INTRODUCTION TO NLP

Prof. Dr. Eşref ADALI

Chapter – I

E-mail : adali@itu.edu.tr

www.adalı.net or www.xn--adal-oza.net

Language

People were able to communicate and understand each other by speaking. Thanks to language, advances were made in culture, art and science. Languages take a long time to develop and their development is continuous.

In time, new words are added to the language or some words are faded out depending on the developments in the society. In fact, the way a language constructs sentences can change over time.

According to the United Nations data, it is understood that more than 4000 languages are spoken today. However, some of these languages are widely spoken, while others are used by a few people.

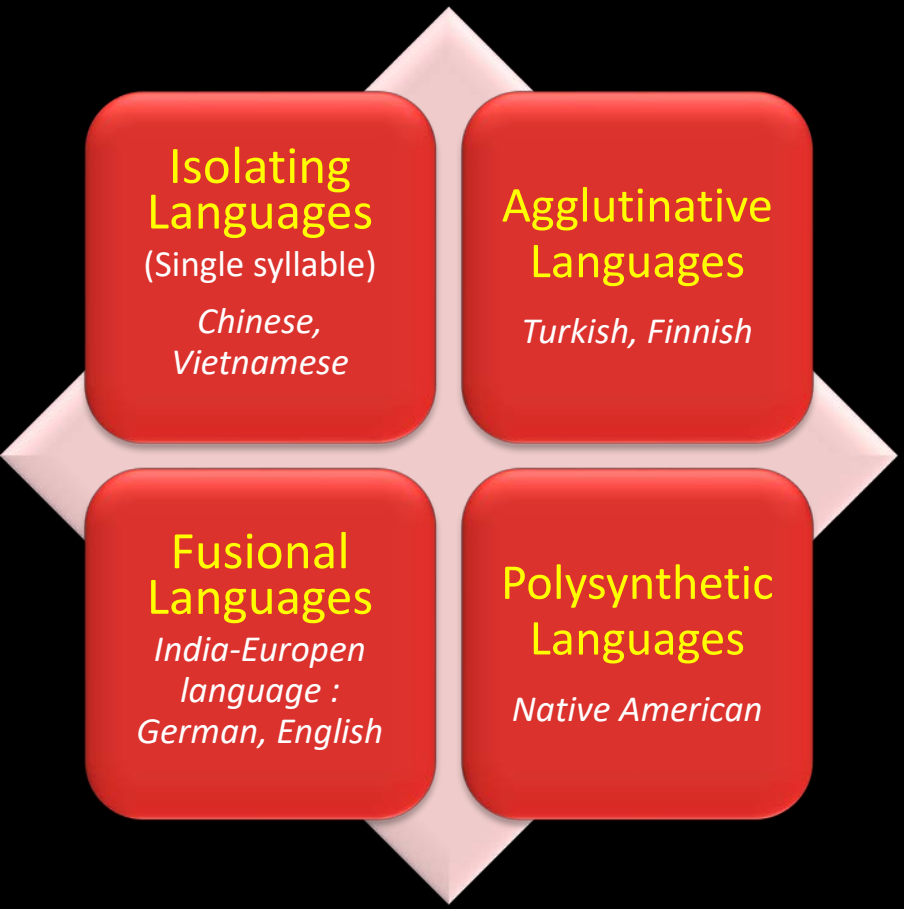
Language

There is a close connection between the sophistication of languages and the development of the societies speaking that language in the field of culture, art and science. In other words, communities with insufficient language cannot be expected to be successful in the fields of culture, art and science.

Confucius (551-474 BC)

"… If I were to take charge of a country, the first thing I would do would undoubtedly be to review its language. Because if the language is imperfect, words cannot express thought well. Duties and services cannot be performed properly if the thought cannot be expressed well. In places where duty and service cannot be performed properly, customs, rules and culture are broken. If custom, rule and culture are corrupted, justice will go astray. If justice goes astray, the people who are bewildered do not know what to do or where the business will lead. That's why nothing is as important as language!...

Language Classification-I



Isolating language: Words do not take suffixes. Words gain meaning depending on their order and emphasis in the sentence. Chinese, Tibetan, Vietnamese, and Himalayan are counted in the set of isolating languages.

