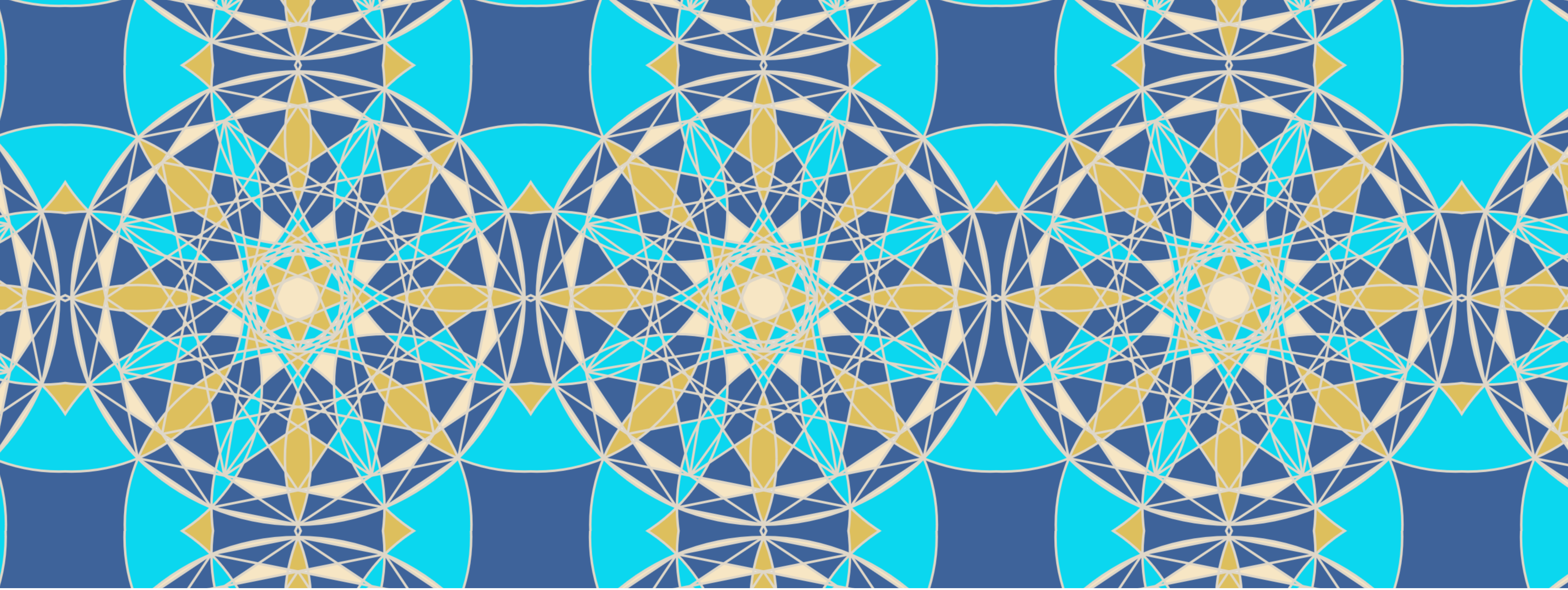


ARABIC LEARNER CORPUS CONSIDERATIONS

Anthony Verardi
University of Pittsburgh
LING 2340 Spring 2020
Data Science for Linguists
Dr. Jevon Heath

Full project available at https://github.com/Data-Science-for-Linguists-2020/Arabic-Learner-Corpus-Considerations

OVERVIEW

I. Introduction

- Motivation
  - Why Arabic?
  - Why a learner corpus?
- Literature Review
  - Arabic as an L2

II. The Arabic Learner Corpus

- Availability and Licensing
- Composition

III. Areas of Investigation

- Learner subgroups
- Support Vector L1 Classifier

IV. Data Organization

- XML and BeautifulSoup
- Building a DataFrame
- Cleaning, Exploratory Data Analysis (EDA)
- Collapsing L1 Data

