Enhancing grammatical structures in web-based texts

Leonardo Zilio[1], Rodrigo Wilkens[2], and Cédrick Fairon[3]

Abstract. Presentation of raw text to language learners is not enough to ensure learning. Thus, we present the Smart and Immersive Language Learning Environment (SMILLE), a system that uses Natural Language Processing (NLP) for enhancing grammatical information in texts chosen by a given user. The enhancements, carried out by means of text highlighting, are designed to draw the users' attention to specific grammatical structures and thus help them to notice their occurrence in authentic contexts. To assess the quality of the enhancements, we carried out an evaluation of 48 structures in terms of precision in different text genres. This diversity approximates the contexts in which a language learner should immerge.

Keywords: NLP, SLA, input enhancements, syntactical highlighting, SMILLE.

1. Introduction

Computer-Assisted Language Learning (CALL) systems have recently started to use NLP applications for aiding in reading activities (Azab et al., 2013). Those systems base their approach to Second Language Acquisition (SLA) on the findings that the presentation of raw input to a language learner is not enough for ensuring that something will be learned (Meurers et al., 2010). So, the learner may not notice the grammatical content that is present in a text and, therefore, not convert the input into intake, as stated by Schmidt (1990, 2012). To address the lack of salience of information in input, the notion of input enhancements was created (Smith, 1993; Smith & Truscott, 2014).

Among the CALL systems that use NLP for identifying relevant SLA information in texts, the Smartreader (Azab et al., 2013), the FLAIR (Chinkina & Meurers, 2016), and the WERTi (Meurers et al., 2010) systems employ syntactic highlighting

1. Université Catholique de Louvain, Louvain-la-Neuve, Belgium; leonardo.zilio@uclouvain.be
2. Université Catholique de Louvain, Louvain-la-Neuve, Belgium; rodrigo.wilkens@uclouvain.be
3. Université Catholique de Louvain, Louvain-la-Neuve, Belgium; cedrick.fairon@uclouvain.be

How to cite this article: Zilio, L., Wilkens, R., & Fairon, C. (2017). Enhancing grammatical structures in web-based texts. In K. Borthwick, L. Bradley & S. Thouësny (Eds), *CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017* (pp. 345-350). Research-publishing.net. https://doi.org/10.14705/rpnet.2017.eurocall2017.738

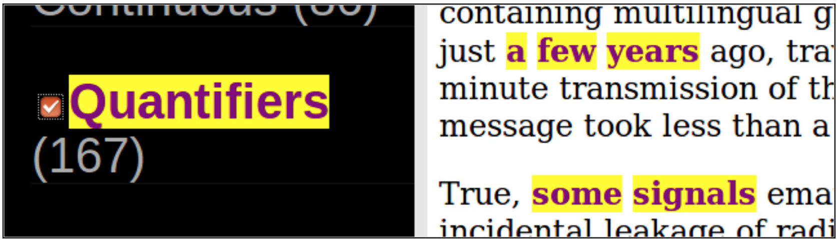
as a means of enhancing raw texts. All of them preprocess texts using the Stanford parser (Manning et al., 2014) and then apply rules with different granularities to get grammatical information and present them to the language learner[4].

This paper presents SMILLE, a system that automatically enhances grammatical structures in English texts chosen by the user. Since the detection of pedagogically relevant grammatical structures cannot be only based on parser information, we developed rules to cover them. In this study, our main focus was to evaluate the precision of these rules in different genres.

2. SMILLE

Using a similar approach to the systems already presented, SMILLE[5] uses input enhancements to draw the reader's attention to specific language structures. The user is free to choose any web-based text that will then be processed with Stanford parser and submitted to a rule-based processing that lists the existing grammatical structures for the language learner, who can choose on the fly which structures are to be enhanced. The enhancements are made by means of color-coding, highlighting, and boldface formatting (based on Simard, 2009), as illustrated in Figure 1. The system can enhance various types of grammatical structures that are based on Common European Framework of Reference for languages (CEFR) recommendations and are pedagogically organized according to Altissia's English curriculum (www.altissia.com). The system also provides access to grammar explanations, which are automatically linked to Altissia's course, and to word definitions from online dictionaries (e.g. Merriam-Webster's dictionary, at https://www.merriam-webster.com).

