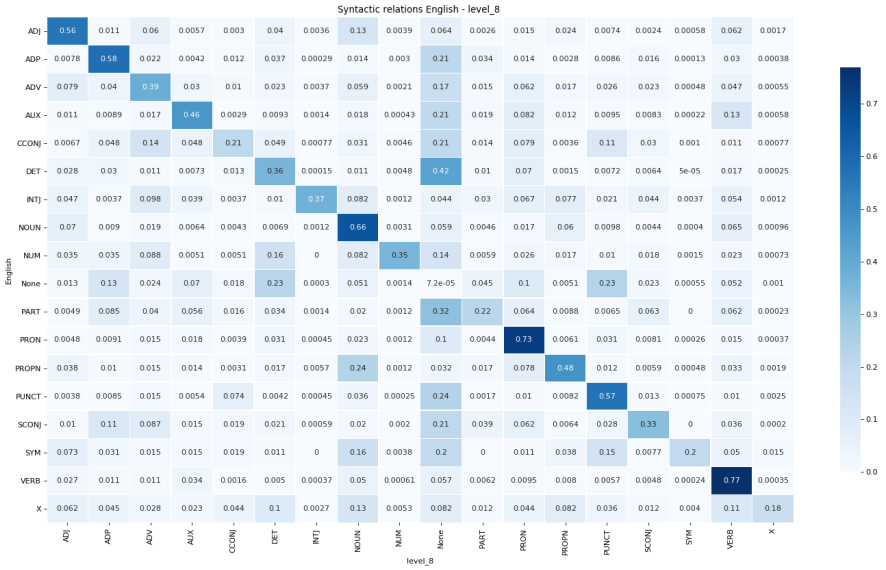

Syntactic relations English - de (minus avg)
English
ADJ 0.026 -0.00096 0.035 -0.0025 0.0078 -0.012 -0.0017 -0.011 -0.0016 0.00076 -0.00036 -0.0043 -0.019 -0.0071 0.0011 -0.00056 -0.0087 -0.0012
ADP 0.00077 0.023 0.021 -0.0018 0.01 -0.0098 0.00019 0.00063 0.003 -0.033 -0.011 -8.5e-05 -0.00031 -0.0027 -0.0004 0.00036 0.00029 -5.7e-05
ADV 0.016 0.0016 0.0028 0.0053 0.0043 -0.0064 -0.00091 -0.013 0.00032 0.02 -0.0016 -0.0054 -0.012 -0.014 0.0062 -0.00067 -0.0027 0.00087
AUX -0.0021 -0.0038 0.0028 0.025 -0.0014 -0.0033 -7.1e-05 -0.0026 0.001 0.0033 0.0014 -0.0082 -0.0085 -0.0035 0.00012 -0.0008 0.00012 -2.2e-05
CCONJ -0.00046 0.0046 0.054 0.029 -0.012 -0.0089 0.0015 -0.0012 0.0014 -0.029 0.0029 0.018 -0.00088 -0.074 0.018 0.00036 -0.0019 -0.00071
DET 0.0016 -0.0057 -0.00054 0.003 0.044 0.019 -0.00019 -0.0021 0.0016 -0.068 0.00012 0.011 0.00067 -0.0048 0.0011 -3.9e-05 -0.00075 0.00011
INTJ 0.016 0.0029 0.025 0.021 0 -0.0013 -0.044 0.0032 0.0062 -0.0068 0.0053 9.7e-05 -0.035 -0.0056 -0.00045 0.013 -0.0011 0.0027
NOUN 0.003 7.4e-05 9.8e-05 0.0065 0.00021 -0.002 -0.00041 0.028 7.9e-06 -0.009 0.00053 -0.00024 -0.016 -0.0042 -0.00052 -0.00049 -0.0049 -0.00058
NUM 0.0083 -0.00049 0.0052 -0.0038 -0.0026 -0.0054 0.00088 -0.012 0.057 -0.025 0.0018 -0.0019 -0.0034 -0.0082 -0.0033 0.00044 -0.0067 -0.00057
