A Re-examination of Syntactic Complexity by Investigating the Internal Structure Variations of Adverbial Clauses across Speech and Writing

Mingyu Wan
Dept. of Linguistics and Translation
City University of Hong Kong
mywan4-c@my.city.edu.hk

Alex Chengyu Fang
Dept. of Linguistics and Translation
City University of Hong Kong
acfang@cityu.edu.hk

Abstract

This paper re-examines the debatable issue about adverbial clause (A,CL) whether it indexes a complex discourse of contemporary English and how it distributes across Speech (S) and Writing (W) with different subtypes of structures by investigating the internal structure variations. A Finite-State-Machine model is adopted for retrieving the internal structures. Empirical results show that A,CL prevails in W than in S with a higher occurrence rate (W: 31.20% vs. S: 14.80%), which confirms its function of indexing a complex discourse; but the standard token-type-ratio of its internal structures shows an opposite distribution (W: 12.89% vs. S: 16.66%), which suggests a higher structural density/variation of the spoken mode; besides, five subtypes of internal structures are identified with various distributions across S and W: S employs a higher proportion of subordinator-overt subordinating A,CL, while W adopts more infinitive A,CL, including to-infinitives, present and past participles; coordinated embeddings and subordinator-covert finite A,CL are commonly found in both modes. Despite of the individual variance of internal subtypes, statistical test indicates a less noticed fact that the overall structural variation of A,CL between S and W is not significant (p-value = 0.5245).

1 Introduction

The definition of language complexity is generally quite broad and has not reached much consensus among researchers of various fields, such as in linguistics, cognitive linguistics, machine translation and L2 learning. People may interchangeably refer to it as text readability (Just and Carpenter, 1980; Rayner, 1998; Vajjala *et al.* , 2016; Charles *et al.* , 2007), or language preparedness/elaboration (Fang, 2006; Fang and Cao, 2015), or writing quality (Crossley *et al.* , 2010; Crossley and Danielle, 2014), or language proficiency (Piln *et al.* , 2016; Ortega, 2003; Larsen-Freeman, 2006; Housen and Bram , 2012).

Pallotti (2015) underlines the polysemy of the term complexity in the linguistic literature and summarizes the different notions of complexity in this field by referring to three main meanings:

- Structural complexity, a formal property of texts and linguistic systems having to do with the number of their elements and their relational patterns.
- Cognitive complexity, having to do with the processing costs associated with linguistic structures.
- Developmental complexity, the order in which linguistic structures emerge and are mastered in second (and, possibly, first) language acquisition.

Not only the definitions are debatable, but also the measurements. Biber (1988) did an influential work on reporting 67 linguistic variations across Speech and Writing based on multi-dimensional factor analysis. Heylighen and Dewaele (1999) developed a simple but effective formula to calculate the fuzziness/complexity of language in terms of formality as in"*F=(noun freq.+adjective freq.+preposition*

freq.+article freq.-pronoun freq.-verb freq.-adverb freq.-interjection freq.+100)/2". Lu Xiaofei (2010) developed an automatic analyzer of measuring syntactic complexity based on 14 indexes, such as mean length of clause (MLC), mean length of sentence (MLS), and mean length of T-unit (MLT), with the aim of identifying the measures that best index second language learners' developmental levels. Among the existing studies, the measurements are found either basing on lexico-grammatical features or on syntactic indexes, with the adverbial clauses receiving most argues and debates.

For example, in terms of the distributional proportion of adverbial subordinate clauses in Speech and Writing, Biber (1988) claimed that it is more common to see it occurring in informal language, as he put it "*that*-clauses, *WH*-clauses, and adverbial subordinators co-occur frequently with interpersonal and reduced-content features such as first and second person pronouns, questions, contractions, hedges, and emphatics. These types of subordination occur frequently in spoken genres, both interactional (conversation) and informational (speeches), but they occur relatively infrequently in informational written genres". Similarly and most famously, Halliday (1979; 1985) indicated that speech and writing are both complex systems but in different ways: speech is more complex in terms of sentence structures while writing in terms of high lexical density. He believes that the structural complexity found in speech is characterized by a relatively higher degree of hypotaxis which involves subordination of various kinds such as adverbial clauses. Most recently and notably, Fang (2006) studied adverbial clauses across three sets of complex systems (i.e., speech vs. writing, spontaneous speech vs. prepared speech, and timed vs. untimed essays) and it consistently produced results that are contrary to Biber and Halliday's observations, as he found that the proportion of adverbial clauses are consistently much lower in speech than in writing and that adverbial clauses are a significant characteristic of planned, elaborated discourse.

