Towards Automatic Multilingual Readability Assessment

THESIS PROPOSAL – APRIL 2016 – BOISE STATE UNIVERSITY

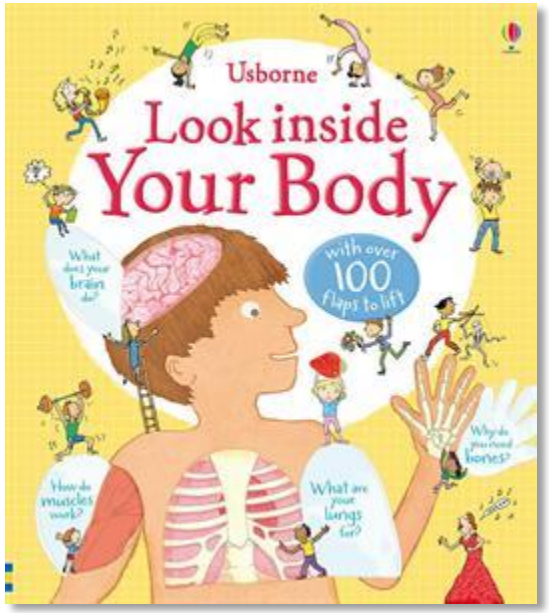
ION MADRAZO AZPIAZU

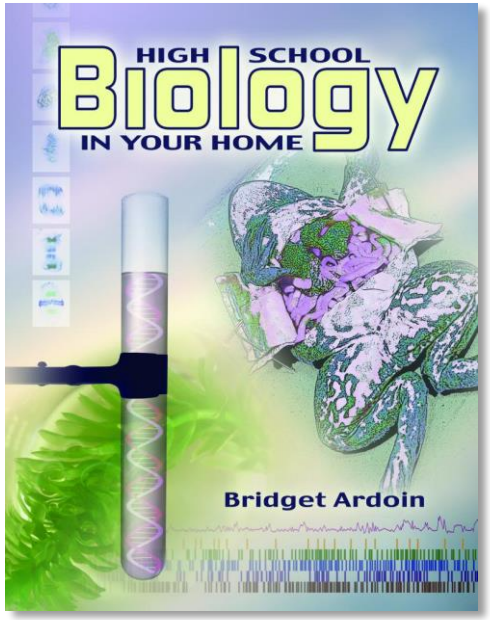
ADVISOR: DR. SOLE PERA

Motivation

Reading for learning versus reading for understanding

What is Readability?





OPEN ACCESS Freely available online PLoS BIOLOGY

A Role for Parasites in Stabilising the Fig-Pollinator Mutualism

Derek W. Dunn[1,2,3], Simon T. Segar[1,2], Jo Ridley[3], Ruth Chan[3], Ross H. Crozier[4], Douglas W. Yu[5], James M. Cook[1,2,3]*

1 Division of Biology, Imperial College London, Ascot, United Kingdom, 2 School of Biological Sciences, University of Reading, Reading, United Kingdom, 3 School of Biological Sciences, University of East Anglia, Norwich, United Kingdom, 4 School of Marine and Tropical Biology, James Cook University, Townsville, Queensland, Australia, 5 Natural Environment Research Council (NERC) Centre for Population Biology, Imperial College London, Ascot, United Kingdom