None -0.00053 -0.013 0.0038 0.0051 -0.009 -0.079 -0.00017 0.00091 0.00012 -9.1e-05 -0.013 -0.02 0.0039 0.11 0.002 -8.9e-05 0.0083 -0.0006
PART 0.0011 -0.016 0.082 -0.0074 -0.0043 -0.021 -0.0024 -0.013 0.0032 -0.06 0.031 -0.0061 -0.0084 -0.0073 0.035 0.00057 -0.0062 0
PRON 0.001 0.0024 0.0012 0.002 -0.0011 0.0086 0.00079 0.0007 0.00027 0.017 0.0017 -0.024 -0.0031 -0.012 0.0035 -0.00019 0.0014 -9.1e-05
PROPN -0.0015 -0.0019 0.0033 -0.00077 -0.0015 -0.0048 0.00018 0.097 -0.0027 -0.0066 6.3e-05 -0.0064 -0.055 -0.01 0.00037 -0.002 -0.0073 -0.00025
PUNCT 0.00081 0.00057 -0.00069 -0.001 0.004 -0.0018 -0.00044 0.0044 0.00023 0.05 -0.00016 -0.0021 0.0048 -0.061 -0.0022 -0.00016 0.0034 0.0019
SCONJ 0.0019 -0.02 0.063 0.0016 0.0038 -0.0013 -0.00063 0.00046 0.0022 -0.02 -0.015 -0.00026 -0.0024 -0.0053 0.0012 0 -0.0095 0
SYM -0.0046 0.0017 -0.0086 -0.0034 -0.0047 0.00011 -0.004 -0.056 0.006 0.14 0 -0.0054 -0.035 -0.014 0 -0.011 0.0074 -0.0078
VERB 0.0054 -0.002 0.0029 0.00057 -0.00064 -0.0029 -0.00024 -0.0061 0.00015 -0.0059 0.00042 -0.0023 -0.0067 -0.0061 -0.00082 -0.00022 0.025 -0.0002
X 0.012 0.011 -0.0066 -0.0064 -0.022 0.059 -0.0036 0.026 -0.0026 0.015 -0.00083 -0.0053 -0.023 -0.014 0.00047 0.0028 -0.025 -0.017
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
de (minus avg)
0.125
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075

