



L1ML Final Presentation

By: Yuzhe Chen, Jessica Huynh,
and Ryan Nicoll



The task: Native Language Identification using L2 written samples

- Goal: Identify L1 of writers from a known universe of languages by salient (language-specific?) morphosyntactic errors made in the text through annotation of said features
- Improve on baseline results from NB classifier
- Extrapolate salient features from annotation to train an improved NB classifier



The corpus: TOEFL11 Corpus of L2 Academic Written English

- TOEFL11 Corpus -- Answers to free response, open-ended questions (8 prompts):
 - *Do you agree or disagree with the following statement? Young people enjoy life more than older people do. Use specific reasons and examples to support your answer*
- 7 L1 languages: Arabic, Chinese, French, Hindi, Japanese, Spanish, Telugu
- Essays average about 348 words in length and scored between 2.0 - 3.5 on a 5.0 scale (i.e., an intermediate English proficiency)
- Data was randomly chosen to control for prompt and for score
- Approximately 16 documents per language were annotated by two annotators (10 for Telugu)
-



Tag set

Extents

- Noun
- Determiner
- Adj
- Preposition
- MissingCopula
- Verb
- AwkwardPunctuation
- Misspelling
- UsageError

Links

- DetNounLink
- SVDisagreement
- PrepositionLink

Annotation Specification

- Four major domains: *Nouns, adjectives, verbs, miscellaneous*
- Nouns:
 - Determiner agreement (extra, mismatch, missing)
 - Preposition agreement (extra, mismatch, missing)
 - Error in plurality (Needed to be singular or plural)
 - Gender
- Predicative adjectives:
 - Preposition agreement (extra, mismatch, missing)
- Verbs:
 - Subject/verb agreement
 - Tense/form errors
 - Subject presence/absence
 - Preposition agreement (extra, mismatch, missing)
- Miscellaneous
 - Misspellings: Vowel issues, homophones, segment voicing, cognate, missing space
 - Awkward punctuation: Punctuation inappropriate for English discourse
 - UsageError: Catch all, NOS error



Problems during data set collection

- Difficulty of the task
 - High cognitive load in relatively long documents with many errors
 - High cognitive load for specification – requires familiarity with conventions and understanding of linguistic phenomena
 - Multi-step processes for certain tags (I.e. tag Subject as "Noun", Verb as "Verb", and SVD disagreement with link for both)
 - One extent can have multiple toggled attributes and be needed as an argument for several links (I.e., domains within our four global domains)
- Time constraint and other annotator issues
 - Difficulties in keeping up with the document workload
 - Reduction of annotator group from 3 to 2



Solutions (current and future)

- Solutions during project:
 - Revamp the DTD and the specification to be more user-friendly and to include more linguistic background information
 - E-mail (and, unofficially, Facebook) for questions
- Future solutions:
 - Training sessions and work sessions
 - Standardized forum for Q & A
 - On-the-fly IAA calculation and verification
 - Emphasize attention to detail and time-consuming nature of task
 - Pare down task to fewer tag domains (or smaller, more workable documents)



