METU, Department Of Computer Engineering

Graduation Project Proposal Form

# Project Information

## Title

Enişte English

## Target

Public [x] Restricted [ ]

## Proposer Information

**Name(s)** Ayşenur Birtürk, PhD; Jason Steinberg

**E-Mail(s)** [birturk@ceng.metu.edu.tr,](mailto:birturk@ceng.metu.edu.tr) [steinber@metu.edu.tr](mailto:steinber@metu.edu.tr)

## IP (Intellectual Property) Information

The end product will be the property of and fully owned by a private corporation that has yet to be officially established (or officially named).

# Project Description and Background Information

## Description

The product will be a software package that uses NLP and machine learning to guide Turkish speakers of English to improve their performance along two dimensions: vocabulary development and error reduction. The scope of the project does not include a complicated GUI, but rather a simple GUI to test functionality. If the software itself is successful, it will be used as the basis of a mobile application.

## Similar Products/Projects

There are currently no products that are attempting to teach specifically Turkish speakers of English how to speak English. However, there are many products that attempt to teach English digitally to an international audience. Some of these are: DuoLingo, Voscreen, and English Ninjas (the latter two of which are Turkish companies). Other products that exist that attempt to automatically assess error correction are Avalingua and ETS’s *Criterion* writing assessment tool, but these are only for formal written text assessment.

## Justification of the proposal

The purpose of the product is to provide the vast majority of Turkish citizens, including the majority of METU students, with the opportunity to actually make genuine, dramatic improvement in their spoken English. The current suite of digital English learning products is insufficient to this task, as evidenced by the miniscule market share that they command (and can be confirmed by asking yourself: how many of these products am I aware of? Of those of which I am aware, how many are actually able to improve someone’s spoken performance (not just pronunciation)?). The failure of these products to improve Turkish speakers’ performance in English speaking is caused by: 1) their pedagogical focus on instruction of decontextualized grammatical rules (though some of these products, for example DuoLingo, do an OK job contextualizing vocabulary development), and 2) their inability to successfully integrate the following technologies – automated speech recognition and computational linguistic analysis tools. This integration is what will allow us to achieve our goal.

## Contributions, Innovation and Originality Aspects of the Project

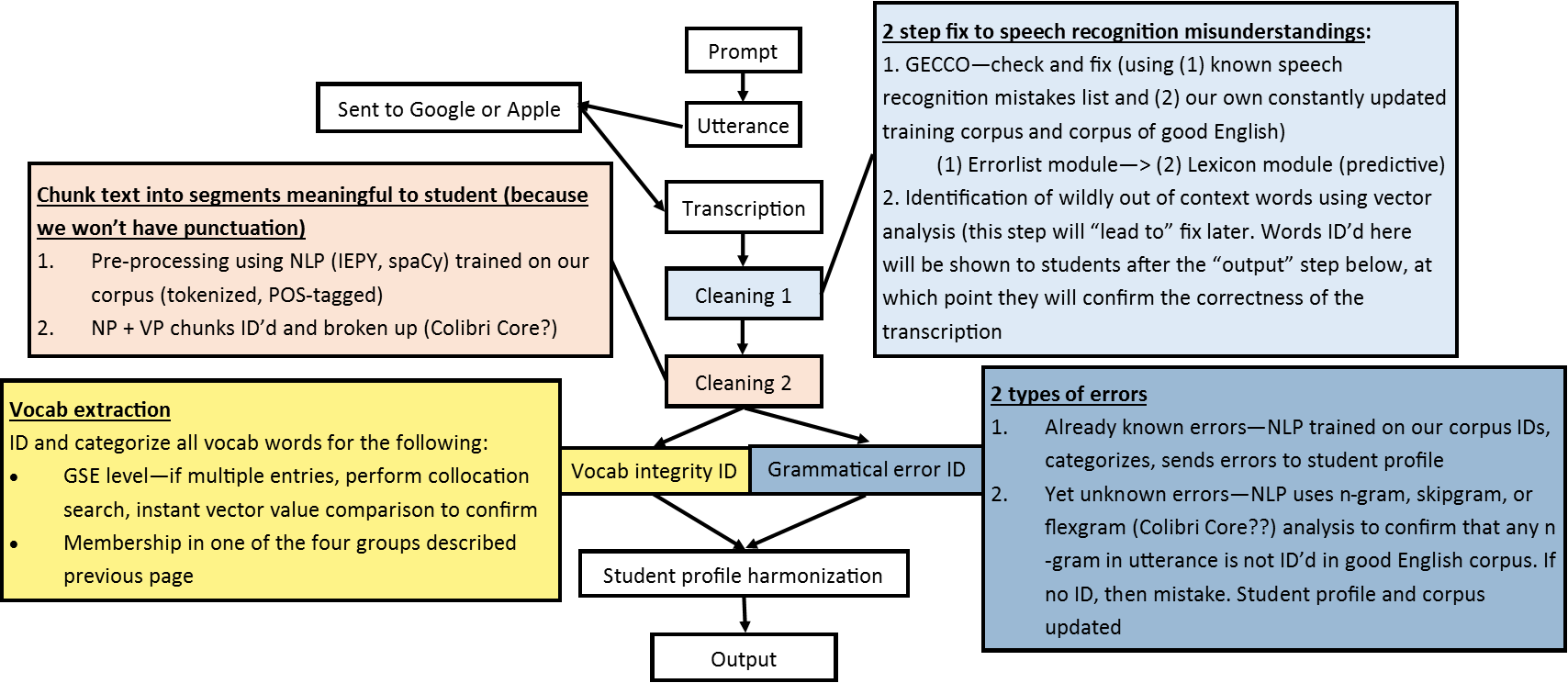
Using available technologies in Natural Language Processing (NLP) and machine learning, we plan to create a product that is, essentially, a computerized version of a one-on-one private language tutor. One specific innovation is that we will modify an existing vocabulary size test (the Nation test, URL below) to include a “depth” component, so that we will be able to more accurately model our students’ linguistic performance ability and more efficiently deliver to them developmentally appropriate vocabulary prompts. The second innovation is

