University of Ljubljana
FACULTY OF ARTS

Acta Linguistica Asiatica

Volume 10, Issue 2, 2020

ACTA LINGUISTICA ASIATICA
Volume 10, Issue 2, 2020

Published by: Znanstvena založba Filozofske fakultete Univerze v Ljubljani
(Ljubljana University Press, Faculty of Arts)

Issued by: Department of Asian Studies

For the publisher: Dr. Roman Kuhar, Dean of the Faculty of Arts

Journal's web page:
http://revije.ff.uni-lj.si/ala/
The journal is published in the scope of Open Journal Systems

ISSN: 2232-3317

Abstracting and Indexing Services:
Scopus, COBISS, dLib, Directory of Open Access Journals, MLA International Bibliography,
Open J-Gate, Google Scholar and ERIH PLUS.

Publication is free of charge.

Address:
University of Ljubljana, Faculty of Arts
Department of Asian Studies
Aškerčeva 2, SI-1000 Ljubljana, Slovenia

E-mail: nina.golob@ff.uni-lj.si

TABLE OF CONTENTS

RESEARCH ARTICLES

FOREWORD

In these strange days of a limited physical and social contact due to the worldwide pandemic we are especially grateful for the existence of the parallel virtual world, which goes beyond human shortcomings. Our work continued without any obstructions and we are pleased to announce the summer ALA issue of the year 2020. In it we offer six research articles that extend over a broad linguistic area and include languages of the far East Asia, namely Mandarin Chinese, Japanese, Korean, and Vietnamese.

The issue opens with the **WU Jiayi**'s article "Contextual Conditions and Constraints in Chinese Dangling Topics: Syntax-Discourse Interface Analysis", in which the author revisits dangling topics in Mandarin Chinese from the semantic and syntactic view, and evolves their findings to the hypothesis concerning language typology.

The second article on Chinese is **Tina ČOK**'s "Lexical Aspect Classification for Unrelated Languages: A Case Study on Slovenian and Chinese Lexical Aspect", in which the author analyzes Chinese and Slovenian verb aspect to show that deeper cognitive differences effect our perception of reality, and upon her findings proposes an upgraded general classification of verb types.

The following article entitled "The New Chinese Corpus of Literary Texts Litchi" by **Mateja PETROVČIČ**, **Radovan GARABÍK**, and **Ľuboš GAJDOŠ** presents a newly launched corpus of Chinese literary texts Litchi, and exemplifies the variety of its benefits.

Furthermore, **Petra JAKLIN** "The Many Meanings of the Japanese Causative: Widening the Pragmatic Take on the -(sa)seru Causative Sentence" is an article in which the author revisits the range of possible interpretations and meanings carried by Japanese causative sentences, and supports her conclusions with comparisons to English and Croatian structures.

HWANG Yoong Hee's article "Normative Forms and Synthetic Structure of Japanese in the Incubation Period of L2: Subject to Sentence-final Forms in Longitudinal Discourse Data of Korean Returnee Sisters' Japanese" focuses on L2 Japanese sentence-final forms and their change mechanism in case of Korean returnees.

Last but not least, "Examining the Part-of-speech Features in Assessing the Readability of Vietnamese Texts" is an article by **An-Vinh LUONG**, **Diep NGUYEN**, and **Dien DINH** that discusses the present state of research on text readability in Vietnamese and proposes an improved model on estimating readability of texts and consequently their classification.

Editors and Editorial board wish the regular and new readers of the ALA journal a pleasant read full of inspiration.

Editors

EDITORS' NOTE REGARDING LI (2016) AND BROSIG (2009)

On 13 June 2020 we have received Brosig's email stating that the examples (1a), (1b) and (2) in paper

> Li, W. (2016). Adjective distribution in Modern Mongolian. *Acta Linguistica Asiatica, 6*(2), 9-22.

resemble examples (20), (59) and (60) of the paper

> Brosig, B. (2009). Depictives and resultatives in Modern Khalkh Mongolian. *Hokkaidō gengo bunka kenkyū*, 7, 71-101.

The ALA journal invites all its readers to pay attention to both of the above papers for their reference.

Editors

RESEARCH ARTICLES

CONTEXTUAL CONDITIONS AND CONSTRAINTS IN CHINESE DANGLING TOPICS: SYNTAX-DISCOURSE INTERFACE ANALYSIS

WU Jiayi
Fu Jen Catholic University, Taiwan
wujiayi.fju@gmail.com

Abstract

This study verifies the aboutness condition proposed by Chao (1968), Chafe (1976), Li and Thompson (1981) Xu and Langendoen (1985) and many others as a relation that holds between sentence-initial NPs and comment clauses of alleged gapless topic-comment constructions in Mandarin Chinese. Four often-cited types of dangling topics in Chinese are revisited and the arguments from both the semantic (Pan & Hu 2002, 2008, 2009) and syntactic (Shi, 2000; Huang & Ting, 2006) views are examined. A critical scrutiny from contextual perspective, in particular Nomi Erteschik-Shir's (2007) notions of topic, reveals that there exist dangling topics in Chinese where the sentence-initial NPs at issue are either contrastive or old topics which need not be syntactically or semantically licensed. The contextual contrast set and commentative conditions are proposed as the licensing conditions with constraints dictating that proper selection and relevance must be observed. At the end the study raises a hypothesis regarding language typology. It is proposed that discourse configurational languages exhibit phenomena of dangling NP topics which are properly licensed by the aboutness relation under which either one of the contextual conditions is satisfied.

Keywords: dangling topics; aboutness relation; licensing conditions; language typology

Povzetek

Ta študija preverja pogoj »govorjenja o« (angl. aboutness), ki ga predlagajo Chao (1968), Chafe (1976), Li in Thompson (1981) Xu in Langendoen (1985) in drugi kot povezavo med samostalniškimi frazami na začetku stavkov in komentarji domnevno neprekinjenih konstrukcij tema – komentar v mandarinščini. Članek ponovno prouči pogosto obravnavane štiri vrste visečih tem v kitajščini in pregleda tako pomenski (Pan & Hu 2002, 2008, 2009) kot tudi skladenjski vidik (Shi, 2000; Huang & Ting, 2006). Kritični pregled z vidika konteksta, še zlasti opredelitev pojma teme v Nomi Erteschik-Shir (2007), razkriva, da v kitajščini obstajajo viseče teme, pri katerih so sporni začetki stavkov bodisi primerjalni bodisi stare teme, ki so pomensko oziroma skladenjsko pogojene. Kontekstualne primerjave in pogoji ocenjevanja so predlagani kot licenčni pogoji z omejitvami, ki narekujejo, da je treba upoštevati pravilno izbiro in ustreznost. Na koncu študija postavlja hipotezo o tipologiji jezika in poda predlog, da v jezikih konfiguracijskega diskurza pride do pojava visečih tem v obliki samostalniških fraz, ki jih pravilno pogojuje »govorjenje o«, s čimer je izpolnjen eden od od kontekstualnih pogojev.

Ključne besede: viseče teme; odnosi pri »govorjenju o«; pritrditveni pogoji; jezikovna tipologija

Acta Linguistica Asiatica, 10(2), 2020.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.10.2.9-48

1 Introduction

Dangling topics in Mandarin Chinese have drawn considerable attention in recent decades. Six type of utterances presented below have often been cited and debated as to the questions of whether there exist dangling topics in Chinese and how they are properly licensed (Shi, 2000; Pan & Hu, 2002; Huang & Ting, 2006).

(1) 他們 大魚吃小魚。
Tamen da-yu chi xiao-yu.
They big-fish eat small-fish
'They act according to the law of the jungle.'

(2) 他們 誰都不來。
Tamen shei dou bu lai.
They who all not come
'They (topic), none of them are coming.'

(3) 那場火 幸虧 消防隊來得快。
Na-chang huo xingkui xiaofang-dui lai-de-kuai.
that-CL fire fortunately fire-brigade come-DE-fast
'As for that fire, fortunately the fire brigade came quickly; (otherwise)…'

(4) 這件事情 你不能 光麻煩一個人。
Zhe-jian shiqing ni bu neng guang mafan yi-ge ren.
this-CL matter you not can only bother one-CL person
'This matter (topic), you can't just bother one person.'

(5) 那種豆子 一斤 三十塊錢。
Na-zhong douzi yi-jin san-shi-kuai qian.
that-CL bean one-catty 30 CL money
'That kind of bean (topic), one catty is thirty dollars.'

(6) 物價 紐約 最貴。
Wu-jia Niuyue zui gui
thing-price New-York most expensive
'Speaking of the price of things, New York is the most expensive.'

1.1 Literature review

To begin with, the alleged gapless constructions above are often interpreted as dangling topics by aboutness relation: the comment clause says something about the topic (Chao, 1968; Chafe, 1976; Li & Thompson, 1981; Xu & Langendoen, 1985).

Another line following semantic approach is a theory of formal semantics proposed by Pan and Hu (2002). A handful of semantic relations are proposed to account for the occurrence of the sentence-initial noun phrases (NPs). Shortly after the semantic analysis, Pan and Hu (2008; 2009) go one step further to propose a semantic-pragmatic account that unifies both the moved and base-generated topics.[1]

Opposite proposals are syntactic views primarily from Shi (2000), and Huang and Ting (2006). On the basis of the evidence in (7), Shi rejects the aboutness relation and proposes a structural account to argue against the semantic view that there exist dangling topics. And in favor of Shi's structural approach Huang and Ting hold that the so-called dangling topic NPs are in fact subjects, NP topics, NP adverbials, and prepositional phrase (PP)-reduced forms.

(7) a. *這件大事 我知道 張校長辭職了。
*Zhe-jian da-shi wo zhidao Zhang-Xiaozhang cizhi le.
this-CL big-issue I know Zhang-Principal resign ASP
'As for this big issue, I know that Principal Zhang has resigned.'

b. 這件大事 就是 張校長辭職了。
Zhe-jian da-shi jiu shi Zhang-Xiaozhang cizhi le.
this-CL big-issue exactly be Zhang-Principal resign ASP this-CL big-issue
'This big issue is that Principal Zhang has resigned.'

1.2 Motivation, Methodology, and Purpose

The motive for conducting this research arises as follows. First, arguments raised within previous literature, be it the syntactic or semantic account of Chinese topics, have their share of problems. Second, the issue has rarely been looked at from contextual perspective. They suffer from difficulty and controversy in identifying topics solely based on the traditional criterion of (in)definiteness. Some of them support the existence of dangling topics in Chinese without a holistic justification that explains all the possible examples, while others reject the existence based on ad hoc assumptions without much contextual consideration. Third, the puzzle of whether the aboutness relation is a valid condition under which the so-called dangling topics can be properly licensed remains open for debates. This paper thus aims to explore the contextual licensing conditions in Chinese and what it implies about language typology.

[1] In their paper, a semantic-pragmatic interface account of dangling topics in Mandarin Chinese, Pan and Hu (2008) argue that topic structures be accounted for at the semantic–pragmatic interface in Mandarin Chinese. Specifically, in Chinese a topic is licensed if there is a variable in the comment and the set generated by this variable produces a non-empty set when intersecting with the set represented by the topic.

This paper adopts a syntax-discourse analysis approach. The notions of topic are primarily drawn from the work of Nomi Erteschik-Shir (2007), *Information Structure: Syntax-Discourse Interface*. Her definitions of topic are relevant and instrumental in shedding light on the puzzles when identifying the Chinese topics in question throughout this paper. As a subfield of linguistics dealing with how sentences are structured with respect to topic and focus, Information structure (IS for short) fares better in identifying topics since it put topics in a broader context.

Topics are given or old, and, as Erteschik-Shir argues, are identified by context.[2] Pronouns, defintes, specific indefinites, generics, and contrastive elements qualify as topics.[3] All sentences must have a topic and a sentence can have multiple topics. They can occur anywhere in a sentence. This paper argues that the initial NPs at issue are dangling topics because they are not only what the utterances are about in discourse but also occupy syntactically in left-peripheral position higher than subject. Specifically, they are either contrastive or old topics that dangle syntactically or semantically unrelated to the following clauses. The relation is simply pragmatic.

Four of the often-debated gapless topics in Chinese, specifically the material/instrument type, as in (4), the nominal-predicate type, as in (5), the event type, as in (3), and finally the (non-)idiom chunk type, as in (1), will be re-visited respectively in section 2, 3, 4, and 5. A rich variety of examples in Chinese will be provided to challenge the previous semantic and syntactic views. And a pragmatic hypothesis will be postulated. There two context-induced licensing conditions that guarantee the well-formedness of dangling topics in Chinese and their constraints will be investigated in section 6 and 7. Specifically, they are contrast set and commentative conditions with constraints dictating that proper selection and relevance have to be observed. Finally, the allegedly vague aboutness relation will thus be justified and claimed cross-linguistically as a valid licensing mechanism under which either one of the contextual conditions is satisfied.

[2] Erteschik-Shir argues that the choice of topic is context dependent. Consider the examples below:

(1) Q: What happened?
A: John washed the dishes.
(2) Q: What happened to the dishes?
A: John washed them.
(3) Q: Why does John look so pleased with himself?
A: He washed the dishes.

The question in (1) forces a reading in which the sentence must be predicated of a stage topic, the current spatio-temporal parameter of the sentence. The question in (2) not only refers to "the dishes" but also asks what happened to them. In (3), "John" is the only given element in the answer, hence the topic.

[3] Erteschik-Shir provides examples of Danish topicalization to exhibit readers a variety of topics. Readers can refer to page 8 of the book.

2 The Material/instrument type

This section deals with the examples, as in (8), which are analyzed as dangling topics by Pan and Hu, whereas sentences involving PP-reduction by Shi (2002) and then NP-movement by Huang and Ting (2006). Their analyses are not without their problems. A careful examination reveals that the sentence-initial NP in (8b) is in fact a contrastive topic that need not be syntactically licensed.

(8) a. 這件事情 你不能 光麻煩一個人。
Zhe-jian shiqing ni bu neng guang mafan yi-ge ren.
this-CL matter you not can only bother one-CL person
'This matter (topic), you can't just bother one person.'

b. 西紅柿 我 炒了雞蛋。
Xihongshi wo chao (le) ji-dan.
Tomato I fry ASP chicken-egg
'I fried eggs with tomatoes.'

2.1 The semantic account

Standing in opposition to Shi's preposition-dropping analysis, Pan and Hu postulate an asymmetry between (9b) and (9c) and suggest that the difference between them lies in whether the initial NP is subcategorized.

(9) a. 為這件事情 你不能 光麻煩一個人。
Wei zhe-jian shiqing ni bu neng guang mafan yi-ge ren.
for this-CL matter you not can only bother one-CL person
'For this matter, you cannot just bother one person.'

b. 這件事情 你不能 光麻煩一個人。
Zhe-jian shiqing ni bu neng guang mafan yi-ge ren.
this-CL matter you not can only bother one-CL person
'This matter (topic), you can't just bother one person.'

c. 為這件事情 張三 打架了。
Wei zhe-jian shiqing Zhangsan dajia le.
for this-CL matter Zhangsan fight ASP
'For this matter, Zhangsan fought.'

d. *這件事情 張三 打架了。
*Zhe-jian shiqing Zhangsan dajia le.
this-CL matter Zhangsan fight ASP
'For this matter, Zhangsan fought.'

Arguing that the initial NP in (9b) is subcategorized by the verb *mafan* 'bother' in the thematic structure, Pan and Hu claim that the subcategorized NP is a dangling topic semantically licensed by bearing the theta role of Instrument or Material of the predicate. Conversely, the initial NP in (9d) cannot be a dangling topic since it is not subcategorized and semantically licensed by the predicate *dajia* 'fight'. The analysis prompts Pan and Hu to further suggest that the sentence in (8b) is also a construction of dangling topic. Nevertheless, their semantic analysis incurs a fundamental problem that, given subcategorized, the initial NPs in (8) should have syntactic relations with their predicates and hence should not be considered as dangling topics per se.

2.2 The syntactic account

Following Shi's structural perspective, Huang and Ting propose an NP-movement solution. However, their analysis only half solves the issue.

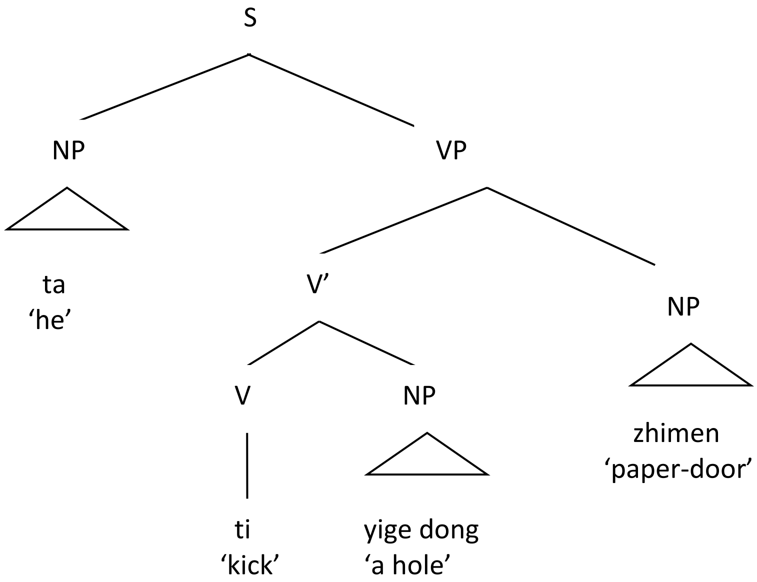
Providing sentences such as (10), they first argue that (8a) involves a double-object construction.

(10) 我 想麻煩你 一件事。
Wo xiang mafan ni yi-jian shi.
I want bother you one-CL matter
'I want to bother you with something.'

The initial NP *zhejian shiqing* 'this matter' in (8a) is analyzed as moving from the direct-object position subcategorized by the three-place predicate *mafan* 'bother', and therefore should not be treated as a dangling topic. We concur with the NP movement analysis to (8a) thus far.

