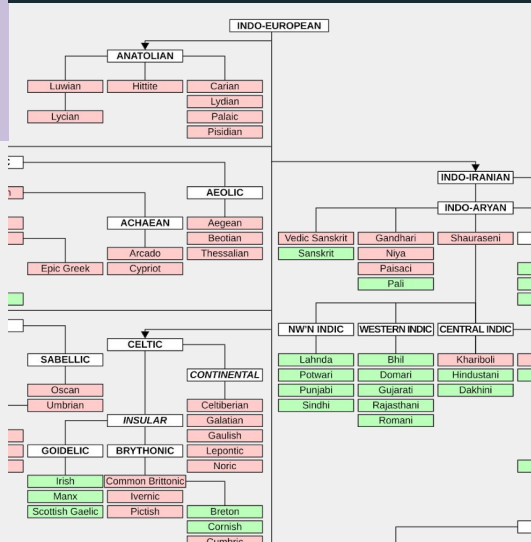
“The P.I.E. Program”

CSI 2300 Final Project

By Sadie Millsap

04.18.2025

- Introduction
- Project Goals
- What is PIE?
- Target Languages
- UML Diagram
- Past Timeline
- Visual Guides
- Unique Features
- Data Summary
- Future Timeline
- “QnA”



Introduction- About the Author

- Senior, Linguistics major w/ concentration in Comp Sci.
- Undergrad Consultant at the OU Writing Center.
- Current/future research in speech habits of online speech communities, focused on identifying linguistic features unique to internet English.

Project Goals:

I. To demonstrate the validity of the Proto-Indo-European (PIE) family tree of languages via user-controlled, octolingual comparisons of corpus data.

II. To further dispel myths and misconceptions of a prescriptive and/or xenophobic nature regarding languages and the cultures that speak them.

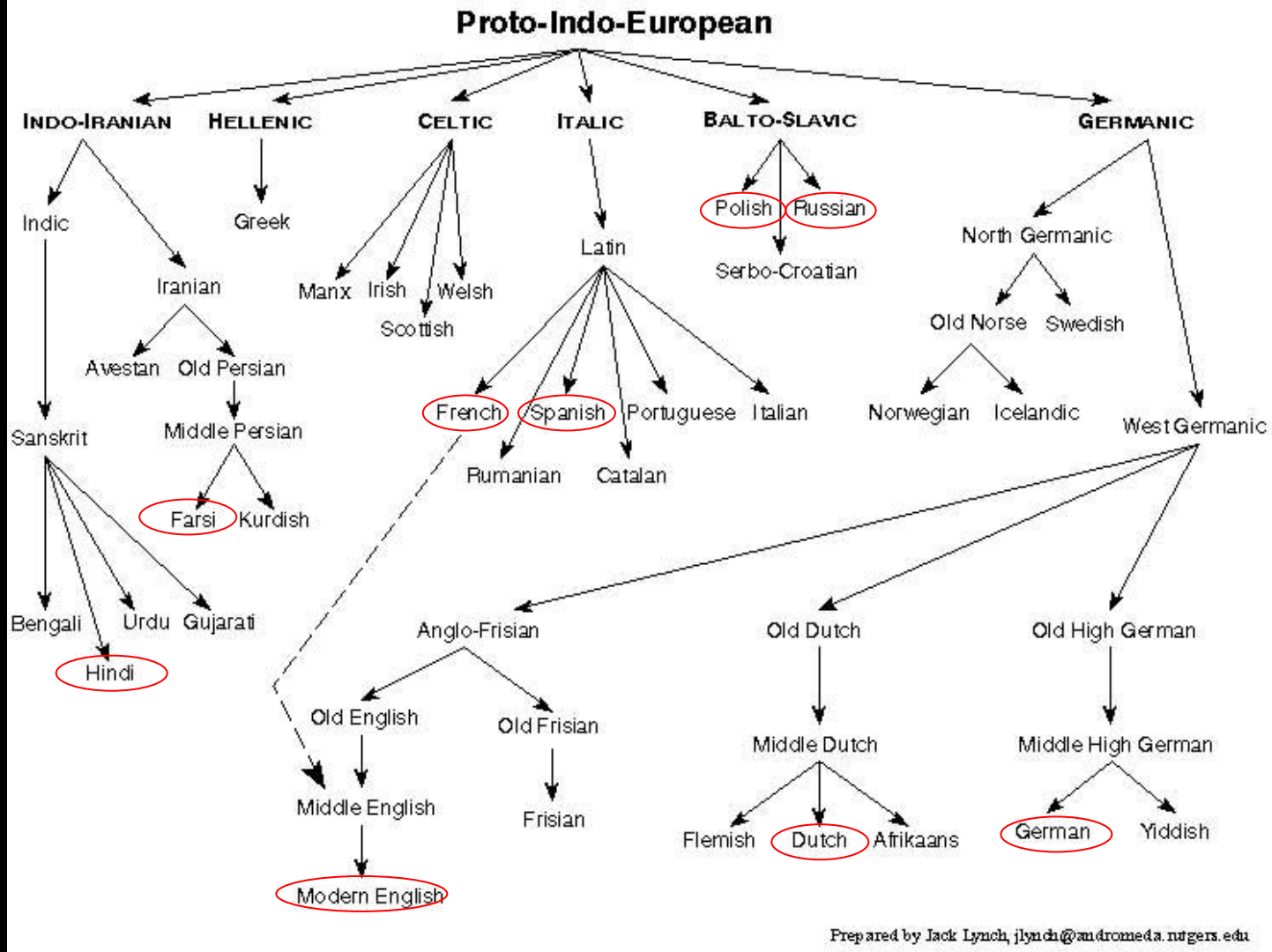
What is P.I.E. ?