It is interesting and intriguing to see from the above studies that the completely opposite conclusions are derived from the same category of syntactic feature under the same language system (speech and writing of contemporary English). The reasons for such opponent findings can be manifold: the definition of adverbial clauses might be different in terms of the grammar they are referring to; the base for calculating the proportion of such features are of high chance different, in terms of the data size, the number of words, clauses and sentences, and so on, and it is almost impossible and unscientific to make a direct comparison; moreover, as it is commonly accepted that complex language systems are highly compressed/condensed in terms of both lexical density and structural embeddings, it is not so persuasive to simply consider only the frequency of adverbial clauses without looking into its embedded structural variations. The inner structures might play a crucial rule in rendering the various distribution and can further account for its syntactic behaviors across Speech and Writing.

Therefore, this paper aims to re-examine adverbial clauses in terms of syntactic complexity through the investigation on the internal structure variations across speech and writing, with a refined data analysis and a statistical test. The remaining parts of the paper are organized as follows: in section 2, we will introduce the corpus and the methodology; in section 3, we will show the empirical results with data analysis and discussions; in section 4, we will conclude our work based on the empirical observations and look forward to future work in further defining syntactic complexity.

2 Data and Methodology

2.1 The ICE Corpus

This paper has adopted the Great Britain Component of the International Corpus of English (ICE-GB) as the database. The ICE-GB corpus belongs to the ICE family. The project for building the ICE family was proposed by Sidney Greenbaum in 1988, with the purpose of providing a language resource for comparative studies of English worldwide. The language varieties in ICE refer to the English language used in 24 nations or regions, where English is the first language or an official additional language. All the corpora in the ICE family follow a general design:

- The overall size of each corpus is one million words of English produced after 1989.

- Each corpus consists of 500 texts of about 2000 words each.
- Each corpus covers 300 spoken and 200 written English texts

Like all the other ICE components, ICE-GB comprises 300 spoken and 200 written texts from 32 categories, amounting to one million words. This corpus is not only POS tagged but also fully parsed and manually validated. The following example and its parse tree is shown below to illustrate parsing scheme.

(1) *If you rang her now, she'd say yes Louis.* ⟨S1A-020-138⟩

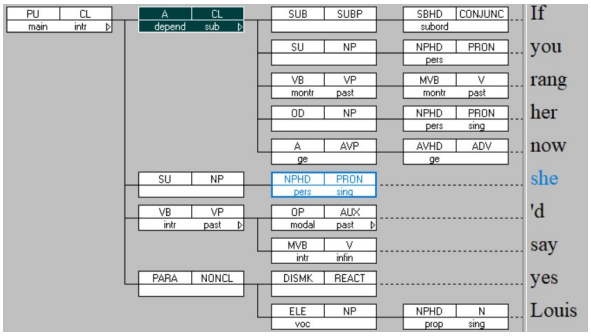

Figure 1: The ICE parse tree for (1).

Each node in the tree is labeled with up to three types of information: word class, syntactic category, syntactic function, as well as grammatical features (see the grammar reference in Quirk (2010)). For example, A,CL represents a clause (CL: syntactic category) with the adverbial syntactic function (A), and it is grammatically regarded as a dependent subordinate clause (*depend* and *sub).* Within the adverbial clause, it is structural realized as a subordinator-headed finite clause with a linear structure of "SUB,SUBP-SU,NP-VB,VP-OD,NP-A,AVP". More embedding structures are made possible because they can all be extracted and converted to such linear forms to capture the comprehensive internal structures of all the adverbial clauses by the Finite-State-Machine derived extraction model (Wan, 2017).

2.2 Method and Tools

In this paper, the methodology is both empirical and statistical, of which the findings are driven by an objective corpus data analysis. The statistical Welch independent t-test, as well as the visualization plots, are implemented in the open source statistical software–R [1] in Rstudio [2].

