Graded lexicons: new resources for educational purposes and much more

Núria Gala[1], Mokhtar B. Billami[2],
Thomas François[3], and Delphine Bernhard[4]

Abstract. Computational tools and resources play an important role for vocabulary acquisition. Although a large variety of dictionaries and learning games are available, few resources provide information about the complexity of a word, either for learning or for comprehension. The idea here is to use frequency counts combined with intra-lexical variables to account for the difficulty of a word. By using such predictors, we have built a lexical resource for French, ReSyf, where words are available with their word senses and their synonyms according to a complexity level. In this paper, we will present the methodology used to build the resource and we will discuss possible applications, from enhancing vocabulary acquisition to text simplification.

Keywords: iCALL, vocabulary acquisition, graded synonym lexicon, lexical complexity, text simplification.

1. Introduction

Computational tools and resources play an important role for vocabulary acquisition. Encouraged by the extensive use of mobile devices, recent intelligent Computer-Assisted Language Learning (iCALL) applications and platforms propose a large variety of learning games that offer challenging possibilities (see

[1] Aix Marseille Université & LIF-CNRS, Marseille, France; nuria.gala@lif.univ-mrs.fr

[2] Aix Marseille Université & LIF-CNRS, Marseille, France; mokhtar.billami@lif.univ-mrs.fr

[3] CENTAL Université Catholique de Louvain, Louvain-la-Neuve, Belgium; thomas.francois@uclouvain.be

[4] LiLPa, Université de Strasbourg, Strasbourg, France; dbernhard@unistra.fr

How to cite this article: Gala, N., Billami, M. B., François, T., & Bernhard, D. (2015). Graded lexicons: new resources for educational purposes and much more. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds), *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 204-209). Dublin: Research-publishing.net. http://dx.doi.org/10.14705/rpnet.2015.000334

Cornillie, Thorne, & Desmet, 2012) compared to more traditional exercises which emphasize repetition (explicit learning[5]). Recent educational tools are These features were on modern pedagogical criteria, offering among other things hyperlinks to electronic dictionaries or concordancers. The information a student can find is related to word forms (morphology), word meanings (semantics, usage) and word patterns (syntax, collocations). These electronic resources may even offer information concerning the origins of the word, its particular usage (constructions), and typically related words (semantically or thematically, word-families).

However, very few tools provide information about the complexity of a word, either for learning or for comprehension (showing for example that 'monster' is a simpler term than its hyponyms 'phoenix' or 'behemoth', or that 'to walk' is easier than its synonyms 'to stroll' or 'to ramble'). Yet, the idea of using frequency counts as a proxy for word difficulty is not new: frequency word lists were built in the past, see for instance Thorndike (1921) and Gougenheim (1958).

The principle behind these lists is that the frequency of a word affects its recognition, thus its acquisition. Based on that principle, which basically relies on corpus-based features, some resources have been built where words are classified across difficulty levels: the English Profile Wordlists (Capel, 2010), or for French, Manulex (Lété, Sprenger-Charolles, & Colé, 2004) and FLELex (François, Gala, Watrin, & Fairon, 2014).

In this paper we introduce a new graded lexical resource for French synonyms, ReSyf, that relies on a large set of intra-lexical and psycholinguistic features to assign a grade level to each synonym. The resource was first introduced in Gala, François, Bernhard, and Fairon (2013), but the current version has been enhanced, especially as regards the possibility to discriminate between synsets of synonyms.

2. Methodology to build ReSyf

ReSyf[6] is a lexical database with graded word-senses containing 31,141 entries, extracted, filtered and annotated using different resources available for French. A

[5] Explicit and implicit learning depends on the users' attention paid to the words (Ma & Kelly, 2006); exercises specifically focused on vocabulary (explicit learning) or activities where lexical acquisition rather occurs implicitly, as a side-effect: the student is repeatedly exposed to words, like in reading.

[6] Retrieved from http://resyf.lif.univ-mrs.fr

predictive model based on lexical and psycholinguistic features related to lexical complexity (Gala et al., 2014) was used to assign grade levels to the entries and synonyms.