Mutualisms are interspecific interactions in which both players benefit. Explaining their maintenance is problematic, because cheaters should outcompete cooperative conspecifics, leading to mutualism instability. Monoecious figs (*Ficus*) are pollinated by host-specific wasps (Agaonidae), whose larvae gall ovules in their "fruits" (syconia). Female pollinating wasps oviposit directly into *Ficus* ovules from inside the receptive syconium. Across *Ficus* species, there is a widely documented segregation of pollinator galls in inner ovules and seeds in outer ovules. This pattern suggests that wasps avoid, or are prevented from ovipositing into, outer ovules, and this results in mutualism stability. However, the mechanisms preventing wasps from exploiting outer ovules remain unknown. We report that in *Ficus rubiginosa*, offspring in outer ovules are vulnerable to attack by parasitic wasps that oviposit from outside the syconium. Parasitism risk decreases towards the centre of the syconium, where inner ovules provide enemy-free space for pollinator offspring. We suggest that the resulting gradient in offspring viability is likely to contribute to selection on pollinators to avoid outer ovules, and by forcing wasps to focus on a subset of ovules, reduces their galling rates. This previously unidentified mechanism may therefore contribute to mutualism persistence independent of additional factors that invoke plant defences against pollinator oviposition, or physiological constraints on pollinators that prevent oviposition in all available ovules.

Citation: Dunn DW, Segar ST, Ridley J, Chan R, Crozier RH, et al. (2008) A role for parasites in stabilising the fig-pollinator mutualism. PLoS Biol 6(3): e59. doi:10.1371/journal.pbio.0060059

Introduction

In a biosphere driven by selection at the level of the individual gene [1], explaining the existence of cooperation, such as mutualism, is a major scientific challenge. Mutualisms are interspecific ecological interactions characterised by reciprocal benefits to both partners [2] that usually involve costly investments by each. What factors thus prevent one partner from imposing unsustainable costs onto the other to enable mutualism stability [3–7]? In some mutualisms, the larger, more sessile partner, manipulates the other by directing benefits to cooperative individuals and costs to cheaters [4–7]. However, a general consensus on mutualism persistence has only recently been formulated, and this clearly shows that a high benefit-to-cost ratio of cooperating is one important factor [8,9].

Fig trees (*Ficus*) and their host-specific agaonid pollinator wasps are a classic example of an obligate mutualism [10,11]. The wasps pollinate the trees, and the trees provide resources for wasp offspring. In monoecious *Ficus*, female wasps push their way through a specialised entrance into receptive syconia (colloquially, "figs"), which are enclosed inflorescences. The wasps then pollinate the tree while depositing their eggs individually into ovules. Thus, each egg laid costs the tree one seed, but upon emergence, the female wasp offspring disperse that tree's pollen. Trees need to produce both wasps and seeds for the mutualism to persist, but natural selection should favour wasps that exploit the maximum number of fig ovules in the short term, resulting in a conflict of interest between wasp and tree. However, the mutualism has persisted for at least 60 million years and has radiated into more than 750 species pairs [12]. The mechanisms preventing wasps from overexploiting figs remain unknown, despite intensive study over four decades.

Within receptive syconia, the lengths of floral styles are highly variable [13,14], and ovipositing pollinators (foundresses) favour flowers with shorter styles for their offspring [15–18]. Style and pedicel lengths of flowers are negatively correlated. Short-styled ovules develop into seeds or galls (when a wasp is present) near the syconium inner cavity, while most long-styled ovules develop into seeds near the outer wall [19,20] (Figure 1). These patterns have been shown to reflect the oviposition preferences of foundresses, and are unlikely to be the result of greater elongation of pedicels containing eggs during syconial maturation, because in receptive syconia, pollinators' eggs are mainly present in short-styled inner ovules [16]. These widespread observations have been tied to four, not necessarily mutually exclusive, mechanisms that have been proposed to stabilise the fig-pollinator mutualism: (1) Unbeatable seeds—outer ovules may be defended biochemically or physically against oviposition or larval development [21]. However, no mechanism has yet been identified. (2) Short ovipositors—pollinators' ovipositors may be too short to fully penetrate the long styles of

Academic Editor: Anurag A. Agrawal, Cornell University, United States of America

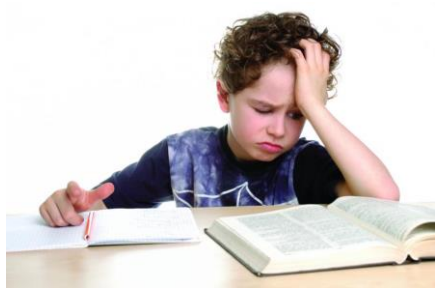
Received September 14, 2007; **Accepted** January 22, 2008; **Published** March 11, 2008