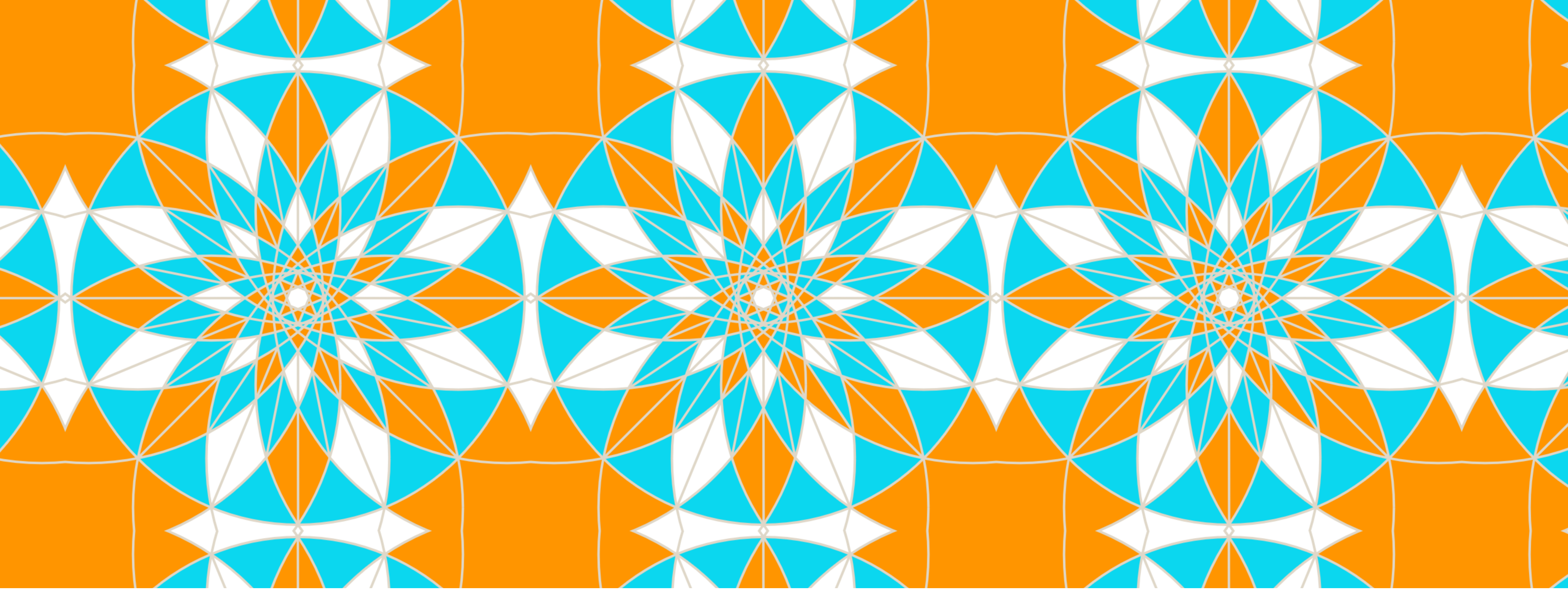
V. Preliminary Analysis

- Subgroup Visualizations (WIP)
- SVC L1 Classifier

VI. Conclusion

- Limitations
- Preliminary Findings
- Future Directions

VII. Works Referenced



INTRODUCTION

Section I

MOTIVATION: WHY ARABIC?

- Modern Standard Arabic (MSA) and dialectal varieties of Arabic remain understudied in both Computational Linguistics and Second Language Acquisition (SLA)
- Personal background as both an L2 Arabic learner and instructor
- While Arabic-language corpora exist, their quality is often dubious or they are confined to a highly specific domain (Zaghouani, 2014)

MOTIVATION: WHY A LEARNER CORPUS?

- Learner corpora are even more rare in the grand scheme of freely-available Arabic corpora
- The curators of the present corpus make a number of claims about its potential for use in SLA research and pedagogical applications
- Seems like good science in general to explore and evaluate a resource before going ahead with using it to inform classroom or curricular interventions

LITERATURE REVIEW: ARABIC AS AN L2

- Dearth of reputable, peer-reviewed research on Arabic as a second language
- Raish (2015) examined Arabic variation from a traditional, phonological standpoint (acquisition of sociophonetic variation during study abroad)
- Alhawary (2009) did look at Arabic learner morphosyntactic acquisition in MSA
- In general (see Alhawary (Ed.), 2018), works on Arabic as an L2 disproportionately investigate MSA as opposed to spoken dialect

THE ARABIC LEARNER CORPUS

Section II

AVAILABILITY AND LICENSING

Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

- Downloadable in both XML and .txt file formats
  - English or Arabic metadata
- Audio data also available as well as handwritten sheets
- POS-tagged .txt and XML files also available
- Freely available under a Creative Commons License Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

COMPOSITION

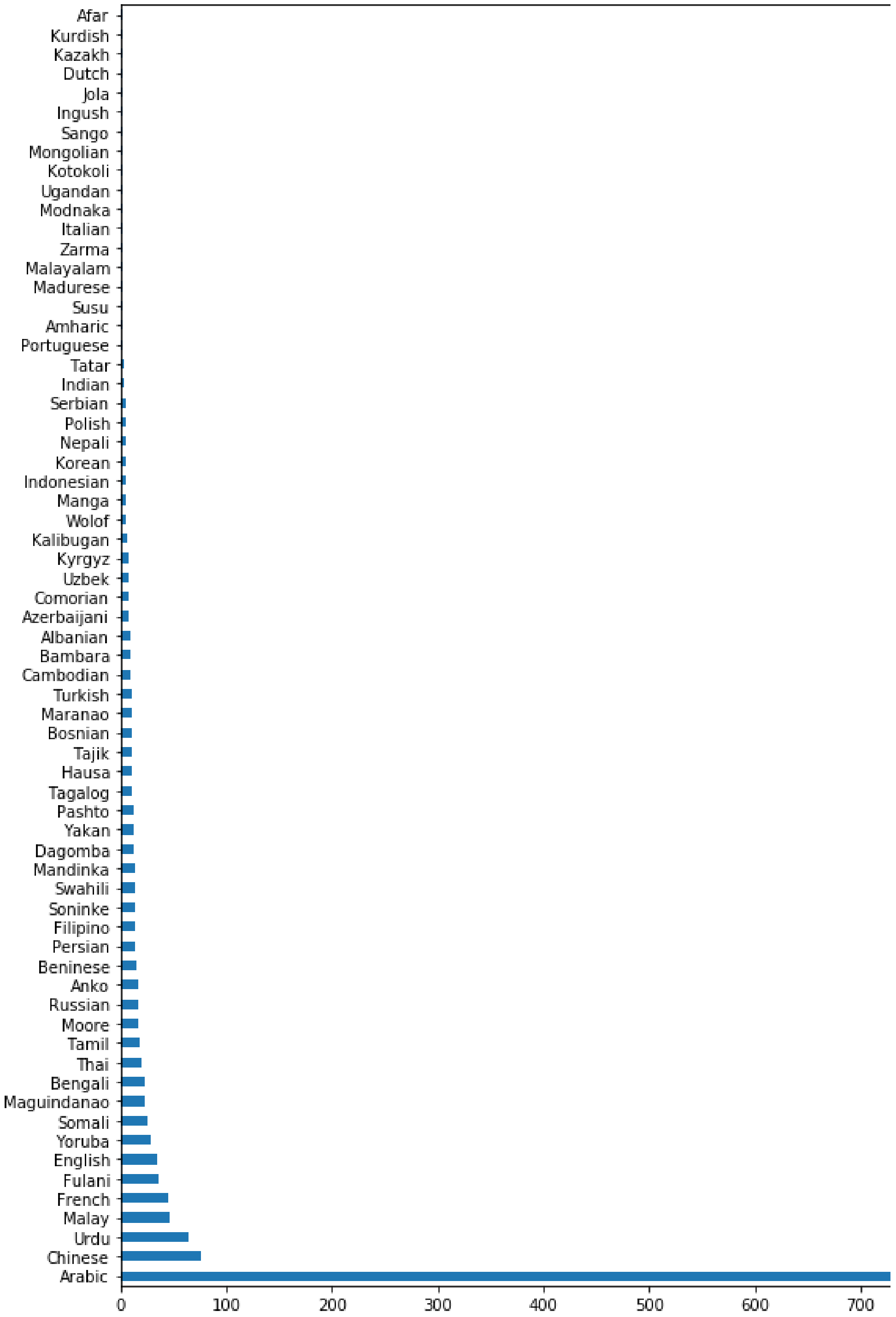
- 1,585 XML files and a README file about the corpus
- “The ALC data has been captured in 2012 and 2013. It includes **282,732** words, **1585** materials (**written** and **spoken**), produced by **942** students from **67** nationalities, and **66** different L1 backgrounds. Average length of a text is **178** words” (Alfaifi, Atwell, & Hedaya, 2014)
- Two portions to each file:
  - Metadata about the participant
  - Text of the response to a prompt (either narrating a vacation trip or describing their studies) and metadata about the response (time on task, dictionary use, etc.)

