Parsing Reviews with Link Grammar - Use of R for Sentiment Analysis

George B. Chorozoglou[1], Nikolaos Z. Zacharis[2], George Mikros[3] and Evangelos C. Papakitsos[4]

[1] Department of Informatics and Computer Engineering, UNIWA, Egaleo, Greece, gchorozoglou@uniwa.gr

[2] Department of Informatics and Computer Engineering, UNIWA, Egaleo, Greece, nzach@uniwa.gr

[3] Department of Italian Language and Literature, NKUA, Athens, Greece, gmikros@isll.uoa.gr

[4] Department of Industrial Design and Production Engineering, UNIWA, Egaleo, Greece, papakitsev@uniwa.gr

Abstract. This paper uses Link Grammar to parse reviews for Moodle learning management system and R Project to draw useful conclusions about Sentiment Analysis. Our goal is to describe the whole procedure in simple terms.

Keywords: *parsing; Link Grammar; R Project; Sentiment Analysis;*

1. Introduction

Syntax Analysis is a procedure to analyse sentences to their constituents. It is achieved by using rules of a grammar. The result of parsing can be either a tree or a dependency tree. We use 93 reviews of Moodle Learning Management System users for parsing, dating from Dec 5, 2017 to Jul 29, 2020. Link Grammar is our grammar to parse all these reviews. From all the data, we can extract useful statistics that can be evaluated. In twelve (12) cases we have corrected common spelling mistakes. A key element that interests us is the percentage of complete syntactic analysis.

Then we use the statistical package R to conduct Sentiment Analysis. We want mainly to investigate how many reviews are positive and how many negative. In addition, we try to distinguish the various sentiments. In the conclusion we report our comments as well as the positives or negatives of Link Grammar.

2.   Link Grammar

Link Grammar (LG) is a context-free Grammar created by Sleator & Temperley (1991). A sentence is part of the grammar if we can draw labeled links between words satisfying four rules: *planarity*, *connectivity*, *ordering & exclusion.* Each word has linking requirements: a formula involving connectors (letters or pairs of letters) and special characters like "+", "–", "&", "or" and/or "@". The use of "+" and "–" in a formula means that the word must join with another one on its right or its left, respectively.

All the words create a Lexicon (dictionary). To connect two words, they must have the same connectors and a plus & minus sign. Planarity is when links are not crossed. All the words must have a link (connectivity). Two words must have one and only one link, according to the ordering of the requirements (ordering & exclusion). A full parsing is called *linkage*. The dictionary can be very easily expanded. An example of a small lexicon and a linkage can be seen in Table 1 and Fig. 1 accordingly.

TABLE 1

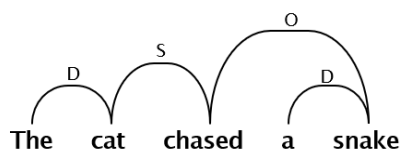| words	formula
the, a	D+
cat, snake	D– & (O– or S+)
chased	S– & O+

Figure 1

2.1 Link Parser

To parse data we use a Link Parser. There are two versions, 4 and 5. The former is online (https://www.link.cs.cmu.edu/link/submit-sentence-4.html), the latter can be used in Linux and the current stable version is Link Grammar 5.10.2 (September 2021). Our goal was to compare the two versions in terms of the percentage of correct syntactic analysis.

Link Parser extracts in a different way the results of the syntactic analysis. In Fig. 2, 3 & 4 we can see the differences in Stanford Parser, spaCy Dependency Parser and Link Parser. It is noteworthy that in the first parser we see a tree, in the

second one there are dependencies, while in the third one we observe the connections between words. We can also see the multiple results in Link Parser (142 linkages) with the remark that it is a Robust Parser, as well.
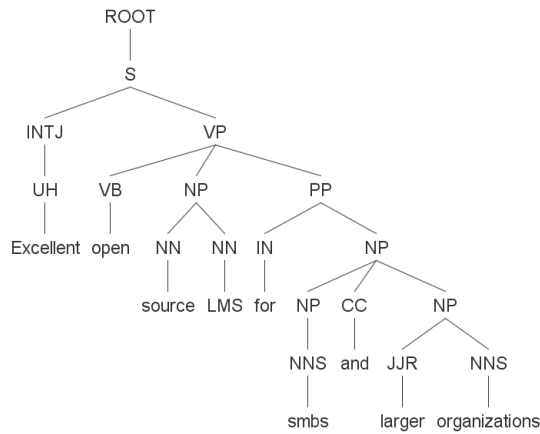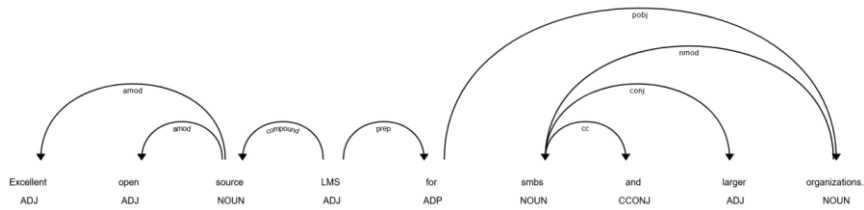
Figure 2

