Lexical Simplification and MWE : how to deal with complexity ?

Núria Gala

Joint AIM-WEST PARSEME-FR Workshop

Grenoble, October 2016

How difficult a text is... and for whom ?

"The question of whether a text is easy to read and understand depends very much on the abilities and experience of the reader." [Saggion et al., 2011]

Readability :

The sum total (including the interactions) of all those elements within a given piece of printed material that affect the success of a group of readers have with it. The success is the extent to which they understand it, read it at a optimal speed, and find it interesting. [Dale and Chall, 1949]

- Litterature ++ for second language acquisition (SLA), but very recent domain of study in NLP : computational readability, text simplification.

Table of Contents

1. Introduction
   - Computational readability
   - Subtask : detecting complex vocabulary
2. Identifying complex words
3. Identifying complexity in MWE
4. Conclusions

Why assess text difficulty ?

- EU recent report (2009) : $19,6\%$ of 15 year old teenagers are "low achievers" in reading
  [De Coster, I. and Baidak, N. and Motiejunaite, A. and Noorani, S., 2011]
- Reading issues can be critical (form for unemployment benefit, drug instruction, living in a foreign country, etc.)
- Assessing (readability prediction) or manipulating text difficulty (automatic text simplification) are seen as :
  - Useful as reading aid systems (improve text accessibility)
  - Helpful for language instructors or readers (improve adaptability of learning).

Computational readability

- Statistical approaches : first formulae, i.e. linear regression with two variables (lexical and syntactic) [Flesch, 1948], [Dale and Chall, 1949]
- Cognitive approaches : integration of cognitive factors, i.e. coherence and cohesion of the text
- Computational approaches :
  - integration of previous paradigms (Statistical and cognitive approaches)
  - automatic extraction of different variables
  - statistical algorithms for text classification (text grades), i.e. Flesch-Kincaid, Gunning-Fog, Coleman-Liau Index, SMOG Index

The output of a readability model

- Readability formulas output a global unique score !
- Example : the Lexile scale “measuring reading ability and the text demand of reading materials” :

Title of work	Lexile
Twilight	720L
Harry Potter and the Sorcerer’s Stone	880L
The Hobbit	1000L

Problems :

- Useful for information retrieval-type applications
- BUT... no information about the passages or lexical forms that are difficult to read or comprehend

Research questions

- What do we mean by 'complex' vocabulary ?
- Is it (only) a matter of frequency and/or length ? *"Size does not matter. Frequency does"* $\longrightarrow$ Frequency is better predictor than length [Wilkens, R. and Dalla Vecchia, A. and Zanon Boito, M. and Padró, M. and Villavicencio, A., 2014]
- Other hypothesis :
    - Form : consistency phoneme-grapheme, morphological structure
    - Meaning : polysemy, compositionnality, abstractness
- Is it possible to automatically identify complex words ? To grade/rank synonyms ?
- Is the methodology employed for single lexical items applicable to MWE ?

Aims of our research

1. Better comprehend the different characteristics that make lexical items difficult to be read or understood (“intrinsic difficulty”)
2. Relate these characteristics with those of a given population (“extrinsic difficulty”)
   $\longrightarrow$ e.g. L1 with developmental disorder (i.e. dyslexia), deafs, low education level, L2, etc.
3. Design models able to automatically predict word difficulty and lexical resources including information on complexity
4. Design reading aids, i.e. automatic text simplification

Joint work with :

- Thomas François (Cental, Univ. catholique de Louvain)
- Delphine Bernhard (LiLPa, Université de Strasbourg)
- Carlos Ramisch, Mokhtar Billami (LIF, Aix Marseille Univ.)

ANR ALECTOR (2017-2020)
Aide à la LECTure pour améliORer l'accès aux documents pour enfants dyslexiques
(Reading Aids to leverage Document Accessibility for Children with Dyslexia)

1 Introduction

2 Identifying complex words
- Assessing Lexical complexity
- Using graded corpora
- Automatically assigning complexity ranks

3 Identifying complexity in MWE

4 Conclusions

Assessing Lexical complexity, related work

Methods to identify complex forms are mainly based on :

- frequencies
- classification
- ranking

The aim is to identify complex lexical items in context and simplify them according to the reader needs.