that we will incorporate speech recognition – but in order to assess grammar and vocabulary, not just pronunciation (as currently commercially available software like DuoLingo do). The third innovation is that we will use machine learning to predict which, among all the speaking prompts we have created, is the best suited to assist student *x* in his or her vocab development and grammatical development at time *y*.

Because we will be building an actively updated corpus, our project has the potential to be the baseline data for computational linguistic studies regarding, for example, the degree to which Turkish students’ L1 (or any non- native speakers of English’s L1) influences their production in English. More broadly, we could use deep learning algorithma to analyze our corpus in order to generate a grammar of Turkish-English. This could have wide ranging academic and industrial applications, including regarding simple Google search attempts in a bilingual environment.

## Technical Aspects of the Project

Here is a workflow/pipeline that could be the basis for project development. In the backend side NodeJS will be used as a server side technology. In addition ExpressJS Framework will be used to ease the development process.



## Targeted Output, Targeted User/Domain Profile

The end product will be software code that operates with users via a simple GUI. Thıs software will explicitly not be linked to a GUI written in Swift or Java (i.e., we are not producing an application right now), but could be linked later, once this software has been proven to work. Testing will be done on the software via the simple GUI (likely web-based), which Turkish speakers of English will use for testing and improvement purposes.

Tangible success measures are: the creation of a functioning prototype, which can be used for testing and improvement; successful solution of problems identified in the testing phase (but which are unknown now). The users who this product are targeting are 50% of the Turkish population and 80% of METU students. If the software is developed to the degree that it works correctly (identifies correctly 80% of errors, produces prompts that support the student’s vocabulary development 100% of the time), the software will be used as the core of a native iOS application and Android application that will be sold commercially.

## Project Development Environment

We will use the folowing technologies: spaCy, NLTK (written in Python); Siri and Google Assistant speech recognition APIs; mmax2 corpus annotation software; NoSQL(MongoDB database of 35,000 annotated English vocabulary words (taken from Pearson’s Global Scale of English). The methods are described above in the technical aspect workflow.

## External Support

The external support that we will provide to the student team is an external programmer’s knowledge. This programmer has experience in the specific technologies we are attempting to use and will work in conjunction with the team as much as is necessary for the team to achieve its goals.

## References

Products similar in some way [https://www.duolingo.com](https://www.duolingo.com/) https://[www.voscreen.com](http://www.voscreen.com/) https://englishninjas.com/tr <http://cilenis.com/en/avalingua/> https://[www.ets.org/criterion](http://www.ets.org/criterion)

Software to be used

spaCy (NLP): [https://spacy.io](https://spacy.io/) NLTK (NLP): [http://www.nltk.org](http://www.nltk.org/)

mmax2 (corpus annotation): [http://mmax2.net](http://mmax2.net/)

English vocabulary source <https://www.english.com/gse/teacher-toolkit/user/lo>