Towards Automatic Multilingual Readability Assessment

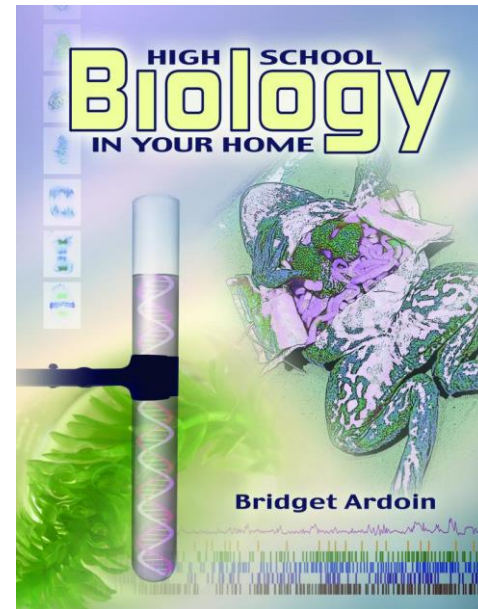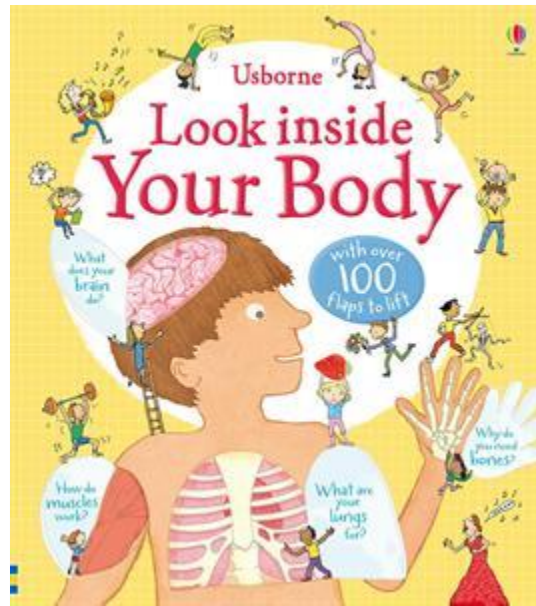ION MADRAZO AZPIAZU

