Natural Language Processing Applications

Week 8: Readability

fit@hcmus
KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

- Introduction to Readability
- Related research
- Corpora
- Assessing Readability

Applied NLP - Readability

INTRODUCTION TO READABILITY

fit@hcmus

Introduction to Readability

- Reading plays a crucial role in information exchange and acquiring knowledge
- Developing proficient reading skills is an important component of success, not only in academic settings but also in business and social settings (Geiser & Studley, 2002; Powell, 2009).
- Participation in cognitively stimulating activities such as reading has been associated with reduced late-life cognitive decline (Wilson, Boyle et al. 2013)

fit@hcmus

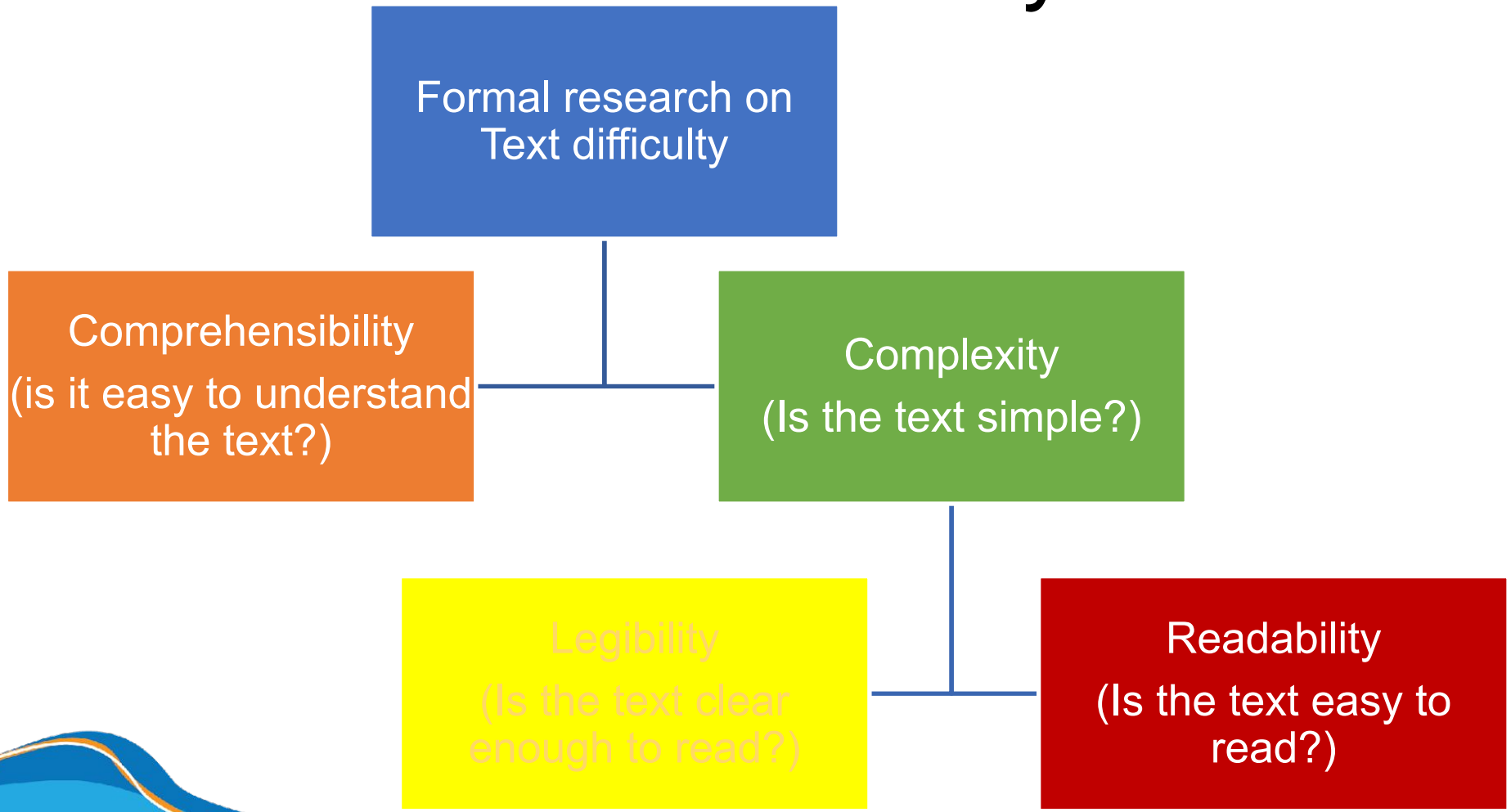
Introduction to Readability

- The Internet underwent an explosive growth since it's inception
- Increased readership globally
- A large amount of text is created daily