The ICE-GB has provided well-annotated syntactic and grammatical tags for this work, so the identification of the adverbial clauses (A,CL) is straightforward and unambiguous by the automatic program run on PyCharm Community Edition 2018.1.3[3] with python3.5. The identification and extraction process of the internal structures of adverbial clauses is based on the Finite State Machine (FSM) which is suitable for retrieving the embedding internal structures for any specified syntactic constituent in the corpus. The detailed description to this model can be found in Wan (2017)'s study on using it for preparing syntactic features for automatic genre classification, which will be briefed in the following section.

2.2.1 The FSM-based Extraction Model[4]

The idea of identifying the internal structures of a constituent was inspired by the Finite State Machine model (Selic *et al.* , 1994), which is an abstract machine that can be defined by a list of an initial state, a finite set of changed states and the conditions for each transition at any given time, as in "A FSM is a mathematical model of computation. It is an abstract machine that can be in exactly one of a finite number of states at any given time. The FSM can change from one state to another in response to some external inputs; the change from one state to another is called a transition. An FSM is defined by a list of its states, its initial state, and the conditions for each transition."

The FSM can change from one state to another in response to some external inputs; the change from one state to another is called a transition. The transition patterns of a FSM are found to resemble the structural transition states of a target constituent within a parse tree. The following diagram illustrates how the internal structures of A,CL can be identified and extracted based on the FSM-derived method as shown below.

[1] https://www.r-project.org/
[2] https://www.rstudio.com/
[3] https://www.jetbrains.com/pycharm/
[4] This section is added to describe the method of extracting the internal structures, as concerned by one reviewer.

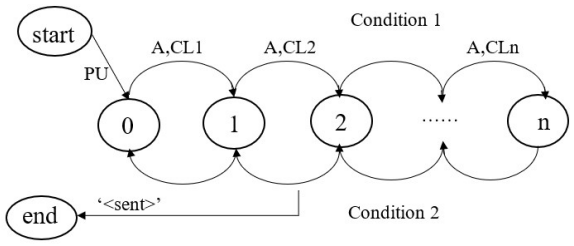

Figure 2: The FSM-based Model for Internal Structure Extraction.

In Figure 2, condition 1 needs to be satisfied when the status increases by one: the current A,CL structure is embedded with a new A,CL structure until reaching a terminal node; condition 2 needs to be satisfied when the status reduces to a certain value: the current node's preceding white spaces are equal or smaller than the latest mother node's. As for which status to jump to, it depends on which previous mother nodes white spaces are closest and smaller than the current nodes. More details of the extraction of the internal structure and the conversion of them into linear structural representations can be described by the following flow chart, with a stimulated status 'i' as an example for illustration:

The flow chart in Figure 3 shows the basic steps of the identification, conversion and construction of structural representations of a target constituent within a status 'i' as an example. The ellipses in Figure 3 means the state jumps out from the current status into a new status (i+1 or 0 to i) and a same process is iteratively conducted. After all the syntactic trees have been searched through, with the status dancing freely among 0 to n, the extracted structural nodes will be converted into linear representations and their frequency distribution will be calculated automatically by using the FreqDist module of NLTK [5]. These frequencies are very important numbers for constructing the feature tables for classification tasks. The linear forms of the internal structures are exemplified in Table 2 below.