ADVISOR: DR. SOLE PERA

Motivation

Reading for learning versus reading for understanding

What is Readability?
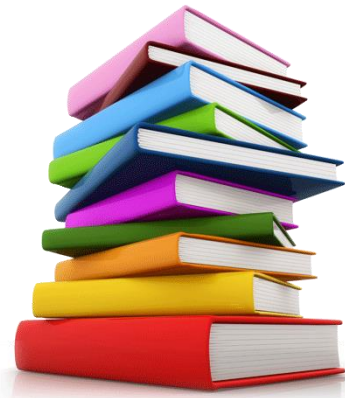

Text complexity/Readability

Applications of Readability

Text simplification for people with reading difficulties

Book recommendation

Literacy assessment

Ensuring medical and legal document understanding

Related Work: Languages
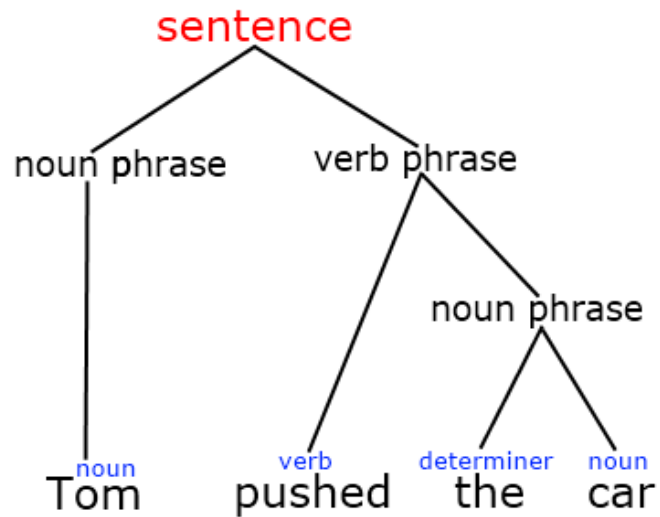
Related Work: Features
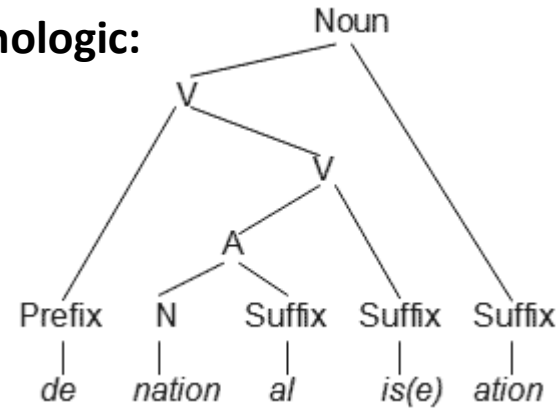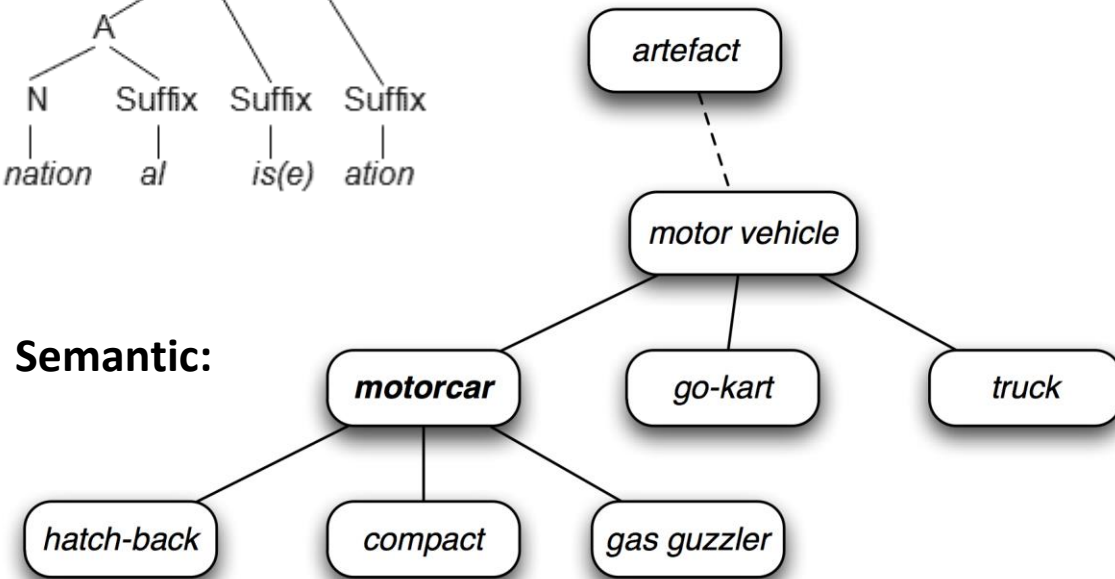
Traditional:
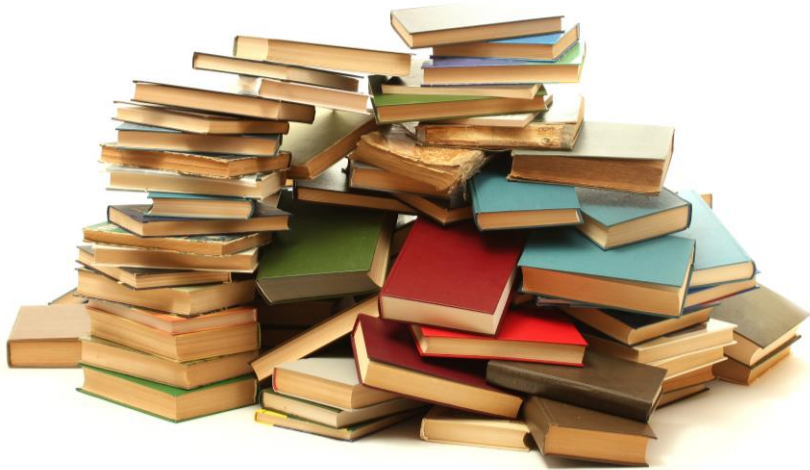
shal·low