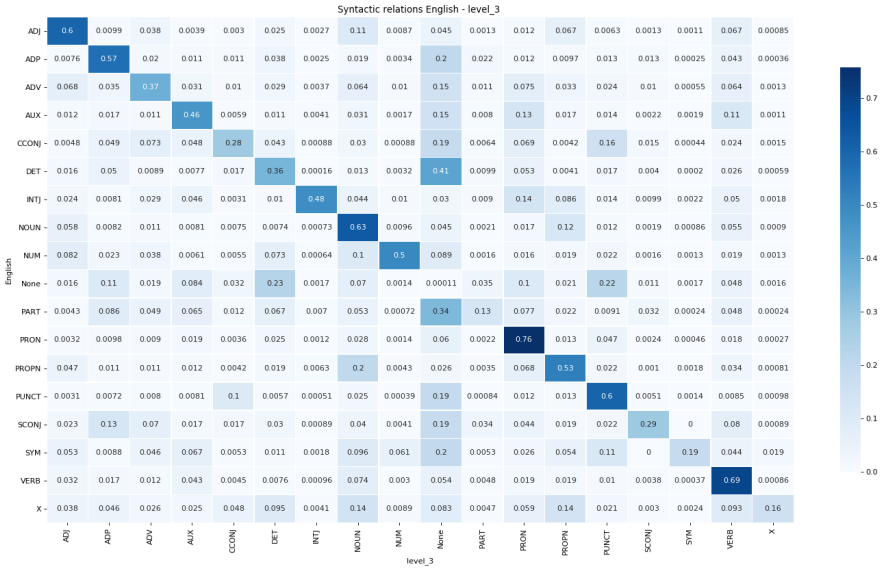

Syntactic relations English - mx (minus avg)
English
ADJ 0.053 0.0014 -0.017 0.00077 -0.0015 -0.0062 0.0014 -0.023 0.0013 -0.0068 0.00041 0.0031 -0.0083 0.007 -0.00024 0.0006 -0.0057 -0.00082
ADP -0.00084 0.0079 -0.0013 -0.00023 0.0065 -0.0038 -0.00042 -0.0015 -0.00072 -0.02 0.0078 0.0006 0.00057 0.0064 -0.00057 0.0005 -0.0019 0.00052
ADV -0.0099 -0.00064 -0.0096 0.0043 0.0043 -0.0064 0.0039 -0.005 -0.00019 -0.027 0.0089 0.017 -0.0034 0.019 -0.00046 2.2e-05 0.0047 0.0003
AUX 0.0057 0.0028 -0.0018 -0.041 0.0027 -5.8e-05 -0.00073 0.011 0.00018 -0.036 -0.0017 0.043 0.0048 0.012 -0.00068 -2.1e-05 -0.0013 0.00084
CCONJ 0.0018 0.014 -0.061 -0.018 0.0051 0.0034 0.00018 -0.0052 -0.001 0.013 0.0045 -0.008 -0.0028 0.058 0.00089 -0.00025 -0.0046 0.00062
DET -0.00059 0.0085 0.0006 -0.0011 0.012 -0.038 8.1e-05 -0.00083 0.00077 0.0073 0.0016 -0.0019 -0.00037 0.0097 0.0001 0.00024 0.0022 0.00016
INTJ 0.0017 -0.0011 -0.025 -0.012 0.00016 0.00062 0.071 -0.025 -0.0053 0.0028 0.0041 -0.0045 -0.0056 0.0048 -0.0016 -0.00033 -0.005 -0.00027
NOUN 0.0092 0.0035 0.0014 0.0021 -0.00015 0.0015 0.00029 -0.04 -0.0016 0.00051 -0.00037 0.0022 0.014 0.0043 0.00061 -2.3e-07 0.0022 -0.00013
NUM -0.00094 -0.00083 9.8e-05 0.0051 0.0012 -0.0032 0 -0.0038 0.015 0.0011 -0.0031 -0.0037 -0.0061 0.00047 -0.0012 -0.0034 0.0016 0.0015
None 0.0012 -0.0071 -0.0021 -0.017 0.0079 -0.088 -5.3e-05 0.0023 -0.00024 7.5e-05 0.012 0.031 0.0015 0.066 -0.0037 0.00081 -0.0056 0.00016
PART -0.0001 -0.0055 0.00014 0.034 -0.00096 0.0052 0.0035 0.0037 8.6e-05 -0.013 -0.018 0.003 0.0033 0.0075 -0.028 0 0.0054 -0.00032
PRON -0.00084 -0.0013 -0.0022 0.0038 -1.9e-05 -0.0075 -0.0004 -0.0068 -0.00049 -0.019 -0.00034 0.0067 0.0016 0.029 -0.0022 -0.00028 -2.9e-05 -0.00016
PROPN 0.0083 0.0055 -0.0048 0.0047 0.002 0.012 -0.0013 -0.0051 0.002 0.0048 -0.0022 0.0059 -0.036 -0.00029 -0.00056 -0.0015 0.0062 9.6e-05
PUNCT 0.0012 -0.0021 0.0011 -0.0016 -0.026 -0.00072 0.0011 -0.0046 -0.00051 -0.053 -0.00011 0.00098 -0.0024 0.099 -0.01 -0.00043 -0.0012 -0.0013
SCONJ 4.7e-05 0.087 -0.033 -0.0026 -0.005 0.003 0.00092 0.0058 0.00053 -0.036 -0.0075 0.0059 0.0026 0.006 -0.033 0 0.0056 0.00061
SYM -0.0013 -0.0014 0.012 -0.0043 -0.0032 0.0016 0 0.0046 -0.034 -0.11 0 0.0027 0.021 0.075 0 0.026 0.02 -0.0001
VERB -0.0028 0.0048 -0.00024 0.012 0.0027 -3.9e-05 0.0011 0.00085 0.00085 0.01 0.0003 0.002 0.00036 0.0094 0.00072 0.00016 -0.043 -2.8e-05
X 0.014 0.02 -0.0028 0.0032 -0.0067 0.021 0.00075 -0.0067 -0.0031 0.0062 -0.00045 0.02 0.054 -0.0065 0.0013 -0.0038 -0.055 -0.057
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
mx (minus avg)
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100