How do you tell **if a piece of text is difficult?**

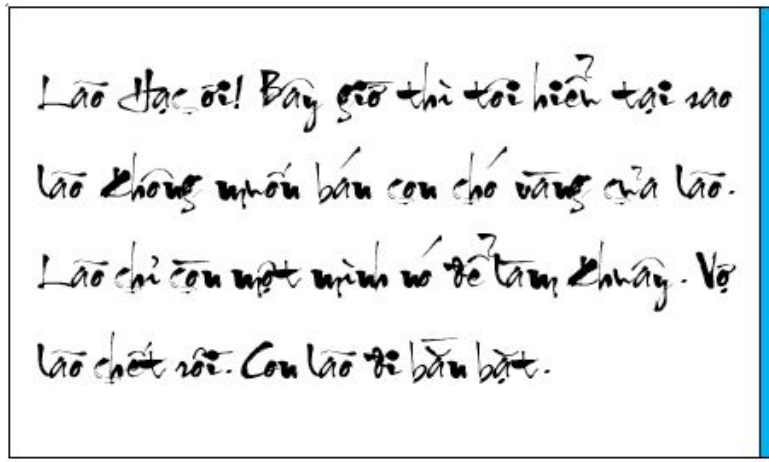
fit@hcmus

Introduction to Readability

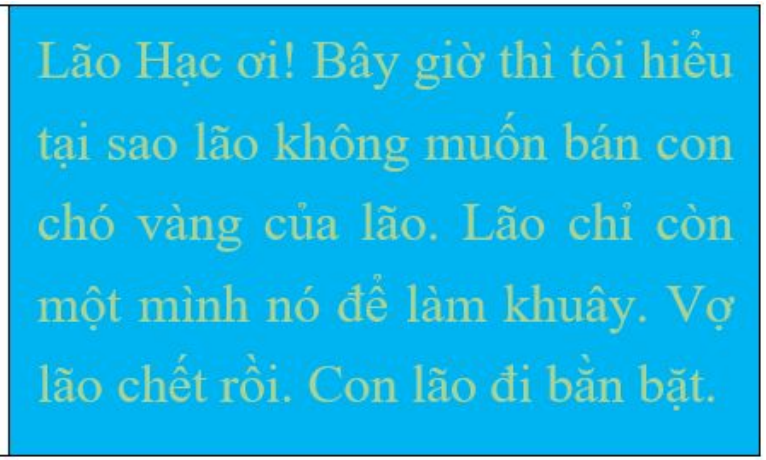


fit@hcmus

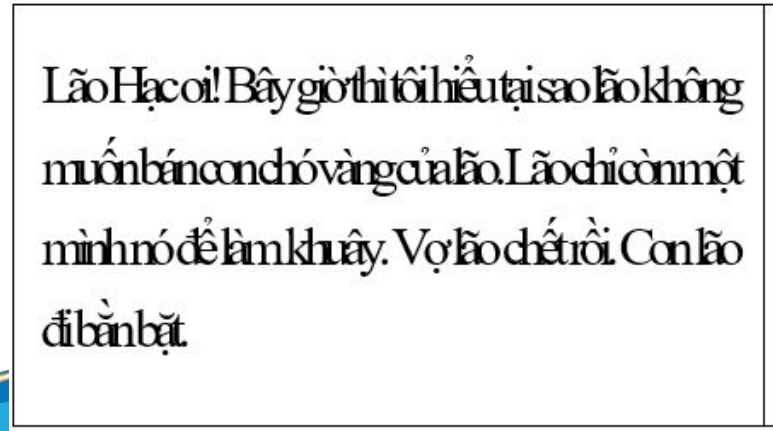
Introduction to Readability



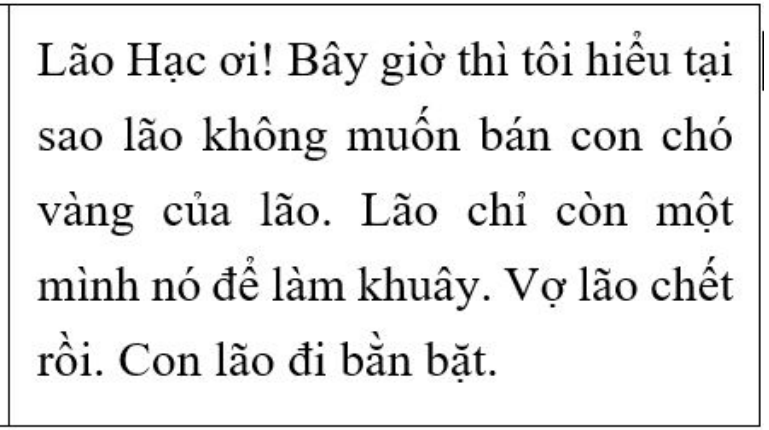
(a)



(b)



(c)



(d)

fit@hcmus

Introduction to Readability

VB1: This news is about two children. They are from Russia. They become *stranded* on a piece of ice. This piece of ice is in a river. People try to save them, but it is not easy. They call a hovercraft. The *hovercraft* saves the children. They are fine.

VB2: Emergency services in Russia came to rescue two children *stranded* on an *ice floe*.

Emergency services dispatched a helicopter and a *hovercraft* to the scene near the eastern city of Khabarovsk after they received a *distress call* about the incident on the Osinovaya River.

Initially, the rescuers were unable to reach the children who stood at the edge of the ice, so the rescuers ended up using the hovercraft to pick them up and bring them to safety. Both children were unharmed. Officials have *issued* a warning that the winter ice on the river had not yet fully formed and was still dangerously thin.

(Nguồn: https://www.newsinlevels.com/)

fit@hcmus

Introduction to Readability

- **VB1**: Đường: chất kết tinh vị ngọt, thường chế từ mía hoặc củ cải đường. (trích Hoàng Phê, "Từ điển Tiếng Việt", 1988)

- **VB2**: Đường: chất có vị ngọt, chế từ mía hoặc củ cải đường. (trích Từ điển Lạc Việt)

fit@hcmus

Introduction to Readability

- Problem statement:

Given a set of documents D = $\{d_1, d_2, \ldots, d_n\}$ and a set of readability measures (classifiers) L = $\{l_1, l_2, \ldots, l_m\}$

Find function f:

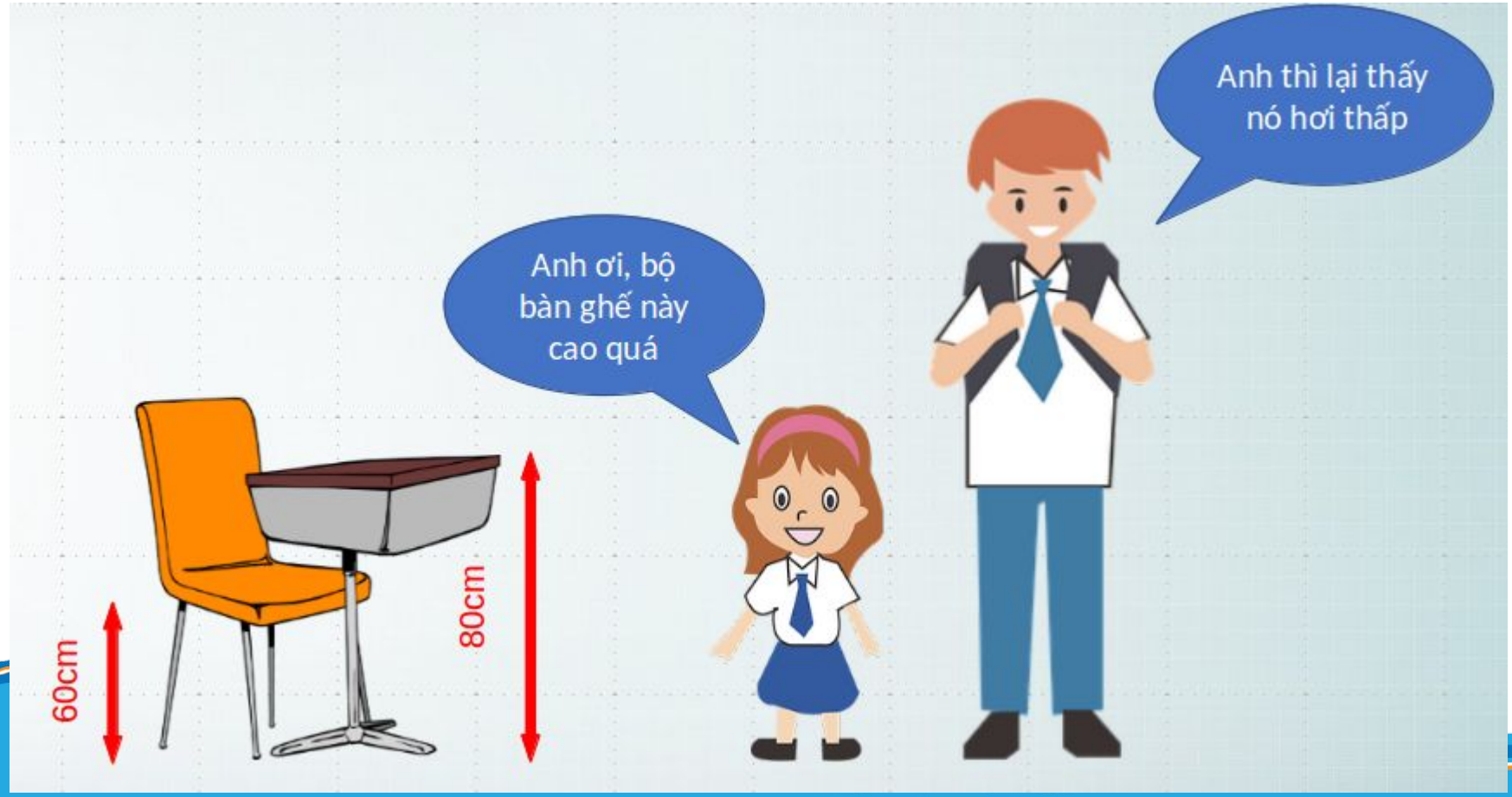
f: D x L -> Boolean

$f(d_i, l_j)$ = true/false

fit@hcmus

Introduction to Readability

- Problem statement:



fit@hcmus

Introduction to Readability

- Applications of Readability:
  - Writing reports
  - Writing textbooks and course materials
  - Publishing
  - Drafting legal documents
  - Writing user manuals
  - Choosing curriculum and language teaching materials for learners
  - ...