AREAS OF INVESTIGATION

Section III

LEARNER SUBGROUPS

- Certain assumptions need to be met to determine what kinds of inferential statistical tests can be safely used
- Initial run: response counts by L1
- Balanced? I think not. How can we improve this?



SUPPORT VECTOR L1 CLASSIFIER

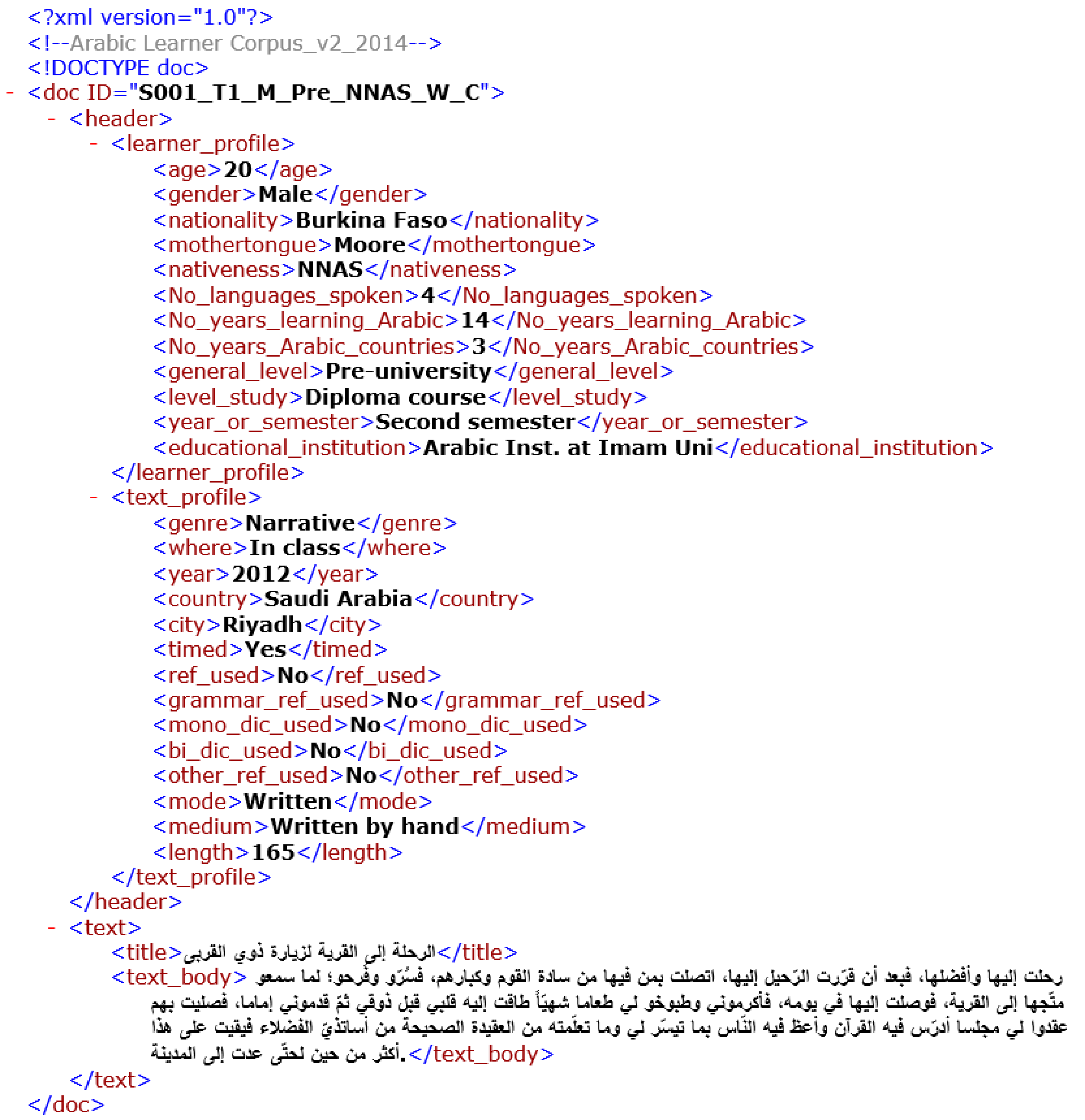
- Can these data be used to train a classifier to identify L1?
  - Is there enough data for Contrastive Interlanguage Analysis (CIA)?
  - Are the responses unique enough to tell apart, given that their average length is relatively short even for Arabic (~178 words)
  - What effect do the imbalances in group size have on classification training? We've mostly worked in class with data that is relatively balanced