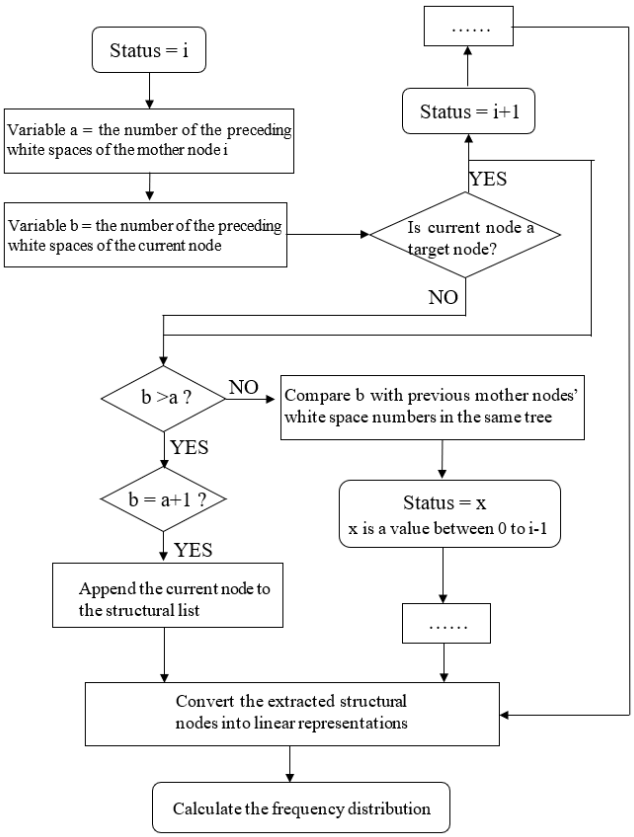

Figure 3: The Flow Chart of Constructing the Internal Structures.

3 Results and Discussions

3.1 General Results

The general statistics about the occurrence of adverbial clauses in the two modes is shown in Table 1 below. All the necessary bases (text, word, and sentences) for normalizing the frequencies are included.

As can be seen in Table 1, the number of texts, words, and sentences between the Speech and Writing modes are different, so it would be unreasonable to compare the figures based on the raw frequencies. The "Token of A,CL", "Type of A,CL" and TTR (token-type-ratio) are hence not much meaningful unless they are divided by the same base.

By contrast, the emphasized numbers (STTR: standardized TTR of A,CL and the calculation method of Fang (2006)) are far more revealing, but they represent two different interpretations: the Fang (2006) index can indicate the overall number of adverbial clauses occurring across speech and writing on the basis of the same amount of sentences and the result generated by using his method in this work conforms to Fang's discovery that adverbial clauses are indeed more prevalent in Writing than Speeching (Wring: 31.20% vs. Speech: 14.80%); but one

[5] https://www.nltk.org/

Items	Speech	Writing
No. of texts	300	200
No. of words	0.6m	0.4m
No. of sentences	59486	23944
Token of A,CL	8778	7485
Type of A,CL	1377	951
TTR of A,CL	15.69%	12.71%
STTR of A,CL	**16.66%**	**12.89%**
Fang (2006)	**14.80%**	**31.20%**

Table 1: General Statistics of A,CL.

important metric is missing in his work, as the STTR is a very useful index for measuring lexical density, it could also be applicably used to index the density of structural variations.

Only with the measurement of both lexical and structural variations can we better index language complexity. Therefore, the "STTR of A,CL" basically implies that the speech mode is structurally more varied than the writing mode, despite that the distribution proportion of adverbial clauses in writing is higher than in speech, and this finding conforms to Halliday's claim of speech being structural complex. With the two indexes adopted in the paper, the argue between Fang and Biber/Halliday can be well explained in that both of the claims are correct but from two different perspectives (one from distribution proportion, the other from structural density).

On the basis of the generalized observation of the adverbial clauses across speech and writing, this paper looks further into the structural variations and has identified five main types of internal structures, as show in the following section.

3.2 The Five Types of Internal Structures

The five identified types of internal structures of adverbial clauses are shown and exemplified in Table 2 below.

The five structural types are clustered automatically based on the similarity of the linear structural forms of all the adverbial clauses and the exemplified forms are the most frequent linear structure among the five types. It has to be noted that the five types also allow for inner structural variations within each subtype.

As can be seen in Table 2, Type 1 structures are

Structure Types	Representative Linear Forms	Corpus Examples
Type 1	VB,VP-A,PP	"So what's he doing **writing about India**"
Type 2	TO,PRTCL-VB,VP-OD,NP	"So you've got more **to bring to the course**"
Type 3	SUB,SUBP-SU,NP-VB,VP-OD,NP	"Uhm uhm is there a possibility **if I express an interest**"
Type 4	SU,NP-VB,VP-CS,NP	"I am not surprised **I've caught from you being all day here**"
Type 5	CJ,CL-COOR,CONJUNC-CJ,CL	"It is a conference performance **fully specifying the expanded orchestration...and creating a life-enhancing surge...**"

Table 2: The Five Inter-structural Types of Adverbial Clauses.

infinitive adverbial clauses including the present participle (ingp) and past participle (edp) infinitives; Type 2 structures are the to-infinitives; Type 3 are the subordinator-headed finite clauses; Type 4 structures are the subordinator-free finite clauses; and Type 5 structures are coordinated adverbial clauses. The structural distribution of the five types across Speech and Writing is shown in Figure 3 below.