To further justify that (8b) also involves movement, they adopt Huang's (1982) complex predicate analysis of *ba*-construction in Chinese, as exemplified in (11), and reason that the initial NP in (8b) is in fact syntactically licensed by being the outer complement of the complex predicate *chao jidan* 'fry egg'.

(11) a. 他 [把紙門] 踢了一個洞。
Ta [ba zhi-men] ti le yi-ge dong.
He [BA paper-door] kick ASP one-CL hole
‘He kicked a hole in the paper-door.’
b.



This account that analogizes (8b) to the complex predicate structure is compelling, especially when we realize that both *xihongshi* ‘tomato’ and *ji-dan* ‘chicken-egg’ in (8b) are roles of material which can equally be the objects of the predicate *chao* ‘fry’. One more similar example is given below:

(12) 玉米 我 炒(了)火腿。
Yumi wo chao (le) huotui.
Corn I fry ASP ham
‘I fried hams with corns.’

However, challenges emerge when it comes to the example in (13) where the sentence-initial NP is not role of material but that of instrument.

(13) 平底鍋 我 炒(了)青菜。
Pingdi-guo$_{\text{Instrument}}$ wo chao (le) qingcai.
flat-pan I fry ASP vegetable
‘I fried vegetable with flat-pan.’

Imagine that the sentence can be uttered when a speaker was asked to list the instruments he/she had used when preparing a meal. Note that *pingdi-guo* ‘flat-pan’ here cannot be the object of the predicate *chao* 'fry'. This suggests that the extension of the complex predicate analysis to (13) would not be a satisfactory state of affairs. There should be a unified solution to accommodate both cases.

2.3 The pragmatic proposal

Following a similar line to Erteschik-Shir's notion of contrastive topic which indicates that contrast is contextually constrained to occur only if a contrast set is available,[4] we argue that prior to the utterances of (8b), (12) and (13) there should be contrast sets discoursually available. In other words, the initial NP is a contrastive dangling topic licensed by a pragmatic condition requiring that one or each member of a contrast set be selected. Take (14) for further illustration.

(14) a. 烤箱 我 烤(了)雞。
Kaoxiang wo kao (le) ji.
Oven I bake ASP chicken
'I baked chicken with oven.'

b. 電鍋 我 燉(了)肉。
Dian-guo wo dun (le) rou.
Electric-cooker I stew ASP meat
'I stewed meat with electric cooker.'

The sentences above can be uttered when an apprentice cook was asked by his/her chef to clarify how he/she prepared a meal using the instruments (utensils) available in the kitchen.

Adopting Erteschik-Shir's notion of contrastive topic, we are therefore convinced that the sentences in (8b) and (12) involve a contrast set concerning the topic of culinary material, as represented in (15). And the sentences in (13) and (14) involve a contrast set regarding culinary instrument, as in (16). The occurrences of the initial NP are apparently not syntactically but pragmatically licensed.

[4] Erteschik-Shir explained that this is the case both when the focus is contrasted as in (1) and when the topic is contrasted as in (2).

(1) a. Which laundry did John wash, the white, or the colored?
b. He washed the white laundry.
(2) a. Tell me about your brothers John and Bill.
b. John is the smart one.

In both responses, one member of the contrast set provided in the context is selected. In (1b), the contrasted element answers a wh-question and is therefore a contrasted focus. In (2b), the contrasted element is diagnosed as a topic because it is one member of the topic set in the discourse.

(15) [Material$_1$, Material$_2$, Material$_3$,…] Contrast set: the topic of culinary material

(16) [Instrument$_1$, Instrument$_2$, Instrument$_3$,…] Contrast set: the topic of culinary instrument

To sum up, some might adopt a syntactic recovery approach and speculate that the material or instrument NPs in the sentence-initial position are syntactically licensed because they can be found in the recovered sentences. We disagree with this theory because to recover those NPs one has to employ additional phases and often ends up with anomalous sentences that people do not usually use. And the recovered sentences do not necessarily guarantee the availability of the un-recovered sentences. For example, both instrument and location NPs can be found in a recovered sentence but it is instrument rather than location that is able to be syntactically licensed in the sentence initial position. And some might even adopt a semantic role approach and argue that they are semantically licensed because both material and instrument are semantic roles denoted by the predicates. Again, the question boils down to why other adjunct roles such as location cannot be licensed in the same position. The answer is clearly a pragmatic requirement and only material and instrument can be contrastive in such context.

3 The nominal-predicate Type

This section addresses another type of topic which involves quantity phrase (QP), as in (5). A discreet review of the problems from both the semantic and syntactic sides indicates that the initial NPs of this type are dangling topics which also involve the pragmatic requirement that one or each member of a contextually available contrast set be selected for evaluation.

3.1 The semantic account

Disagreeing with Shi's (2000) structural analysis that the sentence-initial NP of the so-called nominal-predicate construction in (17) is a subject, Pan and Hu (2002) provide the unacceptability of the sentence in (18) and contend that the initial NP in (17) must be a topic on the basis of a well-known criterion that a topic must be definite while a subject need not be (Li & Thompson, 1979/1981; Tan, 1990; Shyu, 1995). The analysis leads them to further suggest a set-member account[5] that the NP at issue is a dangling topic semantically licensed by the set-member relation between the NP and the QP *yijin* 'one catty'.

[5] Pan and Hu propose a set-member account in which the dangling topic *nazhong douzi* 'that kind of bean' is analyzed as a set among which *yijin* 'one catty' is delimited as the member of the dangling topic.

(17) 那種豆子 一斤 三十塊錢。
Na-zhong douzi yi-jin san-shi-kuai qian.
that-CL bean one-catty 30 CL money
‘That kind of bean (topic), one catty is thirty dollars.’

(18) *一種豆子 一斤 三十塊錢。
*Yi-zhong douzi yi-jin san-shi-kuai qian.
one-CL bean one-catty 30 CL money
‘One kind of bean (topic), one catty is thirty dollars.’

However, their statement appears unsatisfactory, especially when we consider the sentence in (19) where the initial NP is construed with an existential marker *you* ‘have’. That is, the sentence in (18) is saved when the NP in (18) is construed with the marker such as *you* ‘have’. Along a similar line to Hsin’s (2002) analysis that the initial NP of a sentence in Chinese can be indefinite when preceded by the existential marker,[6] we argue that it is fairly possible that the unacceptability of (18) is simply because it is not preceded by the existential marker *you* ‘have’ to produce an indefinite specific reading.

(19) 有一種豆子 一斤 三十塊錢。
You yi-zhong douzi yi-jin san-shi-kuai qian.
have one-CL bean one-catty 30 CL money
‘One kind of bean (topic), one catty is thirty dollars.’

3.2 The syntactic account

Conversely, in light of the fact that there is a strong tendency for clause-initial NPs in Chinese to be definite,[7] Huang and Ting (2006) argue for Shi’s structural theory and

[6] Hsin notes that, if we assume the indefinite NPs in the predicate is assigned the existential meaning by the VP enclosure, then all indefinite NPs outside VP enclosure would need a syntactic marker to be licensed.

[7] Huang and Ting note that there is a strong tendency for clause-initial NPs in Chinese to be definite. The definiteness constraint on a subject can only be relaxed in a limited number of contexts. For example, when the nominal at issue denotes a quantity (Li 1998) as in (1a), expresses a unit reading (Shi 2000) as in (1b), or occurs with stage-level predicates in a root context (Shyu 1995) as in (1c), they can occur in the grammatical subject position.

propose that the initial NP in (17) is a subject rather than a topic. Citing the unacceptability of (20) without the alleged QP *yijin* 'one catty', they argue that the problem for the unacceptability of (18) lies in the indefiniteness of the subject NP rather than that of the topic NP.

(20) *一種豆子 三十塊錢。
*Yi-zhong douzi san-shi-kuai qian.
one-CL bean 30 CLmoney
'One kind of bean is thirty dollars.'

As they further contend, Pan and Hu's set-member account is in fact untenable because it is not applicable to other similar constructions such as (21). Namely, there is clearly no such a set-member relation between the initial NP *yiliang che* 'one car' and the QP *yitian* 'one day' or *sanqian* 'three thousand (dollars)'.

(21) 一輛車 一天 三千。
Yi-liang che yi-tian san-qian.
one-CL car one-day 3000
'One car is three thousand dollars a day.'

Their challenge to the set-member account based on the evidence in (21) appears perspicacious; however, the argument is not without problems. First, in their analysis it seems that all sentence-initial NPs are subjects regardless of whether they are definite or indefinite. This indicates that there should be no topic at all, contrary to the fact. Topic and subject are not mutually exclusive. It has been widely observed that subjects are unmarked topics across languages (e.g., Li &Thompson 1976; Reinhart 1981; Andersen 1991; Lambrecht 1994; Winkler & Göbbel 2002; Erteschik-Shir 2007). It follows that distinguishing subject from topic, and vice versa, based on the traditional criterion of (in)definiteness is inadequate. Second, the syntactic argument again does not take into account the possibility that the unacceptability of (18) in fact results from the fact that the initial NP is not preceded by the existential marker *you* 'have' either. It should be fairly possible that the reason why (18) is unacceptable is simply because

(1) a. San-ge ren chi liang-wan fan.
three-CL person eat two-bowl rice
'Three people eat two bowls of rice.'
b. Yi-zhi qingwa si-tiao tui.
one-CL frog four-CL leg
'A frog (has) four legs.'
c. Yi-wei yi-sheng xiang wo jieshao ta-de bing-ren.
one-CL doctor toward I introduce he-GEN sick-person
'A doctor introduced me to his patients.'

Therefore, if the sentence-initial NP is a subject, the sentence may be ruled out because the NP does not meet the requirement for an indefinite subject. Please refer to their work "Are There Dangling Topics in Mandarin Chinese?" on page 139 for more details.

it is not construed with the existential marker, rather than because it does not meet the requirement for an indefinite subject.

3.3 The topichood of initial NP

Along Huang's (1984) claim that Chinese is one of the topic-drop languages,[8] we argue that the sentence-initial NPs are in fact topics due to their contextual droppability. This is especially true when we compare the examples without the QPs and those without the initial NPs. As shown in (22-23), the droppability of the initial NPs in (22b) and (23b) entails their topichood. (22a) and (23a) are equally acceptable only when QPs are mutually understood as basic units of measurement and thus are often omitted in such context. A distinction then has to be made that sentence b involves topic drop while sentence a QP drop due to common ground (CG).

(22) a. 那種豆子 (一斤) 三十塊錢。
Na-zhong douzi (yi-jin) san-shi-kuai qian.
that-CL bean one-catty 30 CL money
'That kind of bean (topic) is thirty dollars per catty.'

b. (那種豆子) 一斤 三十塊錢。
(Na-zhong douzi) yi-jin san-shi-kuai qian.
that-CL bean one-catty 30 CLmoney
'One catty is thirty dollars.'

[8] Huang (1984) distinguishes subject drop from topic drop as follows: subject drop is dependent on the availability of rich inflectional agreement morphology; topic drop does not exhibit such a dependency. Instead, topic is recoverable from discourse. Topic trop is illustrated in the Chinese examples in (1) below (*e* stands for the omitted topic pronoun).

(1) a. *e* lai-le.
come-LE
'(He) came.'

b. Lisi hen xihuan *e*.
Lisi very like
'Lisi likes (him) very much.'

c. Zhangsan shuo *e* bu renshi Lisi.
Zhangsan say not know Lisi
'Zhangsan said that he did not know Lisi.'

d. Zhangsan shuo Lisi bu renshi *e*.
Zhanfsan say Lisi not know
'Zhangsan said that Lisi did not know (him).'

(23) a. 一輛車 (一天) 三千。
Yi-liang che (yi-tian) san-qian.
one-CL car one-day 3000
‘One car is three thousand dollars a day.’

b. (一輛車） 一天 三千。
(Yi-liang che) yi-tian 3000.
one-CL car one-day three-thousand
‘One day is three thousand dollars.’

According to the paradigm exemplified in (22-23), we further argue that the sentence-initial NPs at issue are by no means subjects. They are topics because they are not the actual units for the prices to be evaluated. The subjects are the elements elsewhere, such as the QPs themselves or price NPs, e.g. *jia-qian* ‘price’, *ding-jia* ‘fixed price’, etc. The former assumption that the subjects could be the QPs themselves is made from the conversation below:

(24) a. 那種豆子 一斤 三十 兩斤 五十。
Na-zhong douzi yi-jin san-shi liang-jin wu-shi.
that-CL bean one-catty 30 two-catty 50
‘That kind of bean, one catty is thirty dollars and two catties fifty dollars.’

b. 我 要買 一斤。
Wo yao mai yi-jin.
I want buy one-catty
‘I would like to buy one catty.’

Both *yi-jin* ‘one catty’ and *liang-jin* ‘two catties’ are the actual units for selling and buying in (24). In other words, The QPs can be viewed as genuine NPs or quantificational units for real price evaluation while the initial NP *Na-zhong douzi* ‘that kind of bean’ is the topic being talked about.

Another observation is that it is fairly plausible that the subject is price NP such as *jia-qian* ‘price’ which is often omitted in discourse since price negotiation is a CG in such context. Its occurrence is cumbersome. To wit:

(25) a.

鳳梨	一箱		五百塊錢。
Fengli	yi-xiang		wu-bai-kuai qian.
pineapple	one-CL		five-hundred-CL money

‘As for pineapple, one box is five hundred dollars.’

b.

?鳳梨	一箱	價錢	五百塊錢。
?Fengli	yi-xiang	jia-qian	wu-bai-kuai qian.
pineapple	one-CL	price	five-hundred-CL money

‘As for pineapple, One box is five hundred dollars.’

c.

鳳梨	一箱	定價/特價	五百塊錢。
Fengli	yi-xiang	ding-jia/te-jia	wu-bai-kuai qian.
pineapple	one-CL	fixed-price/bargain-price	five-hundred-CL money

‘As for pineapple, the fixed/bargain price is five hundred dollars per box.’

As exemplified in (25b-c), unless it is contextually necessary to make a distinction between a fixed price, *ding-jia* ‘fixed price’ and a bargain price, *te-jia* ‘bargain price’, the price NP is often dropped to avoid redundancy.

Having seen the assumptions concerned with the candidates for subject above, we are convinced at this point that the QPs at issues are the subjects with the price NP *jia-qian* being presupposed. The price NP claims the subjecthood only when there is a particular price to make.

3.4 The contrast set proposal

Taking into account Erteschik-Shir’s notion of contrastive topic we have already elaborated in the previous section, we propose that the sentences at issue are in fact contrastive dangling topics. Specifically, the occurrences of the initial NPs are in fact licensed by the pragmatic condition requiring that one or each member of a contextually available contrast set be selected. Consider the following exchanges.

(26) a.

你們	機車	怎麼賣？
Nimen	jiche	zeme mai?
your	motorbike	how sell

‘How much is your motorcycle?’

b.

? 輕型	(一台)	五萬	重型	(一台)	六萬。
?Qingxing	(yi-tai)	wu wan	zhongxing	(yi-tai)	liu wan.
Light-type	one-CL	50000	heavy-type	one-CL	60000

‘Scooter is fifty while motorcycle is sixty thousand dollars.’

c. ? 三陽機車 (一台) 五萬。
? Sanyang jiche (yi-tai) wu wan.
Sanyang motorbike one-CL 50000
‘Sanyang motorbike is fifty thousand dollars.’

d. ?? (一台) 五萬。
?? (yi-tai) wu wan.
one-CL 50000
‘One motorbike is fifty thousand dollars.’

e. 三陽機車 輕型 一台 五萬；
Sanyang jiche$_{\text{topic1}}$ qingxing$_{\text{topic2}}$ yi-tai$_{\text{subject}}$ wu wan;
Sanyang motorbike light-type one-CL 50000
兩台 特價 七萬。
liang-tai$_{\text{topic3}}$ te-jia$_{\text{subject}}$ qi wan.
two-CL bargain-price 70000
‘As for Sanyang motorbike, scooter is fifty thousand dollars. Two scooters are seven thousand dollars as discount.’

(26a) is a question while (26b-e) are possible responses. (26b-d) are not ungrammatical. They are contextually awkward because the topics for example motor brand in (26b) and motor type in (26c) which are expected to be selected are missing. (26d) is even worse because both topics are absent. (26b-d) can be saved only when the topic selections are carried out either in the responses themselves or in the question (26a) prior to the responses. One the other hand, (26e) is satisfactory because it exhibits multiple topics necessary with the motor brand *Sanyang jiche* ‘Sanyang motorbike’ being the main while the others such as the motor size and quantity secondary. The subscripts indicate the identifications of topic and subject. The multiple topics involved are shown below:

(27) [Motor Brand]$_{\text{topic1}}$; [Motor Type]$_{\text{topic2}}$; [Motor Quantity]$_{\text{topic3}}$

Along a similar line to this analysis, sentence (21) that Huang and Ting use to argue against Pan and Hu’s set-member theory can also be accounted for. That is, in (28) the occurrence of the initial NP in fact forces a contrastive reading, entailing that there exists a contrast topic set regarding quantity for evaluation: the initial NP *yi-liang che* ‘one car’ is a contrastive dangling topic licensed by a contrast set regarding quantity and the quantity *yi-liang che* ‘one car’ is first selected, providing as a basic unit for the hearer(s) to infer the rental values of other quantities. Based on this unit, the speaker continues to utter another sentence in which the quantity *liang-liang che* ‘two cars’ is subsequently selected as a discountable unit. This echoes with Erteschik-Shir’s analysis that even (non-specific) indefinites can be topicalized if they are contrastive. What is old and given here is not a particular quantity of car but rather the contrast set discoursally available.

