Utilization of Dependency Type per Sentence to Identify Differences among Genres of English Texts

Masanori Oya
Mejiro University
m.oya@mejiro.ac.jp

Abstract

This study utilized the concept of type per sentence (TPS) with regards to dependency types to identify differences among genres of English texts obtained from an English language corpus. This study also attempted to suggest that TPSs can indicate some of the differences between different genres of texts. TPSs can be used as metrics to indicate how certain genres of texts are more likely to contain certain dependency types compared to others, and the higher TPSs of some dependency types can be shown to be related to higher TPSs of other dependency types.

1. Introduction

The basic assumption of Dependency Grammar (henceforth DG) is that we can find a dependency relationship between each word in a sentence and another word in the same sentence, and these dependency relations among words are labelled under certain categories called "types". The dependency between two words represents their hierarchical relationship, and the "type" of their dependency differentiates it from others. For example, the sentence "The dependency between two words represents their hierarchical relationship" is represented in the following dependency tree. The dependency types are indicated with capital letters:

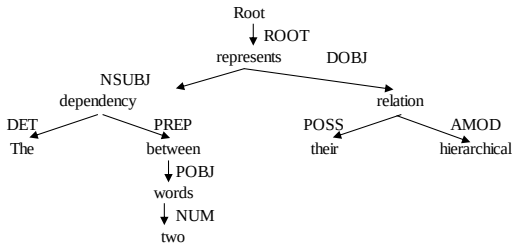

Figure 1. The dependency tree for the sentence "The dependency between two words represents their hierarchical relationship."

The sentence, "The type of a dependency differentiates it from others" is represented in the tree below:

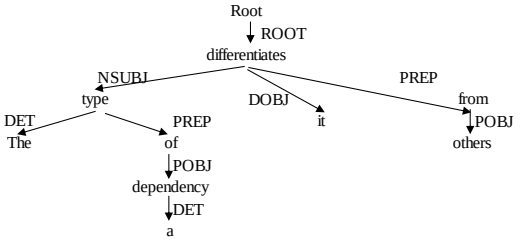

Figure 2. The dependency tree for the sentence "The type of a dependency differentiates it from others."

Along with these assumptions, this study assumes that texts of different genres show differences in frequency of use of certain dependency types; this is because each genre's distinct use of sentence structures can contribute to building differences between various genres, and the differences in frequency of use of dependency types can be partly used as one of the metrics, along with sentence lengths, to indicate the differences among genres. In order to verify this assumption, this study investigates whether the different corpora of different genres actually show different frequencies of use for different dependency types, using type per sentence (henceforth TPS), introduced by Oya (2016), as a metric. The definition of TPS is as follows: "TPS" is the number of examples of a dependency type in a corpus (DT) divided by the number of sentences in the same corpus (S).

TPS = DT / S

Intuitively speaking, the TPS of a given dependency type in a given corpus indicates the average occurrence of that dependency type in a sentence in the same corpus. Thus, it is expected that the TPS of the same dependency type can vary across corpora of different genres. For example, if the sentences of a certain genre tend to have deeper embedding because of the frequent use of subordinate clauses, this fact may be reflected in the TPS of dependency types related to subordinate clauses; furthermore, this TPS may be higher than the TPS of the same dependency types found in the sentences of other genres. The higher TPSs of certain dependency types in some subcorpora indicate that these types are used more often in these subcorpora than in others. This also reflects the structural tendencies of the sentences in the subcorpora, such as deeper embeddedness for clauses or a larger number of modifying elements for content words.

For example, the higher TPS of the dependency type *acl* (adjectival clause) in some subcorpora means that nouns modified by verbs in present participles, past participles, or *to*-infinitives are used more often in these subcorpora than in others. In addition, the higher TPS of the dependency type *advcl* (adverbial clauses) in some subcorpora means that verbs modified by other verbs in present participles, past participles, or *to*-infinitives, or by subordinate clauses introduced by conjuncts (e.g., because, however, nevertheless) are used more often in these subcorpora than in others. Both of these cases indicate that the sentences in these subcorpora show a tendency toward deeper embeddedness compared to the other subcorpora.

For another example, if the TPSs of the dependency type *amod* (adjectival modification) in some subcorpora are found to be higher than the TPSs of *amod* in other subcorpora, it means that the nouns in the former tend to be modified by adjectives more often than those in the latter. The same will be true if the TPSs of other noun-modifying elements (e.g., articles, prepositional phrases or relative clauses) are higher in certain subcorpora than in others; this indicates that the nouns in these subcorpora are more likely to be modified by certain elements than those in other subcorpora.