DATA ORGANIZATION

Section IV

XML AND BEAUTIFULSOUP

- ❖XML markup structured as shown on the right
- ❖Docs marked with a unique ID, then separated into **learner profile**, **text profile**, and **text**
- ❖Structure cuts down the amount of work needed to grab relevant data
- ❖BeautifulSoup can be used to import XML and access tagged content



```
<?xml version="1.0"?>
<!--Arabic Learner Corpus_v2_2014-->
<!DOCTYPE doc>
- <doc ID="S001_T1_M_Pre_NNAS_W_C">
  - <header>
    - <learner_profile>
        <age>20</age>
        <gender>Male</gender>
        <nationality>Burkina Faso</nationality>
        <mothertongue>Moore</mothertongue>
        <nativeness>NNAS</nativeness>
        <No_languages_spoken>4</No_languages_spoken>
        <No_years_learning_Arabic>14</No_years_learning_Arabic>
        <No_years_Arabic_countries>3</No_years_Arabic_countries>
        <general_level>Pre-university</general_level>
        <level_study>Diploma course</level_study>
        <year_or_semester>Second semester</year_or_semester>
        <educational_institution>Arabic Inst. at Imam Uni</educational_institution>
      </learner_profile>
    - <text_profile>
        <genre>Narrative</genre>
        <where>In class</where>
        <year>2012</year>
        <country>Saudi Arabia</country>
        <city>Riyadh</city>
        <timed>Yes</timed>
        <ref_used>No</ref_used>
        <grammar_ref_used>No</grammar_ref_used>
        <mono_dic_used>No</mono_dic_used>
        <bi_dic_used>No</bi_dic_used>
        <other_ref_used>No</other_ref_used>
        <mode>Written</mode>
        <medium>Written by hand</medium>
        <length>165</length>
      </text_profile>
    </header>
  - <text>
      <title>الرحلة إلى القرية لزيارة ذوي القربى</title>
      <text_body> رحلت إليها وأفضلها، فبعد أن قرّرت الرّحيل إليها، اتصلت بمن فيها من سادة القوم وكبارهم، فسُرّو وفَرحو؛ لما سمعو
        متّجها إلى القرية، فوصلت إليها في يومه، فأكرموني وطبوخو لي طعاما شهياً طاقت إليه قلبي قبل ذوقي ثمّ قدموني إماما، فصليت بهم
        عقدوا لي مجلسا أدرّس فيه القرآن وأعظ فيه النّاس بما تيسّر لي وما تعلّمته من العقيدة الصحيحة من أساتذيَ الفضلاء فبقيت على هذا
        أكثر من حين لحتّى عدت إلى المدينة.</text_body>
    </text>
</doc>
```

BUILDING A DATAFRAME

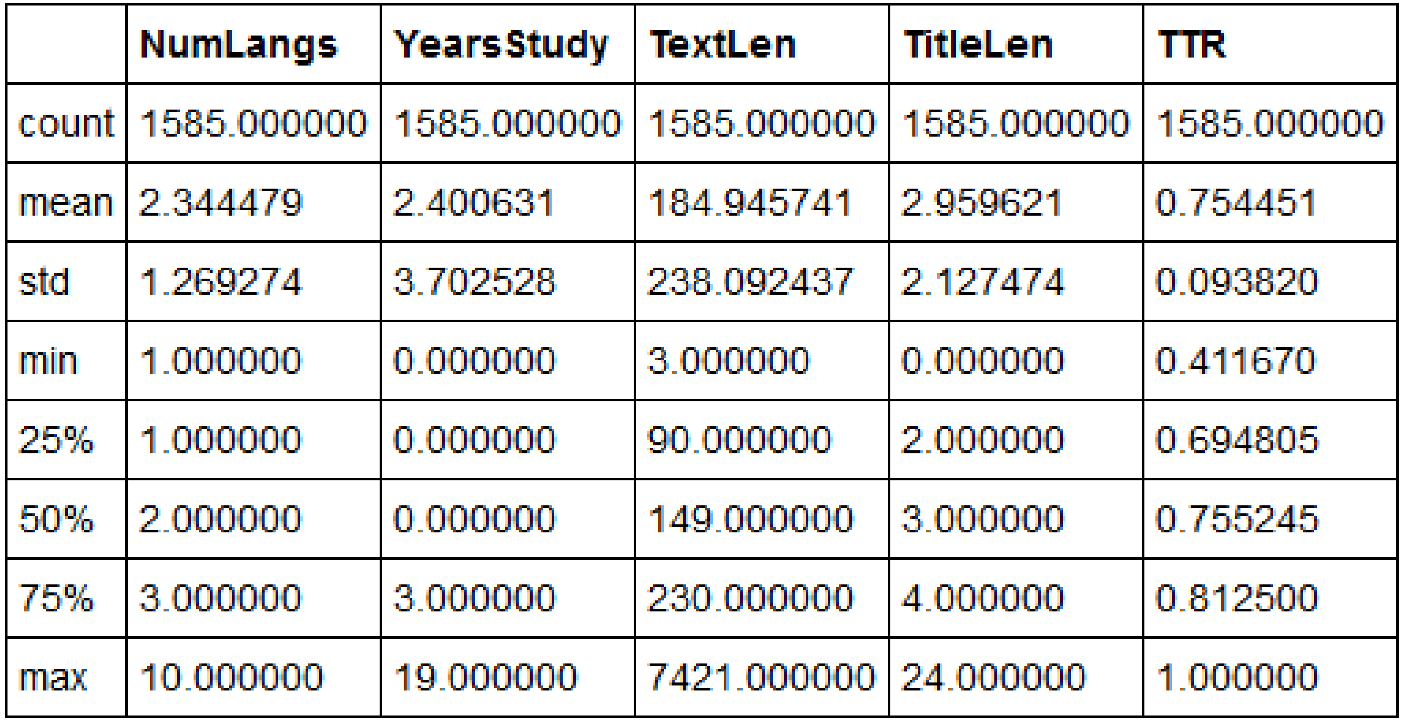
- Wrote a script that iterated through each file, pulling the following info from the XML and inserting it into a DataFrame:
  - DocID (the name of the original response document, to be used as an index value later)
  - L1 (renamed from "Mothertongue" in the original markup; comprised of 66 L1s total)
  - NumLangs (number of languages known by the participant, ranging from 1-10)
  - Nationality
  - Age
  - Gender
  - YearsStudy (years studying Modern Standard Arabic)
  - GenLvl (whether a participant's academic career was pre-university or university)
  - LvlStdy (Secondary school, language course, diploma course, BA, or MA)
  - Title (of the response)
  - Text (of the response)
  - Genre (narrative or discussion)
  - Mode (written or spoken)