(28) 一輛車 一天 三千；
Yi-liang che$_{\underline{\text{topic}}}$ yi-tian$_{\underline{\text{subject}}}$ san-qian;
one-CL car one-day 3000
‘One car is three thousand dollars a day.’
兩輛車 (一天) 特價 五千。
liang-liang che$_{\underline{\text{topic}}}$ (yi-tian)$_{\underline{\text{CG}}}$ te-jia$_{\underline{\text{subject}}}$ 5000.
two-CL car (one-day) bargain-price five thousand
‘Two cars are five thousand dollars a day as discount.’

The analysis of contrastive topics in section 2 and 3 above is just to facilitate discussion. Both types of topic are demonstrated to involve topic selection from contrast set. However, it does not matter when and how the contrast set condition is satisfied in discourse. For example, one may easily find seeming counterexamples, as in (29) and (30), and argues that the initial NPs are not necessarily contrastive. They are old topics[9] that can be derived in the discourse. However, these examples do not amount to present a contradiction to our contrast set proposal at all. They only present another context where the contrast set condition can be satisfied. That is, the presence of the NPs which are often dropped in (29b) and (30b) entails the contrast sets discoursally available. The selection requirement of the sets however has been satisfied already in the previous immediate discourse. The presence of the NPs which can be dropped directly without compromising acceptability is simply to repeat and confirm what has been selected. This explains why they are licensed by the condition but less contrastive than the ones discussed extensively above.

(29) a. 玉米呢？
Yumi ne?
corn question-particle (Q)
‘How about the corns?’

b. (玉米) 我 炒(了)火腿。
(Yumi) wo chao (le) huotui.
Corn I fry ASP ham
‘I fried hams with the corns.’

[9] In Erteschik-Shir’s classification of topics, old topics are the referents that must have been mentioned in the immediate discourse, or else they can be derived from a previously mentioned topic.

(30) a. 那種豆子 多少錢？
Na-zhong douzi duoshao qian?
that-CL bean how-much money
'How much is it for that kind of bean?'

b. (那種豆子) 一斤 三十塊錢。
(Na-zhong douzi) yi-jin san-shi-kuai qian.
that-CL bean one-catty 30 CL money
'That kind of bean (topic) is thirty dollars per catty.'

In a nutshell, the initial NPs discussed above are dangling topics properly licensed by the contrast set condition. They are what the sentences are about in discourse and are the loci for further assessment.

4 The event type

This section revisits the event type which is often associated with the connective, *xingkui* 'fortunately'. The often cited is the example below.

(31) 那場火 幸虧 消防隊來得快。
Na-chang huo xingkui xiaofang-dui lai-de-kuai.
that-CL fire fortunately fire-brigade come-DE-fast
'As for that fire, fortunately the fire brigade came quickly; (otherwise)…'

4.1 The debate

In the literature, the initial NPs of this type are analyzed as either syntactically-licensed NPs related to a position in recovered main clause in discourse (Shi, 2000),[10]

[10] Shi claims that the initial NP in such construction is related to a position in the recovered main clause and thus cannot be analyzed as a dangling topic, as exemplified in (1) and (2).

semantically-licensed topic NPs by the cause-effect relation (Pan & Hu, 2002), or NP adverbials which need not be subcategorized (Huang & Ting, 2006). The debate begins with arguing against Shi's recovering analysis by Pan and Hu, and then Huang and Ting.

First, Pan and Hu argue that the recovering analysis is undesirable since the gap of the recovered main clause for the initial NP does not necessarily guarantee the availability of the un-recovered sentence, as exemplified in (32). A semantic account, namely a cause-effect relation, is thus proposed to address the issue.

(32) a. 這件大事 幸虧 張校長來了，
Zhe-jian da shi xingkui Zhang-Xiaozhang lai le,
this-CL big issue fortunately Zhang-Principal come ASP
要不然 我還不知道 如何處理。
yaoburan wo hai bu zhidao ruhe chuli$\emptyset_i$.
otherwise I still NEG know how deal-with
'As for this big issue, fortunately Principal Zhang has come; otherwise I do not know how to deal with it.'

(1) Na-chang huo$_i$ xingkui xiaofang-dui lai-de-kuai,
that-CL fire fortunately fire-brigade come-DE-fast
buran $\varnothing_i$ jiu hui shao-si bu-shao ren.
Otherwise really will burn-die not-few person
'As for that fire, fortunately the fire brigade came quickly,
or (it) would have killed many people.'

(2) Na-chang huo$_i$ xingkui xiaofang-dui lai-de-kuai,
that-CL fire fortunately fire-brigade come-DE-fast
Buran na-ci$_i$ women dou hui shao-si.
Otherwise that-time we all will burn-die
'As for that fire, fortunately the fire brigade came quickly;
otherwise we would all have been burnt to death at that time.'

The sentence-initial NPs *nachang huo* 'that fire' in both (1) and (2) are related to certain positions in the recovered main clauses, so *nachang huo*s 'that fire' are not dangling topics. Shi further continues that if the adverb *xingkui* 'fortunately' in such construction is eliminated, the sentence becomes unrecoverable but is still well-formed.

(3) Na-chang huo$_i$ xiaofang-dui lai-de-kuai,
that-CL fire fire-brigade come-DE-fast
buran $\varnothing_i$ bu-shao ren.
Otherwise burn-die not-few person
'At the time of that fire, the fire brigade came quickly.'

In this case, the initial NP which is not related to any position inside the following clause is a sentential adverbial.

b. *這件大事　幸虧　張校長來了。
*Zhe-jian da shi　xingkui　Zhang-Xiaozhang lai le.
this-CL big issue　fortunately　Zhang-Principal come ASP
‘As for this big issue, fortunately Principal Zhang has come.’

Second, Huang and Ting also consider Shi’s analysis problematic in that there is unnecessarily a gap in the recovered clause for the sentence-initial NP to be syntactically licensed, as illustrated in (33).

(33) 那場火　幸虧　消防隊來得快，
Na-chang huo　xingkui　xiaofang-dui lai-de-kuai,
that-CL fire　fortunately　fire-brigade come-DE-fast
要不然　他　早就死了。
yaoburan　ta　zao jiu si le.
otherwise　he　early really die ASP
‘As for the fire, fortunately the fire brigade came quickly; otherwise he would have been dead.’

Nevertheless, Huang and Ting do not concur with Pan and Hu’s semantic solution because evidence such as (34) suggest that the proposed cause-effect relation does not hold at all.

(34) a. 那場溫布頓網球賽　幸虧　大雨停了。
Na-chang Wenbudun wang-qiu sai　xingkui　da-yu ting le.
that-CL Wimbledon net-ball match　fortunately　big-rain stop ASP
‘As for that Wimbledon match, fortunately the heavy rain stopped.’

b. 台中縣松鶴部落　幸虧　土石流
Taizhong-xian Songhe-buluo　xingkui　tu-shi-liu
Taichung-county Songhe-village　fortunately　soil-stone-slide
已停止氾濫。
yi tingzhi fanlan.
already stop overflow
‘As for the Songhe village in Taichung County, fortunately the landslide already stopped spreading.’

Instead, by analogy to NPs in English such as yesterday, that way, etc., a conclusion is thus made that the initial NPs of this type are NP adverbials.

Thus far, we side with their objection to the semantic solution. However, in what follows, we will argue against any possible structural analyses based on acceptability downgrade. A dangling topic hypothesis will then be analyzed that the initial NPs are in fact nominalized event NPs and, under Erteschik-Shir's framework, old topics. They occur in context where a speaker feels an urge to express a comment relevant to a topic.

4.2 The PP-reduced form account

One structural possibility is to consider the sentence-initial NPs in (31) and those below are PP-reduced form adverbials.

(35) a. 那場比賽 幸虧 大雨停了。
Na-chang bisai xingkui dayu ting le.
that-CL match fortunately big-rain stop ASP
'As for that match, fortunately the heavy rain stopped; (otherwise)…'

b. 那次地震 幸虧 房子堅固。
Na-ci dizhen xingkui fangzi jiangu.
that-CL earthquake fortunately house solid
'As for that earthquake, fortunately the houses were firm and solid; (otherwise)…'

c. 那次搶案 幸虧 行員機警。
Na-ci qiangan xingkui hang-yuan jijing.
that-CL robbery fortunately bank-staff alert
'As for that robbery, fortunately the bank-staff was alert; (otherwise)…'

This assumption is untenable because it would be a brute force deleting the claimed preposition, such as *zai* 'at'. Specifically, when an adverbial marker such as *zai* 'at' is added to all the initial elements in (35), we find that the sentences incur slight infelicity compared to the un-recovered ones, as illustrated in (36). It is the redundancy of the adverbial marker *zai* 'during' that contributes to the acceptability downgrade. Given the undesirable results in (36), it is questionable that the initial NPs are derived from the reduction of the PP adverbials. Put differently, it is suspicious that the sentences in (31) and (35) are derived from the redundancy of the recovered sentences in (36) for such a derivation is not economical.

(36) a. ? 在那場比賽 幸虧 大雨停了。
? Zai na-chang bisai xingkui dayu ting le.
at that-CL match fortunately big-rain stop ASP
'During that match, fortunately the heavy rain stopped; (otherwise)…'

b. ? 在那次地震 幸虧 房子堅固。
? Zai na-ci dizhen xingkui fangzi jiangu.
at that-CL earthquake fortunately house solid
‘During that earthquake, fortunately the houses were firm and solid; (otherwise)…’

c. ? 在那次搶案 幸虧 行員機警。
?Zai na-ci qiangan xingkui hang-yuan jijing.
at that-CL robbery fortunately bank-staff alert
‘During that robbery, fortunately the bank-staff was alert; (otherwise)…’

The untenable of the PP-reduced form analysis is further supported by the examples below in which the connective, *xingkui* ‘fortunately’, is eliminated under the assumption that the sentences without the adverb *xingkui* ‘fortunately’ should be as perfectly acceptable as the un-recovered ones.

(37) a. ?? 在那場比賽 大雨停了。
?? Zai na-chang bisai dayu ting le.
at that-CL match big-rain stop ASP
‘During that match, fortunately the heavy rain stopped; (otherwise)…’

b. ?? 在那次地震 房子堅固。
?? Zai na-ci dizhen fangzi jiangu.
at that-CL earthquake house solid
‘During that earthquake, fortunately the houses were solid; (otherwise)…’

c. ?? 在那次搶案 行員機警。
??Zai na-ci qiangan hang-yuan jijing.
at that-CL robbery bank-staff alert
‘During that robbery, fortunately the bank-staff was alert; (otherwise)…’

When the adverb *xingkui* ‘fortunately’ is dropped, the sentences obtained in (37) are even more downgraded. The sentences in (37) might be acceptable only in a limited contrastive context in which the presence of *zai* ‘at’ is used for locative emphasis.

4.3 The NP adverbial account

Another structural possibility is that the initial NPs are adverbials themselves. However, this analysis does not hold either. Consider the sentences below.

(38) a. *那場比賽 大雨停了。
*Na-chang bisai dayu ting le.
that-CL match big-rain stop ASP
‘During that match, fortunately the heavy rain stopped; (otherwise)…’

b. *那次地震 房子堅固。
*Na-ci dizhen fangzi jiangu.
that-CL earthquake house solid
‘During that earthquake, fortunately the houses were firm and solid; (otherwise)…’

c. *那次搶案 行員機警。
*Na-chang qiangan hang-yuan jijing.
that-CL robbery bank-staff alert
‘During that robbery, fortunately the bank-staff was alert; (otherwise)…’

Again, when the adverb *xingkui* ‘fortunately’ is removed, the sentences obtained in (38) are the least acceptable compared to all the sentences discussed above. The adverbial analysis apparently cannot explain why the sentences without the adverb sound even worst.

4.4 The dangling topic hypothesis

Aside from the issue of acceptability downgrade, there is another critical reason to reject the sentence-initial NPs at issue as PP-reduced forms or NP adverbials. The initial NPs unnecessarily entail the spatio-temporal parameters:

(39) a. 那場比賽 你一定沒有興趣。
Na-chang bisai ni yiding meiyou xingqu.
That-CL match you certainly NEG interest
‘You are certainly not interested in the match.’

b. 那場比賽 李四早就知道結果了。
Na-chang bisai Lisi laozao jiu zhidao jieguo le.
That-CL match Lisi very-early already know result ASP
‘Lisi has already known very early the result of the match.’

The initial NPs above have little to do with the spatio-temporal parameters. They apparently refer to the events per se. They are nominalized event NPs and old topics which have been mentioned in the immediate discourse or can be derived from a previously mentioned topic. They are licensed in context where a speaker feels an urge to express a comment relevant to a topic.

5 The (non-)idiom Chunk Type

This section covers the (non-)idiom chunk dangling topics, as in (40), which are argued first by Shi (2000), and then Huang and Ting (2006) as subject-predicate constructions. Under Erteschik-Shir's framework of topic, we will argue instead that the chunks are fixed clauses predicated of the old topic *tamen* 'they'.

(40) a. 他們 [大魚吃小魚]。
Tamen [da-yu chi xiao-yu].
They big-fish eat small-fish
'They act according to the law of the jungle.'

b. 他們 [我看你，你看我]。
Tamen [wo kan ni, ni kan wo].
They I look you you look me
'They look at each other.'

5.1 Posing the problems

Shi reasons that the idiom chunks are predicates or verbs, given the fact that the preverbal NPs contained in the idiom chunks do not refer to any NP in the sentences. Adverb *zhuanmen* 'specifically' is particularly used to justify the subjecthood of the initial NPs and the predicatehood of the idiom chunks. However, those claims are challenged by Pan and Hu (2006) who provide the following examples below in which the non-idiom chunks have referential subjects and the sentence in (41b) can pass the *zhuanmen* 'specifically' test even though the chunk is obviously a chunk with two full-fledged sentences.

(41) a. 他們 [你指責我不對，我抱怨你不好]。
Tamen [ni zhize wo bu dui, wo baoyuan ni bu hao].
They you blame I NEG right I complain you NEG good
'They blame each other and complain about each other.'

b. 他們倆 專門[小王先來，小李後到]。
Tamen liang zhuanmen [XiaoWang xian lai, XiaoLi hou dao].
they two specifically XiaoWang first come XiaoLi after arrive
'As for them two, XiaoLi always comes after XiaoWang.'

Following Shi's structural approach, Huang and Ting add another aspect marker test, specifically *zai* 'at' test, and argue that the bracketed idioms listed in (40) should be analyzed as predicates that should have undergone a certain kind of reanalysis as unanalyzable chunks, viewed as a complex V-node in the hierarchical representation. Moreover, they go a step further and argue that the examples in (41) are not dangling topics but normal sentences with initial PP-reduced form adverbials: the initial NP *tamen* ‘they’ is reduced from the PP-adverbial *zai tamen zhizhong* ‘among them’.

Appealing as it might be, this account nevertheless has a number of problems. First, the syntactic analysis is again uneconomical to assume because it entails that the sentences in (41) are generated from unnatural sentences, as in (42), and, likewise, the deletion of *zai* ‘at’ and *zhizhong* ‘midst’ would be a brute force.

(42) a. ??在他們之中 [你指責我不對，我抱怨你不好]。
??Zai tamen zhizhong [ni zhize wo bu dui, wo baoyuan ni bu hao].
at they midst you blame I NEG right I complain you NEG good
‘Among them, one blames and complains about another.’

b. ??在他們倆之中 專門[小王先來，小李後到]。
??Zai tamen liang zhizhong zhuanmen [XiaoWang xian lai, XiaoLi hou dao].
at they two midst specifically XiaoWang first come, XiaoLi after arrive
‘Between them, XiaoLi always comes after XiaoWang.’

The second problem concerns the different treatments of syntactic representation between (40b) and (41). It is rather contradictory that the chunk in (40b) should be analyzed as a complex predicate under V-node whereas the ones in (41) as normal sentences. In what follows, the proposal that the chunks are fixed clauses will be reconsidered and the *zai* ‘at’ test, developed by Huang and Ting to support Shi’s subject-predicate viewpoint and argue for their complex V-node analysis, will be refuted. A pragmatic topic-comment structure will be proposed instead to better represent the syntactic structures of the sentences in question.

5.2 The dangling topic hypothesis

In the literature various grammatical functions of Chinese idioms, including subject, predicate, object, determiner, adverbial, complement, etc., have been widely discussed (Li, 1997; Zheng, 2005). Away from the traditional view that analyzes all Chinese idiom chunks as lexicons or phrases, Ma (1998), Wen (2006) and many others propose that some Chinese idioms alone can be sentences themselves. Their viewpoint supposes that there must be at least two kinds of idiom in terms of morphosyntactic characteristics.

One is the predicate type that is verbal in form while the other is the one that is sentential in structure. Both can be predicated of the initial NP *tamen* ‘they’, as exemplified in (43) and (44) respectively.

(43) a. 他們 [以德報怨]。
Tamen [yi de bao yuan].
They use kindness reply resentment
‘They render good for evil.’

b. 他們 [吃裡扒外]。
Tamen [chi li pa wai].
they eat inside claw outside
‘They live on somebody while helping others secretly.’

(44) a. 他們 [毛遂自薦]。
Tamen [Maosui zi-jian].
They Maosui self-introduce
‘They introduce themselves.’

b. 他們 [鶼鰈情深]。
Tamen [jian-die qing-shen].
They lovebirds-flatfish emotion-deep
‘They love each other very much.’

What’s more, there exist another chunks called proverbs in Chinese that can also be predicated of the initial NP *tamen* ‘they’:

(45) a. 他們 [塞翁失馬焉知非福]。
Tamen [sai-weng shi ma yan zhi fei fu].
They frontier-man lose horse how know NEG fortune
‘Their misfortune might be a blessing in disguise.’

b. 他們 [牆頭草兩面倒]。
Tamen [qiang-tou-cao liang-mian dao].
They wall-head-grass two-side fall
‘They bend with the wind.’

Aligning with the classification made above, we argue that the idioms in (43) are predicated of the “subject” topic NP *tamen* ‘they’ while the sentential chunks in (44) and (45) are predicated of the “dangling” topic NP *tamen* ‘they’.

