Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 233 – 241

Procedia
Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Syntactic ngrams as keystructures reflecting typical syntactic patterns of corpora in Finnish

Veronika Laippala[a,b*], Jenna Kanerva[c], Filip Ginter[c] *

[a] School of Languages and Translation Studies, 20014 University of Turku, Finland
[b] Turku Insitute of Advanced Studies, 20014 University of Turku, Finland
[c] Department of Information Technology, 20014 University of Turku, Finland

Abstract

This article studies syntactic ngrams, i.e. little subtrees of dependency syntax analyses, as keystructures reflecting syntactic characteristics of corpora. While traditional keywords correspond to statistically more or less frequent words of a corpus and are often informative on the corpus topic and style, unlexicalized syntactic ngrams applied in this study extend the level of description beyond individual words to sequences of syntactic elements. The article analyzes the utility of these sequences in corpus description and gives first results on the structural characteristics reflected by them in the studied texts, including Finnish literature, Internet forum discussions from the major Finnish social networking website and Internet discussions following the news and editorials of the major Finnish newspaper's website. The syntactic ngrams are produced with the freely available Finnish Dependency Parser and Ngram Builder and the keystructures analyzed with a linear classifier. The results suggest that syntactic ngrams illustrate both topical features, such as names and Internet urls discussed in the corpora, as well as structural characteristics, such as subject-verb combinations, negations and informal sentence structures, thus both generalizing the information given by traditional keywords from individual words to concepts and providing new knowledge about typical constructions not reached by lexemes.

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.

Keywords: Keyness; syntactic ngrams; computer-mediated communication

* Corresponding author..
E-mail address: mavela@utu.fi

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.
doi:10.1016/j.sbspro.2015.07.441

1. Introduction

In corpus linguistics, keyness is a typical concept used to study statistically meaningful differences between texts or corpora (Scott and Tribble, 2006; Bondi and Scott, 2013). The most often, keyness is analyzed via keywords, i.e. words that are statistically more or less frequent in the corpus under study than in a reference one. The resulting keywords are said to be informative on the text topic and style (Scott and Tribble, 2006: 55-72).

As attested by the frequency of their use, keywords offer a very useful method for the study of text contents. However, since they concentrate only on words, keywords cannot fully illustrate the analyzed texts; for instance structural text features reflected by syntax and morphology are not taken into account. These properties can also reveal important aspects of the texts.

This article extends the concept of keyness by applying it to the study of morphological and syntactic characteristics of texts. These structural features are studied using unlexicalized syntactic ngrams (Kanerva, Luotolahti, Laippala, and Ginter, 2014), i.e. little subtrees of dependency syntax analyses where the actual lexical elements have been removed. While a traditional ngram is composed of words following each other, a syntactic one consists of syntactically related elements, their morphological analyses and the dependency types relating them. Words in a syntactic ngram do not necessarily follow each other linearly. By ignoring the actual words, they extend the level of description beyond individual words to sequences of syntactic elements and offer the possibility to concentrate on structural patters typical of certain texts.

Reporting the first results of an ongoing project, this article aims at analyzing syntactic ngrams as constructions reflecting structural text characteristics. Can they be used as keystructures and what do they tell about differences between texts?

The corpora under study include literature, online discussions following the news articles of a major Finnish newspaper and online discussions from one of Finland's largest online social networking websites. The syntactic ngrams are generated with the freely available Finnish dependency parser[†] and syntactic ngram builder[‡] and the differences counted using a linear classifier implemented in the Vowpal Wabbit machine learning package (Agarwal, Chapelle, Dudik, and Langford, 2014)[§].

2. Syntactic ngrams and the study of structural aspects of text

The syntactic analyses attached to the words of syntactic ngrams deepen the information provided about the context of a given word. Therefore, they have been applied in computational linguistics e.g. to methods that would traditionally concentrate on the linear context of a word (see Sidorov, Velasquez, Stamatatos, Gelbukh, and Chanona-Hernández, 2013). Collections of syntactic ngrams have been published for English (Goldberg and Orwant, 2013) and for Finnish (Kanerva, Luotolahti, Laippala, and Ginter, 2014), and the code to produce syntactic ngrams for the dependency scheme used by the Finnish dependency parser is also publically available.