BUILDING A DATAFRAME

- Thankfully, text direction was preserved while building the data frame!
- Arabic is written from right to left, so some concern here
- Specifying UTF-16 encoding worked just fine
- End result: 1585x12 DataFrame ready to rock and roll

CLEANING, EDA

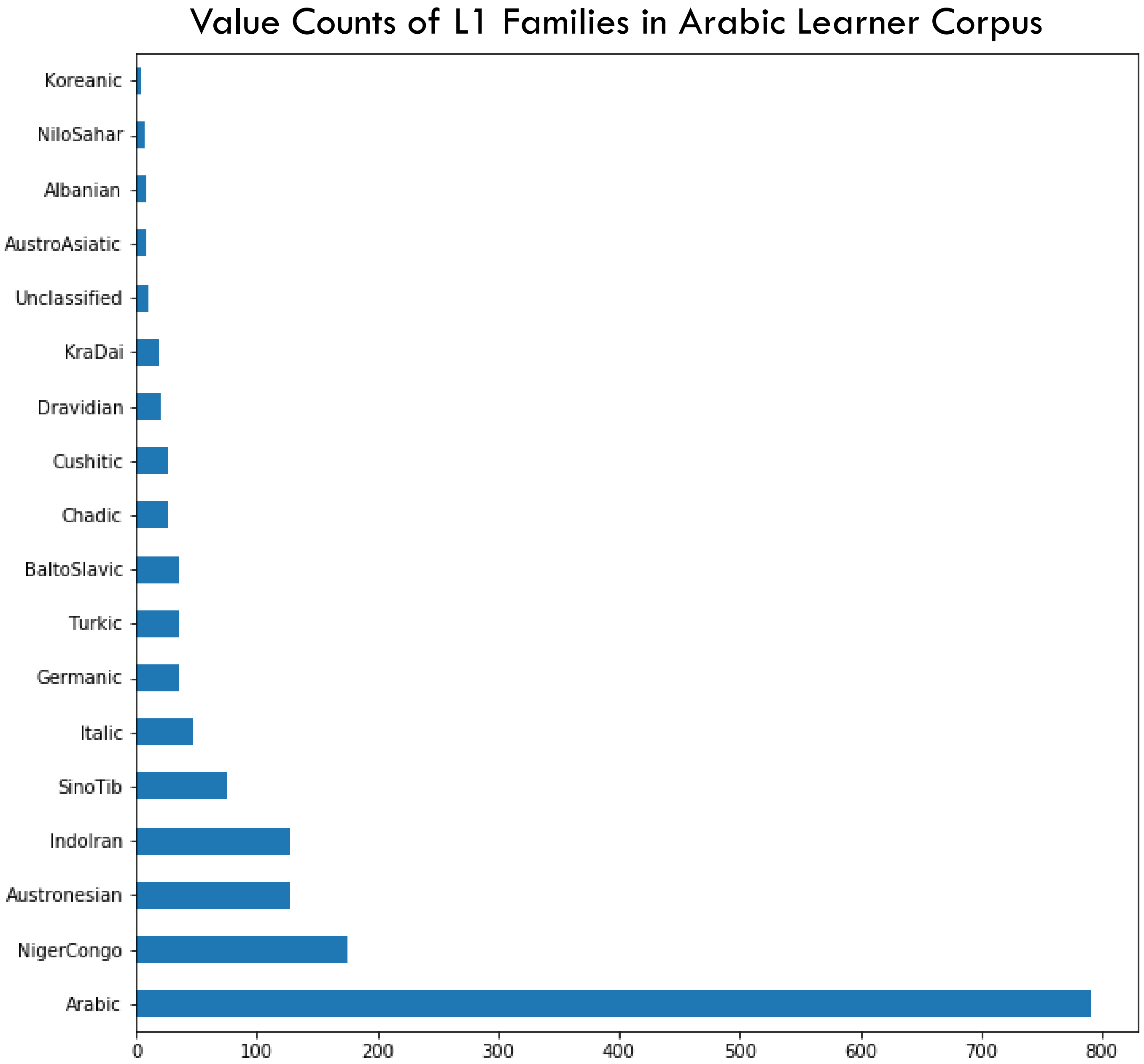
- Cleaning involved:
  - Filling in NaN values in essay titles
  - Renaming some problematic column titles
    - ex. MotherTongue -> L1
  - Adding in my own calculations for text/title length and TTR (chose not to use pre-counted ones)
  - Collapsing L1 data into language families (more on next slide)
- Post-cleaning, ran a .describe() command to get some descriptive statistics on numerical columns