Figure 1. Example of highlighted quantifiers



4. A comparison between SMILLE and other systems is presented in Zilio, Wilkens, and Fairon (2017).
5. For a more complete description of the system, see Zilio and Fairon (2017) and Zilio et al. (2017).

3. Methodology

To assess the system's reliability in showing information to the user, a precision evaluation was conducted with 48 grammatical structures that do not rely solely on parser's information for being detected, requiring complex rules.

For that purpose, we selected four corpora of differing genres (Table 1): BBC: complete news articles from the BBC (2004-2005) corresponding to stories in five topics (entertainment, sports, business, politics, and technology) (Greene & Cunningham, 2005); GUT: selection of books from Project Gutenberg covering different literary genres (The Turn of the Screw, Wastralls, The Picture of Dorian Gray, The Phantom of the Opera, The Certain Hour, Greenmantle, Corpus, The Lair of the White Worm, Animal Ghosts, and The Shunned House); MOV: Cornell Movies, a collection of fictional conversations extracted from 617 raw movie scripts (Danescu-Niculescu-Mizil & Lee, 2011); and SCI: corpus of scientific papers (Jaidka, Chandrasekaran, Rustagi, & Kan, 2016).

Table 1. Description and average precision per corpus

Corpora	Tokens	Types	Sentences	Documents	Average Precision
BBC	978k	38k	41k	2,225	78%
GUT	826k	31k	41k	10	79%
MOV	4,246k	72k	481k	617	81%
SCI	660k	46k	29k	219	72%

All corpora were annotated with SMILLE and then we extracted samples of 25 random instances[6] for each corpus and for the 48 grammatical structures, and evaluated them in terms of precision. This evaluation was carried out by one language specialist[7].

4. System evaluation

The system achieved an overall average precision of 81%, but the median was 91%, indicating that most of the structures (67%) actually scored above the average. After removing outliers (beyond two standard deviations), the actual overall average is 85%. The similar average scores per corpus, as shown in Table 1, also hide the differences among the corpora and the distribution of the phenomena. For

6. If the corpus did not present 25 instances of a given structure, all of them were evaluated.
7. For reasons of space constraints, the complete table of results for all 48 grammatical structures is presented in https://goo.gl/CybVPE.

instance, BBC presented no occurrences for 1/3 of the structures and SCI had less than 10 occurrences for 1/4 of the structures, while GUT and MOV presented at least 25 instances for most of the structures.

Considering the individual structures, we saw that six structures scored below 50% precision, while 26 of them scored above 90%. Some of them had a very low precision score in all corpora, like the connectives of purpose (average 20%), and the connectives of reason and result (average 22%), while others had influence of the genre, like the ellipsed infinitive (0% in SCI, 17% in GUT, and 64% in MOV).

These differences between corpora arise from the preference of distinct forms related with the same grammar structure, as discussed by Roland, Dick, and Elman (2007). This means, for example, that the distribution of the ellipsed infinitive present in SCI are different to those used in MOV.

5. Conclusions

The evaluation of SMILLE showed us where we need to focus our attention for improving the system's performance. While a few of the grammatical structures present low precision scores that need to be addressed before presenting the system to a language learner, most of them had scores above 90%, which is comparable to systems of grammatical labeling (Cer, De Marneffe, Jurafsky, & Manning, 2010).

SMILLE is designed to be used along a regular language course, so, to approximate the variety of texts that a language learner can be in contact with, we used different genres, presenting a broader range of contexts for testing the developed rules, and we observed, for instance, that genre affects the detection of grammatical structures and should be considered for parsing purposes. This result can be used to optimize the system to consider text genres, so that rules could be specialized and applied to certain genres. In general, this allows us to improve SMILLE to better address the user needs.

6. Acknowledgements

We thank the Walloon Region (Projects BEWARE 1510637 and 1610378) for support, and Altissia International for collaboration.