DTD

- Different iterations of the DTD

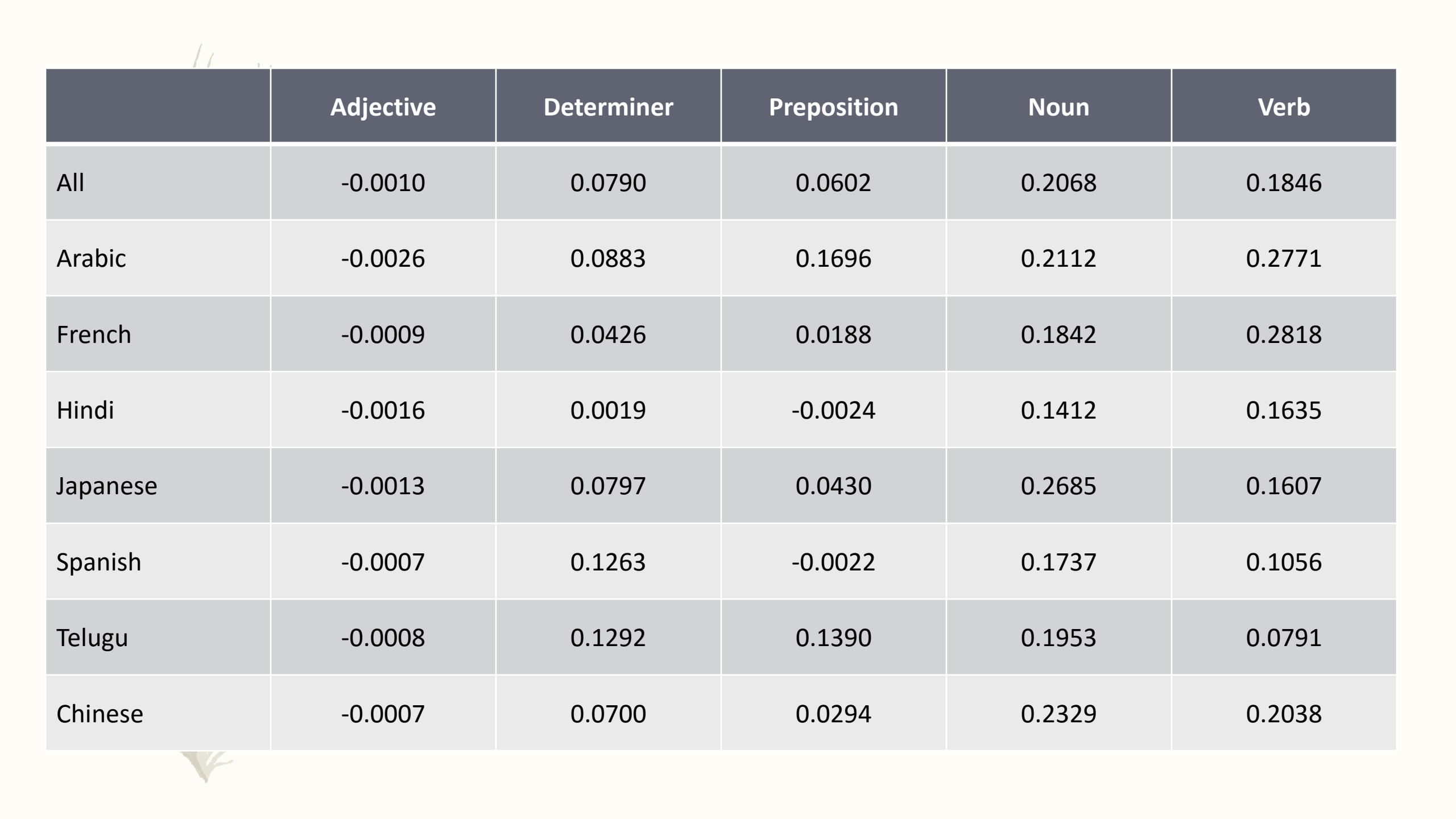
1. Much more complex with 11 extent tags and 4 link tags; package released alongside was just documents in Arabic
2. Pared down to 9 extent tags and 3 much simpler link tags; package consisted of some Arabic/French/Hindi
3. Tweaked (rearranged order of tags, renamed some attributes for clarity); packaged consisted of Spanish/Telugu/Chinese/Japanese



IAA

- Used Krippendorff's alpha, built into MAE
- Data set smaller in size and more sparse than expected
- Looked at both cross-tag and tag-level segmentation across entire gold standard and within languages
- Agreement rather low among extent tags
- As such, didn't measure IAA for link tags

	Overall	Misspelling	Missing Copula	Awkward Punctuation	Usage Error
All	0.4458	0.8239	0.2902	0.0240	0.0027
Arabic	0.3297	0.8508	-0.0001	-0.0006	-0.0087
French	0.4282	0.7147	-0.0001	-0.0003	-0.0042
Hindi	0.4062	0.7610	0.3075	-0.0007	-0.0039
Japanese	0.5044	0.8619	-0.0001	0.1247	-0.0054
Spanish	0.6560	0.9170	0.0000	-0.0002	-0.0075
Telugu	0.3665	0.6209	0.0000	-0.0006	0.3918
Chinese	0.5176	0.9105	0.5832	0.2499	-0.0053



	Adjective	Determiner	Preposition	Noun	Verb
All	-0.0010	0.0790	0.0602	0.2068	0.1846
Arabic	-0.0026	0.0883	0.1696	0.2112	0.2771
French	-0.0009	0.0426	0.0188	0.1842	0.2818
Hindi	-0.0016	0.0019	-0.0024	0.1412	0.1635
Japanese	-0.0013	0.0797	0.0430	0.2685	0.1607
Spanish	-0.0007	0.1263	-0.0022	0.1737	0.1056
Telugu	-0.0008	0.1292	0.1390	0.1953	0.0791
Chinese	-0.0007	0.0700	0.0294	0.2329	0.2038