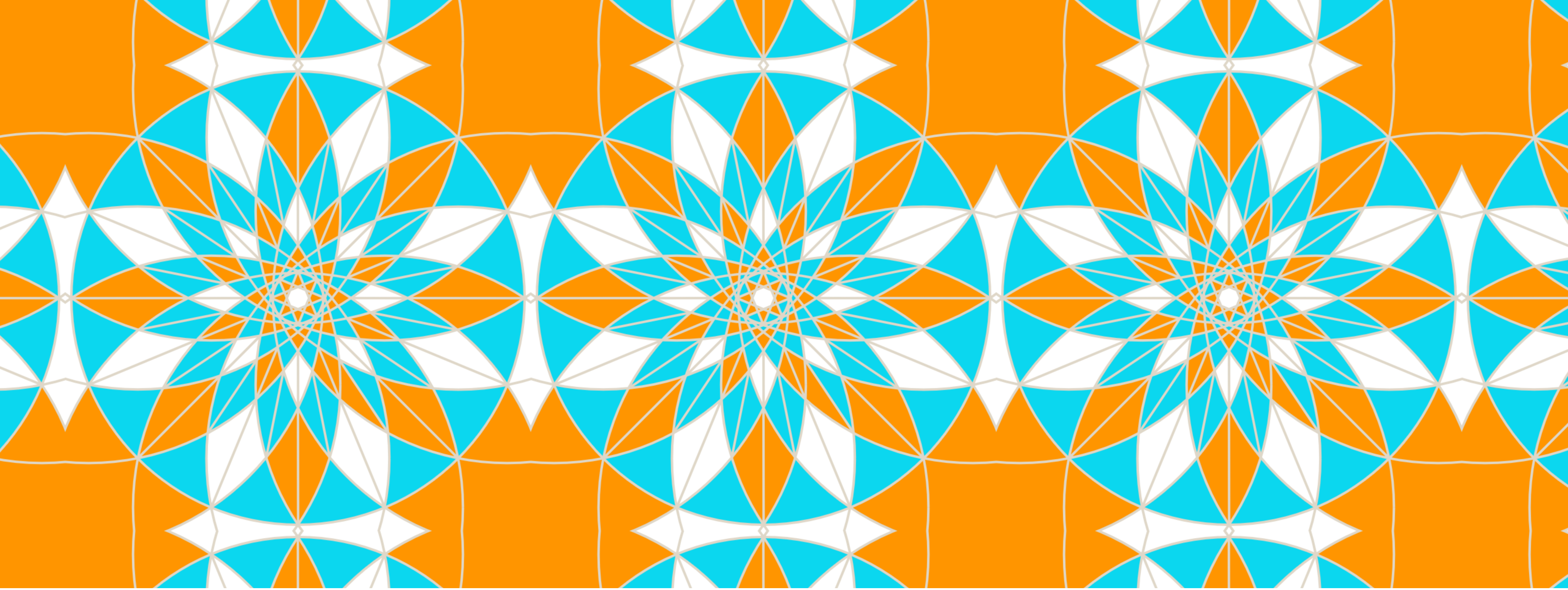


	NumLangs	YearsStudy	TextLen	TitleLen	TTR
count	1585.000000	1585.000000	1585.000000	1585.000000	1585.000000
mean	2.344479	2.400631	184.945741	2.959621	0.754451
std	1.269274	3.702528	238.092437	2.127474	0.093820
min	1.000000	0.000000	3.000000	0.000000	0.411670
25%	1.000000	0.000000	90.000000	2.000000	0.694805
50%	2.000000	0.000000	149.000000	3.000000	0.755245
75%	3.000000	3.000000	230.000000	4.000000	0.812500
max	10.000000	19.000000	7421.000000	24.000000	1.000000

COLLAPSING L1 DATA

- ❖Biggest organization task: collapsing single languages into macro-families
- ❖L1 data alone all over the place (see slide 11)
  - ❖Additionally, some concerns with certain language names (“Moore”, “Ugandan”, “Modnaka”)
  - ❖How to collapse Indo-European? Split into one sub-family down from there
- ❖Consulted Ethnologue for family info
- ❖Turned 66 L1s into 18 families, including one “Unclassified” family for anything that would still only have <5 observations as a category