fit@hcmus

Introduction to Readability

- Common European Framework of Reference for Languages

		A1	A2	B1	B2	C1	C2
UNDERSTANDING	Listening	I can recognise familiar words and very basic phrases concerning myself, my family and immediate concrete surroundings when people speak slowly and clearly.	I can understand phrases and the highest frequency vocabulary related to areas of most immediate personal relevance (e.g. very basic personal and family information, shopping, local area, employment). I can catch the main point in short, clear, simple messages and announcements.	I can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure, etc. I can understand the main point of many radio or TV programmes on current affairs or topics of personal or professional interest when the delivery is relatively slow and clear.	I can understand extended speech and lectures and follow even complex lines of argument provided the topic is reasonably familiar. I can understand most TV news and current affairs programmes. I can understand the majority of films in standard dialect.	I can understand extended speech even when it is not clearly structured and when relationships are only implied and not signalled explicitly. I can understand television programmes and films without too much effort.	I have no difficulty in understanding any kind of spoken language, whether live or broadcast, even when delivered at fast native speed, provided I have some time to get familiar with the accent.
	Reading	I can understand familiar names, words and very simple sentences, for example on notices and posters or in catalogues.	I can read very short, simple texts. I can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus and timetables and I can understand short simple personal letters.	I can understand texts that consist mainly of high frequency everyday or job-related language. I can understand the description of events, feelings and wishes in personal letters.	I can read articles and reports concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints. I can understand contemporary literary prose.	I can understand long and complex factual and literary texts, appreciating distinctions of style. I can understand specialised articles and longer technical instructions, even when they do not relate to my field.	
SPEAKING	Spoken Interaction	I can interact in a simple way provided the other person is prepared to repeat or rephrase things at a slower rate of speech and help me formulate what I'm trying to say. I can ask and answer simple questions in areas of immediate need or on very familiar topics.	I can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar topics and activities. I can handle very short social exchanges, even though I can't usually understand enough to keep the conversation going myself.	I can deal with most situations likely to arise whilst travelling in an area where the language is spoken. I can enter unprepared into conversation on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events).	