-0.0010 (<Tag-level> Alpha-U (Krippendorfs)) Adj::-

-0.0006 (<Tag-level> Alpha-U (Krippendorfs)) Adj::PrepError

0.0240 (<Tag-level> Alpha-U (Krippendorfs)) AwkwardPunctuation::-

0.0790 (<Tag-level> Alpha-U (Krippendorfs)) Determiner::-

0.0142 (<Tag-level> Alpha-U (Krippendorfs)) Determiner::CorrectForm

-0.0003 (<Tag-level> Alpha-U (Krippendorfs)) Determiner::OtherCorrectForm

0.2902 (<Tag-level> Alpha-U (Krippendorfs)) MissingCopula::-

0.1754 (<Tag-level> Alpha-U (Krippendorfs)) MissingCopula::CorrectForm

0.8239 (<Tag-level> Alpha-U (Krippendorfs)) Misspelling::-

0.8123 (<Tag-level> Alpha-U (Krippendorfs)) Misspelling::Error

0.2068 (<Tag-level> Alpha-U (Krippendorfs)) Noun::-

0.1225 (<Tag-level> Alpha-U (Krippendorfs)) Noun::DetError

0.3465 (<Tag-level> Alpha-U (Krippendorfs)) Noun::GenderError

0.1870 (<Tag-level> Alpha-U (Krippendorfs)) Noun::PIError

0.0551 (<Tag-level> Alpha-U (Krippendorfs)) Noun::PrepError

0.0602 (<Tag-level> Alpha-U (Krippendorfs)) Preposition::-

0.0039 (<Tag-level> Alpha-U (Krippendorfs)) Preposition::CorrectPreposition

0.0027 (<Tag-level> Alpha-U (Krippendorfs)) UsageError::-

0.0260 (<Tag-level> Alpha-U (Krippendorfs)) UsageError::ErrorType

0.1846 (<Tag-level> Alpha-U (Krippendorfs)) Verb::-

0.2125 (<Tag-level> Alpha-U (Krippendorfs)) Verb::MissingSubj

0.0009 (<Tag-level> Alpha-U (Krippendorfs)) Verb::PrepError

0.1757 (<Tag-level> Alpha-U (Krippendorfs)) Verb::TenseOrFormError

Agreement on attributes





Looking at expected agreement

- Tricky to determine because the annotators annotated text spans, not documents, and the annotations themselves were sparse
 - Expected that the annotation would be sparse
 - IAA for Misspellings expected to be higher than the others because of preprocessing
 - Expected that Arabic would be the lowest of all the languages



“Those are *really* low numbers” and other observations

- Many instances of partial overlap, particularly with Verb
- Negative and near zero agreement!
 - Tags just not used (Adjective, AwkwardPunctuation, MissingCopula, UsageError)
 - Multiple ways to ‘correct’ something, which affected Preposition and Determiner
- No agreement numbers to quantify this, but inconsistent use of link tags
- Errors were usually caught, but only by one of the annotators (pretty decent gold standard, terrible IAA)



ML: gold standard dataset

- Two or three languages adjudicated per team member
- Data:
 - 16 annotated documents in Arabic, French, Hindi, Japanese, Spanish and Chinese, 10 documents in Telugu
 - Total = $16 \times 6 + 10 = 106$ documents
 - Train : Test = 8 : 2



```
<Noun id="N10" spans="520~529" text="baby bear" GenderError="No" DetError="MissingDeterminer"/>
<Noun id="N11" spans="570~580" text="experience" GenderError="No" PlError="NeedsToBePlural"/>
<Noun id="N12" spans="584~587" text="zoo" GenderError="No" DetError="MissingDeterminer"/>
<Noun id="N13" spans="760~763" text="zoo" GenderError="No" PlError="NeedsToBePlural"/>
<Noun id="N14" spans="805~809" text="mind" GenderError="No" DetError="MissingDeterminer"/>
<Noun id="N15" spans="909~915" text="things" GenderError="No" PlError="NeedsToBeSingular"/>
<Noun id="N16" spans="1086~1092" text="future" GenderError="No" DetError="MissingDeterminer"/>
<Noun id="N17" spans="1193~1201" text="capacity" GenderError="No" DetError="MissingDeterminer"/>
<Noun id="N18" spans="1206~1210" text="mind" GenderError="No" DetError="MissingDeterminer"/>
<Noun id="N19" spans="1340~1344" text="mind" GenderError="No" DetError="MissingDeterminer"/>
<Determiner id="D0" spans="319~322" text="his" CorrectForm="POSSESSIVE" OtherCorrectForm="their"/>
<Verb id="V0" spans="160~164" text="have" MissingSubj="no" TenseOrFormError="Form"/>
<Verb id="V1" spans="402~405" text="are" MissingSubj="no" TenseOrFormError="Tense"/>
<Verb id="V2" spans="429~433" text="have" MissingSubj="no" TenseOrFormError="Tense"/>
<Verb id="V3" spans="716~723" text="will be" MissingSubj="no" TenseOrFormError="Tense"/>
<Verb id="V4" spans="794~798" text="lost" MissingSubj="no" TenseOrFormError="Tense"/>
<Verb id="V5" spans="925~927,937~942" text="is ... stole" MissingSubj="no" TenseOrFormError="Form"/>
<Verb id="V6" spans="1157~1161" text="lost" MissingSubj="no" TenseOrFormError="Tense"/>
<Verb id="V7" spans="1188~1192" text="lost" MissingSubj="no" TenseOrFormError="Tense"/>
<Verb id="V8" spans="1217~1221" text="feel" MissingSubj="no" TenseOrFormError="Form"/>
<Misspelling id="M0" spans="17~27" text="statemaent" Error="OTHER"/>
<Misspelling id="M1" spans="185~193" text="somthing" Error="VowelsIssues"/>
<Misspelling id="M2" spans="463~471" text="suprized" Error="Cognate"/>
<Misspelling id="M5" spans="682~692" text="child food" Error="OTHER"/>
<Misspelling id="M6" spans="1112~1120" text="imprtant" Error="VowelsIssues"/>
<Misspelling id="M7" spans="1177~1184" text="bacause" Error="VowelsIssues"/>
<Misspelling id="M8" spans="1222~1230" text="somthing" Error="VowelsIssues"/>
<Misspelling id="M9" spans="1316~1321" text="plder" Error="OTHER"/>
<UsageError id="U0" spans="180~184" text="feel" ErrorType="Wrong POS"/>
<UsageError id="U1" spans="665~673" text="exciting" ErrorType="Wrong POS"/>
<DetNounLink id="DNL0" articleID="D0" articleText="his" nounID="N14" nounText="life"/>
```