Oya (2016) argued that TPS can be used as a metric to indicate the characteristics of a given corpus based on the frequency of occurrence for each dependency type found in the corpus.

TPS is a metric that can be conceived only within the framework of DG, and it can be calculated only via typed-dependency parsers. In this sense, TPS can be considered as a unique metric that still needs to be investigated from various view-points with different research questions, using sentence data obtained from the large-scale corpora of different genres.

2. Previous Studies

Oya (2016) used the Stanford Parser (Chen and Manning, 2014; de Marneffe et al., 2006) to obtain the dependency trees for English sentences and the TPS of each dependency type in two different corpora, with the intention of explicating the difference between learner English and authentic, academic English in terms of TPS. The corpora used in that study consisted of 881 sentences randomly chosen from a corpus of learner English (The International Corpus of Learner English Version. 2 [Granger et al., 2009]; henceforth *Learners*) and 881 sentences randomly chosen from a corpus of academic English (a manually constructed corpus of some abstracts from seven academic journals; henceforth *Journals*).

The TPSs of some dependency types were found to be higher in *Journals* than in *Learners*; In *Journals*, the TPSs of the top 20 most frequent dependency types were all higher than those in *Learners*. The dependency types *preposition*, *amod* (adjectival modification), and *compound* (noun compound) were the top three types whose TPS were the most deviant across *Learners* and *Journals*. No dependency type in *Learners* had TPS higher than 2. These findings indicate that TPS could be used as an indicator to identify text type.

Not all dependency types were found to be higher in *Journals* than in *Learners*. The TPSs of the dependency type *nsubj* (nominal subject) were almost the same across *Learners* and *Journals*. In *Learners*, there was no inclination of the use of particular dependency types such as *preposition*, *amod*, or *compound*, unlike in *Journals*.

3. Data and Method

3.1 Data

32nd Pacific Asia Conference on Language, Information and Computation
Hong Kong, 1-3 December 2018

This study used the Manually Annotated sub-corpus of American National Corpus (henceforth MASC 500K) (Ide et al., 2008). The MASC 500K contains about 500,000 words of contemporary American English, which are drawn from the Open American National Corpus (OANC) (Ide and Suderman, 2004). Originally, the MASC 500K contained various kinds of manually annotated tags such as sentence boundaries, tokens, lemmas, POSs, noun and verb chunks, and named entities. MASC 500K covers a wide range of genres; the written section consists of the following subcorpora: blog, emails, essay, texts from Ficlet (a website for short fictions that is now closed), fiction, government documents, jokes, journal, letters, movie-scripts, news, non-fiction, spam emails, technical reports, tweets on Twitter and travel guides. The spoken section contains texts from speeches and debates.

3.2 Method

The raw texts without tags (downloaded collectively as a data-only file from the website of ANC: http://www.anc.org/MASC/Download.html) were parsed through the Stanford Lexicalized Parser v.3.7.1 after some parts of the texts (titles, headers, dates, unconventional punctuations, etc.) were manually extracted or fixed; then, the number of each dependency type and its TPS was calculated using an original Ruby script.

Not all the subcorpora were used in this study. The reasons for this decision are as follows. The spoken section of MASC 500K was not included because the intention here was to primarily focus on written data. With the same intention, the subcorpora on jokes and movie-scripts were not included because it contained a fair number of conversational sentences. The subcorpora of e-mails and spams were not included because they contain a lot of the repetitions of reply messages. The subcorpus of tweets was also excluded because it contained many HTML tags.

Since the current version of Stanford Lexicalized Dependency Parser employs Universal Dependencies (de Marneffe et al., 2014), the dependency types used in this study were based on them. For the definition of each dependency type in English, see the Webpage of Universal Dependencies (http://universaldependencies.org/#language-).

4. Results and Discussion

4.1 Overall description of the TPSs of main dependency types

The TPSs of the main dependency types are summarized in Table 1 below.