Proto-Indo-European: an unconfirmed language/group of dialects that existed *before writing was invented* that many Linguists believe is the ancestor of most European languages, as well as most spoken languages in India and a small handful in the Middle East.

- Due to centuries of change, trade, and history, many of the languages in this family are related in terms of their word choice, sentence structure, and similarities in meaning, despite being VERY distinct on a sociocultural level.

- Knowledge of families like this (there are others!) allows people to learn languages easier because they're aware of their native language and its "siblings"

QUESTION (we'll come back to it):
How similar do you think Spanish is to English, on a scale of 0-100%?

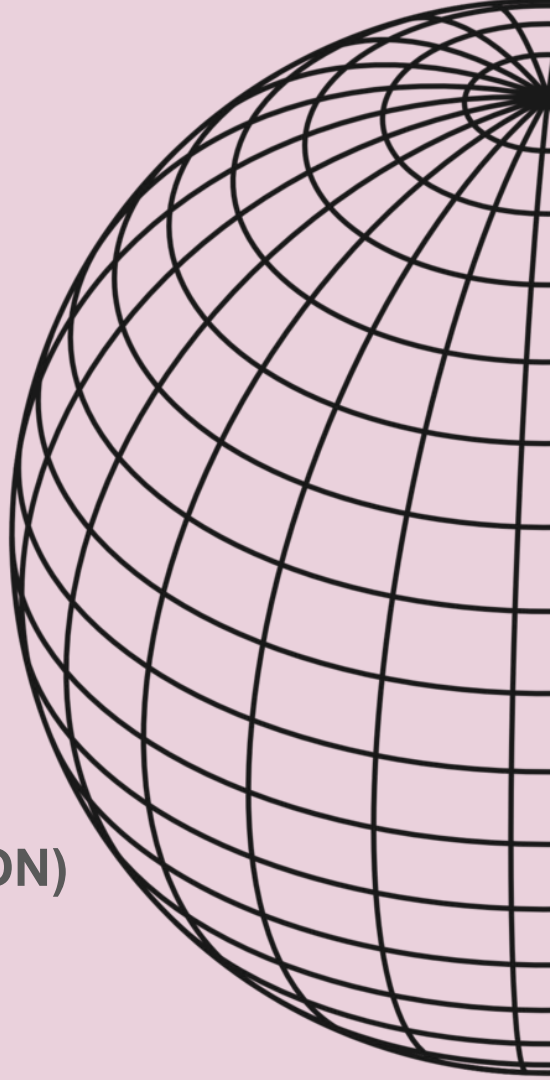

Proto-Indo-European
INDO-IRANIAN
HELLENIC
CELTIC
ITALIC
BALTO-SLAVIC
GERMANIC
Indic
Iranian
Greek
Manx
Irish
Welsh
Scottish
Latin
Polish
Russian
Serbo-Croatian