As already mentioned in the introduction, keyness is traditionally studied via keywords (see articles in Scott and Tribble, 2006; Bondi and Scott, 2013). Keyword analysis offers a useful data-driven perspective to the study of large corpora; no words need to be selected prior to the analysis but the method outputs the important ones for a further study to be done by the researcher. As opposed to methods where the research needs to select the studied features, this data-driven approach ensures that the analysis can concentrate on truly significant aspects of texts.

[†] http://turkunlp.github.io/Finnish-dep-parser/
[‡] https://github.com/jmnybl/syntactic-ngram-builder
[§] https://github.com/JohnLangford/vowpal_wabbit/wiki

Despite their adequacy, data-driven methods are rarely applied to the study of text structure. Ivaska (2014) on combinations of morphological forms, i.e. morphological ngrams in learners' texts, is one of the only ones to have applied keyness to structural text features. Traditional corpus linguistic approaches to text structure would rather concentrate on analyzing the use of preselected features in a number of corpora. For instance, Biber (1995) and the studies reported in Conrad and Biber (2001) apply factor analysis on corpora in English, using combinations of features chosen based on previous research. A part of the features correspond to morphological and syntactic ones. However, as opposed to ngrams, most of them do not include several elements but rather refer e.g. to a feature that could be expressed by a morphological form, such as a noun, or a syntactic dependency type, such as a relative clause. Based on their co-occurrence in certain registers, the analysis places these features to dimensions representing text variation. Depending on the aim of the study, the analysis may include just one dimension, such as "discourse complexity", or several, such as "informational versus involved", "narrative versus non-narrative", etc. Despite this difference of methods, the results reported by these studies point directions for this article as well.

3. Corpora

The corpora included in the study represent various standard and informal text genres:

- Articles from a Finnish newspaper and an online magazine from Turku Dependency Treebank (TDT) (Haverinen, Nyblom, Viljanen, Laippala, Kohonen, Missilä *et al.*, 2013), 23,582 words.
- Finnish Wikipedia from TDT, 41,489 words.
- Finnish literature from Corpus of Translated Finnish (Mauranen, 2000), 98,522 words.
- Internet forum discussions from the Finnish yellow press news site, 57,116 words.
- Internet discussions following the news and editorials published in the news site of the major Finnish newspaper, 102,949 words. Collected by a group of researchers from the University of Turku, Finland.
- Internet forum discussions on exercise and cars from one of Finland's largest online social networking websites 57,116 words.

The three corpora selected for the analysis of typical syntactic patterns are the literature, the Internet discussions following the news and editorials on the major Finnish newspaper news site and the discussion fora from the social networking website. The others are used as a reference corpus. These three are chosen because they include at the same time similarities and differences that allow both the study of syntactic ngrams as a method and the analysis of the typical features of certain genres. Literature offers a perspective to standard, narrative Finnish written by professionals. The Internet discussions are clearly different on this aspect, although they as well include differences: while the discussions from the social networking site are on informal topics, such as cars and sports, those from the news site elaborate on the topic discussed by the news article or the editorial. The social networking site may also attract more conflict-searching writers than the news site. Based on a previous study on the degree of formality of politicians' blogs (Lehti and Laippala, 2014), this difference may be apparent in the language as well. Our hypothesis is that the discussions on the social networking site are informal while the ones on news resemble standard text. In order to get a first idea of the corpora, Table 1 presents the text collections by keywords. These lists describe adequately the topics of the texts.

Table 1. Translations of the 20 most significant keywords of the corpora. Numbers and other non-linguistic features are excluded

Literature	News discussions	Discussions on sports and exercise
me, autumn-prince, from-my-drinkings, of-drinking, Olli, man, announced, we, he-said, I-said, me, he-asked, mother, when, girl, you, he-yelled, Jari, of-the-mistake	Russia, Obama, of-Russia, of-Obama, banks, USA, McCain, latest, now, Bushes, Palin, Bush, U.S., banks', to-Russia, president, McCain's, USA's, Iceland, Georgia	of-the-car, car, jealous, price, consumption, exercise, liters, fitness, of-the-consumption, of-exercise, km, goes, movement, tank, in-the-car, BMW, at-the-gym, in-good-condition, many, car (prt)

4. Methods

The syntactic ngrams are generated with a freely available pipeline consisting of a Finnish dependency parser (Haverinen, Nyblom, Viljanen, Laippala, Kohonen, Missilä *et al.*, 2013) and ngram builder (Kanerva, Luotolahti, Laippala, and Ginter, 2014). The dependency parser is a state-of-the-art analyzer following the Stanford Dependencies (SD) scheme (de Marneffe and Manning, 2008) and includes an entire natural language processing toolkit with tokenization, sentence splitting, morphological and syntactic analyses. With 46 dependency types, the level of detail presented by the parser offers a fruitful basis for linguistic analysis.