5.3 Justifying the hypothesis

The hypothesis can be justified by the fact that the preverbal NPs contained in the chunks can be as referential as normal subjects and need not hold any syntactic and semantic relations with the sentence-initial NPs. The complex V-node analysis made from the *zai* ‘at’ test will also be rejected as an additional support.

5.3.1 The subjecthood of the preverbal NPs

To begin with, the preverbal NPs of the chunks in (44) and (45) often idiomatically refer to the sentence-initial NPs, as shown in (46). This shows that there is no need to reject them as non-subjects since they are as referential as normal subjects.

(46) a. 他們 [鶼鰈情深]。
Tamen$_{i}$ [jian-die$_{i}$ qing-shen].
They lovebirds-flatfish emotion-deep
'They love each other very much.'

b. 他們 [牆頭草兩面倒]。
Tamen$_{i}$ [qiang-tou-cao$_{i}$ liang-mian dao].
They wall-head-grass two-side fall
'They bend with the wind.'

Second, one may also suspect that they are cases of left-dislocation or topic promoting construction. However, consider the paradigm in (47-48):

(47) a. 他們 [虎視眈眈]。
Tamen$_{i}$ [hu$_{i}$ shi dandan].
They tiger see covetously
'They glare like a tiger eyeing its prey.'

b. 他們 視眈眈。
*Tamen shi dandan.
They see covetously
'They glare like a tiger eyeing its prey.'

(48) a. 他們 [塞翁失馬焉知非福]。
Tamen$_{i}$ [sai-weng$_{i}$ shi ma yan zhi fei fu].
They frontier-man lose horse how know NEG fortune
'Their misfortune might be a blessing in disguise.'

b. *他們 失馬焉知非福。
*Tamen shi ma yan zhi fei fu.
They lose horse how know NEG fortune
'Their misfortune might be a blessing in disguise.'

In (47b) and (48b) the co-referential preverbal NPs contained in the chunks are replaced by the alleged left-dislocated NPs. The unacceptability reveals that the sentence initial NPs are not left-dislocated elements and therefore does not bear any syntactic relation with the preverbal NPs. This analysis is further supported by the examples below where there is by no means any co-referential relationship between the sentence-initial and the preverbal NPs.

(49) a. 他們 [鬼打牆]。
Tamen$_i$ [gui$_{*i/j}$ da qiang].
They ghost hit wall
‘They are in a loop.’

b. 他們 [禍不單行]。
Tamen$_i$ [huo$_{*i/j}$ bu dan xing].
They misfortune NEG alone go
‘Their misfortunes never come alone.’

c. 他們 [罪不可赦]。
Tamen$_i$ [zui$_{*i/j}$ bu ke she].
They crime NEG can remit
‘Their sentences cannot be remitted.’

5.3.2 The unreliable of *zai* test

To further support their claim that the idiom chunks at issue belong to one particular type of predicate, viewed as an unanalyzable complex V-node in the syntactic representation, Huang and Ting (2006) continue that the predicatehood of the idiom chunks can be further verified by the fact that they can be preceded by aspect markers such as progressive *zai* :

(50) a. 他們 一直在[大魚吃小魚]。
Tamen yizhi zai [da-yu chi xiao-yu].
They continually ASP big-fish eat small-fish
‘They have been bullying the weaker.’

b. 他們 正在[我看你，你看我]。
Tamen zheng zai [wo kan ni, ni kan wo].
They right ASP I look you you look me
‘They are looking at each other.’

However, the assumption that the idiom chunks should be predicates syntactically under V-node is questionable. We provide three major reasons below to reject the *zai* test. First, the applicability of the *zai* test is highly restrictive and, as shown below, the constructions at issue cannot be construed with other aspect markers, such as *le*, *zhe* and *guo*, etc.

(51) *他們 [大魚吃小魚] 了/著/過。
*Tamen [da-yu chi xiao-yu] le/zhe/guo.
They big-fish eat small-fish ASP
'They already bullied the weaker.'

Second, the *zai* test is unreliable since, as illustrated in (52), it would predict that sentences such as (41a) are acceptable subject-predicate utterances while similar constructions such as (41b) are only barely acceptable. This prediction in turn contradicts Huang and Ting's initial position that both are normal subject-predicate constructions with initial PP-reduced form adverbials:

(52) a. 他們 常常在 [你指責我不對，我抱怨你不好]。
Tamen changchang zai [ni zhize wo bu dui, wo manyuan ni bu hao].
They often ASP you blame I NEG right I complain you NEG good
'They often blame and complain each other.'

b. ? 他們倆 專門在 [小王先來，小李後到]。
?Tamen liang zhuanmen zai [XiaoWang xian lai, XiaoLi hou dao].
they two specifically ASP XiaoWang first come XiaoLi after arrive
'As for them two, XiaoLi is always coming after XiaoWang.'

Third, the test is misleading, especially when we consider that the progressive marker *zai* in fact modifies the verb phrase (VP) of the chunk:

(53) a. 他們 一直在 [黑吃黑]。
Tamen yizhi zai [hei chi hei].
They continually ASP black eat black
'They have been always bullying others.'

b. *他們 一直在 [媳婦欺負婆婆]。
*Tamen yizhi zai [xifu qifu popo].
They continually ASP daughter-in-law bully mother-in-law
'They have been always bullying their mother-in-laws.'

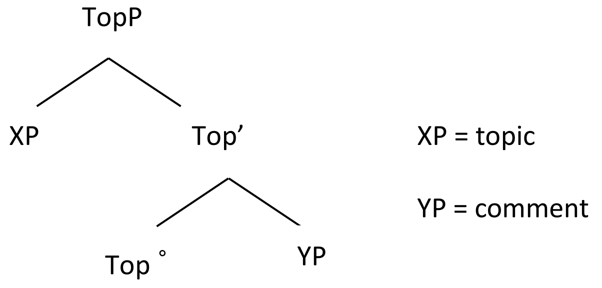
c. 他們 一直在 [欺負婆婆]。
Tamen yizhi zai qifu popo.
They continually ASP bully mother-in-law
'They have been always bullying their mother-in-laws.'

The paradigm above reveals not only that the sentence following the progressive marker *zai* in (53a) is a chunk while the one in (53b) is not, but also that the progressive marker *zai* is merely used to modify the VP, as exemplified in (53c). As it stands, an explanation with respect to the acceptability contrast between (53a) and (53b) proceeds as follows. Given that the subject of the subordinate clause in (53a) is indispensable of the “unanalyzable” clause, there should be no surprise that the sentential chunk as a whole can be construed with the progressive marker *zai* that is in general used to modify the VP of the chunk. Conversely, the unacceptability of (53b) results from the interference of the subject in the “analyzable” clause.

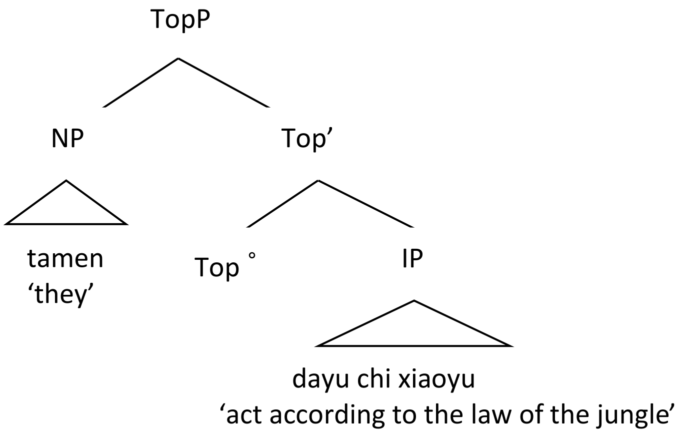
5.3.3 The topic-comment structure

In view of the issues we have elaborated above, we are urged to reject the complex V-node analysis and argue that Rizzi’s (1997) TopP proposal, specifically a topic-comment structure, better accounts for the relation between the sentence-initial NPs and the following clauses. Consider the topic-comment structure in (54) and see how the construction at issue is represented in (55) where IP stands for a sentence:

(54) TopP (Rizzi, 1997):
The specifier is the topic and its complement is the comment.



(55) Top-Comment Structure



The topic-comment analysis has its explanatory power in that it adequately captures the relations among the syntactic units of “double-chunk” constructions.

Consider the double-chunk constructions in (56) where the second chunk is adjoined. Unlike (40b) or (41), the second chunk is adjoined merely for reinforcement:

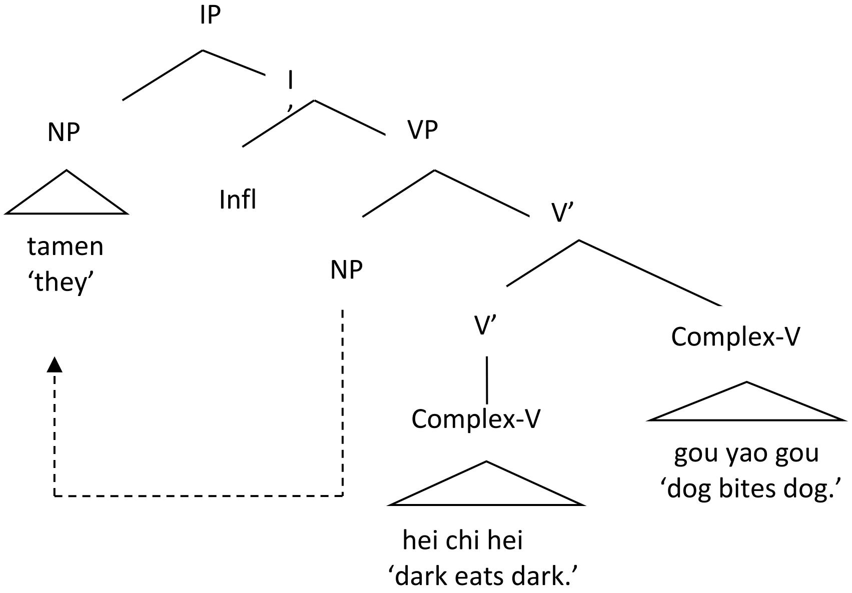
(56) a. 他們 [黑吃黑] [狗咬狗]。
Tamen [hei chi hei] [gou yao gou].
They black eat black dog bite dog
'They act according to the law of the jungle.'

b. 他們 [花開富貴] [子孫滿堂]。
Tamen [hua kai fu-gui] [zi-sun man tang].
They flower bloom fortune offspring full house
'They grow with prosperity and have plenty of offspring.'

c. 他們 [好漢不吃眼前虧] [識時務者為俊傑]。
Tamen [hao-han bu chi yan-qian kui] [shi shi-wu zhe wei jun-jie].
They good-man NEG eat eye-front loss know reality person is hero
'They are the wise men who suit their actions to the time.'

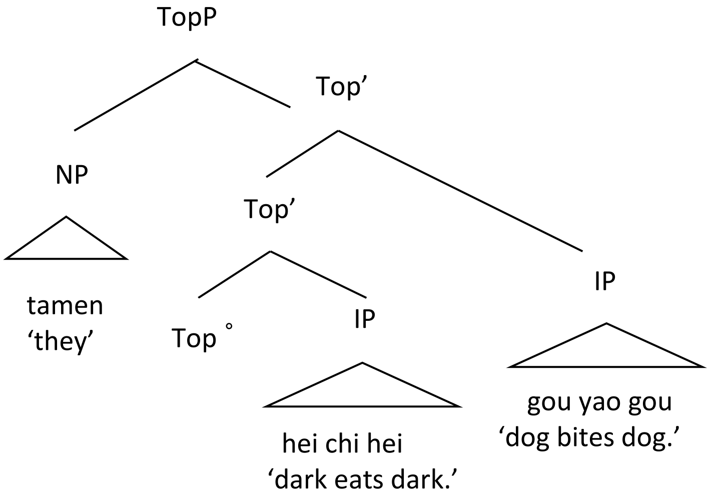
Under the subject-predicate framework, the second unanalyzable complex V-chunk *gou yao gou* 'dog bites dog' in (56a) is adjoined to the VP, as represented in (57). However, the structure in (57) is inadequate since the representation is structurally ambiguous in that either one of the complex V-chunks could be the head of the VP. Specifically, the geometry shows that the second chunk is an adjunct predicate; however, the categories represented in the structure are incompatible with the geometry since the categories indicate that either one of the complex-Vs could potentially be the head of the VP. This representation thus suggests that there should be an interpretation burden, contrary to the fact.

(57) Subject-Predicate Structure



By contrast, analyzed under the topic-comment geometry, the IP-chunk gou yao gou 'dog bites dog' is only adjoined to TopP, as shown in (58). This representation clearly distinguishes the difference between complement and adjunct.

(58) Topic-Comment Structure



The top-comment structure above properly specifies the legal relations among the syntactic units of the sentence: the head, Top ° , subcategorizes the topic *tamen* 'they' as its specifier, the IP-chunk *hei chi hei* 'dark eats dark' as its complement, and finally the IP-chunk *gou yao gou* 'dog bites dog' as its adjunct. This analysis is also flexible in that it allows for a distinction between verbal and sentential chunks and a possible construction where the first chunk is an IP while the other a VP, or vice versa.

Chunks for most language users are opaque and not commonly used. They are used only when a speaker is confident in usage and feels an urge in context to express a comment relevant to a topic in a succinctly figurative or idiomatic way.

6 The licensing conditions and constraints

The present paper has thus far analyzed that there exist dangling topics in Chinese. They are either contrastive or old topics. This section elaborates the licensing conditions and constraints that guarantee the well-formed dangling topic structures.

6.1 The literature review

In light of the fact that the acceptability contrast presented below cannot be adequately accounted for by the traditional aboutness condition, Pan and Hu (2009) postulate set intersection as the topic licensing condition.[11]

(59) a. *幼兒園的小孩 張三 教兒子畫畫。
*Youeryuan de xiaohai Zhangsan jiao erzi huahua.
Kindergarten MOD children Zhangsan teach son draw-poctures
'As for the children in the kindergarten, Zhangsan teaches his son to draw pictures.'

b. 幼兒園的小孩 張三 只教兒子畫畫。
Youeryuan de xiaohai Zhangsan zhi jiao erzi huahua.
Kindergarten MOD children Zhangsan only teach son draw-poctures
'As for the children in the kindergarten, Zhangsan only teaches his son to draw pictures.'

As Pan and Hu contend, (59a) is unacceptable because no set intersection occurs in it. (59b) is acceptable because the intersection of the topic set with the set projected from the focus element *erzi* 'son' is not empty.

In addition, aware of the fact that the licensing condition alone cannot guarantee the well-formedness of all topics, they further add the interpretation condition (predicate-subject condition),[12] whereby topics can be properly interpreted because

[11] Topic Licensing Condition (Pan & Hu, 2008, 1970). A topic can be licensed iff (i) there is a set Z induced by a variable *x* in the comment, and (ii) the set Z thus generated does not produce an empty set when intersecting with the set T denoted by the topic.

[12] Topic Interpretation Condition (Pan & Hu 2009). In a configuration $\Sigma = [_{\mathrm{TopP}}\ X\ [_{\mathrm{IP}} \ldots Y \ldots]]$, the topic X is properly interpreted if it can form a subject-predicate relation with an element Y in the comment clause, where Y is the subject and X the predicate.

they can form a predicate-subject relation with elements contained in comment clauses. This explains why (60a) is well-formed while (60b-c) is ill-formed.

(60) a. 水果 我喜歡蘋果。
Shuiguo wo xihuan pingguo.
Fruit I like apple
'As for fruits, I like apples.'

b. *蘋果 我喜歡水果。
*Pingguo wo xihuan shuiguo.
Apple I like fruitfascinating
'*As for apples, I like fruits.'

c. *蘋果 我喜歡香蕉。
*Pingguo wo xihuan xiangjiao.
Apple I like banana
'*As for apples, I like bananas.'

Specifically, in (60a), the initial NPs is able to form a predicate-subject relation with the element in the comment clause. However, the relation does not hold in (60b-c). Consider the contrast below:

(61) a. 蘋果 是水果。
Pingguo shi shuiguo.
Apple be fruit
'Apples are fruits.'

b. *水果 是蘋果。
*Shuiguo shi pingguo.
Fruit be apple
'*Fruits are apples.'

c. *香蕉 是蘋果。
*Xiangjiao shi pingguo.
Banana be apple
'*Bananas are apples.'

The analysis is fascinating; nevertheless, the conditions would be too strict to accommodate other potential topics. It for instance may not provide an adequate explanation to some of the nominal-predicate topics such as (21), repeated in (62). Neither the set intersection nor the interpretation condition is able to account for the well-formedness of the topic.

(62) 一輛車 一天 三千。
Yi-liang che yi-tian san-qian.
one-CL car one-day 3000
‘One car is three thousand dollars a day.’

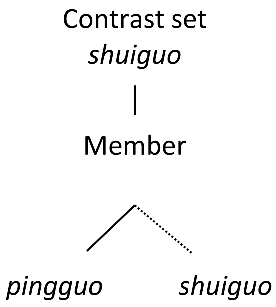
6.2 The contrast set condition

In the previous sections we have proposed that the contrast set condition is one of the topic licensing conditions in context which requires that one or each member be selected. This is when a speaker assumes that a hearer expects her/him to select from a discoursally available set or provide an assessment to a topic which has been previously selected. The acceptability contrast in (59) is thus explicable. (59a) is unacceptable because two potential set members are selected simultaneously from the contrastive topic set regarding the potential candidates to teach drawing; on the other hand, (59b) is acceptable because only one member is selected from the contrast set *youeryuan de xiaohai* ‘the children in the kindergarten’. The same holds true for the acceptability contrast in (60). (60a) is acceptable because the requirement is satisfied: there exists a contrast set which is explicitly expressed by *shuiguo* 'fruit’, from which one member of the set is selected. On the other hand, (60b) and (60c) are unacceptable because the requirement is not met.