North Germanic
Old Norse
Swedish
Norwegian
Icelandic
Avestan
Old Persian
Middle Persian
Sanskrit
Farsi
Kurdish
French
Spanish
Portuguese
Italian
Rumanian
Catalan
West Germanic
Bengali
Urdu
Gujarati
Hindi
Anglo-Frisian
Old Dutch
Old High German
Old English
Old Frisian
Middle Dutch
Middle High German
Middle English
Frisian
Flemish
Dutch
Afrikaans
German
Yiddish
Modern English
Prepared by Jack Lynch, jlynch@andromeda.rutgers.edu

What Languages does the Project Use?

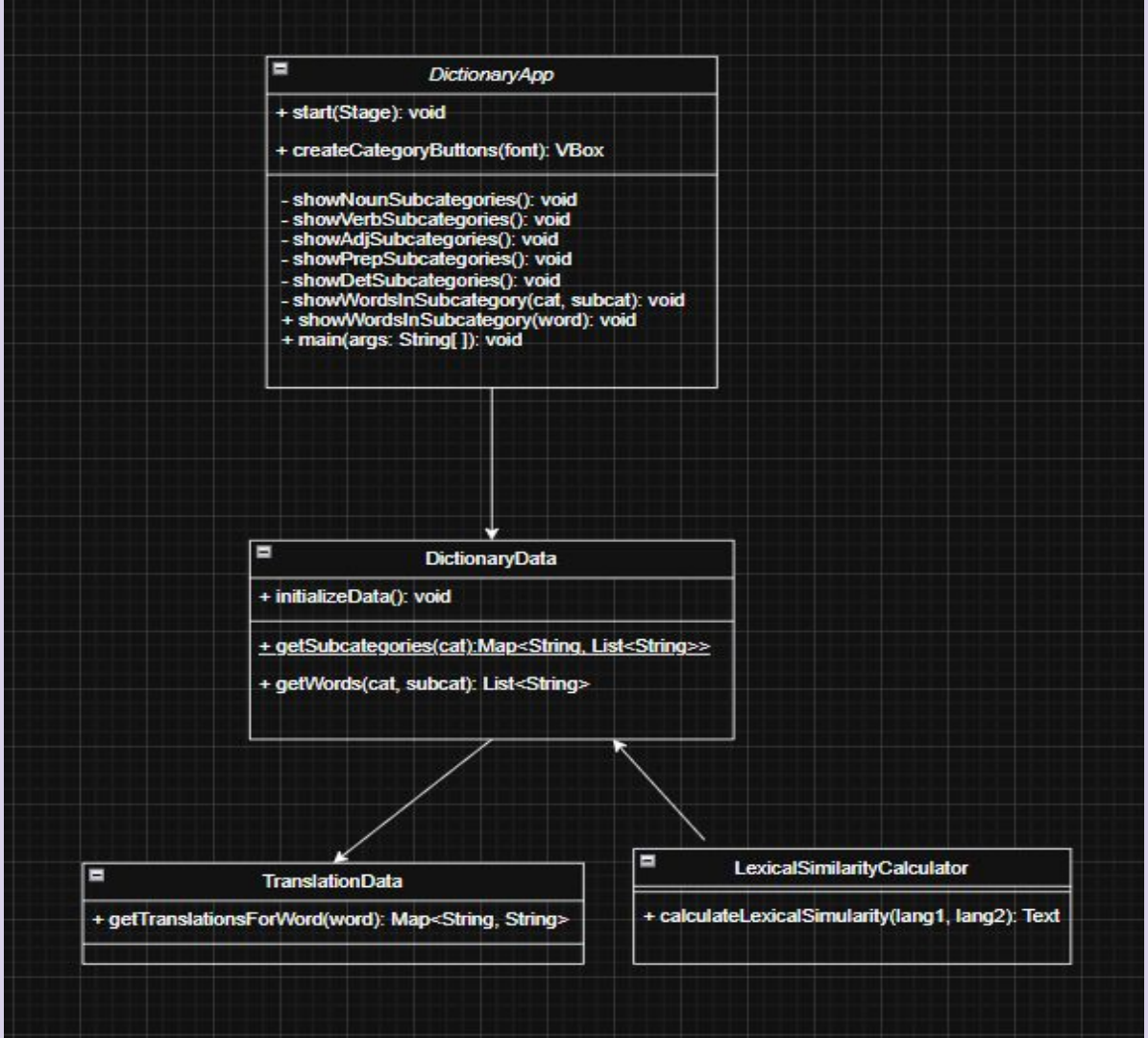
a. English- Spoken by ~1,350,000,000 people worldwide
b. Hindi- Spoken by ~609,000,000 people worldwide
c. Spanish- Spoken by ~600,607,800 people worldwide
d. French- Spoken by ~320,000,000 people worldwide
e. Russian- Spoken by ~258,000,000 people worldwide
f. German- Spoken by ~133,000,000 people worldwide
g. Farsi- Spoken by ~110,000,000 people worldwide
h. Dutch- Spoken by ~28,000,000 people worldwide
i. Polish- Spoken by ~40,000,000 people worldwide