ML: baseline

- Naïve Bayes with bag-of-words
- Trained on both the entirety of the TOEFL11 corpus and just the subset of documents in our gold standard

ZH0: 73 files in test set

FRA: 0/ 73 0.00000%

HIN: 5/ 73 6.84932%

ARA: 0/ 73 0.00000%

ZH0: 51/ 73 69.86301%

SPA: 1/ 73 1.36986%

TEL: 1/ 73 1.36986%

JPN: 15/ 73 20.54795%

ARA: 51 files in test set

FRA: 2/ 51 3.92157%

HIN: 12/ 51 23.52941%

ARA: 29/ 51 56.86275%

ZH0: 1/ 51 1.96078%

SPA: 2/ 51 3.92157%

TEL: 1/ 51 1.96078%

JPN: 4/ 51 7.84314%

SPA: 47 files in test set

The confusion matrix of the test results:

	a	f	h	j	s	t	z
	r	r	i	p	p	e	h
	a	a	n	n	a	l	o
ara	<.>1
fra	.	<.>.	.	.	1	.	.
hin	.	.	<.>.
jpn	1	.	.	<.>.	.	.	.
spa	.	.	1	.	<1>.	.	.
tel	.	1	1	1	.	<1>2	.
zho	1	.	.	1	.	.	<.>

(row = reference; col = test)

Accuracy: 0.154

ML: with features added from annotation

- Experimented with multiple feature extraction methods, but not much improvement.

The confusion matrix of the test results:

	a	f	h	j	s	t	z
r	r	r	i	p	p	e	h
a	a	a	n	n	a	l	o

	<2>	2	1	2	1	1	2
ara	<2>	2	1	2	1	1	2
fra	.	<.>
hin	.	.	<.>
jpn	.	.	1	<.>	1	.	.
spa	<.>	.	.
tel	<.>	.
zho	<.>

(row = reference; col = test)

Accuracy: 0.154

The confusion matrix of the test results:

	a	f	h	j	s	t	z
r	r	r	i	p	p	e	h
a	a	a	n	n	a	l	o

	<.>	1	.	.	1	.	.
ara	<.>	1	.	.	1	.	.
fra	.	<.>	.	.	1	.	.
hin	.	.	<.>
jpn	1	.	.	<1>	.	.	.
spa	.	.	1	.	<.>	.	.
tel	.	1	1	1	.	<1>	2
zho	1	<.>

(row = reference; col = test)

Accuracy: 0.154

ML: with features added from annotation

- So far, found two feature extraction methods that produce better results

The confusion matrix of the test results:

	a	f	h	j	s	t	z
	r	r	i	p	p	e	h
	a	a	n	n	a	l	o
ara	<2>	1	1
fra	.	<.>	.	.	1	.	1
hin	.	.	<.>
jpn	.	1	2	<2>	.	.	.
spa	<.>	.	.
tel	1	<.>	.
zho	.	1	<.>

(row = reference; col = test)

Accuracy: 0.308

tag_attr_name_counts

The confusion matrix of the test results:

	a	f	h	j	s	t	z
	r	r	i	p	p	e	h
	a	a	n	n	a	l	o
ara	<1>
fra	.	<.>
hin	.	.	<1>
jpn	.	1	.	<.>	.	.	.
spa	<1>	.	.
tel	.	.	1	.	1	<.>	1
zho	1	1	.	2	.	1	<1>

(row = reference; col = test)

Accuracy: 0.308

noun_verb_misspelling

MAE the Force be with
you!

Questions?