Abbreviations: s.e., standard error

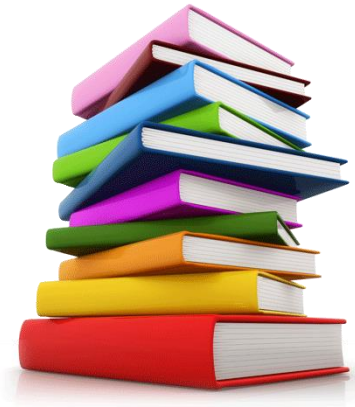
* To whom correspondence should be addressed. E-mail: james.cook@reading.ac.uk

Text complexity/Readability

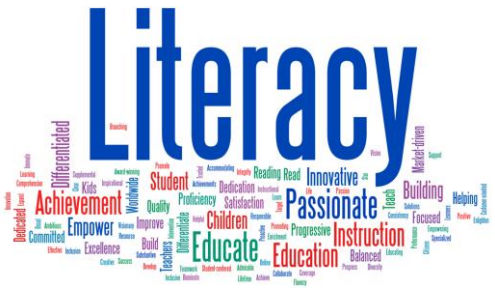
Applications of Readability



Text simplification for people with reading difficulties



Book recommendation



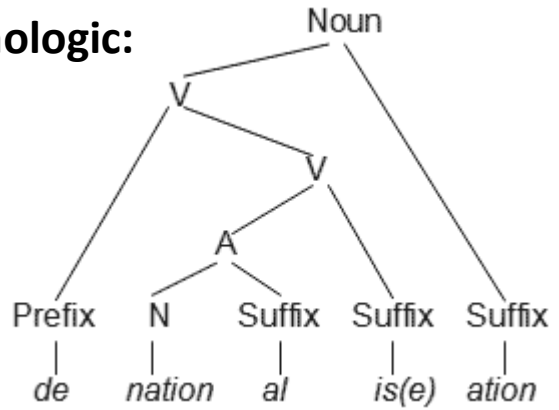
Literacy assessment

Ensuring medical and legal document understanding

Related Work: Features

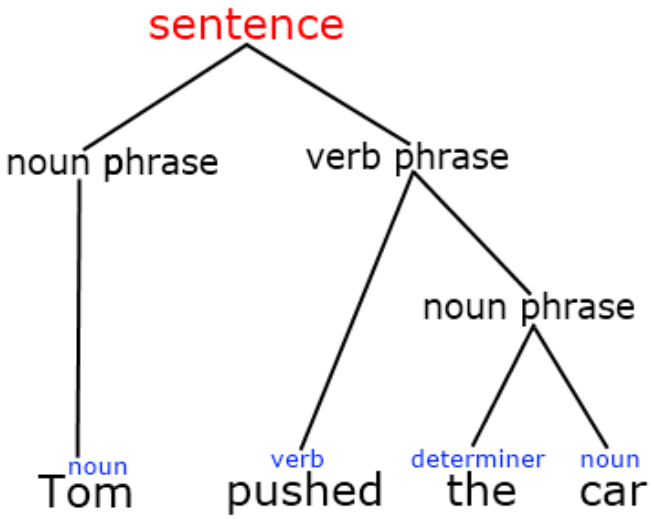
Traditional: shal·low

Morphologic:

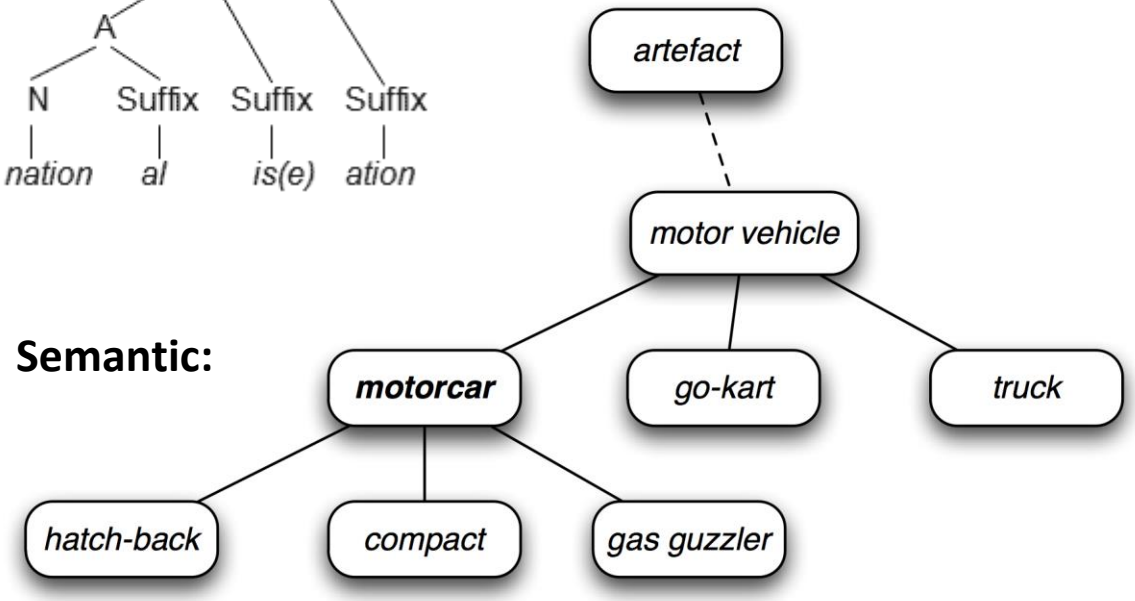


Metadata: <category/>

Syntactic:

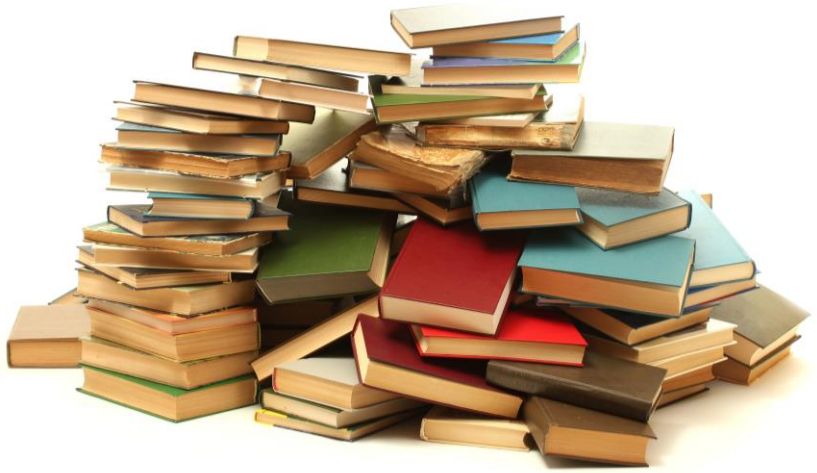


Semantic:



Related Work: Languages

Related Work: Document Types



Kaži 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

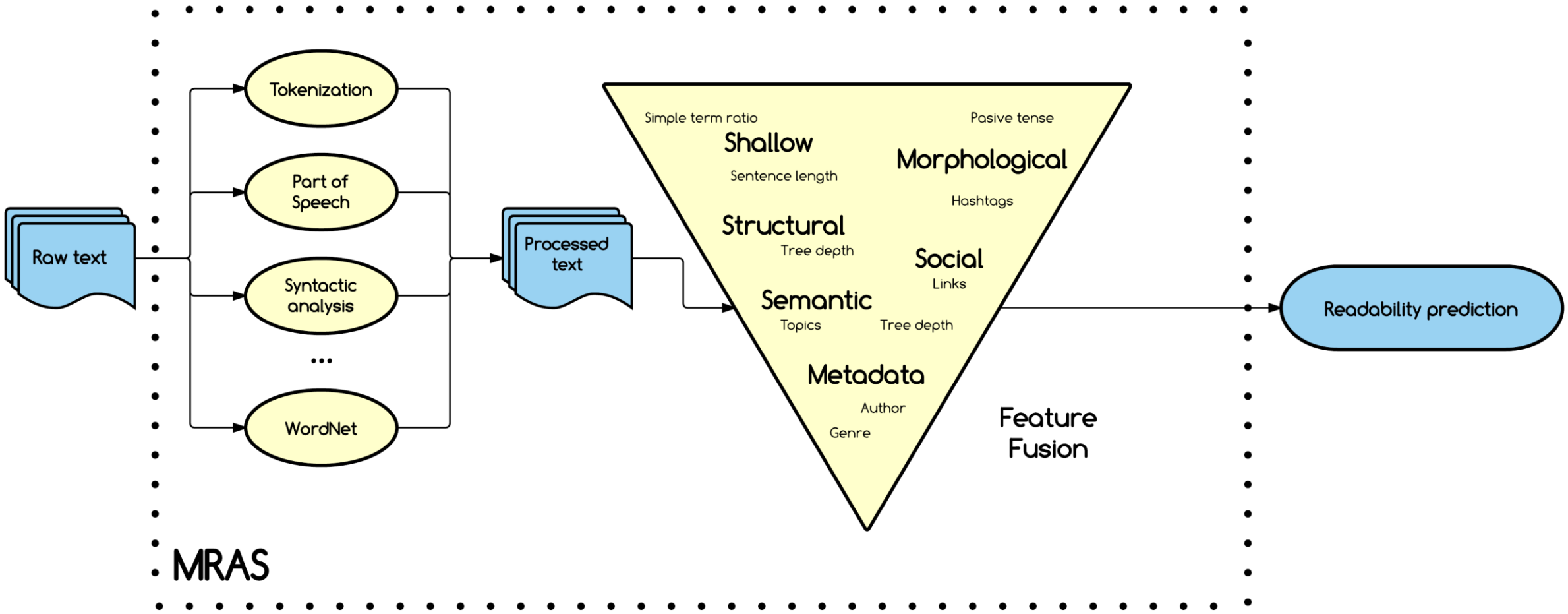
Lucy is going to the park and she is taking the dog for a walk.

Thesis Statement

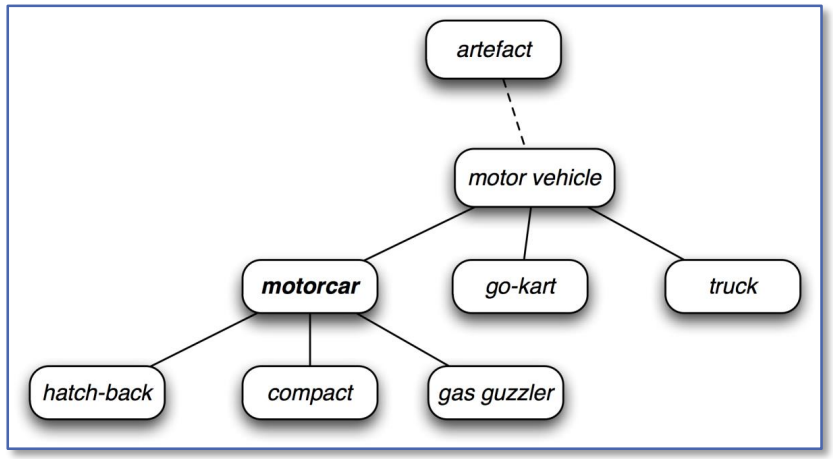
- Multilingual Readability Assessment System (MRAS)
  - Multilingual
  - Multi-document
  - Open-source
  - Adaptable