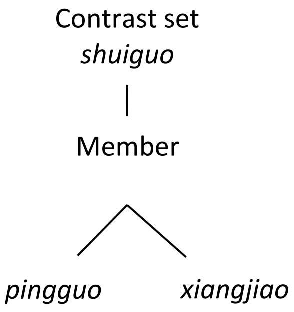
In fact, two discoursal questions can be re-constructed below to potentially elicit responses (60b-c). However, whichever question they respond to, the requirement for the contrast set condition is not met. Consider the questions and how the contrast set condition works to eliminate both (60b-c).

(63) a. 你喜歡什麼水果？
Ni xihuan sheme shuiguo?
you like what fruit
‘What fruit do you like?’

b. *



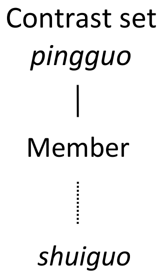
c. *



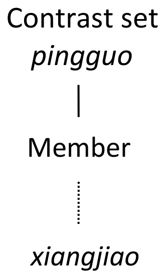
The unacceptability of (60b-c) proceeds as follows. As represented in (63b), there exists a contrast set, namely *shuiguo* 'fruit'; however, one member and the set itself, which is not one of the members, are selected in (60b). As also shown in (63c), two members are selected simultaneously and this leads to a selection conflict in (60c).

(64) a. 你喜歡什麼蘋果？
Ni xihuan sheme pingguo?
you like what apple
'What kind of apple do you like?'

b. *



c. *



Likewise, as represented in (64b-c), there exists a contrast set *pingguo* 'apple' in (60b-c); however, neither *shuiguo* 'fruit' nor *xiangjiao* 'banana' is the member of the contrast set. Both exemplify an improper selection of contrast set.

6.3 The commentative condition

Another condition we are convinced to propose is a commentative relation held between the sentence-initial NPs and the following clauses, a condition when a speaker assumes that a hearer expects her/him to express a contextually relevant comment (point of view) to a topic. The well-formedness of the event and (non)idiom chunk types is licensed by this condition, with the difference being whether the condition is satisfied directly or indirectly. So far as the event type is concerned, the sentence-initial NPs and the comment clauses form an indirect commentative relation in which the latter is indirectly commentative to the former. Take robbery event for example. The acceptability contrast in (65) strongly supports this idea.

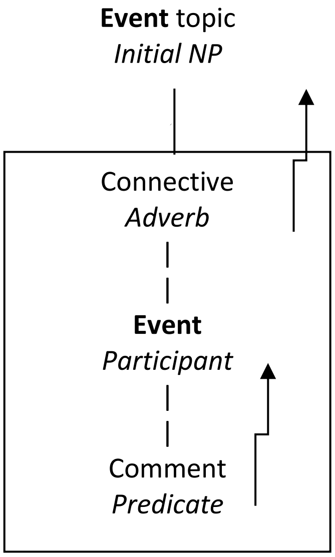
(65) a. 那場搶案 幸虧 行員機警。
Na-chang qiangan xingkui hang-yuan jijing.
That-CL robbery fortunately bank-staff alert
'As for the robbery, fortunately the bank-staff was alert; otherwise…'

b. *那場搶案 行員機警。
*Na-chang qiangan hang-yuan jijing.
That-CL robbery bank-staff alert
‘As for the robbery, the bank-staff was alert; otherwise…’

c. *這個小偷 (幸虧) 警察來得早。
*Zhe-ge xiaotou (xingkui) jingcha lai-de-zao.
this-CL thief (fortunately) police come-DE-early
‘As for this thief, (fortunately) the police came early.’

In (65a) the commentative relation is connected indirectly via the speaker’s comment on one of the salient participants’ actions in the event. The participant is often considered as having an impact on the event. The connective adverb *xingkui* is often required in order to connect the topic NP and its following comment clause. In other words, by making use of the connective, the two are somehow commentatively related. The indirect relation can be captured schematically as below.

(66) Indirect commentative relation



On the other hand, the relation in (65b) is unclear because, without the adverb *xingkui* ‘fortunately’, the meaning expressed in the comment clause is ambiguous: the situation described in the clause is most likely to be interpreted as a statement about the participant’s personality traits irrelevant to the event itself. The relation in (65c) is not connected at all: the speaker’s view on one participant’s performance bears no relevance to another participant in the event. It is simply illogical to comment a person by commenting another. Relevance does play an essential role.

The same goes for the (non-)idiom chunk type. The commentative relation in this type is however direct: the following chunk is directly commentative to the sentence-initial topic NP. Relevance has to be observed. Consider the examples below.

(67) a. 他們 [罪不可赦] /*[公道自在人心]。
Tamen [zui bu ke she] */*[gongdao zi zai renxin]*.
They crime NEG can remit justice naturally in man-heart
'Their crimes are not forgivable.'

b. 這件事 [公道自在人心] /*[罪不可赦]
Zhe-jian shi [gongdao zi zai renxin] */*[zui bu ke she]*.
this-CL matter justice naturally in man-heart crime NEG can remit
'As for this matter, facts speak louder than words and justice will prevail.'

The paradigm in (67) shows that the chunks used to comment on the topics have to be relevant. The second chunk that follows in each sentence is simply irrelevant.

7 Concluding remarks

To recap, in this paper four alleged types of topic have been analyzed as dangling topics in Chinese. The sentence-initial NPs at issue are either contrastive or old topics. The analysis of this study suggests that pragmatically licensing NPs in left-peripheral position is one of the ways that Chinese employs to reflect information structure of topic. The conditions and constraints elaborated are the linguistic and contextual knowledge that a speaker possesses to calculate and package her/his utterances.

Two contextual licensing conditions that guarantee the well-formedness of dangling topics in Chinese have been identified. Specifically, the material/instrument and nominal-predicate types are properly licensed under the contrast set condition. And the event and (non-)idiom chunk types are licensed via the commentative condition. The following table shows the topic types, the licensing conditions, and the constraints.

Table1: Dangling topics and licensing conditions

Dangling type	Licensing condition	Constraint
Material/Instrument	Contrast set	Proper selection
Nominal-predicate	Contrast set	Proper selection
Event	Commentative	Relevance
(Non-)idiom chunk	Commentative	Relevance

This table provides insights to the other two types of topic that we have not yet discussed about. That is, the sentence in (2) is uttered when ‘neither’, one of the possible options, is selected. Likewise, (6) selects New York as the topic for further assessment. Note that there is no rigid one-to-one relation that each condition is exclusively corresponds to specific types of topic. The relation is rather defined as context-dependent in that one condition is more relevant and decisive than the other. This is especially true as one may find examples where the first two types of topic occur as old topics. Similarly, the last two types of topic appear as contrastive topics. They are contextually licensed regardless of contrastive or old topics they appear in discourse. The bottom line is neither the contrast set nor the commentative condition can be violated.

The analysis thus decomposes and lends a further support to the traditional aboutness relation purportedly held between the sentence-initial NPs and the following clauses. The vague aboutness relation can be further justified as a valid licensing mechanism in general, under which either one of the conditions is satisfied. Shi’s refusal to the aboutness relation on the basis of the evidence in (7a), repeated in (68b), is therefore misleading in this regard. Consider the following discourse exchanges:

(68) a. 你知道　　　張校長辭職了？
Ni zhidao　　　Zhang-Xiaozhang cizhi le?
you know　　　Zhang-Xiaozhang resign ASP
‘Do you know Zhang-Xiaozhang has resigned?’

b. *這件大事　　我知道　張校長辭職了。
*Zhe-jian da-shi　wo zhidao　Zhang-Xiaozhang cizhi le.
this-CL big-issue　I know　Zhang-Principal resign ASP
‘As for this big issue, I know that Principal Zhang has resigned.’

b’. 這件大事　　我知道　李四受了很大的委屈。
Zhe-jian da-shi　wo zhidao　Lisi shou-le hen-da de weiqu.
this-CL big-issue　I know　Lisi encounter-ASP very-big MOD grievance
‘As for this big issue, I know that Lisi felt so wronged and unjustly treated.

In (68b-b’) the contrast set condition is irrelevant and not violated. However, (68b) is ill-formed because the commentative condition is not observed. What was brought up in the following clause is not a relevant comment but simply an unnecessary reiteration of the topic itself. The occurrence of the topic NP even causes a licensing conflict with the object NP since they refer exactly to the same thing. On the other hand, the response in (68b’) is well-formed because what was brought up is the speaker’s comment on the perception of the participant in the event, an indirect comment to the topic.

Finally, the result of this study poses an intriguing inquiry that pertains to language typology. Erteschik-Shir discusses language typology in terms of word order and concludes that discourse configurational languages, e. g. Chinese, feature a ranking hierarchy where information structure dominates syntax in determining word order; on the contrary, in configurational languages, e. g. English, syntax dominates information structure. Along that line, we suppose that, in contrast to configurational languages, discourse configurational languages exhibit phenomena of dangling NP topics which need not be syntactically (or semantically) but pragmatically licensed by the aboutness relation. For example, the sentence in (5) suggests that as a discourse configurational language Chinese allows multiple NPs in sentence-initial position as long as they are properly licensed: the topic NP is pragmatically licensed in the left-most position while the QP is syntactically licensed in the subject position. However, as a configurational language English tends to not allow pragmatically licensed NPs. Therefore, similar topic NP in English has to be in subject position while QP in prepositional phrase, or the other way around, as long as they are syntactically licensed.

References

Chafe, W. L. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. N. Li (ed.), *Subject and Topic* (pp. 25--55). New York: Academic Press.

Chao, Y.-R. (1968). *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

Erteschik-Shir, N. (2007). *Information structure: The syntax-discourse interface*. Oxford, UK: Oxford University Press.

Gregory, M. L., & Michaelis, L. A. (2001). Topicalization and Left-Dislocation: A Functional Opposition Revisited. *Journal of Pragmatics*, 33, 1665-1706.

Hsin, A.-L. (2002). On Indefinite Subjects in Chinese. *Chinese Studies, 20*(2), 353-376.

Hu, J., & Pan, H. (2009). Decomposing the aboutness condition for Chinese topic constructions. *The Linguistic Review, 26,* 2-3.

Huang, C.-T. J. (1982). *Logical relations in Chinese and the theory of grammar*. Cambridge, Massachusetts: MIT Press.

Huang, C.-T. J. (1984). On the Distribution and Reference of Empty Pronouns. *Linguistic Inquiry, 15*(4), 531-574.

Huang, R.-h., & Ting, J. (2006). Are There Dangling Topics in Mandarin Chinese? *Concentric: Studies in Linguistics*, *32*(1), 119-146.

Kiss, E. (2004). The EEP in a Topic-Prominent Language. In P. Svenonius (Ed.) *Subjects, Expletives, and the EEP* (pp. 107-124). Oxford: Oxford University Press.

Larson, R. K. (1985). Bare-NP Adverbs. *Linguistic Inquiry, 16*(4), 595-621.

Li, C., & Thompson, S. (1976). Subject and topic: A new typology of language. In Li, C. N. (1976). *Subject and topic: [papers]* (pp. 457-489). New York: Academic Press.

McCawley, J. D. (1988). Adverbial NPs: Bare or Clad in See-Through Garb?. *Language, 64*(3), 583-590.

Pan, H., & Hu, J. (2002). Licensing dangling topics in Chinese. Paper presented at the 2002 LSA Annual Meeting in San Francisco, CA, USA.

Pan, H., & Hu, J. (2008). A semantic-pragmatic interface account of (dangling) topics in Mandarin Chinese. *Journal of Pragmatics, 40*(11), 1966-1981.

Rizzi, L. (1997). The Fine Structure of Left Periphery. In L. Haegeman (Ed.) *Elements of Grammar* (pp. 281-337). Dordrecht: Kluwer.

Shi, D. (1992). *The Nature of Topic Comment Constructions and Topic Chains*. Los Angeles: University of Southern California dissertation.

Shi, D. (2000). Topic and topic-comment constructions in Mandarin Chinese. *Language, 76*(2), 383-408.

Shyu, S. (2011). *The syntax of focus and topic in Mandarin Chinese*. Los Angeles: University of Southern California.

Strawson, P. E. (1964). Identifying Reference and Truth-Values. *Theoria,* 30, 86-99.

Tsao, F.-F. (1990). *Sentence and Clause Structure in Chinese: A Functional Perspective*. Taipei: Student Book Co.

Xu, L., & Langendoen, D. T. (1985). Topic structures in Chinese. *Language,* 61, 1-27.

LEXICAL ASPECT CLASSIFICATION FOR UNRELATED LANGUAGES: A CASE STUDY ON SLOVENIAN AND CHINESE LEXICAL ASPECT

Tina ČOK
Science and Research Centre Koper, Slovenia
tina.cok@zrs-kp.si

Abstract

The present paper presents a comprehensive analysis of the verbal aspect in general and with special emphasis on the comparison of Chinese and Slovenian lexical aspect. Recognised discrepancies between the conceptualisation and verbalisation of actions in unrelated languages indicate that deeper cognitive differences affect our perception of reality, which is something that should be more widely recognized when learning and teaching foreign languages. The contribution of this article is a comparative analysis of available studies by authoritative linguists, based on which we have formulated a new and more comprehensive proposal that will help classify verb types in unrelated languages, and can be further exploited in the field of applied linguistic research.

Keywords: verbal aspect; lexical aspect classification; verb types; unrelated languages; Standard Chinese

Povzetek

Pričujoči prispevek ponuja podrobno analizo glagolskega vida na splošno in s posebnim poudarkom na primerjavi kitajskega in slovenskega leksikalnega glagolskega vida. Prepoznana razhajanja med konceptualizacijo in upovedovanjem glagolskih dejanj pri nesorodnih jezikih so pokazala, da globlje kognitivne razlike vplivajo na naše dojemanje stvarnosti. Dejstvo, ki ga je potrebno pri učenju in poučevanju tujih jezikov bolj upoštevati. S pričujočo študijo želimo k temu prispevati s primerjalno analizo obstoječih raziskav uglednih jezikoslovcev, s pomočjo katere smo oblikovali nov, bolj obsežen predlog klasifikacije glagolske vrstnosti za nesorodne jezike, ki omogoča nadaljnjo uporabo na področju raziskav iz aplikativnega jezikoslovja.

Ključne besede: glagolski vid; klasifikacija leksikalnega vida; glagolska vrstnost; nesorodni jeziki; standardna kitajščina

Acta Linguistica Asiatica, 10(2), 2020.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.10.2.49-64

1 Introduction

This article is one of the several studies on verbal aspect, however, it is unique in the respect that it focuses on the category of verbal aspect in languages that are very rarely compared, not only according to the principles of universal grammar (Smith 1991; Orešnik 1991), but also according to the contributions of linguistic relativity. The latter makes this study relevant, because it can throw a new light on how we conceive and understand verbal aspect. Starting from this small but meaningful linguistic cell, we continue to explore more general linguistic and pragmatic questions, such as how lexical and grammatical aspects are verbalised in different languages and how they reflect the respective models of conceptualisation, assuming that verbs help us understand the way the world around us is conceived.

This paper presents a small portion of a broader empirical study on verb conceptualisation and verbalisation in three unrelated languages (Čok, 2019), i.e. Chinese, Italian and Slovenian[1]. The objective of the study was to draw on the well-established theory of foreign language learning, which supports the idea of relying on already spoken languages when learning foreign languages, including the first language, and on the well-explored assumptions that bilingual and multilingual speakers develop an increased awareness of language that helps them acquire additional languages. However, there is a lack of research focusing on cross-language and metalinguistic awareness in unrelated languages, which are usually taught as a third or even fourth foreign language. Accordingly, it has been suggested (Ibid.) that there are significant conceptual differences in the understanding of reality between native speakers of different languages, and that these differences are greater for unrelated languages, making it difficult to bridge these diverging points.[2]

2 Approaches and method

Part of the previously mentioned empirical study was to thoroughly examine the already existing verb classifications. By doing this, we found that universal grammar can only be partially applied in this matter (because it helps to identify the category for all existing languages), but that different language combinations and comparisons require different approaches to classification. For this, linguistic relativity has proven to be very insightful, especially for the cross-linguistic perspective.

[1] We use the term unrelated languages for languages that have not evolved from the same language family (i.e. Indo-European), the pairs of comparison are thus Slovenian – Standard Chinese and Italian – Standard Chinese.

[2] By diverging points, we mean those critical discrepancies between languages, especially of a syntactic nature, which, without adequate explanation, warning, comparison between languages and metalinguistic consideration, lead to a negative transfer.

The study is based on the hypothesis that the nature of the Chinese verb, in its most basic monomorphemic form, allows a broad freedom of interpretation of the degree of completion, i.e. it is more oriented towards the process than towards the result of the action performed. Based on our knowledge of aspectual systems and their functioning in the three languages, we assumed that the Chinese verb emphasises more the processual phase of an action than the same verb in Slovenian (as well as in Italian), causing the speaker to respectively transfer this emphasis in the process of verbalisation of foreign languages as well as in conceiving reality. The empirical study and its results are not the focus of this paper, so we will not go into further detail on this matter. For details see Čok (2019).

This paper focuses on the theoretical analysis of the lexical aspect for two syntactically very different languages: Standard Chinese and Slovenian. For this purpose, we examined the established aspectual classifications of verbs as proposed by Vendler (1967), Smith (1991), Xiao and McEnery (2004), Peck et al. (2013) and Koenig and Chief (2008). Based on these classifications, we developed and proposed a refined verb classification that could encompass all features of the two investigated languages. The selected studies on verb aspect were analysed with a comparative method, by which verb features and proposed classifications were reviewed and integrated into a new classification according to the objectives of this study.