	Blog	*Essays*	*Ficlets*	*Fiction*	*Govt*	*Journal*	*Letters*	*News*	*Nonf*	*Tech*	*Travel*	*All*	*M*	*SD*
acl	0.173	0.256	0.060	0.096	0.284	0.252	0.174	0.200	0.157	0.324	0.181	0.167	0.196	0.079
acl:relcl	0.233	0.270	0.080	0.099	0.229	0.292	0.187	0.232	0.223	0.216	0.147	0.177	0.201	0.067
advcl	0.350	0.431	0.180	0.248	0.382	0.404	0.321	0.310	0.276	0.320	0.214	0.291	0.312	0.078
advmod	0.926	1.139	0.597	0.682	0.728	1.209	0.719	0.668	0.767	0.855	0.742	0.778	0.821	0.196
amod	1.026	1.844	0.344	0.566	1.898	1.541	1.235	1.500	1.954	2.640	1.631	1.225	1.471	0.655
aux	0.587	0.556	0.299	0.431	0.552	0.513	0.516	0.525	0.380	0.360	0.258	0.438	0.452	0.113
auxpass	0.143	0.341	0.051	0.060	0.235	0.261	0.103	0.229	0.240	0.477	0.218	0.173	0.214	0.125
case	1.879	2.989	0.766	1.187	2.913	2.925	1.899	2.663	2.460	3.503	2.622	2.016	2.346	0.828
ccomp	0.359	0.320	0.204	0.272	0.301	0.435	0.244	0.451	0.206	0.268	0.093	0.276	0.287	0.104
compound	0.946	1.028	0.358	0.229	1.974	1.250	1.492	2.106	1.215	2.192	1.740	1.090	1.321	0.661
conj:and	0.465	0.754	0.216	0.299	0.947	0.624	0.742	0.585	0.739	0.932	0.775	0.554	0.644	0.237
dep	0.439	0.504	0.398	0.191	0.280	0.552	0.355	0.353	0.449	0.673	0.323	0.380	0.411	0.134
det	1.553	2.607	0.665	1.068	2.344	2.362	1.423	2.116	2.162	2.089	2.240	1.648	1.875	0.611
dobj	0.987	1.050	0.538	0.734	1.246	1.064	1.171	1.079	0.811	0.861	0.798	0.879	0.940	0.211
iobj	0.017	0.008	0.005	0.014	0.009	0.021	0.027	0.023	0.007	0.003	0.005	0.012	0.013	0.008
nmod: of	0.351	0.816	0.135	0.245	0.753	0.748	0.489	0.579	0.783	0.948	0.641	0.491	0.590	0.258
nsubj	1.820	1.770	1.287	1.597	1.571	2.060	1.443	1.784	1.487	1.317	1.327	1.560	1.587	0.247
nsubjpass	0.118	0.293	0.045	0.051	0.209	0.222	0.091	0.197	0.221	0.445	0.202	0.152	0.190	0.116
xcomp	0.369	0.407	0.235	0.257	0.373	0.414	0.389	0.363	0.223	0.247	0.244	0.306	0.320	0.078

Table 1. The TPSs of the main dependency types across the subcorpora in MASC 500K used in this study

The name of the subcorpus *Govt* is an abbreviation of government documents, *Nonf* is non-fiction, *Tech* is technical reports, and *Travel* is travel guides. *All* is the whole documents used in this study, and the TPSs in the column *All* are calculated with the sum of each dependency type in *All* divided by the number of the sentences in *All*. The TPSs of *All* are not included when calculating the mean and SD.

Table 1 does not contain all the dependency types and their TPSs; in particular, the subtypes of *nmod* (nominal modification) are excluded, with the exception of *nmod:of* (nominal modification by a prepositional phrase with "of"). This is because the number of *nmods* is too large to include in one table. The subtype *nmod:of* is included in this table because it has the highest TPS compared to the TPSs of other subtypes of *nmod*s.

The following section will discuss the findings in Table 1, focusing on some of the high (or low) TPSs of the dependency types found in the subcorpus.

First, it is obvious that the subcorpus *Ficlet* contains smaller TPSs in many of the main dependency types due to its smaller word per sentence (henceforth WPS) compared to the other subcorpora (the WPS of Ficlet is about 9.8, the WPS of Fiction is about 12.32, and that of other subcorpora are all higher than 18). Shorter sentences contain a smaller number of dependency relationships compared to longer sentences; thus, if a given subcorpus contains a large proportion of short sentences (like the subcorpus Ficlets in this study), the TPSs of each dependency type will be low.