References

Azab, M., Salama, A., Oflazer, K., Shima, H., Araki, J., & Mitamura, T. (2013). An NLP-based reading tool for aiding non-native English readers. *International Conference Recent Advances in Natural Language Processing, RANLP.*

Cer, D. M., De Marneffe, M. C., Jurafsky, D., & Manning, C. D. (2010). Parsing to Stanford dependencies: trade-offs between speed and accuracy. In N. Calzolari et al. (Eds), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

Chinkina, M., & Meurers, D. (2016). Linguistically aware information retrieval: providing input enrichment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of {NLP} for Building Educational Applications, BEA@NAACL-HLT 2016, June 16, 2016, San Diego, California* (pp. 188-198). https://doi.org/10.18653/v1/W16-0521

Danescu-Niculescu-Mizil, C., & Lee, L. (2011). Chameleons in imagined conversations: a new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics* (pp. 76-87). ACL.

Greene D., Cunningham, P. (2005). Producing accurate interpretable clusters from high-dimensional data. In A.M. Jorge et al. (Eds), *Knowledge Discovery in Databases: PKDD 2005. Lecture Notes in Computer Science, vol 3721*. Springer. https://doi.org/10.1007/11564126_49

Jaidka, K., Chandrasekaran, M. K., Rustagi, S., & Kan, M. Y. (2016). Overview of the CL-SciSumm 2016 Shared Task. In *BIRNDL@ JCDL* (pp. 93-102).

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55-60). Association for Computational Linguistics.

Meurers, D., Ziai, R., Amaral, L., Boyd, A., Dimitrov, A., Metcalf, V., & Ott, N. (2010). Enhancing authentic web pages for language learners. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 10-18). ACL.

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: a corpus analysis. *Journal of memory and language, 57*(3), 348-379. https://doi.org/10.1016/j.jml.2007.03.002

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied linguistics, 11*(2), 129-158. https://doi.org/10.1093/applin/11.2.129

Schmidt, R. W. (2012). Attention, awareness, and individual differences in language learning. In W. M. Chan, K. N. Chin, S. Bhatt & I. Walker (Eds), *Perspectives on individual characteristics and foreign language education* (pp. 27-50). De Gruyter Mouton. https://doi.org/10.1515/9781614510932.27

Simard, D. (2009). Differential effects of textual enhancement formats on intake. *System, 37*(1), 124-135. https://doi.org/10.1016/j.system.2008.06.005

Smith, M. S. (1993). Input enhancement in instructed SLA. *Studies in second language acquisition, 15*(2), 165-179.

Smith, M. S., & Truscott, J. (2014). Explaining input enhancement: a MOGUL perspective. *International Review of Applied Linguistics in Language Teaching, 52*(3), pp. 253-281. https://doi.org/10.1017/S0272263100011943

Zilio, L., & Fairon, C. (2017). Adaptive system for language learning. In *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)* (pp. 47-49). IEEE. https://doi.org/10.1109/ICALT.2017.46

Zilio, L., Wilkens, R., & Fairon, C. (2017). Using NLP for enhancing second language acquisition. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2017) in Varna, Bulgaria, 2-8 September 2017* (pp. 839-846).

Published by Research-publishing.net, not-for-profit association
Contact: info@research-publishing.net

CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017
Edited by Kate Borthwick, Linda Bradley, and Sylvie Thouësny

Typeset by Research-publishing.net

Cover design based on © Josef Brett's, Multimedia Developer, Digital Learning, http://www.eurocall2017.uk/, reproduced with kind permissions from the copyright holder.

Cover layout by © Raphaël Savina (raphael@savina.net)
Photo "frog" on cover by © Raphaël Savina (raphael@savina.net)

Fonts used are licensed under a SIL Open Font License

ISBN13: 978-2-490057-04-7 (Ebook, PDF, colour)
ISBN13: 978-2-490057-05-4 (Ebook, EPUB, colour)
ISBN13: 978-2-490057-03-0 (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'.

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit: Bibliothèque Nationale de France - Dépôt légal: décembre 2017.