3 Previous studies

3.1 Verb types and features

In many languages, the verb is inflected and encodes tense, aspect, mood and voice. It often also helps convey person, gender and number of the subject or object. Nevertheless, not all of the languages in the world present these features. It has been previously proposed that different languages take different perspectives on activities and events. Scholars like Ikegami (1985) have worked on the difference between Japanese and English, arguing that Japanese is a process-oriented and English is a result-oriented language. Basically, the perspective on how we understand an action being focused on the process, which might or might not end up in a change of state or towards a result made possible by this change, depends on how this action is expressed through the use of the verb. Nevertheless, not only the verb can be the carrier of this perspective. Language is full of more or less subtle mechanisms, i.e. implicit references, word order, pauses between words etc. which even subconsciously convey what is the conveyer’s standpoint or what segment of the action she or he is focusing on.

Verbs in the Indo-European linguistic tradition have been, following Vendler’s classification (1967), divided into four main types according to their inherent property of (semantic) eventuality; verbs which express state, activity, achievement, accomplishment and semelfactive as a separate category, subsequently added to Vendler's classification by Bernard Comrie in 1976.

This four (plus one) folded categorization represents a long-standing linguistic standard in the scope of classification of verbal aspect. Only recently, the study of verbal aspect has flourished due to the increased interest in the subject by linguists with different linguistic backgrounds. New research and studies on the features of the Chinese lexical aspect, especially in the cross-linguistic perspective, have proposed new verb features and consequently new classifications (Smith, 1991; Koenig & Chief, 2008; Peck et al., 2013; Thepkanjana & Uehara, 2009, 2010; Xiao & McEnery, 2004). For Chinese especially, the standard division seemed not to fit entirely, especially in terms of the categories of achievements and accomplishments. On the other hand, when we try to classify verbs in Slovenian, we are dealing with a great interconnectivity between verb class and aspectual pairs, which makes it difficult to directly apply Vendler's classification. Slovenian presents a pretty elaborate system made complex by the grammatical aspect of the verb, for which every verb has two forms, the perfective (*dovršnik*) and the imperfective (*nedovršnik*). To translate Vendler's test phrases used to classify verbs (in English) to Slovenian, we would need to switch from the perfective to the imperfective form and vice versa. In Chinese, a single verb is very often not enough to express completion of an action. In these cases, the Chinese language employs resultative and compound causative constructions, which are, simply put, either a combination of two verbs or a verb and an adjective, where the first one conveys the activity and the second one the realisation that transfers the focus from the activity in process to its result. We can find more evidence of the lack of real accomplishment verbs in Chinese in Zhao (2005). Besides the resultative and compound causative construction, a maybe even more common way of changing aspect in Chinese is by employing the aspectual marker *le*. But since it sometimes also affects only the tense, its reliability in this concern is, so to speak, weak. What can be deduced from previous studies on the ambiguities and peculiarities in the interpretation of the expressed completion of some Chinese verbs (Koenig & Chief 2008; Peck et al., 2013; Thepkanjana & Uehara 2009, 2010) is that they comprise a very wide and ambiguous scale of degree of completion, which is often open to interpretations.

The methodology employed in our empirical study required a classification of verbal actions able to comprise as much as possible the features of Chinese, Italian and Slovenian language and at the same time preserve as high as possible the degree of universality to keep the features abstract while interpreting the results. We designed our classification of lexical aspect on the basis of Smith (1991), Xiao and McEnery (2004), Peck et al. (2013) and Koeing and Chief (2008). While all of these classifications contributed a great deal to ours, they did not entirely fit the language combinations of our choice. Hence the need for a newly adjusted categorisation, which, in our opinion, best comprises the features of verbal actions in general.

Before presenting the revised classification, we will examine the features and approaches of previous proposals, for it is a revealing and useful diachronic developmental process, that can help to better understand the linguistic implications of lexical aspect and justify our adjustments.

3.2 Approaches to classifications of verbal actions for unrelated languages

How we understand whether an action focuses on the process that ends in a change of state or with a result that represents that change of state, depends on how that action is verbalised or expressed through another argument. To express this bias with the verb means to do so explicitly, whereas it is not always the verb that defines the perspective of an activity. Language is full of more or less subtle mechanisms, sometimes even unconsciously reflecting our perspective on activities. In addition to verbs and arguments, such mechanisms may include the syntactic structure of the sentence, noun properties, the presence or absence of the subject, and some language-specific mechanisms of expression.

When studying the focus of verbal actions in different languages, we must, first of all, consider the two basic aspectual components of the verb, its verbal aspect (also known as grammatical aspect) and the nature of the verbal action (also known as Aktionsart or lexical aspect)[3], each language having a very specific way of functioning and using the two components. In Miklič, for example, we find that "unlike the situation in the Romance languages, the Slavic aspectual opposition shows certain interdependence with the nature of the verbal action". (Miklič, 2007, p. 92)

In the Indo-European tradition, the inherent semantic property of verbal action (lexical aspect) is most often defined according to Vendler's classification (1957) in four main types, namely verbs expressing state, activity, accomplishment and achievement. Vendler's classification was supplemented with the category of semelfactives[4] in 1967 by Bernard Comrie and further refined by Smith (1991). Although Vendler's classification is perfectly suited to provide the basic framework for classification, it is too loose for cross-linguistic studies, especially when comparing unrelated languages that show large linguistic discrepancies.

In the following pages, we will outline five main classifications of verbal actions, emphasizing aspects that are in our opinion unsuitable for classification in a cross-linguistic perspective.

[3] Smith (1991) uses the terms situation aspect for grammatical aspect and viewpoint aspect for lexical aspect. In the present paper, we will use the terms grammatical and lexical aspect.

[4] Semelfactives comprise one-time atelic actions that last a very short time, a moment. They are also found in some classifications as a subcategory of achievement verbs.

The first of its kind is Vendler's distribution of verbs (1957), which is based on sample questions (for English) and is intended to help classify verb types according to how they occur within a time interval.[5]

(1) I. For how long did he ...? (activity)
II. How long did it take to ... ? (accomplishment)
III. At what time did you ... ? (achievement)

However, when these sample questions are applied to languages other than English, they show certain weaknesses. Vendler's tests for the classification of Slovenian verbs, for instance, are not directly transferable due to the strong interaction between lexical and grammatical aspects. In order to make the sentences meaningful, the verbs must actually shift between the perfective and imperfective form, which also means shifting between unmarked and marked usage.

(2) a) Koliko časa **je potiskal** voziček? (activity)
'For how long did he push IMPRF the cart?'

b) Koliko časa je porabil, da **je narisal** krog? (accomplishment)
'How long did it take him to draw PERF the circle?'

c) Ob kateri uri **je dosegel** vrh gore? (achievement)
'At what time did he reach PERF the top of the mountain?'

As can be seen in the above examples, the Slovenian language is specific because of its elaborated verbal system, in which grammatical properties of the aspect are recognized in two separate verbal forms: *dovršnik* (perfective) and *nedovršnik* (imperfective). Since the verbal aspect is expressed lexically, the same verbal action can occur in several categories of the verbal aspect by transforming the perfective into an imperfective verb form and vice versa. This is also the reason why Slovenian, with the exception of biaspectual verbs (Žele, 2011), does not allow such ambiguities in the interpretation of the verbal aspect.

Chinese verbs can also be applied to Vendler's sample questions (3) but there are some divergencies between how these are employed in Standard Chinese compared to other languages.

[5] We only include the verb types that are relevant to this study.

(3) I. 他看了电视多长时间?
Tā kàn-le diànshì duō cháng shíjiān?
'For how long did he watch TV?'

II. 他花了多长时间画一个圆圈?
Tā huā-le duō cháng shíjiān huà yī ge yuánquān?
'How long did it take him to draw the circle?'

III. 他什么时候到达了山顶?
Tā shénme shíhou dàodá-le shāndǐng?
'When did he reach the top of the mountain?'

To express the difference in the grammatical aspect (between perfective and imperfective verbs) in Standard Chinese, a lexically independent (monomorphemic) verb often does not suffice. In cases where it must be clearly stated that the action has reached its goal or endpoint, in Standard Chinese we must employ resultative verb compounds (RVC), which, simply put, are a combination of two verbs or a verb and an adjective, the former expressing the activity and the latter the result. The problem of the inconsistency of verbs in Standard Chinese with Indo-European languages is also found in Zhao, who states that "Chinese does not have accomplishment verbs (Chu, 1976; Smith, 1997; Sybesma, 1997; Tai, 1984). Since all predicates, except states, are activities, which are dynamic and have an open range, an accomplishment that denotes a bounded event is always a complex consisting of an activity/cause predicate, and a result/state predicate." (Zhao, 2005 pp. 65–66). In Zhao, we see accomplishment verbs being considered as RVC. Similar to Zhao, Tai (1984) does not distinguish between accomplishment and achievement verbs, but considers them as actions expressed by resultative verb compounds. In fact, Tai suggests that all monomorphemic verbs in Chinese are either state verbs or activity verbs, whereas if we want to express a result, we must use RCV, which always conveys a completed or finished action. We argue, however, that there is a fundamental difference between achievement and accomplishment verbs on the one hand and RCV on the other. In addition to what previous studies (Petrovčič, 2009; Xiao & McEnery, 2004) have shared on this topic, namely distinguishing between the two verb types by using them as the complement of stop, which sounds normal with accomplishments but odd with achievements, and that RCV do not function in the same way as accomplishment and achievement verbs because RCV are incompatible with imperfective markers *zai* and *-zhe*, while the use of the two markers for accomplishment and achievement verbs, followed by the perfective marker *le*, is perfectly grammatical, our empirical research has shown that there is a fundamental difference in the degree of completeness expressed by activities and accomplishments or achievements. RCV as a language category should therefore be considered as a feature of the grammatical aspect, just like the perfective marker *le*.

3.3 Verb types and classifications of the lexical aspect

A breakthrough with regard to the aspectual studies following Zeno Vendler is the work proposed by Carlota S. Smith (1991). Her contrastive analysis of the verbal aspect in English, French, Russian, Chinese and Navajo laid the foundation not only for further research on the verbal aspect, but especially for cross-linguistic research. Although Smith based her study heavily on Vendler's (1957) and Comrie's (1976) classification of the lexical aspect, her parallel comparison of several language systems enabled her to put forward new potential interpretations of the functioning of verbal actions, also by applying the principles of linguistic relativism.

Smith takes verbal actions and classifies them into five types or categories according to their inherent semantic nature. Unlike Vendler, Smith analyses not only the bare verb, but also the entire situation in the sentence. Therefore, she divides situation types into states, activities, accomplishments, semelfactives, and achievements, based on the specific characteristics that the corresponding verb types possess. Smith identifies three basic features, by which she classifies the verbs into categories of lexical aspect: [± static], [± durative] and [± telic].

[± Stative] is a property that divides verbs into two major sub-groups: states and others (activities, accomplishments, achievements and semelfactives). "States are the simplest of the situation types. In temporal schema they consist only of a period of undifferentiated moments, without endpoints […]. " (Smith, 1991, p. 28).

[± Telicity] connects a subset of actions that differ according to whether or not they are aimed at achieving a goal or a result. "... when the goal is reached, a change of state occurs and the event is completed [...]. The goal is intrinsic to the event, constituting its natural final point." (Smith, 1991, p. 29). Therefore, telic events can either be only completed or completed and at the same time accomplished.

The [± duration] property divides events into those that start and end at the same time in an internal time structure and those that have at least a minimum unit of duration to make the start and end points stand-alone events.

If we apply these features to the verb situations, states are defined as [+ static], [-telic] and [+ durative]. Activities are actions that are [+ durative], but their duration is homogeneous, so they are [-telic]. Accomplishments are also [+ durative], which unlike activities are [+ telic], as they are defined by a succession of different phases that progress to the end point when the goal is reached and, consequently, a new situation arises. Achievements are instantaneous [- durative] actions, their starting and ending point overlap, which leads to a new situation, so they are also [+ telic]. Semelfectives differ from achievements only in that their realization does not bring about any change or new situation, and therefore they are [-telic] and [-durative] actions.

For studies dealing with the Chinese verbal aspect as is the case of this paper, the findings of Richard Xiao and Tony McEnery (2004) are extremely valuable. The authors base their findings on authentic corpus data, which they interpret using statistical analysis. They also refine the categorization of the lexical verbal aspect. Instead of the three basic features of the verb situation as suggested by Smith (1991), Xiao and McEnery (2004) propose a five-way classification system; in addition to [± dynamic1], [± durative] and [± telic], they also recognise the features of [± bounded] and [± result]. By adding these two features, Xiao and McEnery try to solve ambiguities and class overlaps, because verbs that are telic have at most the potential to elicit the result or not, they therefore suggest a model in which, "[t] he feature [± telic] is associated with the presence or absence of a final spatial endpoint." (Xiao & McEnery, 2004, p. 46).

Moreover, unlike Smith (1991), who addresses lexical aspect through a verb situation, Xiao and McEnery establish it using the so-called two-level model, "in which situation aspect is modelled by ‘verb classes’ at the lexical level and as ‘situational types’ at the sentential level." (Xiao & McEnery, 2004, p. 33).

The main contribution of Xiao and McEnery, which is also relevant for the present study, is the fundamental difference between the accomplishment and achievement verbs. The authors argued that this divergence reflects “mainly in whether they do or not encode a result. [...] By the [± result] criterion, accomplishment verbs place emphasis on the process leading up to a result [...], but verbs themselves do not provide any information concerning the success in the achieving of the result; they imply but do not encode a result. […] In contrast, achievement verbs encode a result themselves.” (Xiao & McEnery, 2004, pp. 55–56). The main difference between the achievement and accomplishment verbs is thus seen in the fact that the temporal and spatial ends are encoded in the verb itself. For achievement verbs, these two points are said to already exist in the verb itself, whereas in the case of the accomplishment verbs, the endpoint is to be defined by verb arguments or complements.

The importance attached to the feature of [± telicity] and how much it is lacking, especially with regard to the properties of the Chinese verb, was again acknowledged in a study proposed by Peck et al. (2013). For this reason, the authors introduce a new feature called [± scalarity], whose characteristics are closely related to those of [± telicity]. The verb has the property of [+ scalarity] when it conveys a scalar change. This scalar property of a verb can be defined as open/closed, which corresponds to the feature of telic/atelic, for it tells us whether an action has an endpoint or not. In addition, they suggest that for durative and punctual actions, these should be defined as multi-point and two-point actions respectively. Finally, they propose four binary features for the Chinese verb (± dynamic, ± scalar, ± telic in ± punctual) and identify six classes, among which the so-called class of multi-point closed scalar verbs, equivalents of accomplishment verbs displaying [+ dynamic], [+ scalar], [+ telic] and [- punctual] features.

For Peck et al. (2013), the need to introduce the feature of [± scalarity] was motivated by the difficulty of defining telicity for verbs that exhibit a measurable scalar change (i.e. cool, darken, lengthen ...), often referred to in the literature as degree achievement verbs. However, we have also encountered similar classification problems with other verbs that cannot be classified with the standard test for telicity, such as *in one hour* for telic actions and *for one hour* for atelic actions.

For us, employing *for*- and *in*- adverbials to test [± telicity] for three very different languages has proven unreliable in several cases, which shows that the analysis of a verbal action alone is sometimes deceptive, so that a broader sentence situation should be examined or that the same verbal actions in different languages comprise some fundamental intrinsic semantic discrepancies. For example, the Italian verb to choose (scegliere) allows the use of both adverbials in the case of achievement verbs.

(4) Caterina ha scelto i vestiti per / in un'ora.
'Caterina chose the dresses in / *for an hour.'

Due to the unreliability of the test with *for*- and *in*- adverbials for cross-linguistic studies, the introduction of the feature of [± scalarity] to distinguish between verbs of activity, accomplishment and achievement has proven to be extremely valuable. In order to use a unified classification for more different languages with divergent syntactic and semantic properties, applying the feature of [± scalarity] resolved the ambiguities that arose during verb type analyses.

However, we did not fully follow what Peck et al. (2013) proposed in their study. Instead of subdividing verbs into scalar closed/open actions and scalar multi-point/two-point actions and replacing the traditional features for the lexical verb aspect, namely [± telic] and [± durative], we propose a compromise, namely the preservation of the two features and the introduction of a new distinguishing feature of [± scalarity]. Unlike Peck et al. (2013), we do not consider scalarity in its strictly mathematical meaning as a series of stages, points or intervals that indicate measurement values on a particular dimension, but as a change that occurs gradually, step by step, over time and causes a certain visible change, even on an object on which an action is performed, or a general change in the situation. By introducing scalar change, we not only solve the classification of degree achievement verbs, but we can also better define and distinguish between activity and accomplishment verbs, since it is in the latter pair that most disagreements and inconsistencies are found in previous studies.

Another important study on scalarity and the change of state for Chinese aspect was proposed by Koenig and Chief (2008). The authors offer an interesting explanation for cases in which cross-linguistic analyses of certain actions did not show semantic correspondence in achieving the result and onsetting a change of state. They explain their findings using examples they found online with the search engines Google and Baidu:

(5) 须眉和孙码字把老罗杀了没杀死。
Xūméi hé Sūn Mǎzì bǎ Lǎo Luó shā-le méi shāsǐ.
'Xu Mei and Sun Mazi killed Lao Luo, but didn't make him die (lit.).'
Intended meaning: Xu Mei and Sun Mazi tried to kill Lao Luo, but he didn't die.

(6) 我盖了新房子，房子还没盖完。
Wǒ gài-le xīn fángzi, fángzi hái méi gài-wán.
'I build a new house, but it is not finished.'

(7) 托尔斯泰的战争与和平我不喜欢，读了几次都没读完。
Tuōěrsītài-de Zhànzhēng yǔ Hépíng wǒ bù xǐhuān, dúle jīcì dōu méi dú-wán.
'I don't like Tolstoy's *War and Peace*, I read it several times, but never finished reading it.'

For the examples above where the verb is used with the aspectual marker *le*, Koenig and Chief note that they are read "as if, in those languages, there are described killings in which no death occurred, repairs in which nothing gets fixed, persuasions in which nobody was persuaded… We call this phenomenon the *Incompleteness Effect* (in short, the IE), meaning that the described killings, repairs or persuasions need not be completed." (Koenig & Chief, 2008, p. 243).