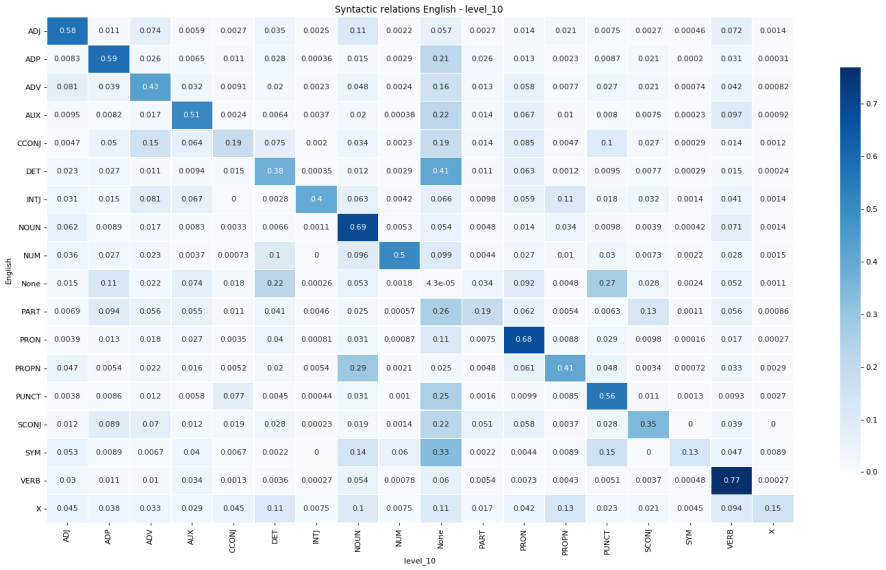
Classified Syntactic Errors by level:

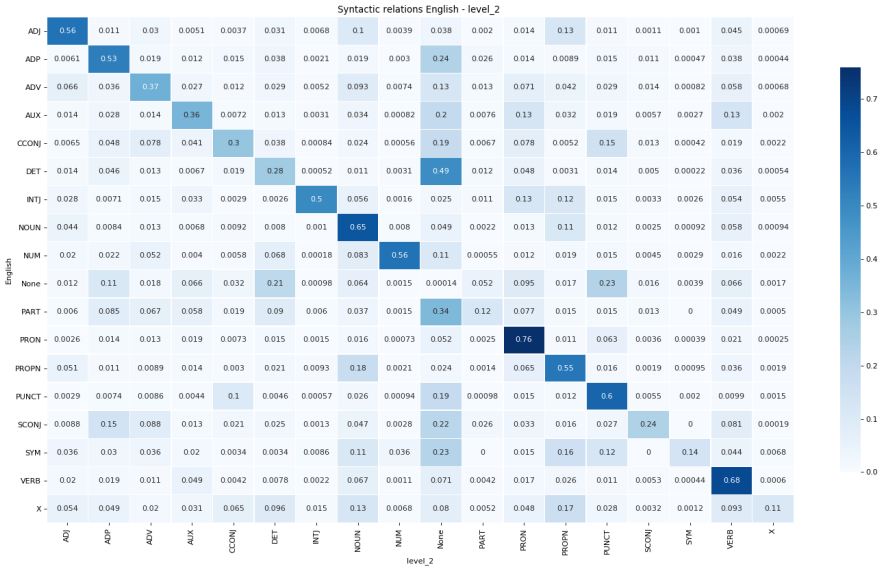





Syntactic relations English - level_8
English
level_8
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7