First approach : complex words are rare words

Using frequencies

- Pioneer study by [Carroll et al., 1999]
$\longrightarrow$ Frequencies of words are looked up in the Kucera-Francis frequency list [Quinlan, 1992]
- A word complexity measure combining word length and word frequency [Biran et al., 2011]

Second approach : classification

Classification methods for complex words

- [Shardlow, 2013] compares a frequency-based approach with a SVM model with 6 variables
  $\longrightarrow$ Frequency, document count, word length, syllable count, sense count, and number of synonyms
  $\longrightarrow$ both methods got a similar F1 value !
- [Gala et al., 2014] used a 24-variable SVM model and reached 63% on 3 classes ($+2\%$ over frequency baseline)
- [Baeza-Yates et al., 2015] added some variables based on spelling patterns to predict complex words for dyslexic children
  $72,3\%$ of accuracy for 2 classes

Third approach : ranking

Approach based on ranking for simplification

- Idea : define a ranking function that can sort a set of synonyms and find the best one to replace a word in context
- [Jauhar and Specia, 2012] use frequencies, syllable count, N-gram model, LSA model, and some psycholinguistics features
  $\longrightarrow \kappa = 0.496$ between predictions and gold-standard ranking
- [Horn et al., 2014] use candidate probability $p(c_i|w)$, word frequency, language models, and context frequency.

Current limitations

- Most approaches work on the form rather than at the sense level
  $\longrightarrow$ no difference between lime ('lemon') and lime ('calcium oxide')
- Most approaches are based only on linguistic characteristics (absolute complexity)
  $\longrightarrow$ generic models : relative complexity or difficulty is overlooked (tailoring to specific users is not foreseen)
- MWE are not taken into account
- As a result, performance remains quite low :
  - [Jauhar and Specia, 2012] : $\kappa = 0,496$
  - [Gala et al., 2014] : accuracy = 63% (3 classes)
  - [Baeza-Yates et al., 2015] : accuracy = 72,3% (2 classes)

Building a lexicon from graded corpora

An alternative to identify complex words : build a graded lexicon from a corpus of texts whose difficulty is known.

- Seminal study by [Lété et al., 2004] : Manulex
  $\longrightarrow$ describes word distributions over 3 grades (primary school)
- [François et al., 2014] adapted the idea for L2 French :
  FLELex $\longrightarrow$ describes word distributions over 6 grades (CEFR scale)

Manulex and FLELex are freely available on the web.

Example of entries in FLELex

lemma	tag	A1	A2	B1	B2	C1	C2	total
voiture (1)	NOM	633.3	598.5	482.7	202.7	271.9	25.9	461.5
abandonner (2)	VER	35.5	62.3	104.8	79.8	73.6	28.5	78.2
justice (3)	NOM	3.9	17.3	79.1	13.2	106.3	72.9	48.1
kilo (4)	NOM	40.3	29.9	10.2	0	1.6	0	19.8
logique (5)	NOM	0	0	6.8	18.6	36.3	9.6	9.9
en bas (6)	ADV	34.9	28.5	13	32.8	1.6	0	24
en clair (7)	ADV	0	0	0	0	8.2	19.5	1.2
sous réserve de (8)	PREP	0	0	0.361	0	0	0	0.03

FLELex-TT has 14,236 entries (no MWEs, but manually cleaned)

FLELex-CRF includes 17,871 entries (MWEs, but not cleaned yet)

Both resources are freely available at
`http://cental.uclouvain.be/flelex/`

Building a graded lexicon with synonyms

Another alternative to identify complex words : automatically assign a rank within a synset of synonyms. Methodology :

1. Get a list of words with difficulty annotations
2. Train a ranking model based on intrinsic linguistic characteristics
   $\longrightarrow$ Pairwise ranking algorithm : SVMRank [Herbrich et al., 2000]
3. Test the model's prediction against human opinions
4. Get a resource of synonyms and apply the model on it
5. At the end, create a resource (ReSyf) in which every set of synonyms is ranked according to reading difficulty (lexical complexity)
   `http://resyf.lif.univ-mrs.fr/ResyfApplication/index.htm`

Features related to lexical difficulty

Features based on the orthographic form [Gala et al., 2014] :

- Letter count, phonemes count, syllable count
- Density and frequency of the orthographical neighbours of the target word
- Syllabic structure (3 classes of frequency)
- Consistancy between the written and oral form :
  - 0 = transparency : 'abruti' `[abRyti]`
  - $<$ 2 characters : 'abrite*r*' `[abRite]`
  - $>$ 2 characters : 'l*e*nte*m*e*nt*' `[l@tm@]`
- Spelling patterns : double vowels (e.g. ée `[e]`), double consonant (e.g. pp `[p]`), digraphes (e.g. ch `[S]`)

Features related to lexical difficulty (2)

Morphological information :

- Morpheme count, prefixation (yes/no), suffixation (yes/no), is compound (yes/no), minimal freq. of the pref./suf., mean freq. of the pref./suf., size of the morphological family
- New variables : most frequent word in the family, mean freq. of the words in the family, etc.