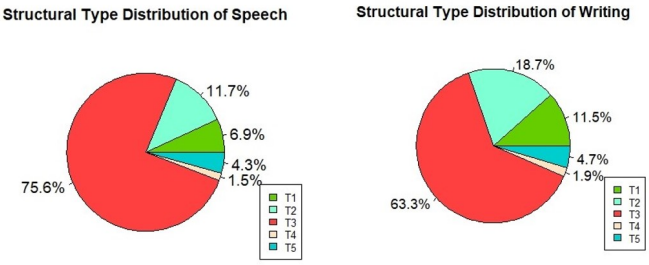

Figure 4: The Structural Distribution of the Five Types across Speech and Writing.

As shown in Figure 4, Type 3 (in red) accounts for the largest proportion of all the structures for

both discourse modes, but the ratio of it in Speech is higher than in Writing (75.6% vs. 63.3%), which basically conforms to Biber's claim of more subordinated adverbial clauses in speech. Type 1 and 2 seem to skew between the two modes with Writing seeing higher proportion of to-infinitives and and participle infinitives than Speech. There is no much difference found for type 4 and 5 structures, which means that the subordinator-free finite clauses and coordinated clauses are basically the same for speech and writing. As a summary, the speech mode tends to adopt more subordinated clauses with overt subordinators; the writing mode tends to adopt more infinitive clauses (including to-infinitives, present and past participles); coordinated embeddings and subordinator-covert finite clauses are commonly found in both discourse types.

3.3 The Statistical Test

The above section showed the syntactic variations of adverbial clauses across speech and writing specific to certain structural types, but it is still not clear yet if the overall structural variance of all the adverbial clauses is significant or not. And this part is statistically important and empirically necessary since it can provide more persuasive evidence for the investigation of the adverbial clausal behavior in different complex language systems. Therefore, in this section, the Welch independent t-test is conducted to find out the statistical difference of adverbial clauses between speech and writing.

The boxplot (Figure 5) below is used to display the original internal structure variance of the adverbials between the two modes from a glance of the data.

From a general view of the boxplot in Figure 5, the mean value of the structural occurrences of the adverbial clauses in the two discourse modes look very similar and the variances of the structures seem also not to be very obvious. The result of the Welch t-test shows that the p-value is 0.5245, which is larger than 0.05. It means H0 hypothesis is not rejected, which implies that the structural variances of adverbial clauses across speech and writing are not significant. The statistical result is straightforward and persuasive to reveal a rarely noticed fact that the adverbial clauses are actually not very much different in terms of its structural complexity across

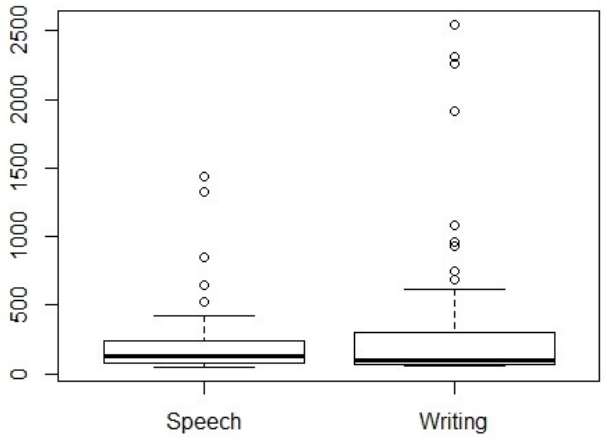

Figure 5: The Boxplot of the Structural Types across Speech and Writing.

speech and writing, but it shows more variations for certain sub-types, such as the subordinator-overt depending clauses as well as the infinitive clauses.

4 Conclusion[6]

In this paper, we have re-examined a debatable topic about adverbial clauses (A,CL) in terms of its index of complex discourse of contemporary English, and has taken an innovative step compared to Biber, Halliday and Fang's work, by studying the internal structure variation of A,CL across speech and writing. The Finite-State-Machine model is adopted for retrieving such internal structures. The ICE-GB corpus is used as the database.