I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible. I can take an active part in discussion in familiar contexts, accounting for and sustaining my views.	I can express myself fluently and spontaneously without much obvious searching for expressions. I can use language flexibly and effectively for social and professional purposes. I can formulate ideas and opinions with precision and relate my contribution skilfully to those of other speakers.	
	Spoken Production	I can use simple phrases and sentences to describe where I live and people I know.	I can use a series of phrases and sentences to describe in simple terms my family and other people, living conditions, my educational background and my present or most recent job.	I can connect phrases in a simple way in order to describe experiences and events, my dreams, hopes and ambitions. I can briefly give reasons and explanations for opinions and plans. I can narrate a story or relate the plot of a book or film and describe my reactions.	I can present clear, detailed descriptions on a wide range of subjects related to my field of interest. I can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.	I can present clear, detailed descriptions of complex subjects integrating sub-themes, developing particular points and rounding off with an appropriate conclusion.	
WRITING	Writing	I can write a short, simple postcard, for example sending holiday greetings. I can fill in forms with personal details, for example entering my name, nationality and address on a hotel registration form.	I can write short, simple notes and messages relating to matters in areas of immediate needs. I can write a very simple personal letter, for example thanking someone for something.	I can write simple connected text on topics which are familiar or of personal interest. I can write personal letters describing experiences and impressions.	I can write clear, detailed text on a wide range of subjects related to my interests. I can write an essay or report, passing on information or giving reasons in support of or against a particular point of view. I can write letters highlighting the personal significance of events and experiences.	I can express myself in clear, well-structured text, expressing points of view at some length. I can write about complex subjects in a letter, an essay or a report, underlining what I consider to be the salient issues. I can select style appropriate to the reader in mind.	which helps the recipient to notice and remember significant points. I can write summaries and reviews of professional or literary works.

I can read very short, simple texts. I can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus and timetables and I can understand short simple personal letters.

fit@hcmus

Introduction to Readability

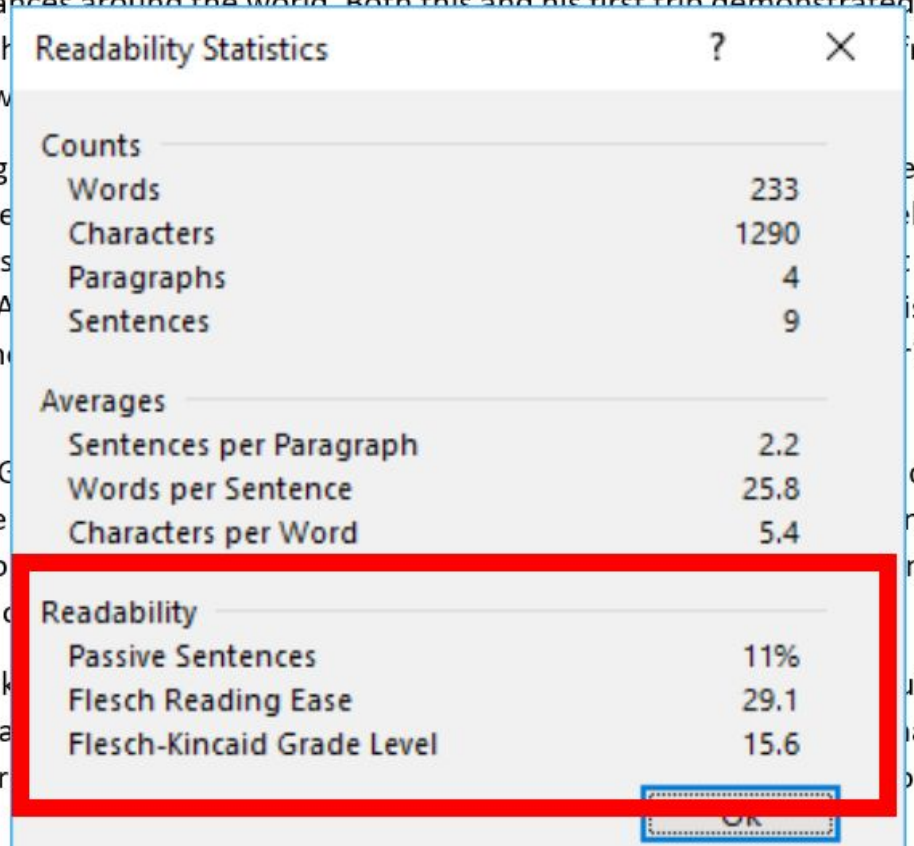
- Evaluating Readability in Microsoft Word