The syntactic ngram builder generates four possible ngram types: biarcs composed of two dependencies between three tokens, triarcs between four tokens and quadrarcs between five tokens. For the purposes of this study, all these four types are included. The analyzed syntactic ngrams are unlexicalized, i.e. the words are excluded. All the other information generated is kept, including the part-of-speech and morphological features.

While keywords can be analyzed with many standard corpus tools such as Antconc, the exploration of keystructures is more challenging especially with larger corpora. In this study, the analysis of the keyness is done with a linear classifier implemented in the Vowpal Wabbit machine learning package (Agarwal, Chapelle, Dudik, and Langford, 2014)[**]. It is trained with the stochastic gradient descent method with a 66% / 33% split on train and test corpora. The classification is done to the corpus under study and to a reference one including all the other ones, each labelled text segment consisting of ten sentences. The advantage of Vowpal Wabbit is that while performing the classification, it generates a list of the most significant features of the analyzed text classes. As these are the ones that distinguish the classes the best, they are at the same time the keystructures.

5. Syntactic ngrams in the corpora

In order to analyze the keystructures, 30 most significant syntactic ngrams of each corpus are listed and grouped to similar groups. The following sections discuss these features.

Table 2. Classifier performances

	Literature	News discussions	Discussions on sports and exercise
F-score	83%	91%	64%
AUC	98,7%	97,9%	90,5%

The classifier performances for all the three corpora are presented in Table 2 with two figures. The F-score is the balanced and harmonic mean of precision and recall, very frequently used in natural language processing tasks. In addition, we report AUC, the area under the receiver operating characteristic curve, which corresponds to the probability that a randomly chosen negative and a randomly chosen positive example will be correctly ranked. As AUC is more convenient for comparing performance across corpora with different class distributions, it suits the present task very well. However, it is important to notice that the figures are presented only to demonstrate the confidence level of the analysis and that the aim is not to build a system performing the classification as well as possible but to analyze if syntactic ngrams can be used to study the text structures and what they tell about this structure. The classification is merely a method to discover the significant features.

[**] https://github.com/JohnLangford/vowpal_wabbit/wiki

5.1. Literature

In the literature corpus, two feature groups are extremely dominant; out of the 30 syntactic ngrams analyzed, 10 correspond to verb-verb co-ordinations and 17 to subject-verb combinations. In addition, the analyzed features include three syntactic ngrams describing verbal complements.

Examples on the two most frequent structures are shown in Figure 1. The lines 1 and 3 illustrate co-ordinations (conj) of finite verbs, the latter one including also a conjunction (cc). The main difference between the examples is the variation of the verb person: while the example on line 1 co-ordinates same verb forms, that on line 3 combines verbs in first person (PRS_Sg1) and third person singular (PRS_Sg3). As can be noticed from the example lexicalizations on the lines 2 and 4, this variation of forms reflects a noticeable difference in practice, on the sentence level. All the other verb-verb co-ordinations structures not included in the Figure are very similar, the most important difference being that one of them co-ordinates first person plural and one third person plural verbs. These all would seem to reflect narrative constructions one could expect to find in literary texts.

The syntactic ngrams on the lines 5 and 7 illustrate subject-verb combinations, another keystructure of the literary corpus. As explicit subjects are not mandatory in Finnish, their frequent use seems to be typical of literature. These structures include both nouns and pronouns as subjects, most of the verbs being in third person singular. Many of the structures include also a co-ordination or a coordinating conjunction departing from the verb. This highlights even more the importance of verbal co-ordinations in these texts. As with verb-verb combinations, also these seem to reflect typical narrative structures.