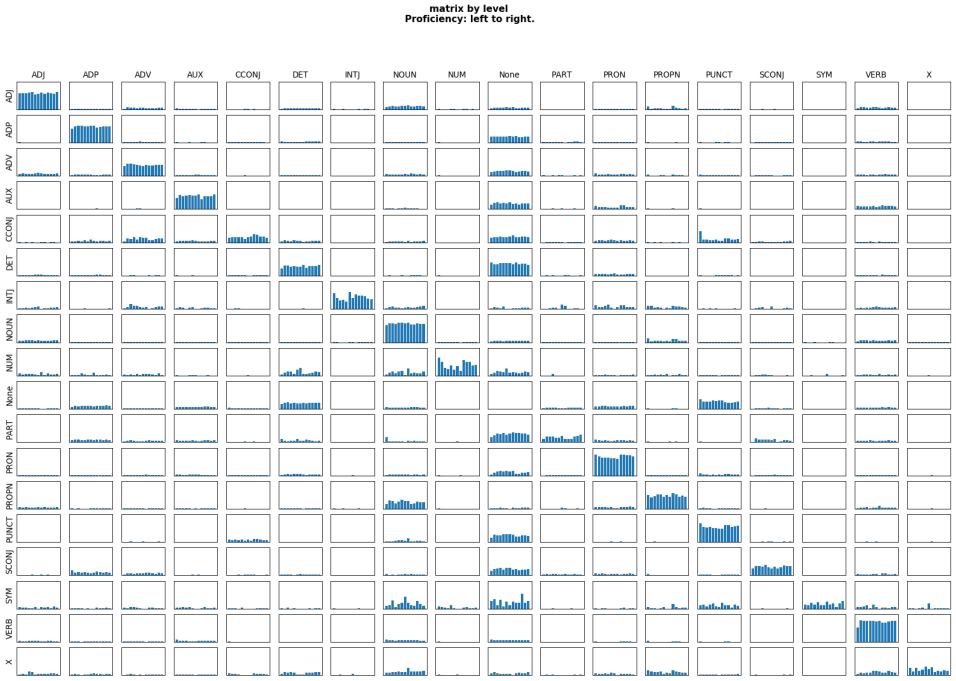

Syntactic relations English - level_3
English
level_3
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7


Syntactic relations English - level_10
English
level_10
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7


Syntactic relations English - level_2
English
level_2
ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM None PART PRON PROPN PUNCT SCONJ SYM VERB X
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7

All Levels together:

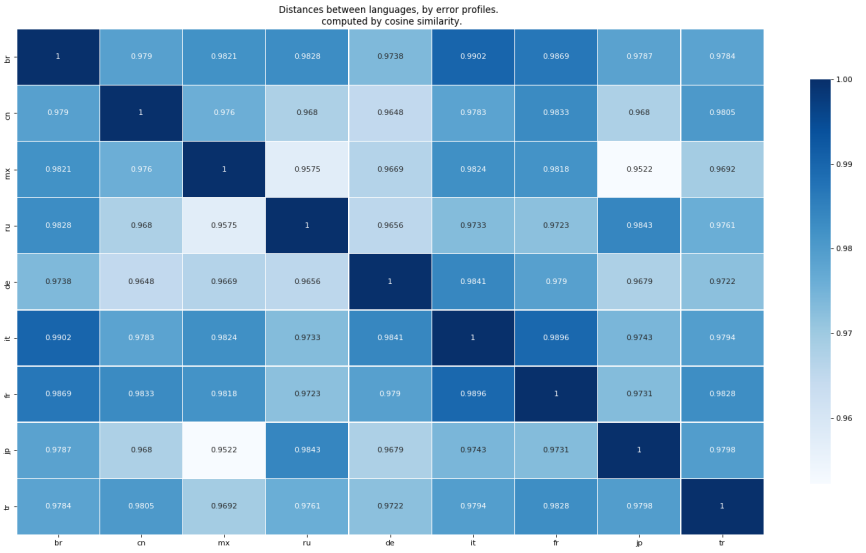
For visualization purposes, We inserted into each cell all levels together.
In each cell the levels are ordered the same, by proficiency (1,2,3,4…16):