President Trump just concluded a second overseas trip to further advance America's interests and values, and to strengthen our alliances around the world. Both this and his first trip demonstrated the resurgence of American leadersh ront mutual threats and achieve renew

Discussions with world leaders hig ergy, untapped markets that can be ope king the chance to build better futures can form the basis for lasting peace. A ision for securing the American hom rican influence.

Meetings in Poland and at the G ding coalitions to get the best possible nnot be a passive member of internatio mon threats, seize mutually beneficial

In Warsaw, President Trump spok utual support and defense of Poland a at a "strong Poland is a blessing to Eur the world."



Applied NLP - Readability

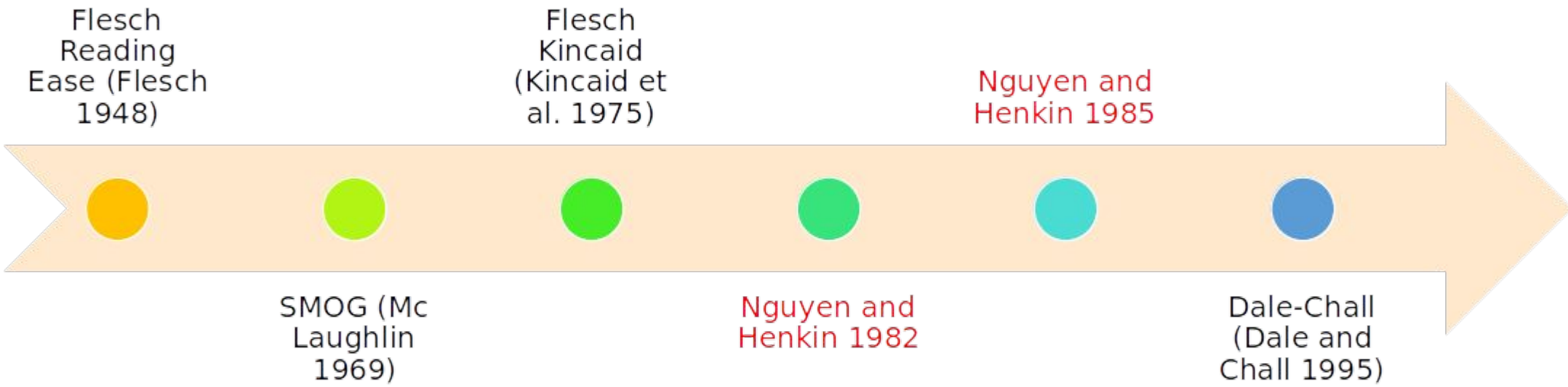
RELATED RESEARCH

Related research

- Statistical approach:
  - Sherman (1893): average sentence length shorten over time
    - Pre-Elizabeth: average 50 words/sentence
    - Elizabeth: avg. 45 w/s
    - Victorian: avg. 29 w/s
    - Sherman: avg. 23 w/s
    - Today: avg. 20 w/s (Dubay 2004)