The morphological decomposition, an unsupervised analysis

- Decomposition into labelled segments (base, prefix, suffix)
- Then, identification of the family [Bernhard, 2010]
- Examples :

  rouille – antirouille ; rouilleux
  dérouiller – dérouillage ; dérouillement ;
  débrouille – brouilleur ; brouilleuse ; débrouilleur ; débrouilleuse

  brouille – brouillerie ; brouilleux

Features related to lexical difficulty (3)

Other features :

- Polysemy :
  - Binary variable indicating whether the word is considered as polysemous in JeuxDeMots [Lafourcade, 2007]
  - Number of synsets in BabelNet [Navigli and Ponzetto, 2010]
- Frequencies :
  - Logarithm of the word frequency in Lexique3 [New et al., 2007]
  - Presence or absence from the Gougenheim's list [Gougenheim et al., 1964]

More significant predictors

1. Nb phonemes, letters, syllables
2. Polysemy
3. Nb orthographical neighbours
4. Nasal vowels
5. Morphological family size
6. Prefixation
7. Nb morphemes
8. Orthographical patterns (double vowels/cons., digraphs)

Algorithm for the pair creation

Using the variables in a ranking algorithm [François, T. and Billami M. B. and Gala, N. and Bernhard, D., 2016]

$\mathbf{v_i} = (f_1, \cdots, f_n)$

$\mathbf{v_j} = (f_1, \cdots, f_n)$

$l_i = 3$

abolition

annulation

$l_j = 1$

abolition vs. annulation

$l_{ij} = sgn(l_i - l_j)$

$\mathbf{v_{ij}} = (f_1, \cdots, f_n) - (f_1, \cdots, f_n)$

If abolition is more difficult than annulation, $l_{ij} = 1$.

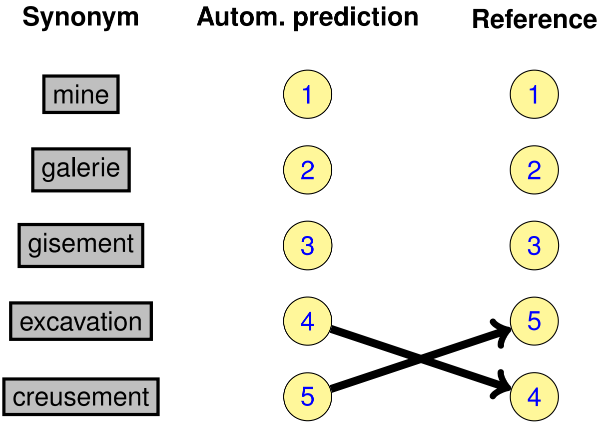
Assessing the model with human judges

- External evaluation : comparing the model predictions with ratings of human judges.
- Data : 40 vectors of synonyms (3.5 words in average) were assessed by 40 judges.
- Average agreement between judges : Krippendorff's $\alpha = 0,4$.
- Ratings vs. predictions : Cohen's quadratic $k = 0,63$ (strong agreement).
- MRR (mean reciprocal rank) = 0,84.

Human evaluation : example

Synonym	Autom. prediction	Reference
associer	1	1
combiner	2	2
assimiler	3	3
entremêler	4	4
amalgamer	5	5

Human evaluation : example



ReSyF

Sens 3:		
	bleu	1
	sombre	2
	morne	3
	déprimant	4
Sens 4:		
	bleu	1
	contusion	2
	ecchymose	3
	meurtrissure	4
Sens 5:		
	bleu	1
	fromage à pâte persillée	2

ReSyf

1. Introduction
2. Identifying complex words
3. Identifying complexity in MWE
   - Related work, psycholinguistic criteria
   - Annotation Campaign
4. Conclusions

Starting point

Are the criteria used for ranking single words applicable to MWE ?
Right now : average of variables of words in a MWE...

Our Aims :

- To identify the criteria that make a MWE complex to read and/or understand, eg. :

  temps mort *vs* pause/arrêt (time out)
  gros mot *vs* injure/juron (swear word)

- To go beyond the notion of frequency by exploring new hints :
  1. compositionnality
  2. opaque/less opaque semantics
  3. abstractness/concreteness

What has been done related to abstractness ?

