HACETTEPE UNIVERSITY

COMPUTER ENGINEERING

BBM479 NATURAL LANGUAGE PROCESSİNG LAB.

PROJECT REPORT

Group Members:	Burcu İSKENDER	21328103
	Kürşat AKTAŞ	21227949

Subject: Language Detection (English,Turkish,German)

CONTENTS

1. Group

We are working as a two-person group on this project. Members of our group are Burcu İskender and Kürşat Aktaş.

2. Task

The program will determine the language of the given text. But there is a restriction on the number of languages. Inputs may be in Turkish, English ,German or mixed up.The output of the program will be given along with the probabilities on which the given input is written.

For example ;

input : “ Bu proje incredible. “
output : Given string is in Turkish language.
Probability Details
Turkish : 99.96 %
German : 0.04 %
English : 0.01 %

2. Solution

First of all our program reads the input files, in this reading step program only accept the words. So we don’t need to store numbers, punctuation marks or other special characters for determining the language of a string. After that, program constructs calculates word frequencies for each language with the help of training datasets.

Because of each of the 3 languages (English, Turkish, German) mostly has distinct words in their language domains, we have only used unigram probability for detection process. This approach makes our program faster than bigram approach (which we thought to use and mentioned in project report). To explain detection progress, program firstly clears the input from numbers, punctuation marks and special characters and splits it to its words. Then it calculates the unigram probabilities of each words for each language separately. After that it shows to the user the language which has the biggest unigram probability.

When we have examining the output of our program with different inputs, we observed that it runs with %90-99 success rate. However, program may produces uncertain output when the given input is ambiguous like “Hello naber ?”. Because of these ambiguous inputs causes the same problems in the most known language detection programs (Google Translate etc.), we thought it's satisfiable for this project.