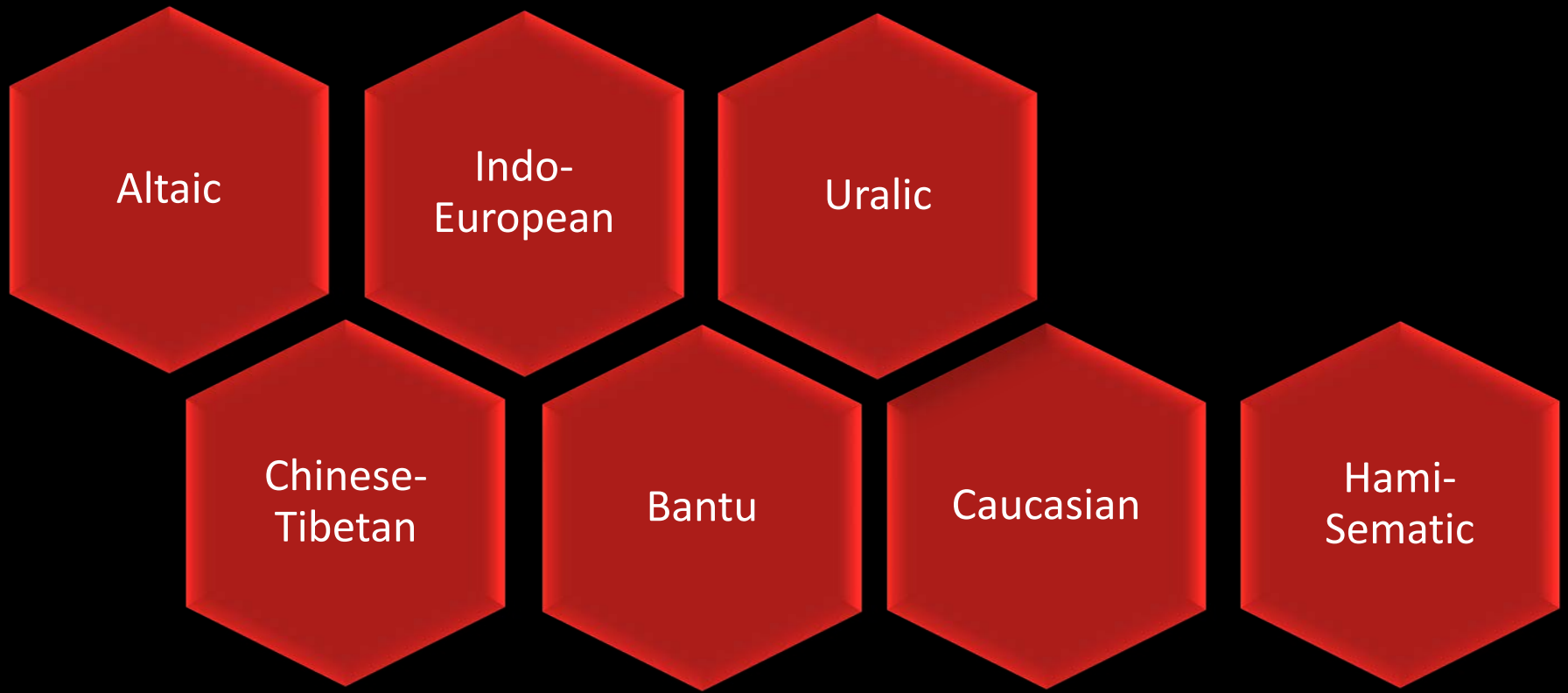
Polysynthetic languages: The verb merges with the other elements of the sentence. Therefore, the action can be the entire sentence. Native American languages are included in this cluster.