The TPSs of *advmod* (adverbial modification) in Essay (1.139) and Journal (1.209) were found to be higher than the TPSs of the same type in other subcorpora; this could be because adverbs tend to modify the texts in these genres more often than texts in other genres.

The TPS of *amod* (adjectival modification) in technical reports (indicated as Tech in the table) (2.64) is higher than the TPSs of the same type in other subcorpora; the explanation for this could be that the nouns in technical reports need more modification by adjectives. The same explanation can be applied to the TPSs of *compound* in Tech (2.192), which is the highest among the subcorpora; the explanation for this could be that the nouns in technical reports are more likely to require modification through noun compounding.

The TPSs of *ccomp* (clausal complement) do not show any drastic difference across the subcorpora (within the range from 0.2 to 0.46), with the exception of Travel (0.093); this suggests that the sentences in this subcorpus have a lesser tendency to use clausal complements compared to sentences in other genres. This can be partly explained by the fact that the verbs used frequently in travel guides do not take clausal complements because one of the purposes of texts in this genre is to avoid expressing the writer's attitude toward a certain statement by using *that*-clauses, and to describe the tourist spots of interest as attractively and objectively as possible.

The TPSs of *iobj* (indirect object) are small across all the subcorpora, and this indicates that the indirect objects are not used as frequently as other core arguments such as the subject or the direct object.

The TPSs of *nsubjpass* (nominal passive subjects) are low across all the subcorpora; however, we observe that those in Ficlet (0.045) and Fiction (0.051) are smaller than those found in other subcorpora. This fact seems to support the argument that writers of fictional stories (including short ones like those in Ficlets) tend to avoid using the passive voice as compared to writers of other genres. This argument can be verified with more data by employing different methods and using different viewpoints.

4.2 Some apparent correlations between TPSs and Writer intention

This section shows that some TPSs appear to be correlated with each other across the different subcorpora. This section also explicates the logic behind such correlations by examining the writer's possible intention.

Each pair of TPSs can fall into one of the following categories: those that appear positively correlated with each other (e.g., *acl* [adjectival clause] and *acl_relcl* [relative clause]; *amod* and *compound*), those that appear negatively correlated with each other (e.g., *dobj* and *nsubjpass*), and those that are not correlated with each other (e.g., *dobj* and *iobj*, *aux* [auxiliary] and *dep* [undefined dependency]).

Acl and *acl_relcl* are the first example of TPS pairs that appear positively correlated with each

other. In Figure 3 below, the subcorpora are plotted with their TPS of *acl* on the x-axis and their TPS of *acl_relcl* on the y-axis.

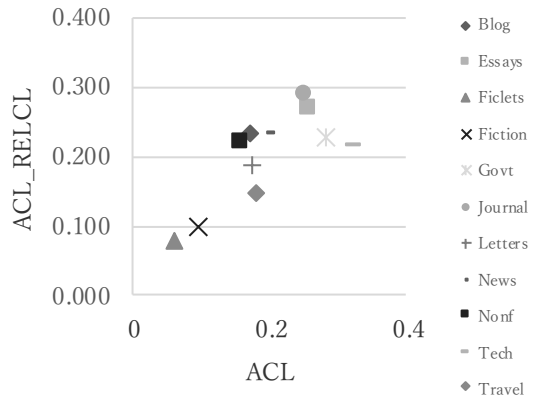

Figure 3. The distribution of TPSs of *acl* and *acl_relcl* across the subcorpora

As the distribution above shows, across the subcorpora, the TPS of the dependency type *acl_relcl* increases in proportion to the TPS of the dependency type *acl*. The type *acl_relcl* is a subtype of *acl*, and it is natural that the TPS of a subtype increases as its metatype increases across the subcorpora.

Amod and *compound* form the second example of TPSs pairs that appear positively correlated with each other. In Figure 4 below, the subcorpora are plotted with their TPS of *amod* on the x-axis and their TPS of *compound* on the y-axis.