2.1. From words to word-senses

The entries were extracted from the list of concepts[7] in French from BabelNet 2.5.1 (Navigli & Ponzetto, 2012). As defined by Navigli and Ponzetto (2012), a "concept in [BabelNet] is represented as a synonym set (called *synset*)[: a] set of words that share the same meaning. For instance, the concept of play as a dramatic work is expressed by the following synset" *drame*, *jeu dramatique*, *pièce (théâtre)*, *pièce de théâtre*, *texte dramatique*, *œuvre dramatique* (p. 218). Each concept was extracted with a list of associated weighted synsets. The weight of a synset corresponds to its semantic connections in BabelNet.

The list obtained at this stage was filtered with two French reference resources: Lexique 3 (New, Pallier, Ferrand, & Matos, 2001) and the Trésor de la Langue Française informatisé (TLFi) (Dendien & Pierrel, 2003). While BabelNet provided an important amount of data, the reference resources were used to validate the lexical items and to remove wrong items (words in languages other than French, orthographic errors, rare or domain-specific terms, etc.). The monosemic words without information from BabelNet were enriched with synonyms extracted from JeuxDeMots (Lafourcade, 2007).

All the words obtained were tagged with TreeTagger[8] (Schmid, 1994). As a result of the Part Of Speech (POS) tagging, we removed wrong lemmas and verified that all the synonyms of the synset had the same POS tag as the target word. When the concept appeared as a multiword expression or collocation, we kept the POS tag of the first item (i.e. noun for *texte dramatique*). Finally, when a concept appeared with a hypernym (i.e. *pièce – théâtre*) we verified that the hypernym was included in the domain list of JeuxDeMots[9] and we kept the hypernym as a synonym. The results obtained are reported in Table 1, while Table 2 presents the number of final words containing at least one synonym.

[7] We are interested in the selection of the concepts without taking into account the presence of the named entities. We consider a concept as being a sense in the lexical network BabelNet (Navigli & Ponzetto, 2012).

[8] TreeTagger, morphosyntactic annotation tool, retrieved from http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[9] There are 3,086 different domains in JeuxDeMots, version 05/2015.

Table 1. Number of target words in ReSyf filtered with Lexique 3 and TLFi

POS	*One sense*	*Polysemic*	*Total*
Nouns	15 482	16 825	32 307
Verbs	1 003	1 890	2 893
Adjectives	2 620	2 032	4 652
Adverbs	451	595	1 046
Total	19 556	21 342	40 898

Table 2. Number of target words in ReSyf for which each sense contains at least one synonym

Nouns	*Verbs*	*Adjectives*	*Adverbs*	*Total*
24 599	2 612	3 094	836	**31 141**

2.2. Word-senses graduation with a level of difficulty

The difficulty of the entries and the synonyms in ReSyf was established as the result of a twofold process. First, we gathered a “gold standard” list of about 19,000 frequent words in French having difficulty level information. This resource was obtained from Manulex by Gala et al. (2014), who transformed the frequency distribution across the three levels in Manulex into a single grade level.

In a second step, we trained a statistical model on this gold standard, using 49 intra-lexical and psycholinguistic features. The set of features is further detailed in Gala et al. (2014) and includes the number of letters, the number of phonemes, the number of syllables, the presence of specific spelling patterns, the syllabic structure, morphemic information, the number of meanings, frequency of the word in Lexique3, etc. These features were combined within a linear Support Vector Machine (SVM) model with L2 regularization[10]. The final model reached an accuracy of 63%, which was 2% better than the baseline relying only on frequency as a predictor, but still was not a satisfactory result for our purpose. As

[10] The only metaparameter of this model is the cost, which was set to 0,5 as the result of a grid search between values of 100 and 0,001.

a consequence, in the building of ReSyf, our model was used only to assign grade level to words absent from Manulex. Otherwise, the level assigned was directly obtained from this resource.

Table 3 presents an example of an entry: the word *jeu* (game) with its POS tag and its grade. The list of graded synonyms has also a weight according to the relevance of the sense.