Language Classification-II

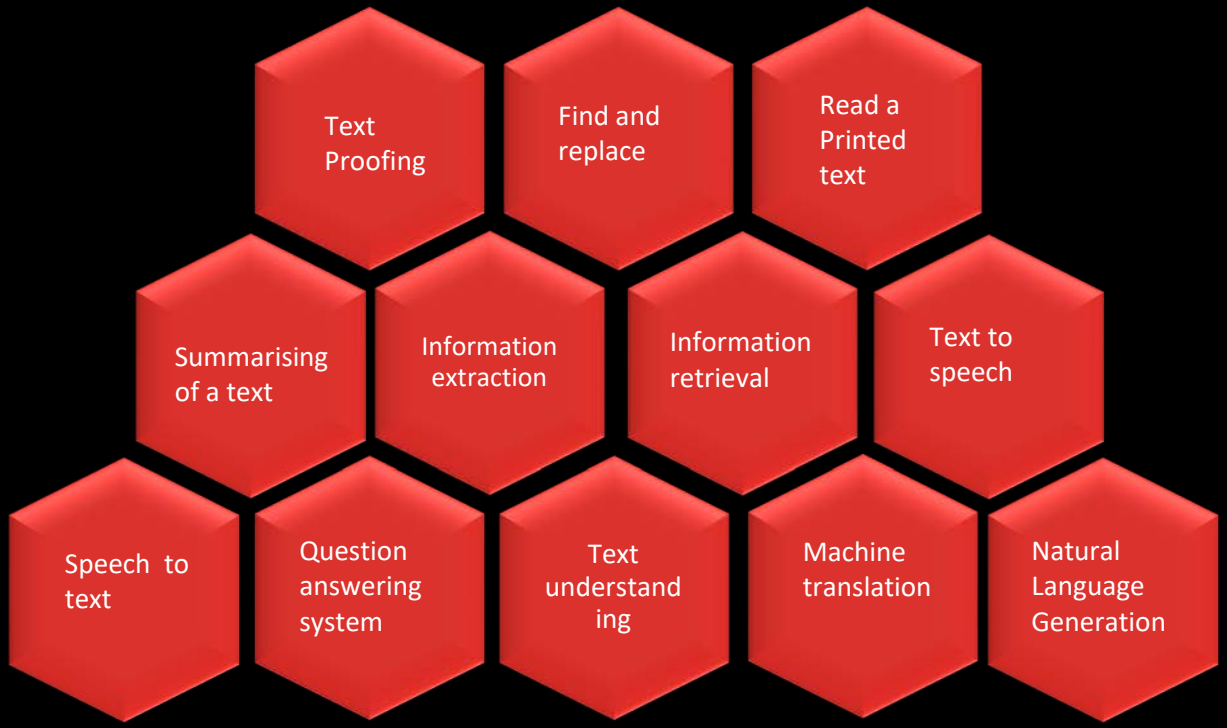
Fusional language: Indo-European and Hami-Semitic languages are evaluated within the fusional language structure. In Indo-European languages, the body word acquires a new meaning by taking prefixes and suffixes. The number of prefixes and suffixes added to a root usually does not exceed one. The fusional language family includes Arabic, Arabic, Jewish, Persian, German, English, French and Spanish.

Agglutinative languages: The basis of the word is the root word. New words are formed by adding derivational and inflectional suffixes to the root word, but the root never changes. There is no limit to the number of suffixes that can be added to the root word. Therefore, many words can be derived from a root word. The agglutinative language family includes Turkish, Hungarian, Manchu, Tungusic, Finnish, Mongolian, Samoyed, Korean, and Japanese.

Language Families



Natural Language Processing - NLP



Understanding and generation of languages that humans use naturally

Text Proofing

Finding and correction of;

- Spelling errors
- Punctuation errors
- Grammatical errors

Indo-European language dictionary required

Rules can be used in Turkish languages:

- Vowel and Consonant harmony
- Syllable structure
- Harmony of suffix
- Fusion sounds
- Sound drop

Dünyanın en güzel kelabekleri Eğirdir'de görülürler.

Kelebakler >> kelebekler

Türtklerin tarihi çok eskidir.

Türtklerin >> Türklerin

V, VC, VCC, CV, CVC, CVCC

Find and Replace

In fusional languages, words take a single suffix, so it's very easy to find and replace;

Dankey >> hourse
Dankeys >> hourses

In agglutinative languages, words can have many suffixes. Suffixes are added in accordance with sound harmony. A letter may be added between root and suffix. So it is not easy as fusional languages find and replace method.

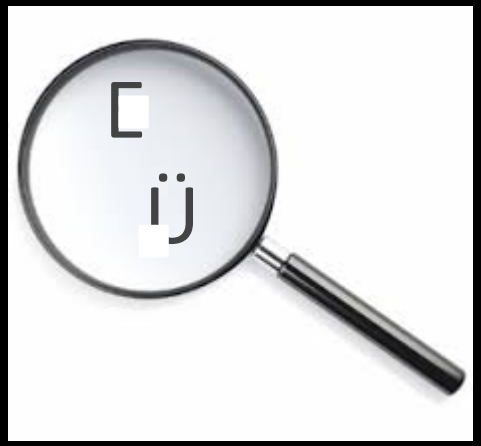
Eşek >> eşekler
At >> atlar
Eşeği >> atı
Eşeğin >> atın
Eşekçik >> atcık

- Adamın biri Timur'a iyi bakımlı bir eşek hediye etmiş. Etraftaki dalkavuklar eşeği öve öve bitirememiş. Eşeğin gözleri kocaman, kulakları dikmiş. Eşekçik devamlı koşuyormuş.
- Adamın biri Timur'a iyi bakımlı bir at hediye etmiş. Etraftaki dalkavuklar atı öve öve bitirememiş. Atın gözleri kocaman, kulakları dikmiş. Atçık devamlı koşuyormuş.

Read a Printed Text

- Texts can be books or documents printed in ancient times, or they can be information entry forms that users fill out with their hands or with a typewriter or computer.
- Reading texts by optical methods and transferring them to the computer is image processing or character recognition.
- Strings that read optically written texts and convert the letters and numbers they read into computer characters produce very successful results (better than 90%).
- However, they cannot be said to have made any mistaken recognition.