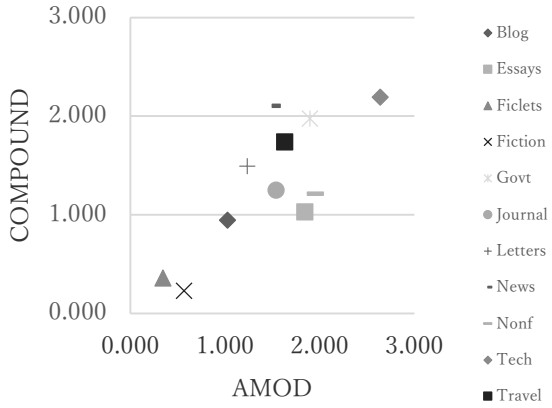

Figure 4. The distribution of TPSs of *amod* and *compound* across the subcorpora

The possible correlation between *amod* and *compound* can be considered a natural result when we take the functions of these two dependency types into consideration; both share the function of modifying nouns, and, if a writer intends to modify more nouns in the text in some way, adjectives and noun compounds are usually able to meet this requirement. As a result, the writer will develop a tendency to use more adjectives and noun compounds. The same logic applies to the pair of *amod* and *det* (in Figure 5).

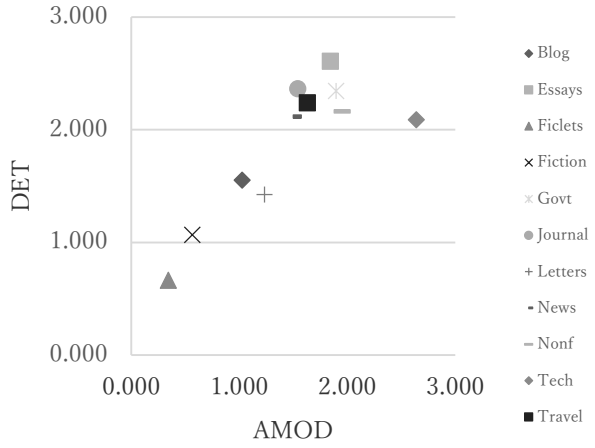

Figure 5. The distribution of TPSs of *amod* and *det* across the subcorpora

In the case of the pair *advcl* and *ccomp* (in Figure 6), the higher usage frequency of these dependency types indicates that the writer intends to modify the clauses in the relevant text with more clauses, and both of these dependency types will meet this requirement.

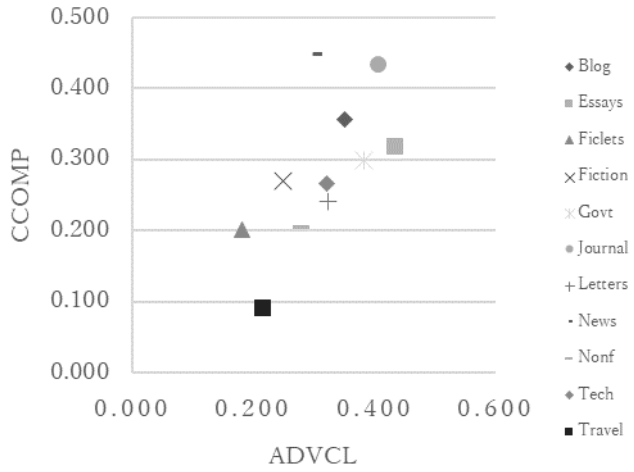

Figure 6. The distribution of TPSs of *advcl* and *ccomp* across the subcorpora

A higher TPS for *advcl* represents the more frequent use of adjuncts, while a higher TPS for *ccomp* represents the more frequent use of verbs that take a clausal complement as their argument; therefore, these two dependency types do not necessarily belong to the same grammatical category. Moreover, as far as clausal complements and adverbial clauses are concerned, it is possible that the distinction between arguments and adjuncts should not be considered as essential as it is conceived to be in the field of syntax. What should be considered essential here is the writer's intention to provide further explanations or modify given clauses; the syntactic elements to realize this intention are clausal complements in some cases and adverbial clauses in others.

In the case of the pair *ccomp* and *xcomp* (in Figure 7), the more frequent use of these dependency types also indicates the writer's intention to express ideas by using verbs that require either clausal complements or *to*-infinitive complements.