Table 3. Example of a ReSyf entry with different lists of synonyms

Target word	*Graded synonyms*
jeu_N_1	[partie_N_1, catch_N_3]+1425 [jeu de hasard_N_1, pari_N_1]+918 [drame_N_2, jeu dramatique_N_2, pièce (théâtre)_N_1, pièce de théâtre_N_1, texte dramatique_N_2, œuvre dramatique_N_3]+392 [casse-tête_N_3, puzzle_N_1]+392 [blague_N_1, bouffonnerie_N_1, farce_N_1, plaisanterie_N_1, tour_N_1]+345

3. Conclusion

In this paper, we have presented a graded lexicon for French synonyms where words account for a level of complexity calculated from frequency counts, intra-lexical and psycholinguistic features. While aimed at text simplification, this graded lexicon can also help learners of French to acquire vocabulary and to improve language acquisition. On the one hand, the lexicon itself can be used for explicit learning of French vocabulary guided by the different grades of the synonyms of a word. On the other hand, it can be used to carry out word substitution within an automatic text simplification system aiming at helping learners and children with reading impairments to get through a text, rediscovering the pleasure of reading (as they can better understand what they read), and thus entering a virtuous circle, whereby reading and decoding skills are trained through reading practice.

References

Capel, A. (2010). Insights and issues arising from the English Profile Wordlists project. *Research Notes, 41*, 2-7. Cambridge: Cambridge ESOL.

Cornillie, F., Thorne, S. L., & Desmet, P. (2012). Digital games for language learning: from hype to insight? *ReCALL*, 24(3), 243-256. doi:10.1017/S0958344012000134

Dendien, J., & Pierrel, J.-M. (2003). Le Trésor de la langue française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatique des Langues, 44*(2), 11-37.

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

Gala, N., François T., Bernhard, D., & Fairon, C. (2014). Un modèle pour prédire la complexité lexicale et graduer les mots. *Actes de Traitement Automatique des Langues Naturelles (TALN 2014), Marseille*.

Gougenheim, G. (1958). *Dictionnaire fondamental de la langue française*. Paris: Didier.

Lafourcade, M. (2007). Making people play for lexical acquisition. In *Proc. SNLP 2007, 7th Symposium on Natural Language Processing, Pattaya, Thailande*.

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). Manulex: a grade-level lexical database from French elementary-school readers. *Behavior Research Methods, Instruments & Computers, 36*(1), 156-166. doi:10.3758/BF03195560

Ma, Q. & Kelly, P. (2006). Computer assisted vocabulary learning: design and evaluation. *Computer Assisted Language Learning, 19*(1), 15-45. doi:10.1080/09588220600803998

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence, 193,* 217-250. doi:10.1016/j.artint.2012.07.001

New, G. A., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur Internet : lexique 3. *L'année psychologique, 101*, 447-462. doi:10.3406/psy.2001.1341

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing, Manchester, UK*.

Thorndike, E. (1921). *The teacher's word book*. New York: Columbia University.

Research-publishing.net

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; info@research-publishing.net

Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy
Edited by Francesca Helm, Linda Bradley, Marta Guarda, and Sylvie Thouësny

Disclaimer: Research-publishing.net does not take any responsibility for the content of the pages written by the authors of this book. The authors have recognised that the work described was not published before, or that it is not under consideration for publication elsewhere. While the information in this book are believed to be true and accurate on the date of its going to press, neither the editorial team, nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, expressed or implied, with respect to the material contained herein. While Research-publishing.net is committed to publishing works of integrity, the words are the authors' alone.

Trademark notice: product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Copyrighted material: every effort has been made by the editorial team to trace copyright holders and to obtain their permission for the use of copyrighted material in this book. In the event of errors or omissions, please notify the publisher of any corrections that will need to be incorporated in future editions of this book.

Typeset by Research-publishing.net
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-28-5 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

ISBN13: 978-1-908416-29-2 (Ebook, PDF, colour)
ISBN13: 978-1-908416-30-8 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2015.