Empirical results show that A,CL prevails in Writing (W) than in Speech (S) with a higher occurrence rate (W: 31.20% vs. S: 14.80%), which confirms its function of indexing a complex discourse; but the standard token-type-ratio of its internal structures shows an opposite distribution (W: 12.89% vs. S: 16.66%), which suggests a higher structural density/variation of the spoken mode; besides, five subtypes of internal structures are identified with various distributions across S and W: S employs a higher proportion of subordinator-overt subordinating A,CL, while W adopts more infinitive A,CL, including to-infinitives, present and past participles;

[6]In response to reviewers' suggestions, we have further developed the conclusion with summary on achievements and discoveries; besides, we have also elaborated the future direction with further steps to take on such a debatable topic.

coordinated embeddings and subordinator-covert finite A,CL are commonly found in both modes. Despite of the individual variance of internal subtypes, statistical test reveals a less noticed fact that the overall structural variation of A,CL between S and W is not significant (p-value=0.5245).

This work has not only answered to the main debatable arguments about adverbial clauses among Biber, Halliday and Fang, but also has discovered that complex language system, such as Writing, may show a higher lexical diversity but a lower structural density. This implies that complex language systems are actually more rigid and templated in terms of structural composition. But this finding needs further verification with more work on other clausal types in future, such as CS,CL (subject complement clause), NPPO,CL (noun phrase postmodifying clause) and CO,CL (object complement clause).

References

Biber D. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge, UK.

Charles J., Adrian S. and Keith R. 2007. *Eye movement research: A window on mind and brain, chapter Eye movements in reading words and sentences*. Oxford:Elsevier Ltd., pages 341–372.

Crossley S. A., Cai Z. and McNamara D. S. 2010. *Syntagmatic, Paradigmatic, and Automatic N-gram Approaches to Assessing Essay Quality*. In G. M. Youngblood and P. M. McCarthy (Eds.), Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference, Palo Alto, California, pp. 214–219. The AAAI Press.

Crossley S. A. and McNamara D. S. 2014. *Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners*. Journal of Second Language Writing. 26:66–79

Fang C. 2006. *A corpus-based empirical account of adverbial clauses across speech and writing in contemporary British English. Advances in Natural Language Processing*, Springer, Berlin, Heidelberg, pp. 32–43.

Fang C. and Cao J. 2015. *Text genres and registers: The computation of linguistic features*. Springer.

Halliday M.A.K. 1979. *Differences between Spoken and Written Language: Some Implications for Literacy Teaching*. Proc. 4th Australian reading conference, Adelaide, Australia, Vol. 2:37–52.

Halliday M.A.K. 1985. *Spoken and Written Language*. Victoria: Keakin University Press.

Heylighen F., and Dewaele J. M. 1999. *Formality of language: definition, measurement and behavioral determinants*. Interner Bericht, Center Leo Apostel, Vrije Universiteit Brssel.

Housen A. and Bram B. 2012. *Defining and operationalising l2 complexity*. In Alex Housen, Folkert Kuiken, and Ineke Vedder, editors, Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA, pages 21–46. John Benjamins, Amsterdam.

Just M.A. and Carpenter P.A. 1980. *A theory of reading: From eye fixations to comprehension* . Psychological Review, 87:329–355.

Larsen-Freeman D. 2006. *The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English*. Applied Linguistics, 27(4):590–619.

Lu X. 2010. *Automatic analysis of syntactic complexity in second language writing*. *International journal of corpus linguistics*, 15(4):474–496.

Ortega L. 2003. *Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing*. Applied Linguistics, 24(4):492–518.

Pallotti G. 2015. *A simple view of linguistic complexity*. Second Language Research, 31:117–134.

Piln I., Alfter D. and Volodina E. 2016. *Coursebook Texts as a Helping Hand for Classifying Linguistic Complexity in Language Learners' Writings*. Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), pp. 120–126.

Quirk Q. 2010. *A comprehensive grammar of the English language*. Pearson Education India.

Rayner K. 1998. *Eye movements in reading and information processing: 20 years of research*. Psychological Bulletin, 124:372–422.

Selic B. 1994. *Real-time object-oriented modeling*. New York: John Wiley and Sons.

Vajjala S., Meurers M., Eitel A. and Scheiter K. 2016. *Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts*. Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), pp. 32–37.

Wan M. 2017. *The Application of Fine-gained Syntactic Features to Automatic Genre Classification*. Ph.D thesis, City University of Hong Kong.