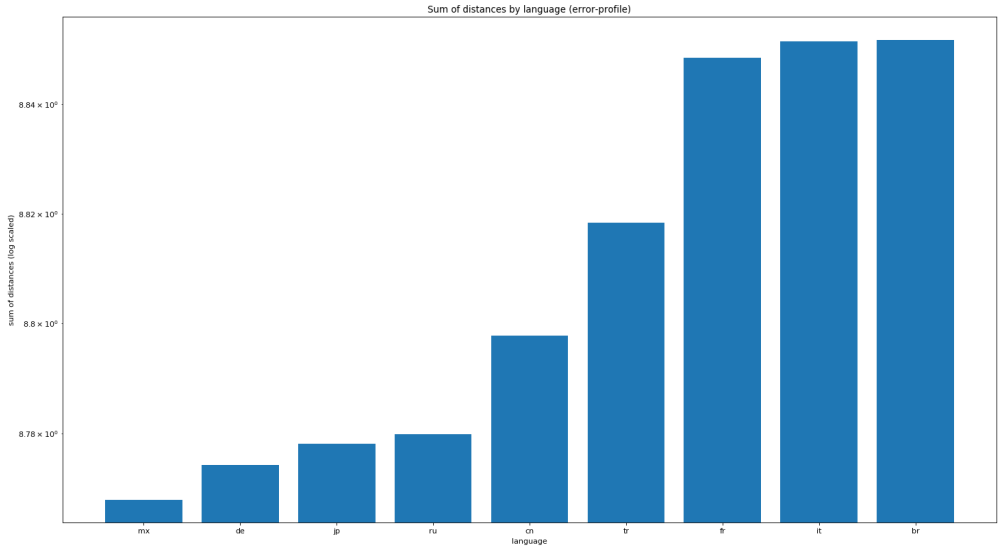
It is hard to see a specific influence or correlation of proficiency level on the error profile, but this gives a first visual guide for anyone wanting to search deeper.

Distances

We computed distances between languages, represented by their error-profile matrix. Distances were computed with cosine similarity
Cosine similarity this was done per line, then averaged:



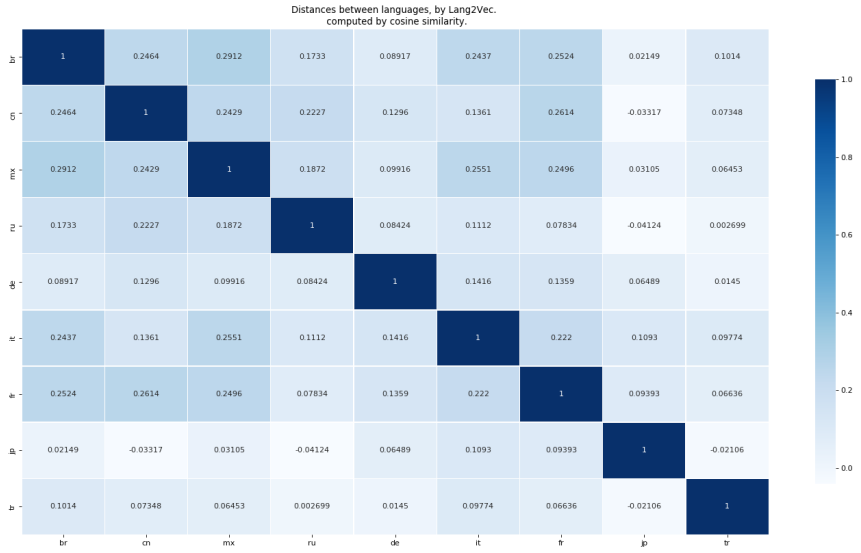
It seems that Jp and de look the least similar to all other languages.
If we sum the distances:
We used a log scale on the y axis, as the numbers are all close