TOTAL SPEAKER POPULATION:

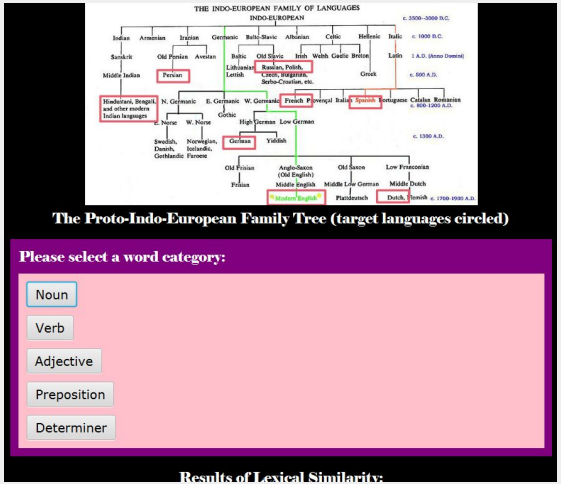
~3,400,000,000 people (42% OF WORLD POPULATION)



UML Diagram

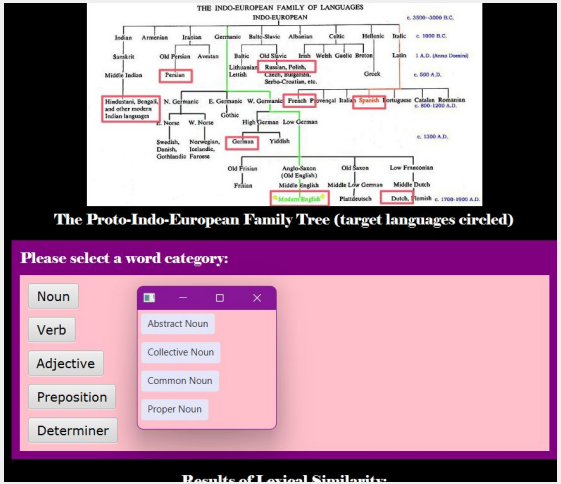

DictionaryApp
+ start(Stage): void
+ createCategoryButtons(font): VBox
- showNounSubcategories(): void
- showVerbSubcategories(): void
- showAdjSubcategories(): void
- showPrepSubcategories(): void
- showDetSubcategories(): void
- showWordsInSubcategory(cat, subcat): void
+ showWordsInSubcategory(word): void
+ main(args: String[ ]): void
DictionaryData
+ initializeData(): void
+ getSubcategories(cat):Map<String, List<String>>
+ getWords(cat, subcat): List<String>
TranslationData
+ getTranslationsForWord(word): Map<String, String>
LexicalSimilarityCalculator
+ calculateLexicalSimularity(lang1, lang2): Text

So How Does It Work? (1/2)