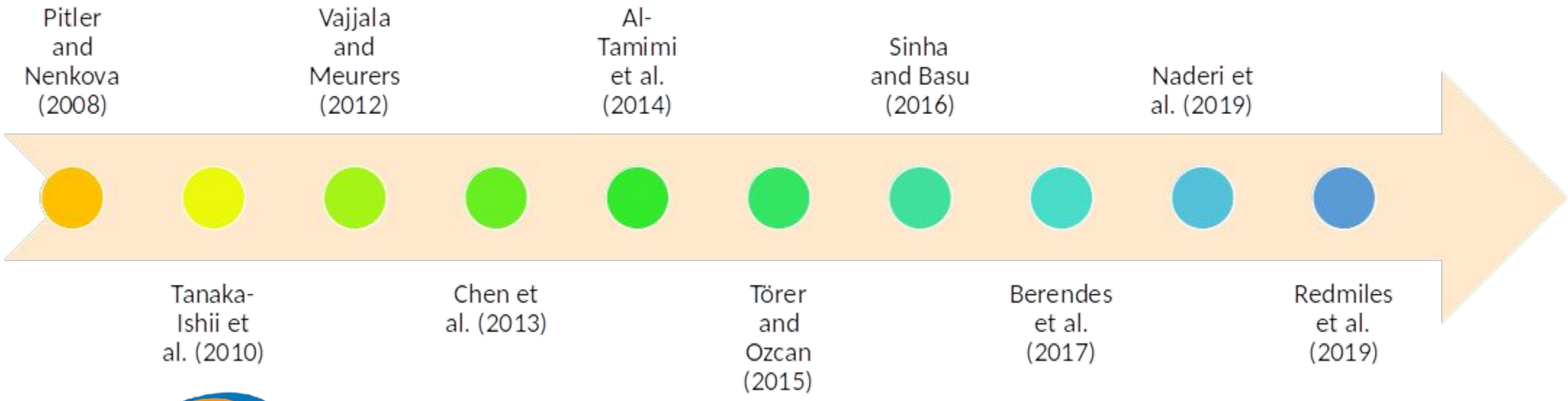
Related research

- Statistical approach: Readability formulas



Related research

- Machine learning approach



fit@hcmus

Applied NLP - Readability

CORPORA

Corpora

- Collecting corpora:
  - 371 reading assignments in Vietnamese (primary school) and Vietnamese Literature (secondary and high school) textbooks
    - 11 grades in 3 school grades
  - 1825 texts belonging to literature and linguistic field
    - 4 readability classifications: Very Easy, Easy, Medium, Hard

fit@hcmus

Applied NLP - Readability

Assessing Vietnamese text Readability

fit@hcmus

Assessing Vietnamese text Readability

- Features: 262 linguistic features

15 đặc trưng bề mặt
14 đặc trưng về tần suất từ và tần suất chữ
150 đặc trưng từ loại
27 đặc trưng cú pháp
20 đặc trưng mô hình ngôn ngữ mức bề mặt
12 đặc trưng mô hình ngôn ngữ mức từ loại
8 đặc trưng đơn giản ở cấp độ ngữ nghĩa
17 đặc trưng riêng của tiếng Việt

Assessing Vietnamese text Readability

- Vietnamese text readability formulas
  - Correlation analysis: analysing features that correlates with readability
  - Regression analysis: analysing feature weights in regression equation

fit@hcmus

Đánh giá độ khó văn bản tiếng Việt

- Correlation analysis:

Đặc trưng bề mặt: tỉ lệ số từ 1 chữ trên tổng số từ	-0,8667
Đặc trưng tần suất: tỉ lệ số từ dễ phân biệt trên tổng số từ	**-0,8823**
Đặc trưng từ loại: tỉ lệ số từ có nhiều từ loại trên tổng số từ	-0,8455
Đặc trưng cú pháp: trung bình độ cao của cây cú pháp	0,8219
Đặc trưng mô hình ngôn ngữ bề mặt: thứ hạng trung bình của các tri-gram mức kí tự	-0,8719
Đặc trưng mô hình ngôn ngữ mức từ loại: tần suất trung bình của các bi-gram mức từ loại	-0,8310
Đặc trưng ngữ nghĩa: tỉ lệ số từ đa nghĩa trên tổng số từ	-0,8398
Đặc trưng của tiếng việt: tỉ lệ số từ mượn phân biệt trên tổng số từ phân biệt	0,8253

fit@hcmus

Đánh giá độ khó văn bản tiếng Việt

- Regression analysis:
  - Experiment 1: all features that correlate highly with readability (r≥0,7) (116 features), no features removed
  - Experiment 2: features that correlate highly with readability, removing features that are highly correlated together
    - Features that correlate highly with readability (r≥0,7) is selected
    - Remove features that correlate highly with eachother (r≥0,8): 4 features

fit@hcmus

Đánh giá độ khó văn bản tiếng Việt

- Evaluating readability using Machine learning approach
  - RFECV algorithm: feature selection method that rank features by fitting a model and removing the weakest feature (or features) recursively
  - Classification and evaluation using SVM classification algorithm
  - K-fold cross-validation

fit@hcmus

Đánh giá độ khó văn bản tiếng Việt

- Mô hình học sâu dựa trên BERT