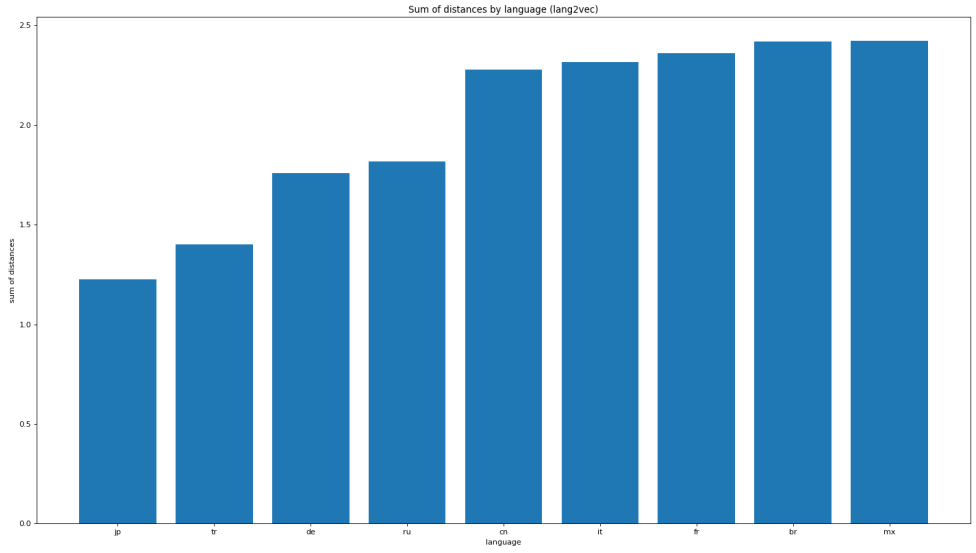
It seems mx, de, jp and ru are dissimilar to other languages, at least in regards to their error-profile.

Next we Downloaded the lang2vec dataset and extracted the features vectors.

We computed distances between languages, represented by their feature vectors.
Distances were computed with cosine similarity
Cosine similarity was done on the whole vector:



It seems that jp and tr look the least similar to all other languages.
Let's sum it:

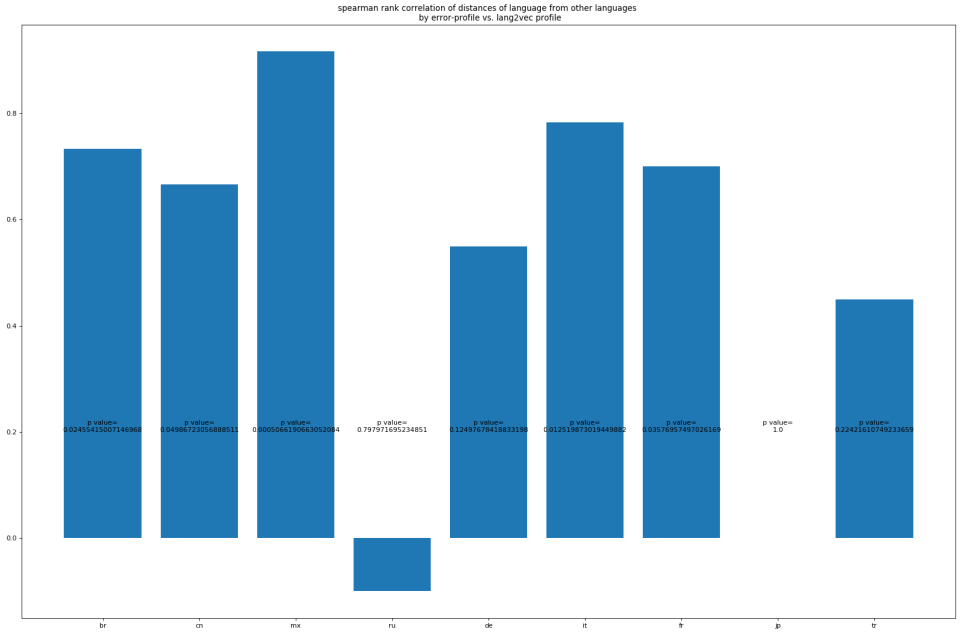


Here we didn't need the log scale..
It seems jp and tr (Turkish) are dissimilar to other languages, at least in regards to their lang2vec vector.

Interestingly, Japanese is dissimilar to other languages in both these comparisons.

Measuring distances between languages with our error-profile compared to lang2vec: Quantifying the correlation between the 2 methods for measuring the distances was done by Spearman-Rank correlation.

Measuring distances between languages with our error-profile compared to lang2vec, shows high correlation for most languages, with the highest being spanish (mexico). Interestingly Japanese and Russian showed negative or zero correlation.