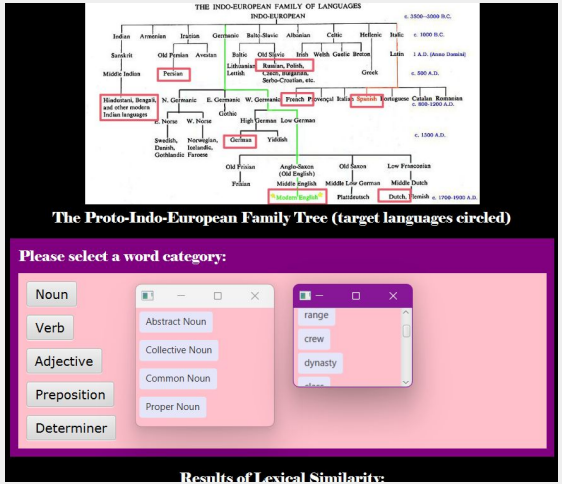
1: Select a word category

-Noun
-Verb
-Adjective
-Preposition
-Determiner



2: Select a sub-category

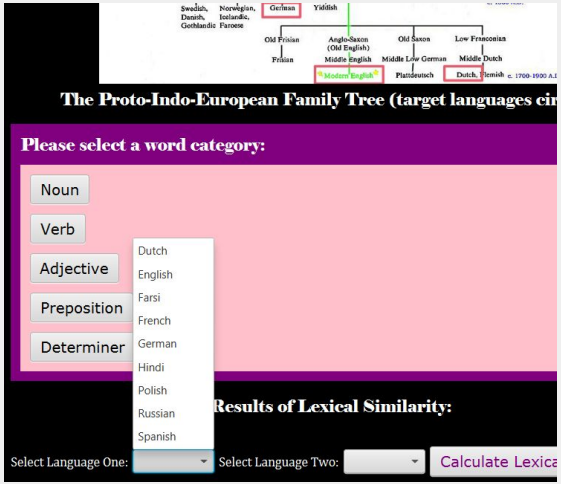
- Four per word category
- plan to integrate descriptions into program over the summer



3: Select a word!

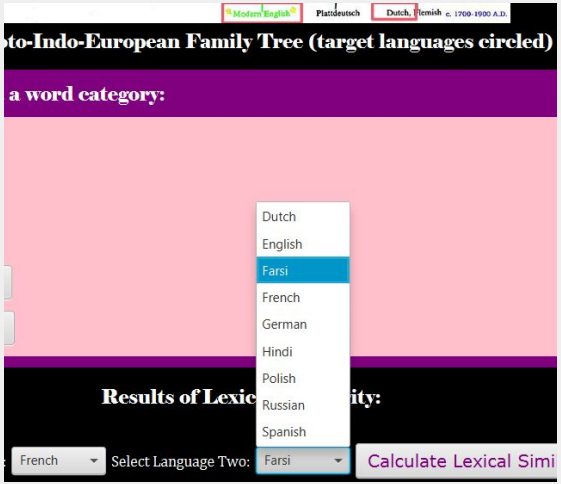
- Each word list is a ScrollPane
- Pressing will pull up terminal listing the word as translated

So How Does It Work? (2/2)



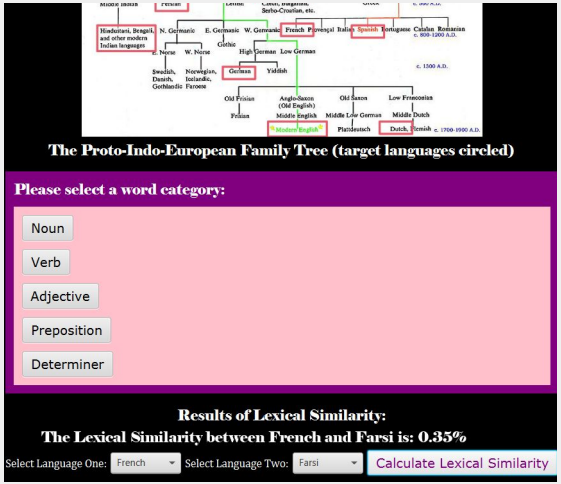
1: Select the first language

- ChoiceBox contains all languages in project for user selection
- Custom message if format is improper



2: Select the second language

- Case switch was used to determine output
- CB1 + CB2 = FREQ



3: Press button to calculate!

- Output is printed as a user-friendly string
- White font/black background for max visibility

PROJECT BENCHMARKS/OVERVIEW

01

GATHER ENG. WORDS

I used Google to find the 100 most common words in each category.

I then saved several copies of the words in several formats to streamline the future coding process.