Figure 1. Examples of the keystructures in the literature corpus, followed by possible lexicalizations to illustrate. The part-of-speech classes are shown in the dark blue boxes and the morphological features under them.

5.2. News site discussions

Out of the 30 first keystructures of the forum discussions from the news site, 19 describe forum participant's names and comment dates. As these are not informative on the linguistic characteristics of the discussion text and could have been deleted already during the corpus preprocessing, we ignored these features and extended the analysis in order to include 30 keystructures even with these dates excluded.

Similarly to the patterns discussed in the previous section, groups of similar syntactic constructions seem to emerge from the keystuctures of the news site discussion corpus as well. Out of the analyzed structures, seven include links to Internet urls and another six describe proper name constructions, as the examples on lines 5 and 6 of Figure 2. These are related rather to the discussion topics than to frequent syntactic constructions. In contrast, the rest of the keystructures seem to reflect linguistically more interesting syntactic patterns. Lines 1 and 2 in Figure 2 illustrate a construction composed of a finite verb and a direct noun object with a derivation that could in English be translated with *–ing*. In addition, there is a dependency relating the direct noun and the genitive object that in the example lexicalization is *Northern Ossetia.* Note that in fact the syntax analysis of line 2 is wrong, as *recognition* should be the sentence subject. As the error however is systematic, this does not have great importance in the analysis. The studied keystructures include seven constructions similar to this one, each including a verb and a direct object with the same derivation.

Further, the keystructures include six constructions illustrated on the lines 3 and 4, composed of an adverb, a proper noun subject, a finite verb and a direct object. In addition to Finland, other frequent proper nouns as subjects in the corpus include Russia, McCain and Palin, and adverbs the Finnish equivalents for *so*, *now* and *then*. Most importantly, these keystructures would seem to describe the discussion topic on the news sites. However, they do also reflect the syntactic structures used in the discussion, although the usages of this structure would need more detailed analyses to be understood completely.

Finally, the patterns on lines 7 and 8 illustrate a sentence complement with *that* and a finite verb in the passive voice (PRS_4+VOICE_PASS). In fact, as verbs in passive do not appear in the keystructures of the other corpora, this seems to be typical of the news site discussion. A closer study of these patterns reveals that they frequently express reported speech and that the passives are used in reference to the news article, often repeating the phrases used in the news text. This supports the previous study by Biber (2001) noting the frequency of passives in press reportage, and explains their appearance as a keystructure in this corpus.

Figure 2. Examples of the keystructures in the news site discussion corpus, followed by possible lexicalizations to illustrate. The part-of-speech classes are shown in the dark blue boxes and the morphological features under them.

5.3. Internet forum discussions in the social networking site

In contrast with the two other corpora, the keystructures of the Internet forum discussions from the social networking site are relatively diverse, there are not as clear subgroups emerging as there was from the news site and literature corpora. Nevertheless, the keystructures found seem to reflect very well forum discussion characteristics. Again, five of the keystructures are Internet urls. From the rest, nine represent co-ordinations and various clausal complement types that would need more detailed analyses to be interpreted. The remaining ones are, however, very interesting since they seem to reflect informality and argumentation, characteristics we already associated with this corpus when presenting the hypothesis in Section 3.

The most common subgroup of the keystructures is formed by eight constructions including a negation (neg). Many of these include a verb in the first person singular and seem to express disagreements or uncertainty. On lines 1 and 2 in Figure 3, there is a simple negated copular structure with an predicative adjective and a first person subject. As the only part varying in the lexicalized form is the predicative adjective, all these keystructures are relatively similar, the example lexicalization on line 2 indicating simply that the subject is not sure about something. Other frequent predicatives in these structures include *better* and *stupid*.

On lines 3 and 4, the example includes a negated first person verb and a direct object (dobj), the sentence in the lexicalization being part of a discussion on using notebooks for counting fuel consumption and the subject articulating his disagreement with the previous participants. This structure includes also an adverb and a quantifying pronoun attached to the object. In addition to the adverb strengthening the argument, the Finnish words *lainkaan (~at all)* and *mitään (~any)* used in the lexicalization indicate a degree of informality; such expression would not be used in standard Finnish. Further, especially the use of a pronoun of this type attached to the noun (det), regardless of its lexicalized form, has an informal effect to the structure. In addition to this keystructure, the 30 most typical ones include four other patterns reflecting informality, the main decisive characteristics in the others being playful writing and abbreviations.