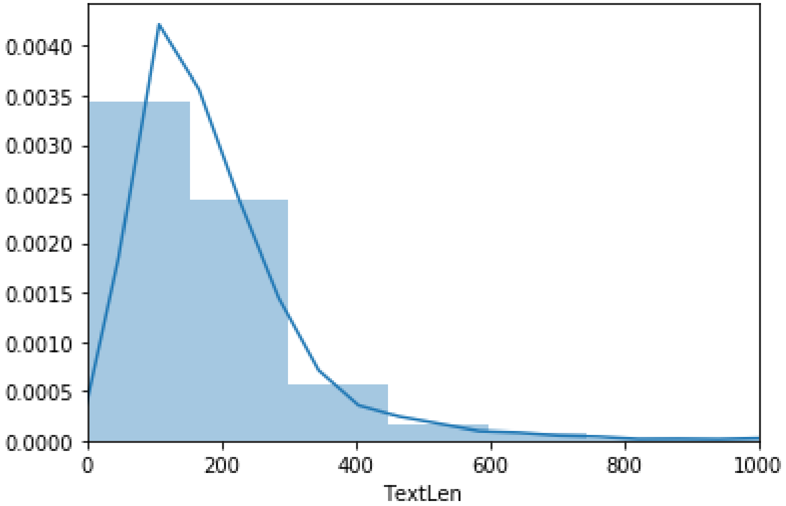
PRELIMINARY ANALYSIS

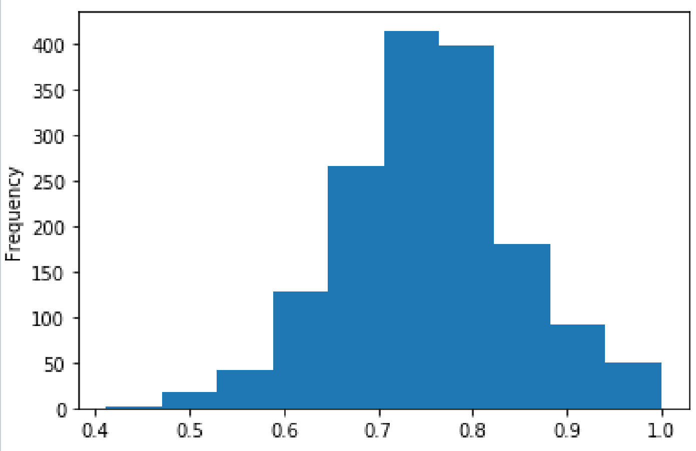
Section V

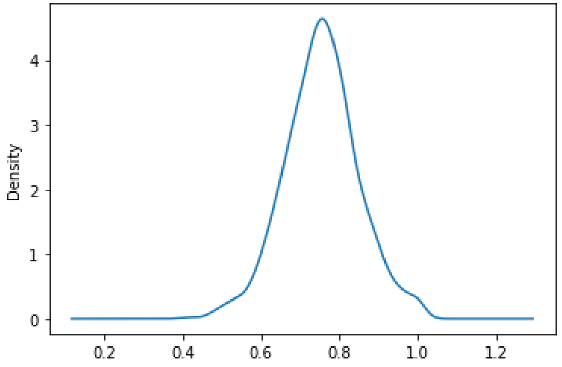
SUBGROUP VISUALIZATIONS (WIP)

- ❖End goal: text length and TTR distributions as boxplots by L1 family
- ❖For now: histogram and density plots of *overall* dataset for both
- ❖Text length skews right
- ❖TTR actually pretty normal
- ❖For inferential stats, non-parametric test for full data TextLength/maybe log transform, parametric probably fine for TTR

Clockwise from top left corner: Text length histogram+ density plot, TTR density plot, and TTR histogram