... not too much

- Some works in psycholinguistics (concreteness and imageability)
- “most of the concrete words are considered imageable, whereas abstract words show higher variability in imageability ratings” [Barber, 2013]
- “Imageability is a semantic variable that measures how easy it is for a word to arouse mental images (...) Imageability is significantly correlated with AoA, familiarity, length, and N, but less clearly with word frequency. Thus, more imageable words tend to be acquired earlier, are more familiar, tend to be shorter, and tend to have more orthographic neighbours than less imageable words.” [Stadthagen-Gonzalez and Davis, 2006]
- $\longrightarrow$ Bristol Norms (1526 English words)

What has been done related to abstractness ?

... not too much

- Concrete and Abstract words are represented, processed and retrieved differently on the brain [Ferré et al., 2015]
- Qualitative difference as regards to the representation in memory
  - Concrete concepts organized in terms of semantic similarity
  - Abstract concepts organized by their association with other concepts

$\longrightarrow$ Are abstract words more complex ? Is abstractness a predictor of complexity ? What about in MWE ?

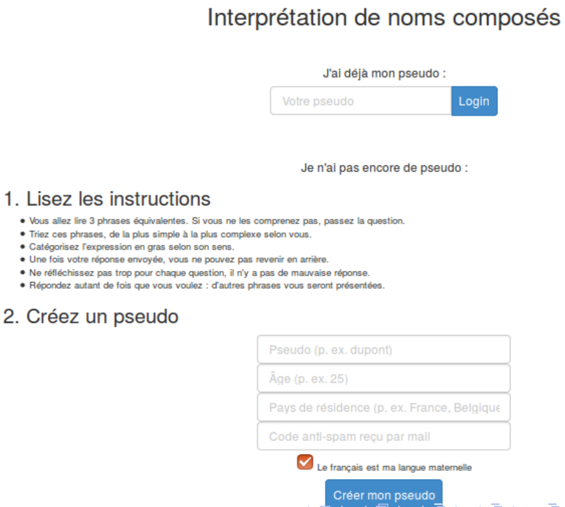
Annotation Campaign

`http://www.inf.ufrgs.br/mwe/simple_fr/`

Code anti-spam : bonnepommebonnepoire

- Ranking synonyms (including MWE)
- Categorizing among Concrete and Abstract concepts
- Choosing a sub-type (object, person, place... *vs* idea, feeling, time...)

Annotation campaign (interface)



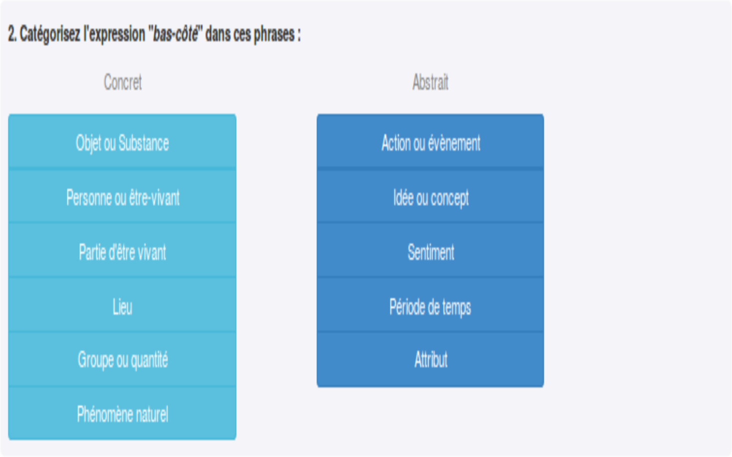
MWE ranking

1. Triez les phrases ci-dessous de la plus simple à la plus complexe :

	+SIMPLE	↔	−SIMPLE
Les véhicules défilent sur le ***bas-côté****.*			
*Les véhicules défilent sur l'****accotement****.*			
Les véhicules défilent sur la ***bordure****.*			

Je ne comprends pas les phrases → Passer

MWE abstractness



(Brand new) results

- 912 annotations in ten days (we aim to obtain n thousands)
- 51 anonymous participants (mastering French, living in a French-speaking country)
- 180 synsets to be annotated (3 items each, including at least 1 MWE)
- About 500 lexical items to categorize (180 x 3 excluding duplicated entries)
- 34 synsets to be compared with ReSyf (where 26 MWE have been ranked in first position)

... work in progress !

Conclusions

- Collaborative work on readability (identifying lexical complexity)
- Aiming to improve the knowledge we have about words in order to create text simplification systems (reading aids)
- General observations to be adapted to specific readers with specific needs
- Work to be done on MWE !

Future work

- Analyze correlations of abstractness / complexity
- Include new measures to better analyze complexity in MWE
- Enrich the lexical resource ReSyf with a better ranking of MWE
- ...