Besides the three verbs in the examples, they gave a full list of verbs which in their opinion display similar properties in regard to the IE: (*jiǎn* 剪 'to cut with scissors', *xiū* 修 'to repair', *quàn* 劝 'to persuade', *shā* 杀 'to kill', *guān* 关 'to close', *niàn* 念 'to read', *chī* 吃 'to eat', *hōng* 烘 'to dry (clothes)', *xǐ* 洗 'to wash', *zhǔ* 煮 'to cook', *dú* 读 'to read', *xiě* 写 'to write', *bèi* 背 'to recite (memorize)', *chàng* 唱 'to sing', *xiàzài* 下载 'to download', *jiāo* 教 'to teach', *gài* 盖 'to build', *zhì* 治 'to cure', *zhuā* 抓 'to catch', *diǎn* 点 'to light up' ...).

In analysing the effect, Koenig and Chief relied on three already established hypotheses about 1) the influence of one or more sentence arguments on the understanding of the verb aspect, 2) the actual meaning and effect of the *le* aspectual marker, and 3) the influence of the inherent meaning of the verbal action. On the basis of online examples related to the three hypotheses, they rejected the first two and confirmed the third one. They identify the third hypothesis as the most plausible, but argue that all existing studies have failed in proving it. What is most troubling is the fact that neither study succeeds in answering two important questions: "(1) How can the class of incomplete stems be defined in Mandarin (or Hindi or Thai)? and (2) Do incomplete stems belong to a natural semantic class?” (Koenig & Chief, 2008, p. 251).

In addition, Koenig and Chief (2008) supplement the confirmed hypothesis of intrinsic semantic differences between languages with the scalar hypothesis, which is very similar to the theory presented by Peck et al. (2013), but is presented in much more detail and extended to all verbs that somehow involve an incremental change of properties.

The identified degrees of change are defined as the highest degree on the scale, such as for the verb 杀 ‘to kill’, where the threshold represents the lowest possible degree of health or the highest possible degree of wounding. In other cases, the degree is defined culturally or individually, such as for the verb 煮 ‘to cook’, which can represent the highest degree at different levels, depending on the type of food, culture or individual taste. Based on the findings, Koenig and Chief propose: "Only those stems that denote the induced normative gradable changes can lead to the IE." (Koenig & Chief, 2008, p. 252).

In the proposed classification of change, the relation between change and the unit of time is crucial. In this, they followed Krifka (1989), who emphasizes the interplay between the change of state and the progression of the event over time, so that the activity can progress to different stages over time, which in turn affects the state of completion of the action. Koenig and Chief distinguish between actions where there is a correlation between the change in degree and the progression of the event over time (the longer we read the book, the more pages have been read) and actions where there is no such correlation (the longer we repair the computer, the more it is repaired). These changes are referred to as "*non-incremental* (non-IC), as the degree of change does not incrementally follow the temporal progression of the event." (Koenig & Chief, 2008, p. 254).

In addition to considering the progression of the event over time, Koenig and Chief (2008) also emphasise the fundamental distinction between the scales used. They also distinguish between three types of scales according to three types of incremental change, depending on whether the degree of change includes the affected object part-whole structure, the distance traversed by the theme since the event’s inception, or the degree to which the affected object has a dimensional property (such as being tall, long, or hot).

In their study, although they allow the existence of incomplete stems in other languages, such as English and French, they conclude that "the main difference would be that in Mandarin, but not English, induced non-incremental gradable change of state stems (e.g. shā ‘kill’), are incomplete.” (Koenig & Chief, 2008, p. 259).

The classifications of verb types studied above and the identification of new features that define the intrinsic nature of the verbal action are by no means exhaustive, but they are most relevant in cross-linguistic research because they help to establish a classification that can encompass, as far as possible, the characteristics of several languages at once, while maintaining the highest possible degree of universality in its interpretation.

4 Proposed classification and conclusions

Below, we present an adapted classification of lexical aspect, a proposal that we consider to be the most optimal encompassing of properties of verbal actions in different languages. We agree with those experts whose classification of the lexical aspect is based on lexemes as the main carriers of the meaning of the verbal action, even in the context of a broader sentence situation, although we are aware that different arguments can, under certain conditions, cause a change in the verb type. In the table below, the proposed features mainly consider the meaning of the verb as a lexeme, although they can also be applied to a broader sentence situation.

Table 1: Proposed features of verb types and their classification

Verb type[6]	[±dynamic]	[±scalar]	[±bounded]	[±telic]	[±result]	Example
activity	+	-	-	-	-	play
semelfact./iter semelfactive	+	-	±	-	-	sneeze
accomplishment	+	+	+	+	-	build
achievement	+	+	+	+	+	find

The proposed classification is mainly based on that proposed by Xiao and McEnery (2004) and adapted to the feature of scalarity as defined in Peck et al. (2013) and Koenig and Chief (2008). Furthermore, we propose to divide the semelfective verbs into punctual and iterative readings according to the duration of the action, but we reject the idea of considering repetitive semelfective actions as activities.

At the sentence level, where arguments must be taken into account, there are two distinctive situations. The first relates to the accomplishment verbs, which in most studies are defined as derived activities when they are not directly related to the object and therefore do not have a final spatial endpoint.

[6] State verbs differ in their features from other classified verbs - they are relatively static and they show no progressive changes through time - for which they are not relevant and have been excluded from the present study.

(8) Mati kuha, oče in sin pa čakata na hrano.
‘Mother is cooking, father and son are waiting for food.’

(9) V kuhinji mama kuha mineštro in peče palačinke.
‘In the kitchen, mum is cooking a minestrone and making pancakes.’[7]

The second relates to achievement verbs, where at the sentence level we have identified two types of viewpoint toward the action. An achievement that we see in its entirety is seen as a punctual action, where the point of onset of the action coincides with the point of completion and the onset of a new state (‘to fall’). However, achievement verbs can also be expressed through the prism of their progression as it takes place (‘falling’). Considering that, in the progressive viewpoint, achievement verbs do not change the feature [+telic] but only the [-result], we have not chosen to follow Xiao and McEnery (2004), who see the change in the telicity and therefore place it among the derived activities at the sentence level.

(10) 爸爸正在杀鸡。
Bàba zhèngzài shā jī.
‘Father is right in the middle of killing the chicken.’

(11) 我在关门。
Wǒ zài guān mén.
‘I’m just closing the door.’

(12) Zaradi njih padajo stvari z mize.
‘Because of them, things are falling off the table.’

(13) Fant pada s stola.
‘A kid is falling off the chair.’ [8]

Much more could have been explored and reported in regard to the conceptualisation and verbalisation of the verbal aspect in unrelated languages, but due to space constraints, we focused mainly on the principles of the lexical aspect in two languages that display very different properties and structural discrepancies, because of which they are very difficult to compare or consider with standard

[7] The examples have been acquired by native speakers with language tests in the empirical study Čok (2019).

[8] See Čok (2019).

classifications. We therefore suggest a reconsideration of the existing classifications, by proposing a revised one, which is particularly helpful when unrelated languages are compared with ambiguous conceptualisation and verbalisation of actions. In our study, the verbal aspect is the main research category, but the results obtained can be applied to broader linguistic and cognitive research, as in our opinion verb is one of the most fundamental language categories and its verbalisation appears essential for understanding the creation and conception of meanings and behaviours as ongoing processes in the person's mind. Also, we identify this category as one of the more semantically abstract categories, which is why mastering semantic discrepancies between a target and a native language represents one of the biggest challenges for foreign language learning.

References

Chu, C. C. (1976). Some semantic aspects of action verbs. *Lingua, 40*, 43–54.

Comrie, B. (1976). *Aspect. An introduction to the Study of Verbal Aspect and Related Poblems.* Cambridge: Cambridge University Press. Https://user.phil-fak.uni-duesseldorf.de/~filip/Comrie.Aspect.pdf.

Čok, T. (2019). Konceptualozacija in upovedovanje glagolskega dejanja v slovenščini, kitajščini in italijanščini. Doktorska disertacija. Koper: Pedagoška fakulteta, Univerza na Primorskem.

Koenig, J.-P., & Chief, L.-C. (2008). Scalarity and state-changes in Mandarin (and other languages). *Empirical Issues in Syntax and Semantics, 7*, 241–262. Http://www.cssp.cnrs.fr/eiss7/koenig-chief-eiss7.pdf.

Krifka, M. (1989). Nominal reference, temporal constitution, and quantification in event semantics. In R. Bartsch, J. v. Benthem, & P. v. Boas (Eds.), *Semantics and Contextual Expressions* (pp. 75–115). Dordrecht: Foris.

Miklič, T. (2007). Metafore o načinih gledanja na zunajjezikovna dejanja v obravnavanju glagolskega vida. *Slavistična revija: časopis za jezikoslovje in literarne vede, 35*(1/2), 85–103. Https://srl.si/ojs/srl/article/view/COBISS_ID-34762850.

Orešnik, J. (1994). *Slovenski glagolski vid in univerzalna slovnica.* Ljubljana: Slovenska akademija znanosti in umetnosti.

Peck, J., Lin, J., & Sun, C. (2013). Aspectual Classification of Mandarin Chinese Verbs: A Perspective of Scale Structure. *Language and Linguistics, 14*(4), 663–700. Http://www.ling.sinica.edu.tw/files/publication/j2013_4_02_9567.pdf.

Petrovčič, M. (2009). *Operator Le in Chinese*. Saarbrücken: VDM Verlag.

Smith, C. S. (1991). *The Parameter of Aspect.* Dordrecht: Kluwer Academic Publishers.

Smith, C. S. (1997). *The Parameter of Aspect (Second Edition).* Dordrecht: Kluwer Academic Publishers.

Sybesma, R. (1997). Why Chinese verb -le is a resultative predicate. *Journal of East Asian Linguistics, 6*, 215–62. Https://www.jstor.org/stable/pdf/20100721.pdf.

Tai, J. H.-Y. (1984). Verbs and Times in Chinese: Vendler's Four Categories. *Papers from the Parasession on Lexical Semantics*, (str. 289–296).

Thepkanjana, K., & Uehara, S. (2009). Resultative constructions with ''implied-result'' and ''entailed-result'' verbs in Thai and English: a contrastive study. *Linguistics, 47*(3), 589–618. Http://content.ebscohost.com/ContentServer.asp?T=P&P=AN&K=39880702&S =R&D=a9h&EbscoContent=dGJyMNLe80Sep7I4yNfsOLCmr06eprBSsqe4SbW WxWXS&ContentCustomer=dGJyMPGqtE6zrrFIuePfgeyx44Dt6fIA

Thepkanjana, K., & Uehara, S. (2010). Syntactic and Semantic Discrepancies among the Verbs for 'kill' in English, Chinese and Thai. *PACLIC 24 Proceedings*, (str. 291–300). Https://www.aclweb.org/anthology/Y10- 1033.

Vendler, Z. (1967). *Linguistics in Philosophy.* Ithaca: Cornell University Press.

Xiao, R., & McEnery, T. (2004). *Aspect in Mandarine Chinese. A Corpus-Based Study.* Amsterdam; Philadephia: John Benjamin Publishing Company.

Zhao, Y. (2005). Causativity in Chinese and Its Representations In English, Japanese and Korean Speakers' L2 Chinese Grammars. Doctoral dissertation. Cambridge: Faculty of Oriental Studies, University of Cambridge.

Žele, A. (2011). Leksemski in skladenjski vpliv na vidskost (na primeru slovenščine). *Opera slavica, XXI, 4*, 22–35. Https://digilib.phil.muni.cz/bitstream/handle/11222.digilib/117455/2_OperaSlavic a_21-2011-4_4.pdf?sequence=1.

THE NEW CHINESE CORPUS OF LITERARY TEXTS LITCHI

Mateja PETROVČIČ
University of Ljubljana, Slovenia
mateja.petrovcic@ff.uni-lj.si

Radovan GARABÍK
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia
garabik@kassiopeia.juls.savba.sk

Ľuboš GAJDOŠ
Faculty of Arts of the Comenius University in Bratislava, Slovakia
lubos.gajdos@uniba.sk

Abstract

The aim of the article is to introduce the corpus of Chinese literary texts and to describe the process and design principles behind the corpus construction. The authors provide information regarding the reasoning behind the chosen structure and annotation of the corpus, and further discuss possibilities the corpus opens for linguistic research and language learning. The article provides several examples of how the corpus can be used at various levels of language research.

Keywords: Chinese; corpus linguistics; building and using corpora; literary texts; Litchi

Povzetek

Namen članka je predstaviti korpus kitajskih literarnih besedil ter opisati procese in principe njegove izdelave. Sledi utemeljitev za izbrano strukturo korpusa in obrazložitev uporabljenega sistema označevanja. V nadaljevanju prispevek predstavi možnosti uporabe korpusa za jezikoslovne raziskave in učenje oziroma poučevanje jezika. Različne zahtevnostne stopnje smo avtorji tudi ponazorili s številnimi primeri.

Ključne besede: kitajščina; korpusno jezikoslovje; izdelava in uporaba korpusov; literarna besedila; Litchi

Acta Linguistica Asiatica, 10(2), 2020.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.10.2.65-81

1 Introduction

Corpus linguistics is a well-established indispensable part of linguistic research in general. We can find the most prominent use of monolingual huge corpora both in scientific research or practical uses, notably in lexicography, language education, natural language processing, and as a valuable data source in machine learning and data mining / data science.

There is a lack of corpora of different text types and genres in general, although it is indisputable that texts have many different functions in social life and result in corresponding differences in form and substance (Council of Europe, 2011, p. 93). Awareness of these features are mentioned several times in the Common European framework of reference for languages, advising the users to consider with which text types the learner will need/be equipped/be required to deal receptively, productively, interactively, and in mediation (Council of Europe, 2011, p. 96). Therefore, specialized corpora are needed in L2 acquisition as complementary learning resources.

Among freely available Chinese corpora, the BCC corpus (http://bcc.blcu.edu.cn) is an exception in this respect, since it distinguishes the following subcategories of texts: literature *wénxué* 文学, press *bàokān* 报刊, multi-domain[1] *duōlǐngyù* 多领域, *Wēibó* 微博, science and technology *kējì* 科技, and ancient Chinese *gǔ Hànyǔ* 古汉语. However, even though the BCC corpus enables users to go beyond simple search, its functions are very limited compared to CQL supported corpora.[2] Therefore, the authors realized a creation of an extra corpus of Chinese literary texts (named *Litchi*) with advanced functionality would be a useful addition to freely available corpora of Chinese language.

Perhaps surprisingly, there are only a few academic institutions in Europe that build and use corpora of Chinese language, even though Chinese as a second language and Chinese exolinguistic research is steadily gaining in popularity. Comenius University in Bratislava (Slovakia) is one of the institutes working on Chinese language corpus linguistics and corpus creation. The first corpus (the web-corpus Hanku)[3] was created in 2016, followed by the corpus of legal Chinese in 2018.[4]

Although the usefulness of language corpora is indisputable, we nevertheless sometimes encounter questions of practical considerations. Why is there a strong need

[1] In the present version, this part is called *duōlǐngyù* 多领域 (multi-domain), and in the previous versions it was called *zōnghé* 综合 (comprehensive). As stated on the website, this section includes texts from the newspapers, literature, Weibo, science and technology. These contents are independent and do not intersect with other sections of the BCC corpus. The goal of this part was to build a "balanced" corpus.

[2] See retrievable examples *jiǎnsuǒshì shìlì* 检索式示例 (http://bcc.blcu.edu.cn/help) for details.

[3] See more in Gajdoš, Garabík and Benická (2016).

[4] See more at http://158.195.113.63/run.cgi/corp_info?corpname=zh-law.

for such corpora and how can they be effectively exploited? We firmly believe that it is indispensable to make further research on registers of modern Chinese to answer such questions. Sadly, the current utilization of available corpus linguistic resources in scientific research in this area still does not reach satisfactory levels in many respects.

2 Availability

The *Litchi* is available via the website of Comenius University,[5] using the NoSketch Engine web-based corpus manager (Rychlý, 2007; Kilgarriff et al., 2014). The process of the building has begun in autumn 2019, and the corpus structure is modelled after our previous implementation to keep user access compatible across different corpora. Main parameters of the corpus are summarized in the following table.

Table 1: Parameters of the corpus Litchi

Parameters	Status	Notes
Type	synchronous	literary texts from the Internet
Language of interface	Slovak, English, Chinese, others	explanatory notes in English
Size (May 2020)	92 613 119	in tokens
Tokenization	into words (*cí* 词)	automatic statistical word segmentation
POS annotation	yes	Penn Chinese Treebank tagset
Bibliographic annotation	yes	title, author’s name, alternative author's name(s), authors’ geographical origin, gender, date of birth indication
Style and genre annotation	no	
Phonetic annotation	yes	Hànyǔ pīnyīn: tones marked by diacritics; tones marked by numerals
Syntactic annotation	yes	Penn Chinese Treebank compatible dependency annotation
Statistic tools	yes	absolute frequency, relative frequency average reduced frequency
Save results directly from the interface	yes	in text or XML format

[5] Available at https://fphil.uniba.sk/katedry-a-odborne-pracoviska/katedra-vychodoazijskych-studii/cinsky-jazykovy-korpus/litchi/.

Parameters	Status	Notes
KWIC	yes	KWIC or sentence view
Collocations search	yes	many collocation measures
Advanced search options	yes	Boolean operators—conjunction, disjunction, negation; possibility to use regular expressions at the character, word, pinyin, and metadata level; full CQL etc.
Sorting by	yes	Multi level sorting hierarchy; left, right, node, references etc.
Availability	registration required	free to use for registered users, registration not restricted

3 Corpus compilation

The Litchi corpus is compiled of freely available literary works in Chinese published on the Internet. The source texts are stored in the GB18030 character encoding (as a national standard of the People's Republic of China) (Lunde, 2009, p. 105). However, all the subsequent processing and annotation are performed in the UTF-8 encoding,[6] to ensure maximum compatibility of processing tools, corpus manager and user access.