Finally, another feature reflecting characteristics of these discussions is imperatives: the 30 keystructures include four composed of an imperative (TENSE_Imprt), as illustrated on lines 5 and 6 of Figure 3. This is a very interesting component of the discussion, as it seems that the participants are if not giving orders, at least instructions to each other. Based on the frequency of studies on advice-giving in different online contexts (e.g. Lindholm, 2010; Morrow, 2006), this seems to be another characteristic of Internet fora.

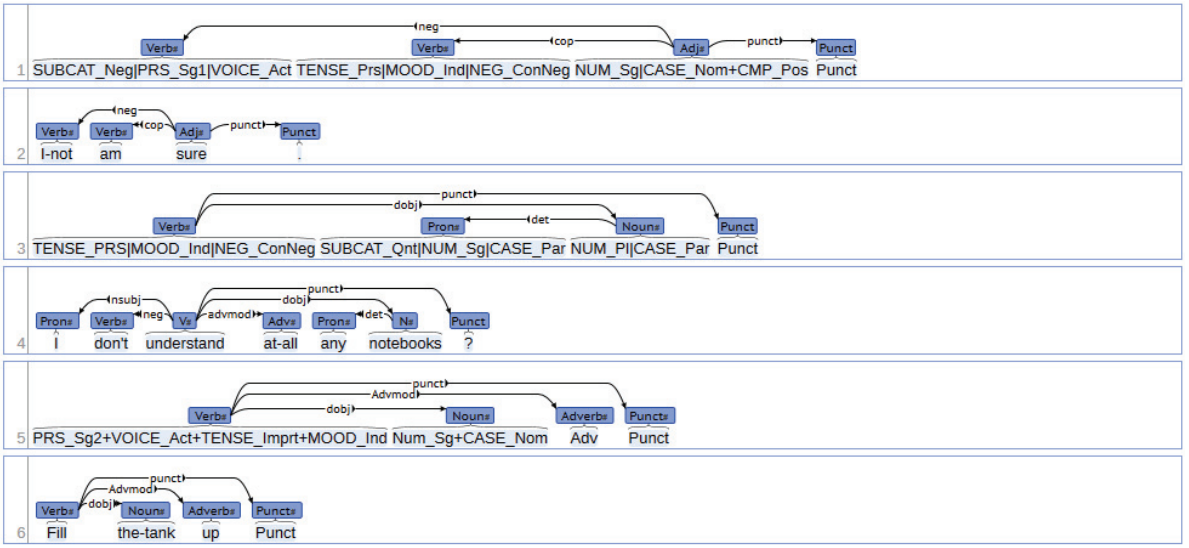

Figure 3. Examples of the keystructures in the Internet forum discussion corpus, followed by possible lexicalizations to illustrate. The part-of-speech classes are shown in the dark blue boxes and the morphological features under them.

6. Discussion and conclusions

As shown in the previous sections, some of the keystructures reflect topical aspects of texts, such as names or Internet urls. Although the level of description is more general and thus perhaps more applicable to larger corpora, this functioning is somewhat similar to that of traditional keywords. However, the majority of the keystructures seem to describe structural characteristics of texts, such as co-ordinations or subject-verb combinations. Naturally, more detailed analyses are needed in order to fully understand the role of these syntactic properties in the corpora and in the text genre they represent. For this objective, the scope of the current article is not enough. However, already the analysis of a modest group of 30 most significant syntactic ngrams of each corpus gives interesting information outside the scope of traditional keywords.

The keystructures illustrating the literature corpus included verb-verb co-ordinations and subject-verb combinations that would seem to be typical narrative constructions one could expect to find in literature. In contrast, the discussions following the news and editorials included names, proper name subjects with an adverb, passive voice in clausal complements and direct objects. While the names and proper name subjects are informative on the text topic and the frequency of the passive voice is supported by previous studies on press reportage, the role of the adverbs and clausal complements requires further studies. Finally, the keystructures of the Internet discussions on the social networking site included imperatives used to give advice to other participants, determiner-noun combinations typical of informal language and different negations illustrating e.g. uncertainty and disagreement. These seem to describe characteristics of online discussion fora, some of them being also discussed in previous studies. Naturally, in order to fully understand the phenomena reflected by the syntactic ngrams, a detailed analysis of the most frequent lexicalizations of the constructions is needed, as we demonstrated with predicative adjectives and proper name subjects.