Thanks

Núria Gala
nuria.gala@univ-amu.fr

LABORATOIRE D'INFORMATIQUE FONDAMENTALE de Marseille

Aix Marseille université

BLRI Brain & Language Research Institute

Baeza-Yates, R., Mayo-Casademont, M., and Rello, L. (2015).

Feasibility of word difficulty prediction.

In *International Symposium on String Processing and Information Retrieval*, pages 362–373. Springer.

Barber (2013).

Concreteness in word processing : ERP and behavioral effects in a lexical decision task.

Brain Language, 125 :47–53.

Bernhard, D. (2010).

Apprentissage non supervisé de familles morphologisques : Comparaison de méthodes et aspects multilingues.

Traitement Automatique des Langues, 51(2) :11–39.

Biran, O., Brody, S., and Elhadad, N. (2011).

Putting it simply : a context-aware approach to lexical simplification.

In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers-Volume 2*, pages 496–501. Association for Computational Linguistics.

Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., and Tait, J. (1999).

Simplifying text for language-impaired readers.

In *Proceedings of EACL 1999*, pages 269–270.

Dale, E. and Chall, J. (1949).

The concept of readability.

Elementary English, 26(1) :19–26.

De Coster, I. and Baidak, N. and Motiejunaite, A. and Noorani, S. (2011).
Teaching Readind in Europe : COntexts, Policies and Practices.
Education, Audiovisual and CUlture Executive Agency.

Ferré, P., Guasch, M., García-Chico, T., and Sánchez-Casas, R. (2015).
Are there qualitative differences in the representation of abstract and concrete words ?
Quarterly Journ. of Experimental Psychology, 68(12) :2402–2418.

Flesch, R. (1948).
A new readability yardstick.
Journal of Applied Psychology, 32(3) :221–233.

François, T., Gala, N., Watrin, P., and Fairon, C. (2014).
FLELex : a graded lexical resource for French foreign learners.
In *Proceedings of International conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik.

François, T. and Billami M. B. and Gala, N. and Bernhard, D. (2016).
Bleu, contusion, ecchymose : tri automatique de synonymes en fonction de leur difficulté.
In *Actes de la conférence Traitement Automatique des Langues Naturelles*, Paris.

Gala, N., François, T., Bernhard, D., and Fairon, C. (2014).
Un modèle pour prédire la complexité lexicale et graduer les mots.
In *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2014)*, pages 91–102.

Gougenheim, G., Michéa, R., Rivenc, P., and Sauvageot, A. (1964).
L'élaboration du français fondamental (1er degré).
Didier, Paris.

Herbrich, R., Graepel, T., and Obermayer, K. (2000).
Large margin rank boundaries for ordinal regression.
chapter 7, pages 115–132. MIT Press, Cambridge.

Horn, C., Manduca, C., and Kauchak, D. (2014).
Learning a lexical simplifier using wikipedia.
In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 458–463.

Jauhar, S. K. and Specia, L. (2012).
Uow-shef : Simplex–lexical simplicity ranking based on contextual and psycholinguistic features.
In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 477–481. Association for Computational Linguistics.

Lafourcade, M. (2007).
Making people play for lexical acquisition with the jeuxdemots prototype.
In *SNLP'07 : 7th international symposium on natural language processing.*

Lété, B., Sprenger-Charolles, L., and Colé, P. (2004).
Manulex : A grade-level lexical database from French elementary-school readers.
Behavior Research Methods, Instruments and Computers, 36 :156–166.

Navigli, R. and Ponzetto, S. P. (2010).
Babelnet : Building a very large multilingual semantic network.
In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225.

New, B., Brysbaert, M., Veronis, J., and Pallier, C. (2007).
The use of film subtitles to estimate word frequencies.
Applied Psycholinguistics, 28(04) :661–677.

Quinlan, P. (1992).
The Oxford psycholinguistic database.
University Press.

Saggion, H., Gomez-Martin, E., Anula, A., Bourg, L., and Etayo, E. (2011).
Text simplification in SImplext : Making texts more accessible.
Procesamiento del Lenguaje Natural (SEPLN), 46.

Shardlow, M. (2013).
A comparison of techniques to automatically identify complex words.
In *ACL Student Research Workshop*, pages 103–109.

Stadthagen-Gonzalez, H. and Davis, C. J. (2006).
The bristol norms for age of acquisition, imageability, and familiarity.
Behavioural Reseach Methods, 38(4) :598–605.

Wilkens, R. and Dalla Vecchia, A. and Zanon Boito, M. and Padró, M. and Villavicencio, A. (2014).
Size Does Not Matter. Frequency Does. A Study of Features for Measuring Lexical Complexity.
Advances in Artificial Intelligence, Lecture Notes in Computer Science, 8864 :129–140.