02

DESIGN THE “BONES”

Next, I created the basic classes that were required to begin testing:

DictionaryApp.java
DictionaryData.java
TranslationData.java

From there, I started researching the most efficient way to store the data, and how to present it to the user.

03

THE TRANSLATIONS

From here, I took my lists and translated them into each respective language, storing the data in separate text files on my pc.

Over the course of ***several*** weeks, I began the arduous process of writing each translation into code.

04

PUT IT ALL TOGETHER

Finally, I created the

LexicalSimilarityCalculator.java

class and integrated it into the original program at the suggestion of one of my Linguistics professors.

Last but not least, I used my sketches to finalize the GUI for the main and sub-windows.

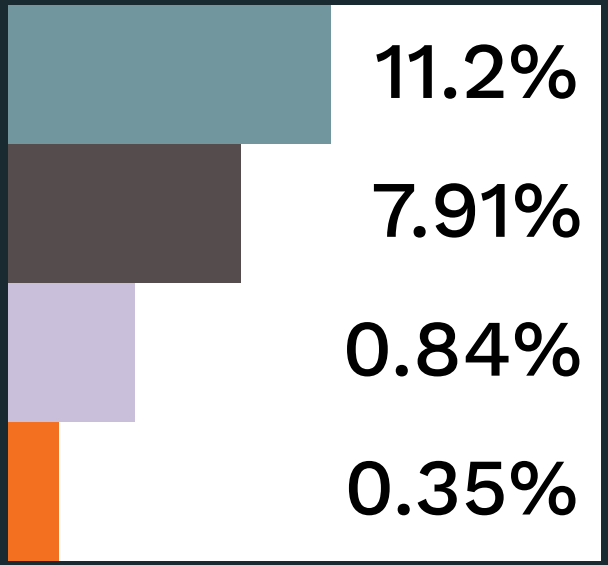
Unique Features/Design Elements:

Hash Map	This object is used in the structure of the translation data; each word is stored as a key, with the related values (translated words) assigned depending on the key selection.
Choice Box	This object is used in the Lexical Similarity Calculator; Since the similarity depends largely on the two variables (languages) selected by the user, all languages in the program needed to be displayed before the user can make a choice
Image	This object is used to display the entirety of the P.I.E. family of languages. Using a free photo editing software, I added red rectangles roughly around each language included in the program so far.
Pane	This object is used *heavily* in the design of the GUI, specifically in organizing each of the three major sections of the project and their positions within the larger project window.

Corpus Data

Here's a summary of some interesting results from the dataset I used!

Source:
http://ukc.disi.unitn.it/index.php/lexsim/#:~:text=According%20to%20the%20definition%20by



11.2%	French = Spanish
7.91%	English = Spanish
0.84%	Hindi = Dutch
0.35%	Farsi = French

Summer Improvement Timeline

“To improve is to change; to be perfect is to change often.”
-Winston Churchill.

“No ifs”	Plan to change if statements to switch-case statements
More Words	Adding more languages and translations (currently at 8/33/4,000, goal is 20/33/10,000)
Phonetic Desc’s	Интернационализировать = *[internatsionaliziro-vat’sya]*
More Control	Plan on learning HTML to design a site for the sake of more creative control & presentation power
Extra Context	Category descriptions, sentence builder, and info pages on language structures (dependent on ability to learn HTML)

Q&A


THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
GOZAIMASHITA
EFCHARISTO
DANKSCHEEN
JUSPAXAR
BOLZÏN
MERCI
BÏYAN
SHUKRIA
TINGKI
TASHAKKUR ATU
YAQHANYELAY
SUKSAMA
EKHMET
MEHRBANI
GRAZIE
MAAKE
KOMAPSUMNIDA
PALDIES
SPASIBO
SHACHALHUYA
NUHUH
CHALTU
WABEEJA
MAITEKA
YUSPAGARÀTAM
HUI
GUI
HATUR
EKOJU
SIKOMO
MAKETAI
MINMONGHAR
SPASIBO
DENKUUJA
NENACHALHIYA
UNALCHEESH
HIERSI
ATTO
ANIHA
DHANYABAAD
SAIKO
MERASTAHHY
GAEJTHO
LAH
TAVTAPUCH
MEDAHAGSE
AGUYJE
FAKAAUE
BAIHA