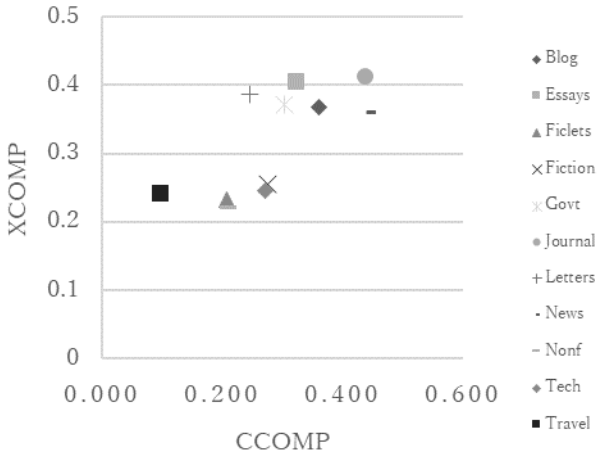

Figure 7. The distribution of TPSs of *ccomp* and *xcomp* across the subcorpora

The graph clearly shows that the distribution of the subcorpora seems to be divided into two groups based on the TPSs of *xcomp*. More corpus data needs to be explored before we can conclude that this is coincidental in the case of MASC 500K or that the TPS of *xcomp* in general can categorize different genres of texts into two groups as indicated in Figure 7.

Not all dependency type pairs show apparent correlations. For example, *aux* and *dep* do not appear to be correlated with each other (Figure 8), and it is difficult to conceive of any factor that both are related to.

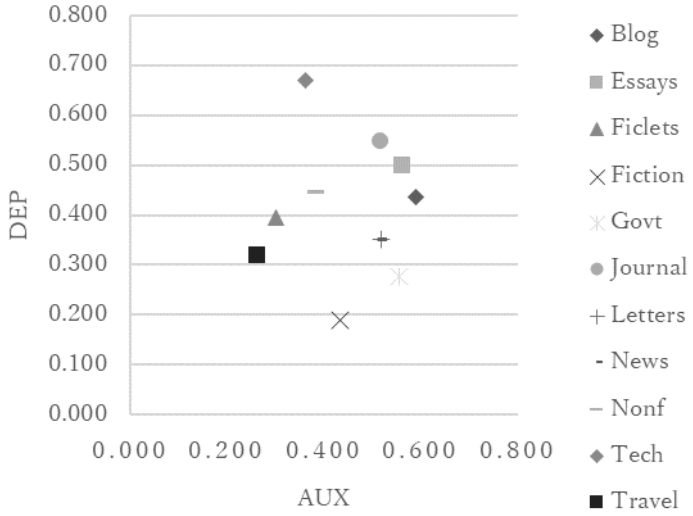

Figure 8. The distribution of the TPSs of *aux* and *dep* across the subcorpora

5. Conclusion

This study utilized the TPS of dependency types, a concept that was first introduced by Oya (2016), to analyze the genres of different English texts obtained from MASC 500K. The study findings show that TPSs can partially explain the differences across different genres of texts. This finding about the TPSs reflects the fact that certain genres of texts are more likely to contain certain dependency types compared to others. In addition, the higher TPSs of some dependency types were shown to be related to the higher TPSs of other dependency types.

This study also suggests a number of new lines of investigation; first, a thorough survey of correlations between all the possible pairs of dependency types should be conducted in order to reveal the structural characteristics that occur across different genres of texts. We have not yet investigated the TPSs of different genres of spoken data, and comparison of the TPSs of spoken data and those of written data is expected to yield interesting results. Lastly, it may be interesting to explore the possibility of employing TPSs as one of the factors for writer identification; that is, TPSs can be used to indicate that Writer A uses more of a certain dependency type than Writer B. In addition, it may be possible to utilize the frequent occurrences of certain dependency types in an unidentified text, along with other factors, to reveal the identity of its writer. All these possibilities can be investigated further in future research studies.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 17K02740.

References

Danqi Chen and Chris Manning. 2014. A Fast and Accurate Dependency Parser Using Neural Networks. Proceedings of EMNLP 2014.

Marie-Catherine De Marneffe, Bill MacCartney, and Chris Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. Proceedings of the Language Resources and Evaluation Conference (LREC) 2006.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford Dependencies: A Cross-Linguistic typology. Proceedings of LREC.

Masanori Oya. 2016. Dependency Types in Learner English and Authentic English. Proceedings of the 21st International Conference of Pan-Pacific Association of Applied Linguistics.

Nancy Ide and Keith Suderman 2004. The American National Corpus First Release. Proceedings of the Fourth Language Resources and Evaluation Conference (LREC), 1681-84.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. MASC: The Manually Annotated Sub-Corpus of American English. Proceedings of the Eighth Language Resources and Evaluation Conference (LREC), 2455-2460.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. International Corpus of Learner English. Version 2. Presses universitaires de Louvain, Louvain-la-Neuve.