Morphologic:

Metadata:

Syntactic:

Semantic:

Related Work: Document Types

K aži mi koji sat nosiš i reći ću ti tko si. Ova već pomalo potrošena fraza primjenjiva na naše brojne navike i vrijednosti koje posjedujemo zapravo je nevjerojatno istinita. Gotovo je nemoguće da na ruci nosite sat poput Glashuette Originala, a da ne znate da iza njega stoji višestoljetna urarska tradicija ili pak da uz najnovije Boss odijelo na ruku stavite novi Swatch popularnoga dizajnera, a da trenutačno ne budete prepoznati kao trendseter. Ma kakav bio sat koji nosite on govori prije svega o tome koliko o satovima znate, kakav vam je ukus, a u konačnici i koliko si možete priuštiti. Uz prilog koji vam je u rukama odabir sata koji želite bit će vjerujem lakši. Posjetili smo dva vodeća sajma koji prate novosti iz svijeta satova, a donosimo i priču o velikome povratku mehaničkoga sata. Ljubitelji automobila iznenadit će se čitajući prilog *Sat za svaki automobil* koliko su ova dva svijeta posljednjih

Lucy is going to
the park and
she is taking the
dog for a walk.

8

Thesis Statement

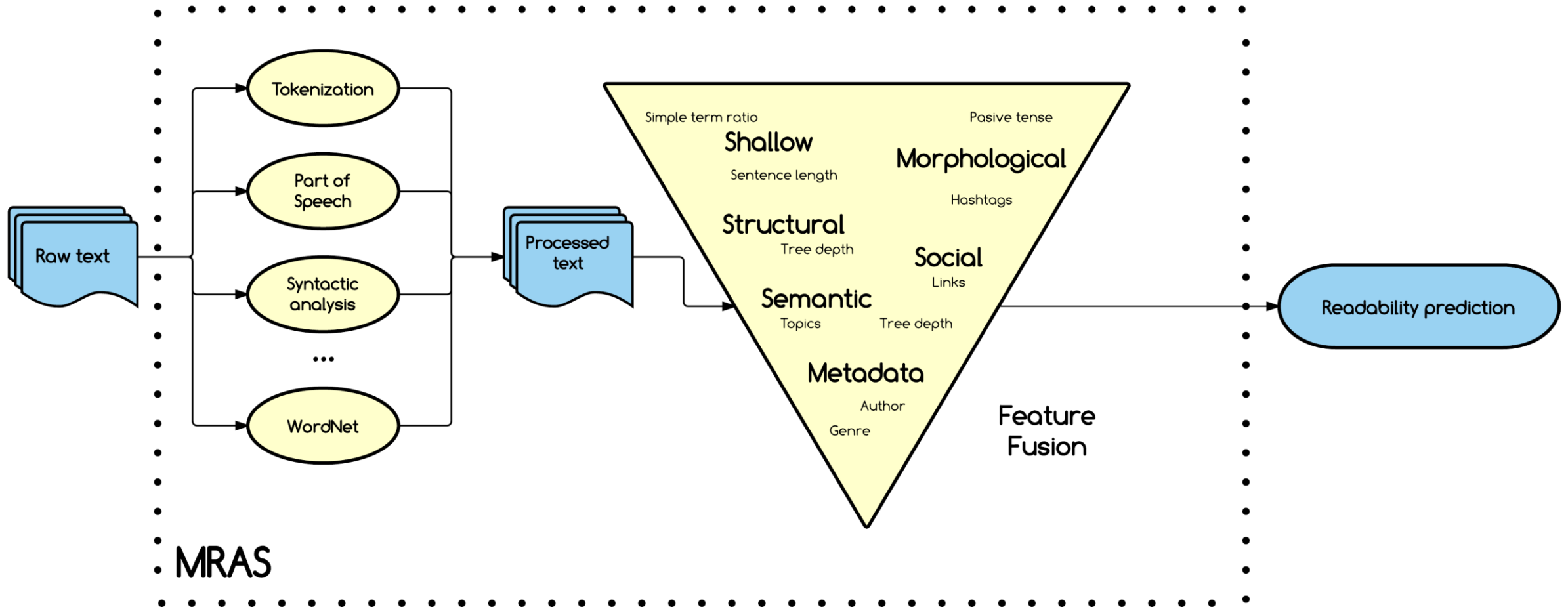o Multilingual Readability Assessment System (MRAS)