Figure 3

```
Found 142 linkages (142 had no P.P. violations)

        Linkage 8, cost vector = (UNUSED=0 DIS= 2.10 LEN=20)

    +--------------------------------------------Xp------------------------------------------------+
    +----------------------------->Wa-----------------------------+                                |
    |           +-------------A------------+----------SJlp---------+---------SJrp---------+         |
    |           |         +---A---+---AN---+--Mp--+---Jp--+        |         +-----Am-----+         |
    |           |         |       |        |      |       |        |         |            |         |
LEFT-WALL excellent.a open.a source.n LMS[!].n for.p smbs[!].n and.j-n larger.a-c organizations.n .
```
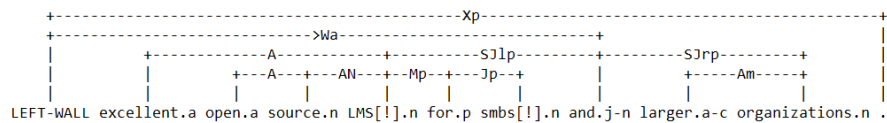
Figure 4

3. Statistical Data

Analysing Moodle reviews, we can see interesting statistical facts about sentences and parsing, by comparing also the two versions of Link Parser. In the following Table 2 we present data like *Number of Words per Sentence*, *Percentage of POS Tagging*, *Average Number of Linkages* or *Percentage of Correct Full Parsing*. It is obvious that version 5 has better results in parsing but numerous. When Link Parser does not produce a complete result, it brings us a partial solution by omitting one or more words from the sentence.

TABLE 2

	Version	
	4	**5**
Average Number of Words per sentence	9.11	
Full Parsing	36.4%	61.6%
Correct Full Parsing	29.3%	48.5%
Average Number of Linkages	7.4	1554.8
Average Number of not full Linkages	3.7	367.9
Percentage of POS Tagging	86.5%	95.11%
Percentage of POS Tagging in spaCy	95.11%	

Besides, sentences vary from 1 word to 21. The maximum number of linkages in version 4 is 54 while in version 5 is 37,800. From all these linkages, we had to distinguish those that were right. Not full linkages are sentences with partial parsing. As for POS Tagging, we compare Link Parser with spaCy, an open-source library of Natural Language Processing (NLP) programs written in Python & Cython.

4. Sentiment Analysis with R

After parsing, the next step is to use R Project for Sentiment Analysis. It is the process of detecting positive or negative sentiment in a text. We use R Version 4.1.1 "Kick Things" (2021-08-10) in combination with RStudio Version 2021.09.0 Build 351 "Ghost Orchid" (2021-09-20). R is a free software environment for statistical computing and graphics while RStudio develops free and open tools for R, and enterprise-ready professional products. In Fig. 5 we can see RStudio.

Before Sentiment Analysis we must transform our data to have the best results. The steps we should take for the best accuracy are:

- Replace special characters (like "/" or "@") with space;
- Convert the text to lower case;
- Remove numbers and punctuations;
- Reduce words to their root form (Text stemming);
- Remove English common stopwords that are words without value in Sentiment Analysis (like "a", "I", "the", "for", "if", …).
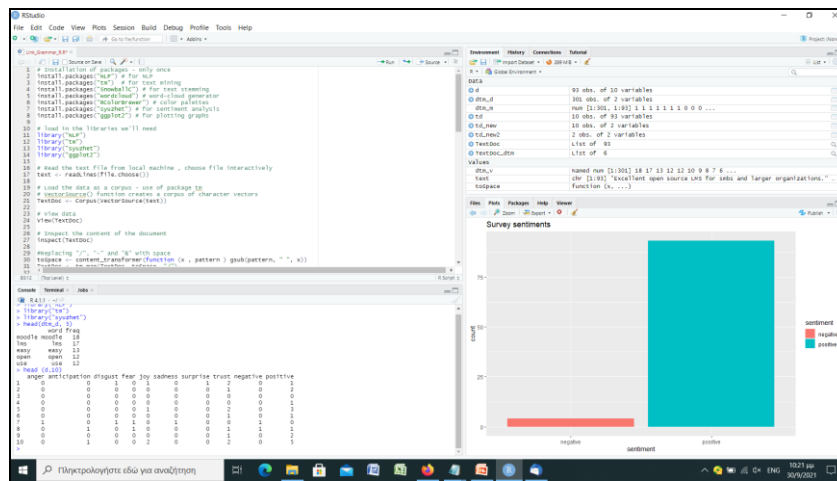

Figure 5

In Fig. 6 & 7 we observe the plots of count of words associated with each sentiment and count of words associated with negative & positive sentiments respectively. We have installed and used the *NLP*, *tm*, *syuzhet* and *ggplot2* packages.
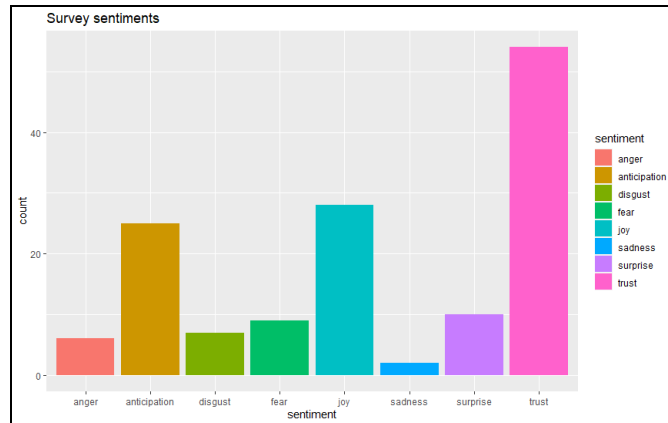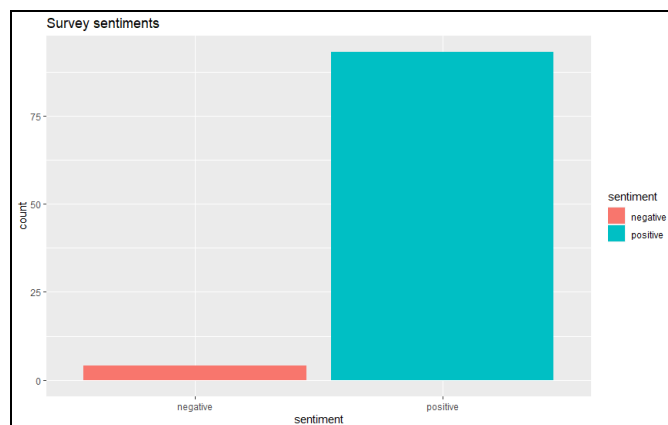
Figure 6

Figure 7

5. Conclusion

Link Grammar offers a (robust) algorithmic structure, but it has a poor user interface. Our data for syntax analysis were Moodle reviews in English language. There were multiple results for full parsed sentences. The percentage of Correct Full Parsing is not so good (48.5% for version 5). In contrary the results of Sentiment Analysis are close to reality, especially about positive and negative reviews.

References

D. Sleator and D. Temperley. "Parsing English with a Link Grammar" in Proceedings of the Third International Workshop on Parsing Technologies, Tilburg, Netherlands and Durbuy, Belgium, 1993, pp. 277-292.

D. Sleator and D. Temperley. "Parsing English with a Link Grammar". Carnegie Mellon University Computer Science technical report CMU-CS-91-196, October 1991.

https://elearningindustry.com/directory/elearning-software/moodle/reviews

http://www.abisource.com/projects/link-grammar/

https://www.link.cs.cmu.edu/link/index.html

https://monkeylearn.com/sentiment-analysis/

https://www.red-gate.com/simple-talk/sql/bi/text-mining-and-sentiment-analysis-with-r/

https://spacy.io/

https://nlp.stanford.edu/software/lex-parser.shtml

https://www.r-project.org/

https://www.rstudio.com/