- How:
  - Analysis of existing features
  - Exploration and design of new features
  - Analysis of fusion methods

Proposed Method



Text Processing



WordNet

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

Freeling

Terms	Documents													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
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	0	1	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

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…

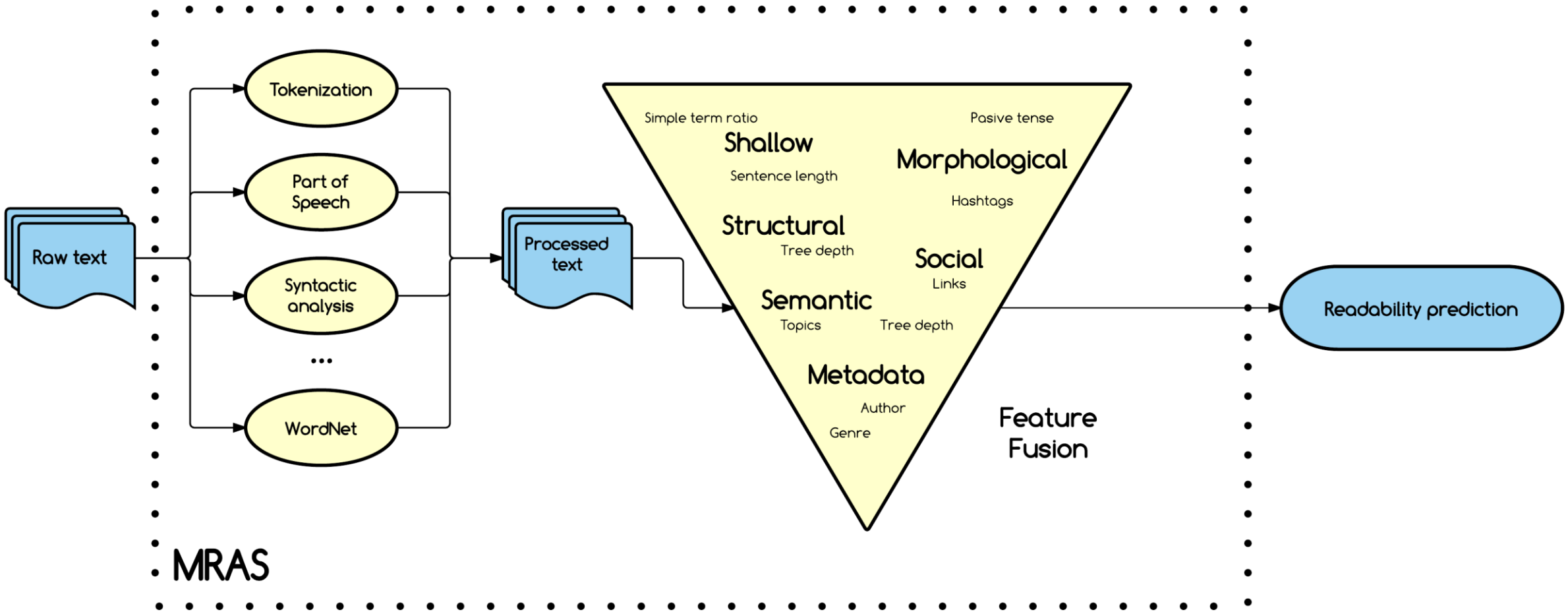
Research Questions

- 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' 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?


Tokenization
Part of Speech
Syntactic analysis
...
WordNet
Raw text
Processed text
Simple term ratio
Pasive tense
Shallow
Sentence length
Morphological
Hashtags
Structural
Tree depth
Social
Links
Semantic
Topics
Tree depth
Metadata
Author
Genre
Feature Fusion
Readability prediction
MRAS