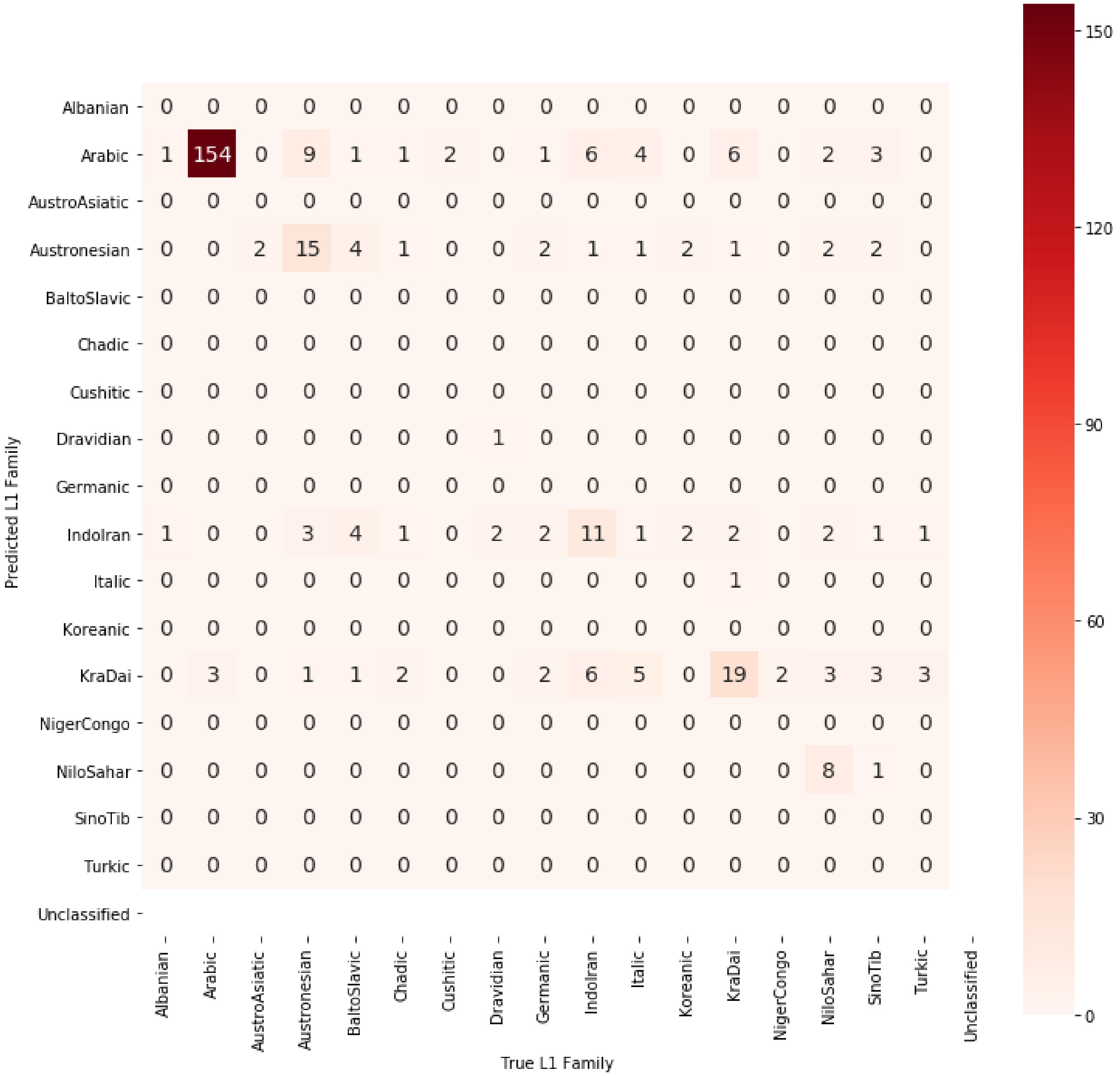


SVC L1 CLASSIFIER

- Using SVC model and TfIdf Vectorizer in a pipeline
- Fine-tuning parameters with Gridsearch CV: trying different max features, min doc frequency, and max doc frequency
- Best parameters:
  - {'tfIdf__max_df': 0.75, 'tfIdf__max_features': 5000, 'tfIdf__min_df': 2}
- Using a 20% random testing split after tuning another SVC model with aforementioned parameters
- Overall accuracy: 65.62%
- How to interpret?
  - Base probability: Arabic L1 “family” has a much higher base probability in the data than anything else (~49%); chance of randomly drawing an “Arabic” sample is about 50/50
  - Doesn’t seem to be doing too great when considering that in an evenly split dataset (ETS), our classifiers got up to around 70-75% with a base probability of ~10%
  - Looks like unevenness of data plays a big role—how to qualify?

SVC L1 CLASSIFIER

- ❖Taking a close look at output of classifier using a confusion matrix
- ❖Classifier correctly labels ALL true Arabic samples as Arabic, good job there
- ❖Other groups not doing so hot: sometimes fails to classify ANY correctly, other times gets maybe 65-75% right in a family



	Albanian	Arabic	AustroAsiatic	Austronesian	BaltoSlavic	Chadic	Cushitic	Dravidian	Germanic	IndoIran	Italic	Koreanic	KraDai	NigerCongo	NiloSahar	SinoTib	Turkic	Unclassified
Albanian	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Arabic	1	154	0	9	1	1	2	0	1	6	4	0	6	0	2	3	0	
AustroAsiatic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Austronesian	0	0	2	15	4	1	0	0	2	1	1	2	1	0	2	2	0	
BaltoSlavic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Chadic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Cushitic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Dravidian	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
Germanic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
IndoIran	1	0	0	3	4	1	0	2	2	11	1	2	2	0	2	1	1	
Italic	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
Koreanic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
KraDai	0	3	0	1	1	2	0	0	2	6	5	0	19	2	3	3	3	
NigerCongo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
NiloSahar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	1	0	
SinoTib	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Turkic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Unclassified																		

CONCLUSION

Section VI

LIMITATIONS

- Have not accounted for other types of grouping factors within my analysis or included other factors in my model besides actual text
  - There are not 1,585 participants, there are 942; not all participants did all tasks
  - This can also impact what inferential tests are appropriate to use
  - Worried that collapsing the data any further into L1 + only written examples would leave too little to work with for ML
  - Lots of additional data to work with, need to figure out feature union
- Some doubts about whether I'm operationalizing "usefulness" fairly in constructing this argument
  - If collapsing further into Arabic native speaker vs. Arabic learner is helpful, is that so bad? On the other hand, what are we losing in generalizability by collapsing so many different L1 families together?

PRELIMINARY FINDINGS

- Limitations considered, I think some serious doubts remain about what kinds of questions this dataset can be used to explore
- Contrastive Interlanguage Analysis (CIA) probably best done on relatively equal groups like the ETS corpus (would love a citation here; need to do some research)
- When considering what questions are fair to even explore with a dataset, knowing what you're actually dealing with and not just taking the set for granted are necessary
- That being said, highlights difficulties of working with under-researched languages, and I'm not blaming the authors for any shortcomings—it's amazing that they've even made this and made it public and I thank them for their work

FUTURE DIRECTIONS

- In short: finishing up!
- Breaking out learner groups and visualizing text length/TTR distributions by L1 family
- Trying another go at a classifier with only native Arabic/non-native Arabic as labels

شكرا كتير كتير!

WORKS REFERENCED

Alfaifi, A., Atwell, E. and Hedaya, I. (2014). Arabic Learner Corpus (ALC) v2: A New Written and Spoken Corpus of Arabic Learners. In the proceedings of the Learner Corpus Studies in Asia and the World (LCSAW) 2014, 31 May - 01 Jun 2014. Kobe, Japan. <>.

Alhawary, M. T. (2009). Arabic Second Language Acquisition of Morphosyntax. New Haven, United States: Yale University Press.

Alhawary, M. T. (Ed.) (2018). Routledge Handbook of Arabic Second Language Acquisition (1 ed.). London: Routledge.

Raish, M. (2015). The Acquisition of an Egyptian Phonological Variant by U.S. Students in Cairo. Foreign Language Annals, 48(2), 267-283. doi:10.1111/flan.12140

Zaghouani, Wajdi. (2014) Critical Survey of the Freely Available Arabic Corpora. Published in the Proceedings of the International Conference on Language Resources and Evaluation (LREC'2014), OSACT Workshop. Reykjavik, Iceland, 26-31 May 2014