To conclude, this article has demonstrated that syntactic ngrams can indeed be used as keystructures to illustrate structural corpus characteristics, some of them being more topical and some more syntactic. In comparison with traditional keywords, syntactic ngrams offer both complementary information generalizing beyond individual words to concepts and information depending on syntax not reached by lexemes. In addition, it is interesting to notice that some of these constructions, such as negations or imperatives, clearly illustrate phenomena related to the discourse at hand, only realized by syntax. If used as keys to the corpora, this can be very useful.

This article has presented only the first results on the use of syntactic ngrams as keystructures, and leaves numerous paths for more detailed studies. Future methodological work will include the application of other classifiers and other levels of feature representation and delexicalization. For instance, the effect of excluding the morphological features would be interesting to study. In addition, an obvious direction is the extension of the corpora used; this could as well lead to results better reflecting genre characteristics. Naturally, more detailed and more extensive studies of the keystructures are also needed in order to yet better understand their potential.

References

Agarwal, A., Chapelle, O., Dudik, M., and Langford, J. (2011). A Reliable Effective Terascale Linear Learning System. *JMLR, 15*, 1111 − 1133.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge.: Cambridge University Press.

Biber, D. (2001). On the Complexity of Discourse Complexity. In S. Conrad, and Biber, D. (Eds.), *Variation in English: Multi-Dimensional Studies* (pp. 215-240). USA: Routledge.

Conrad, S., and Biber, D. (Eds.) (2001). *Variation in English: Multi-Dimensional Studies*. USA: Routledge.

Bondi, M., and Scott, M. (Eds.) (2013). *Keyness in Texts*. Amsterdam, Philadelphia: Benjamins.

Goldberg, Y., and Orwant, J. (2013). A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books. *Second Joint Conference on Lexical and Computational Semantics (*SEM), 1: Proceedings of the Main Conference and the Share d Task: Semantic Textual Similarity* (pp. 241–247). Association for Computational Linguistics.

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., and Ginter, F. (2013). Building the Essential Resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation, 48(3),* 493 - 531.

Ivaska, I. (2014). Edistyneen oppijansuomen avainrakenteita: korpusnäkökulma kahden kielimuodon tyypillisiin rakenteellisiin eroihin. - Key structures in advanced learner Finnish: Corpus approach towards structural differences between two language forms. *Virittäjä, 2*, 161 - 193.

Kanerva, J., Luotolahti, J., Laippala, V., and Ginter, F. (2014). Syntactic N-gram Collection from a Large-Scale Corpus of Internet Finnish. Proceedings of the Sixth International Conference Baltic HLT.

Lehti, L., and Laippala, V. (2014). Style in French Politicians' Blogs: Degree of Formality. *Language at Internet, 11*.

Lindholm, L. (2010). 'A Little Story, for Food for Thought……': Narratives in Advice Discourse. In S-K. Tanskanen, M-L Helasvuo, M. Johansson, and M. Raitaniemi (Eds.), *Discourses in Interaction* (pp. 223-236). Amsterdam & Philadelphia: John Benjamins.

Mauranen, A. (2000). Strange Strings in Translated Language: A Study on Corpora. In M. Olohan (Ed.), *Intercultural Faultlines. Research Models in Translation Studies 1*, (pp. 119–141). Manchester: St. Jerome Publishing.

de Marneffe, M., and Manning, C. (2008). Stanford Typed Dependencies Tepresentation. *Proceedings of COLING'08, Workshop on Cross-Framework and Cross-Domain Parser Evaluation*.

Morrow, P. (2006). Telling about Problems and Giving Advice in an Internet Discussion Forum: Some Discourse Features. *Discourse Studies, 8(4),* 531 - 548.

Scott, M., and Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Philadelphia, PA, USA: John Benjamins Publishing Company.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., and Chanona-Hernández, L. (2013). Syntactic Dependency-Based N-grams as Classification Features. *Advances in Computational Intelligence. Lecture Notes in Computer Science 7630* (pp. 1-11).