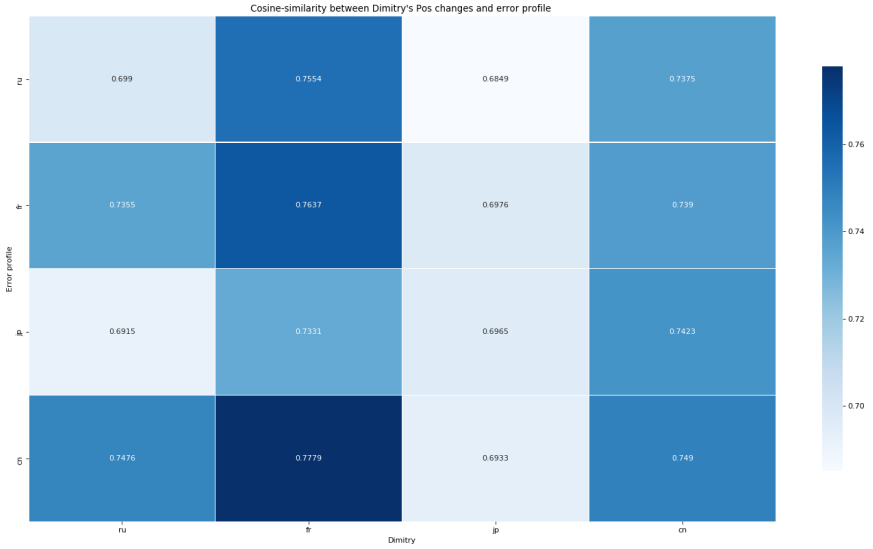


Comparing to other work:

Using data from *'Fine-Grained Analysis of Cross-Linguistic Syntactic Divergences*' (Nikolaev et al. 2020[8]) we compared various languages to our error-profile matrices for the same languages. We measured the distance between our matrices and this data, with cosine-similarity as well.

[8] https://arxiv.org/abs/2005.03436

All distances are similar and we cannot see a preference for the same language across the two datasets. (The diagonal is not emphasized.)



Results

Recap of Some results:

Measuring distances between languages with our error-profile compared to lang2vec, showed high correlation for most languages, with the highest being Spanish (mexico). Interestingly Japanese and Russian showed negative or zero correlation.

It seems jp and tr (Turkish) are dissimilar to other languages, at least in regards to their lang2vec vector. Interestingly, Japanese is dissimilar to other languages in both these comparisons.

It seems mx, de, jp and ru are dissimilar to other languages, at least in regards to their error-profile.

Missed DET POS:

We can see there are 2 prominent words in every language. All the rest are less frequent.

'The' and 'a'. For cn, de, it and tw(taiwan, also chinese) 'the' is the most missed word.

For br, fr, mx, ru, and us 'a' is the most missed word.

All missed POS:

We can see that comma(,) is prominent, except for ru with 'a' and tw with 'the'. Interestingly, 'to' joined the top 5 missed words, except for sa(saudi - arabic) where 'is' joined instead. 'Is' joined for ru as well. '.' (stop) joined the top 5 for all languages except ru.

Future work:

It can be interesting to look at the distribution of specific words in the errors in different nationalities and correlate that with the respective language syntactic use of that word.

Also studying differences in the above matrices and finding correlation to specific languages, or as a learning process, as correlated to differences in proficiency levels, in the DS as a whole, and in specific languages.

For more matrices see the appendix.

Another thing is parsing another parallel corpus the same way and comparing the results.

Link to github

Github repository of this work :
https://github.com/amichw/EFCAMDAT

References:

Works used or cited:

https://github.com/borgr/GEC_UD_divergences:

title = "Classifying Syntactic Errors in Learner Language",
author = "Choshen, Leshem and
Nikolaev, Dmitry and
Berzak, Yevgeni and
Abend, Omri",
booktitle = "Proceedings of the 24th Conference on Computational Natural Language Learning",
month = nov,
year = "2020",
address = "Online",
publisher = "Association for Computational Linguistics",
url = "https://www.aclweb.org/anthology/2020.conll-1.7",
pages = "97--107",

https://github.com/ufal/udpipe
https://github.com/chrisjbryant/errant
https://github.com/antonisa/lang2vec
https://universaldependencies.org/
https://arxiv.org/pdf/2005.03436.pdf
https://github.com/macleginn/exploring-clmd-divergences