Adı	E R E N
Soyadı	M U T L U
Mesleği	M E M U R
Hesap numarası	1 2 3 4 – 9 8 5 7 4
Adresi	
İlçe	Ü S K Ü D A R
İl	İ S T A N B U L



Summarising of a Text

- People who don't have the time may need to have a summary of a book or report.
- NLP provide good solution for summarising process.

There are different types of summaries depending what the summarization program focuses on to make the summary of the text, for example *generic summaries* or *query relevant summaries* (sometimes called *query-biased summaries*). Summarization systems are able to create both query relevant text summaries and generic machine-generated summaries depending on what the user needs. Summarization of *multimedia* documents, e.g. pictures or movies are also possible.
Some systems will generate a summary based on a single source document, while others can use multiple source documents (for example, a cluster of news stories on the same topic). These systems are known as *multi-document summarization* syst
There are different types of summaries depending what the summarization program focuses on to make the summary of the text, for example *generic summaries* or *query relevant summaries* (sometimes called *query-biased summaries*). Summarization systems are able to create both query relevant text summaries and generic machine-generated summaries depending on what the user needs. Summarization of multimedia documents, e.g. pictures or movies are also possible.
Some systems will generate a summary based on a single source document, while others can use multiple source documents (for example, a cluster of news stories on the same topic). These systems are known as *multi-document summarization* systems.

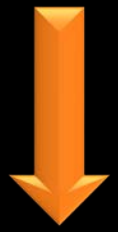
- Title
- Keywords
- Topics
- Frequently used word

Information Extraction - IE

Extract structured information from unstructured documents

Unstructured

Ahmet says he will marry Ayşe on May 15 in Istanbul



Structured

<parties> Ahmet, Ayşe
<Action> marriage
<time> May 15
<location> Istanbul

Information Retrieval - IR

Searching for
- Information in documents
- Documents themselves
- Metadata which describe documents

or searching within databases
or searching within WWW

Precision = relevant documents / all needed documents

Recall = relevant document / retrieved documents

Fall-Out = non-relevant documents / all non-relevant documents

Text to Speech

Converts normal language text into speech.

Add natural sounds

Passengers boarding the *Ankara* flight, please go to gate *212*.

Artificial sound generation

Speech to Text

Converts spoken words to machine-readable input .

- Transcribing the speaker's speech,
- Preparing meeting minutes,

short orders

Bring my slipper
Close the door
Left vein occluded
Two hundred twelve, two hundred eighty-five, thirty-six eighty-six

Question Answering

Answering of any question which are given written or verbal.

Closed-domain

- Answering question in a specific domain (for example, computer, sport, film, travel)

Open-domain

- Answering about nearly everything

Computer : Good afternoon, how can I help you?

Customer : I'm looking for a single room

Computer : Want a single room?

Customer : Yes

Computer : What day will you come?

Customer : For Monday, October 10th and three days after

Computer : Are you going to check in on October 10?

Customer: Yes, I'll check in on October 10 at noon. I told you !

Computer : How many days do you plan to stay ?

Customer : I will leave on October 13

Computer: We have room for you. If you want a room with a lake view, you will pay 10 TL more, but with a view. Do you want?

Customer: Maybe

Computer : I don't understand. did you say yes

Customer: Yes, I said yes.

Computer: I reserved your room. May I know your credentials now?

. . .

Text Understanding

Understanding of one or more sentence exactly.

Please, transfer 123,00 € to Ayşe Tan; account number 123456789 from my saving account,
Thanks,
Eren Çalışkan

Money transfer from Ali's account to Ayşe's account without any operator

Bring my sleeper

Machine Translation

Translation of a text or speech from original language to another language.

I am goint to school today.

Bugün okula gidiyorum

Natural Language Generation

Structured information

<object> 1st Singular
<subject> apple
<action> to eat
<time> now

Generated Sentence

I am eating an apple now

References

[1] http://www.azquotes.com
[2] https://www.ethnologue.com/guides/how-many-languages
[3] https://www.ethnologue.com
[4] B. Sakça, Dil Aileleri, http://www.yenimakale.com/dil-aileleri.html, 2009
[5] B. Atalay, Türk Dilinde Ekler ve Kökler Üzerine Bir Deneme, TDK Yayınları, İstanbul, 355)-378. 1942
[6] E. Adalı, Similarities and Differences of Turkic Languages, Turklang 2016, Bishkek
[7] E. İlgen, Türkçe Sözcük Anlam Belirsizliği Giderme, Doktora Tezi, İTÜ Fen Bilimleri,2015
[8] D. Yüret, Discovery of Linguistic Relations Using Lexical Attraction, Doktora Tezi, MIT, 1998
[9] A. Pirkola, Morphological Typology of Language for IR, Journal of Documentation 57 (3), 330-348