- o Multilingual
- o Multi-document
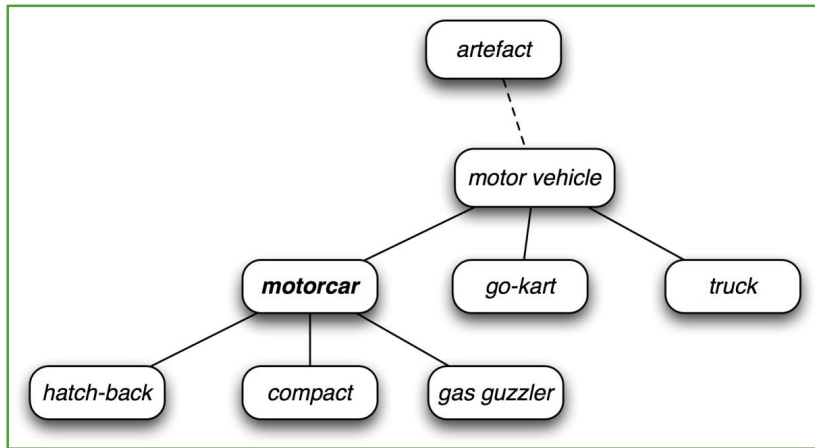- o Open-source
- o Adaptable

o How:

o Analysis of existing features

- o Exploration and design of new features
- o Analysis of fusion methods

Proposed Method

Text Processing

WordNet

- Tokenization
- Sentence Split
- Part of Speech
- Dependency
- Named entities

Freeling

Terms	Documents													
	M.1	M.2	M.3	M.4	M.5	M.6	M.7	M.8	M.9	M.10	M.11	M.12	M.13	M.14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	1	0	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0

LSA

Feature Extraction

Shallow features	Morphological	Structural	Semantic	Social*	Metadata
• Word length • Sentence length	• Tense • Number	• Syntactic • Pragmatic • Dependency ngrams *	• Concept density • Followability* • Ambiguity*	• Hashtags • Links • Mentions	• Author • Description • Genre

*Novel

Feature Fusion

Classification

Regression

Ordinal Classification

Evaluation

❑Dataset

Language	Dataset	Description
English	Lexile	Contains book titles associated with readability level
	Standardized tests	Tests for English level, contain various texts per test
	Other	News for kids, exercises for learning English
Spanish	Lexile	Contains book titles associated with readability level
	Learning resources	Various exercises for learning Spanish
Basque	Learning resources	Various exercises for learning Basque
Multilingual	Parallel corpus	Same documents translated into English and Spanish

❑Metrics: Accuracy, Precision, Recall, RMSE…

Research Questions

❑ Performance: Learning model, Feature subset

❑ Features versus Language

❑ Comparison : Baselines, State of the art

❑ Same text in different language, same readability level?

❑ Can a MRAS be trained for two languages and predict a third one?

❑ Languages with low resources, can get help of other languages's data?

Schedule

Date	Milestone
April 2016	Gather existing datasets for design and development
May 2016	Feature Exploration
June 2016	Feature fusion
July 2016	Experiments
July 2016	Thesis draft
August 2016	Defense

Questions?

MRAS