3.1 Cleanup

Text cleanup consists of removing unwanted characters, collapsing whitespace to a single ordinary space, replacing control characters with a space, and unifying line endings. Tokenization is performed by *ZPar* (Zhang & Clark, 2011), with tokens equal to Chinese words (*cí* 词), but tokens include also numbers, punctuation and other symbols (with the exception of white space and control characters).

Tones are written either using standard diacritics or, for users lacking the means to enter Hànyǔ pīnyīn diacritics, there is a possibility to use digits 1 to 5 (with the neutral tone having the number 5). The transcription into Hànyǔ pīnyīn was performed by the *xpinyin* package.[7]

[6] In practice, we can treat the GB18030 as an alternative ASCII-extending encoding of the Unicode character repertoire (on par with UTF-8).

[7] See more at https://lxneng.com/posts/70.

3.2 Corpus structure

Since the Litchi corpus was compiled with Chinese as foreign language instruction in mind, the annotation has been designed to facilitate queries by inexperienced Chinese speakers (e.g. students of the language).[8]

3.3 Positional attributes

Positional attributes describe token-level annotation – the basic unit of the corpus is a *token*, it usually corresponds to a word, but also punctuation characters and numerals are separate tokens. Given the specifics of written Chinese language, tokens in the Litchi corpus are equal to Chinese words (*cí* 词); tokenization (word segmentation) in Chinese is a nontrivial task and a certain amount of errors is to be expected.

Each token can be assigned several attributes, further describing or specifying the token, its grammatical or lexical features. Following positional attributes are used in the Litchi corpus: *word, lemma, tag, pinyin, npinyin, head, deprel*.

The fundamental attribute ***word*** is the basic unit of the corpus (token). It is in the original form of the word (*cí* 词) in the text as written in Hànzì. Example: 斯洛文尼亚 (Slovenia).

We repurposed the default attribute ***lemma*** to be an all-encompassing default query type. It is a combination of a word written in Hànzì, each individual character (*zì* 字) of a word in Hànzì, Hànyǔ pīnyīn transcription of a ***word***, using both diacritics and numerals to mark tones, a transcription with the tones omitted, as well as a union of transcriptions of individual characters (字) of the word. We aim for inclusiveness – if a user enters a single syllable (in either Hànzì or one of the two Hànyǔ pīnyīn transcriptions, or even in Hànyǔ pīnyīn without tones), the corpus manager will search for all the words containing the syllable. For example, the word *Sīluòwénníyǎ* 斯洛文尼亚 will be assigned the "lemma" luo|luo4|luò|ni|ni2|ní|si|si1|si1luo4wen2ni2ya4|siluowenniya|sī|sīluòwénníyà|wen|wen2|wén|ya|ya4|yà|亚|尼|文|斯|斯洛文尼亚|洛.

The attribute ***tag*** is the part of speech tag, a two or three uppercase ASCII character denoting the part of speech of the *word*. For example, the word 斯洛文尼亚 will be likely part-of-speech tagged as NR (i.e. Proper Noun).

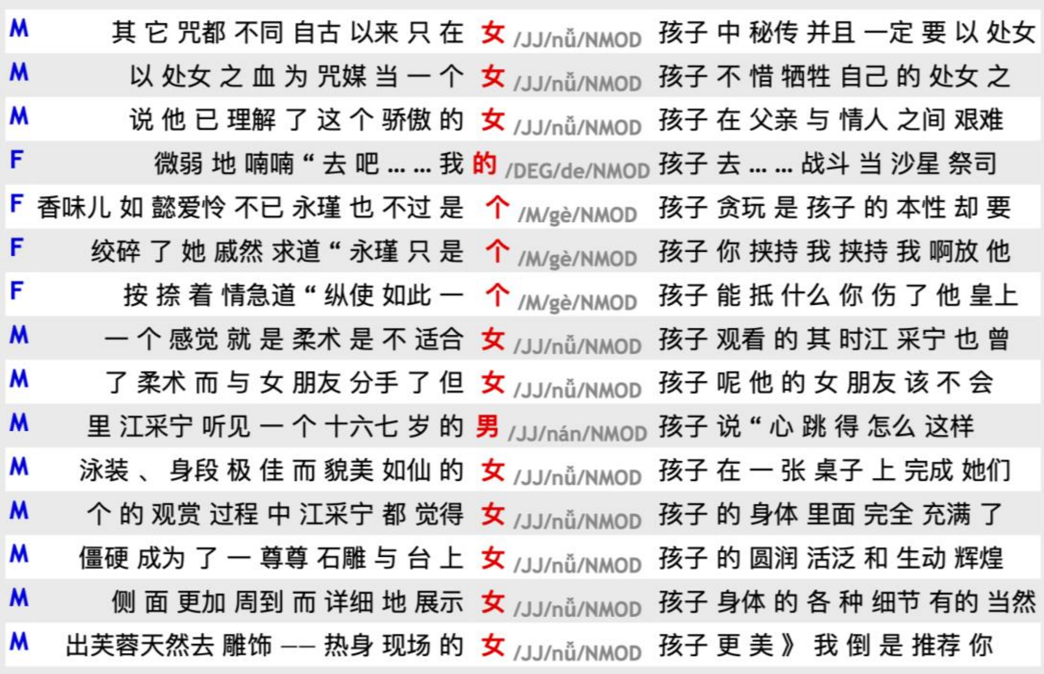
The ***pinyin*** attribute is the transcription of word using the Hànyǔ pīnyīn method, tones are indicated by diacritics. The transcription is in lowercase. For example, 斯洛文尼亚 will be transcribed as *sīluòwénníyà*. Characters with multiple readings will be assigned only the first (in some collation) reading.

[8] For details see Chapter 4.

The ***npinyin*** is again the Hànyǔ pīnyīn transcription, but this time the tones are indicated by numerals 1 to 5 (5 stands for the neutral tone). For example, 斯洛文尼亚 will be transcribed as *si1luo4wen2ni2ya4*.

Tokens in the current sentence are numbered (counted from zero) and the attribute ***head*** is the token number the current word is in relation to.

The attribute ***deprel*** marks the syntactical relation of the word (node) to the governing word (node).



M	其 它 咒都 不同 自古 以来 只 在	女 /JJ/nǚ/NMOD	孩子 中 秘传 并且 一定 要 以 处女
M	以 处女 之 血 为 咒媒 当 一 个	女 /JJ/nǚ/NMOD	孩子 不 惜 牺牲 自己 的 处女 之
M	说 他 已 理解 了 这 个 骄傲 的	女 /JJ/nǚ/NMOD	孩子 在 父亲 与 情人 之间 艰难
F	微弱 地 喃喃 “ 去 吧 … … 我	的 /DEG/de/NMOD	孩子 去 … … 战斗 当 沙星 祭司
F	香味儿 如 懿爱怜 不已 永瑾 也 不过 是	个 /M/gè/NMOD	孩子 贪玩 是 孩子 的 本性 却 要
F	绞碎 了 她 戚然 求道 “ 永瑾 只 是	个 /M/gè/NMOD	孩子 你 挟持 我 挟持 我 啊放 他
F	按 捺 着 情急道 “ 纵使 如此 一	个 /M/gè/NMOD	孩子 能 抵 什么 你 伤 了 他 皇上
M	一 个 感觉 就 是 柔术 是 不 适合	女 /JJ/nǚ/NMOD	孩子 观看 的 其 时江 采宁 也 曾
M	了 柔术 而 与 女 朋友 分手 了 但	女 /JJ/nǚ/NMOD	孩子 呢 他 的 女 朋友 该 不 会
M	里 江采宁 听见 一 个 十六七 岁 的	男 /JJ/nán/NMOD	孩子 说 “ 心 跳 得 怎么 这样
M	泳装 、 身段 极 佳 而 貌美 如仙 的	女 /JJ/nǚ/NMOD	孩子 在 一 张 桌子 上 完成 她们
M	个 的 观赏 过程 中 江采宁 都 觉得	女 /JJ/nǚ/NMOD	孩子 的 身体 里面 完全 充满 了
M	僵硬 成为 了 一 尊尊 石雕 与 台 上	女 /JJ/nǚ/NMOD	孩子 的 圆润 活泛 和 生动 辉煌
M	侧 面 更加 周到 而 详细 地 展示	女 /JJ/nǚ/NMOD	孩子 身体 的 各 种 细节 有的 当然
M	出芙蓉天然去 雕饰 —— 热身 现场 的	女 /JJ/nǚ/NMOD	孩子 更 美 》 我 倒 是 推荐 你

Figure 1: Example of the use of the attribute *deprel* (NMOD – functionally corresponds to an attributive); searching for nominal modifiers of the word *háizi* 孩子

3.4 Structures

The corpus possesses a hierarchical structure – the so-called *structures* describe information about grouping of tokens, or intra-token information. The corpus can be thus seen as a stream of tokens, interrupted by special marks denoting a start of a structure, end of a structure, or a structure between two tokens.

The Litchi corpus uses following structures (compatible with de facto standards in written language corpora):

<doc> stands for one document, which is a logically and conceptually separate standalone unit, typically a book, a short story etc. The structure contains several attributes, providing annotation of the document (metadata).

<p> marks paragraphs, units conveying a sort of coarse-grained segmentation of text; paragraphs are inferred from the structure of the text itself, without resorting to linguistic information

<s> marks sentences, segmented according to heuristic-statistical model of the *ZPar* segmentation.

The structure ***<g>***, often used in other corpora to mark that there was no whitespace between tokens is not used since spaces in written Chinese are mostly irrelevant (and not used).

3.5 Document annotation

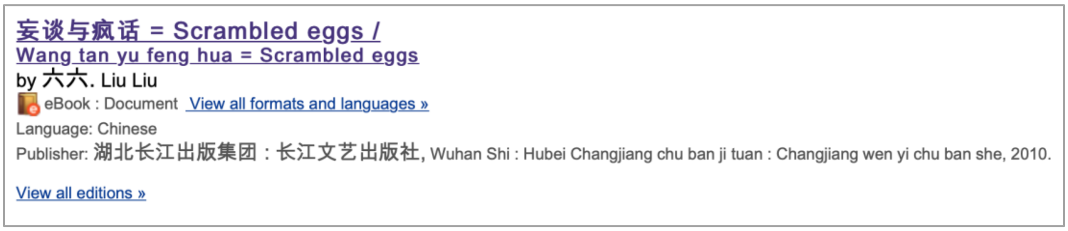
Each document has a certain set of metadata (document annotation) that are kept in the compiled corpus and can be queried or the results can be filtered by the metadata.

doc.title is the name of the document (e.g. book title), written in Hànzì. The Litchi corpus includes 1312 different literary works.

doc.author is the name of the author (pen name, if different from the real name), written in Hànzì.

doc.alter_name comprises alternative author names (either the real name or other pen names). If the author is of non-Chinese origin, this string includes the name (or multiple name variants) either in the original language or in a well-known transcription.

The motivation for this labeling was to maintain the original pair title-author. For example, according to the WorldCat, the work 妄谈与疯话 (Wàng tán yǔ fēnghuà) is written by author 六六 (Liu Liu) (see Figure 2).



妄谈与疯话 = Scrambled eggs /
Wang tan yu feng hua = Scrambled eggs
by 六六. Liu Liu
eBook : Document View all formats and languages »
Language: Chinese
Publisher: 湖北长江出版集团：长江文艺出版社, Wuhan Shi : Hubei Changjiang chu ban ji tuan : Changjiang wen yi chu ban she, 2010.
View all editions »

Figure 2: The results of the query "妄谈与疯话" in WorldCat

However, Liu Liu is a pen name of the author with the real name Zhang Xin 张辛, as provided elsewhere on the Internet, for example the Xiabook.com (https://www.shutxt.com/writer/61/).

Similarly, the author of the work entitled 尼尔斯骑鹅旅行记 (*Níěrsī qí é lǚxíngjì*) is 西尔玛·拉格洛夫 (Xī'ěrmǎ Lāgéluòfū) in the *doc.author* field, and Selma Lagerlöf in the *doc.alter_name* field.

All the bibliographic records in Litchi (including origin, gender, and age range) have been checked manually to verify and complement the meta data. As a result, 539 different authors are included in the corpus.[9] Some tasks were quite intriguing, for example the author "B·N·崔可夫" (N B Cuikefu), which stands for "Vasily Ivanovich Chuikov". In the original Chinese text, "B" is a letter of a Cyrillic script and stands for the letter "V" in Latin script. However, instead of "И" (e.g. letter "I" in Latin script), the mirror form "N" has been used. The Chinese version is therefore a mixture between "Vasily Ivanovich Chuikov" and "Василий Иванович Чуйков".

doc.authors_origin provides information on authors' origin in the geographical or linguistic sense (e.g. China, Korea, Japan, etc.) to enable the user to distinguish originally Chinese texts from the translated works into Chinese. This is a two-letter abbreviation of the region.

doc.gender is the gender of the author, we use the value self-described by the author or the gender the author is commonly considered to be of. This is not strictly a binary valued item – currently, there are three values present in the corpus annotation: M for male, F for female, N/A for unknown gender.

doc.born_in provides information related to the authors' age, in 15-year intervals. Authors born in the previous centuires have only the century of their birth recorded here (written in English, with the numeric part at the beginning).

The value "N/A" has been assigned to all the bibliographic data where no clear and straight expressions were available in the authors' online profiles. For example, if a person's brief presentation avoided the use of 3rd person personal pronouns and used neutral expressions, such as *bǐzhě* 笔者 (writer), *zuòzhě* 作者 (author), *qí* 其 (his/her), *běnrén* 本人 (I/me), it was not possible to assign a clear gender value. Similarly, if the author states to be born "in a small village in Yue nan" (生于粤南一小村), this does not necessarily mean "in the South Guangdong". Unlike the expressions Eastern/Northern/Western Guangdong (粤东/粤北/粤西), the notion *Yuè nán* 粤南 (lit. Southern Guangdong) doesn't seem to refer to any real geographical places.[10] If the description was further masked with blurred expressions such as "graduated from the BA studies at a certain university" (某名校本科毕业), this further justified the use of the "N/A" value.

The final proportion of known and unknown information for fields *doc.gender*, *doc.authors_origin* and *doc.born_in* is presented in Figue 3:

[9] The BCC corpus includes works from 469 different authors.

[10] See http://wap.yuexinet.net/view.php?aid=38

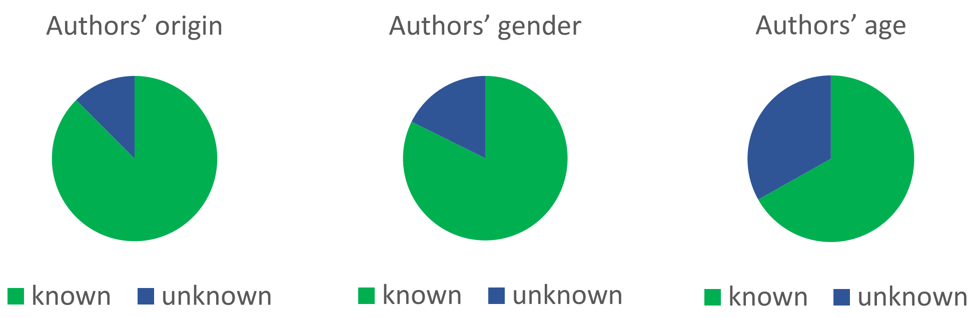


Figure 3: Proportion of known/unknown information in the authors' data

4 Usage in linguistics research

The corpus manager is a very powerful tool when in the hands of an experienced user. originally aimed at scientific research in linguistics and related fields, in the last decades of ever-increasing importance of corpus linguistics the usage of corpora converged to a subfield of descriptive linguistics with its own terminology, approaches, good practices and established rules. Nevertheless, the learning curve is not prohibitively steep, the corpus manager can even be used by completely casual users, if we prepare the corpus adequately and provide sane defaults.

For pedagogical reasons, we arbitrarily divide the corpus usage into these levels:

- basic
- advanced
- expert

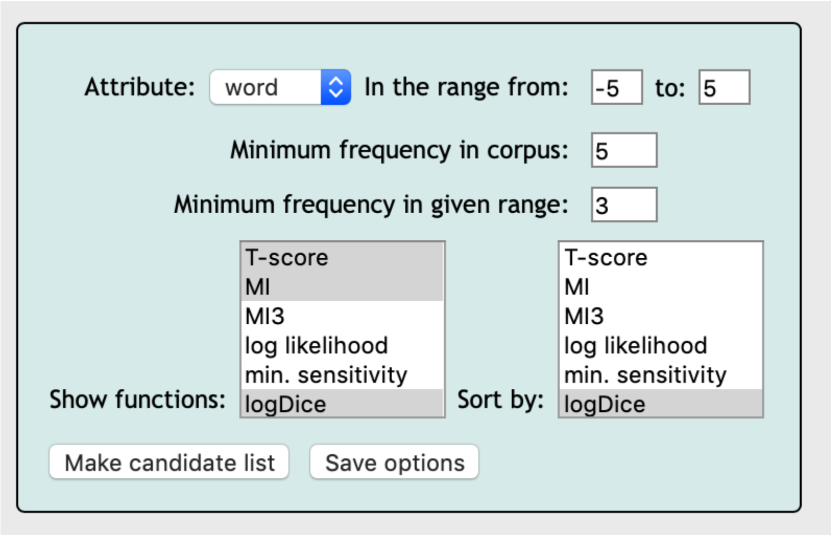
Needless to say, this division is based on our experience and the dividing lines between the levels are not strictly delineated.

4.1 Basic use

At the basic level a user may search for a word as KWIC (Key Word In Context). This is a very basic option when searching for concordance (context) of KWIC and it is very useful for students of foreign languages or translators. This usage usually does not require any additional instructions – users just type the word and get a readable list of occurrences. For Chinese language corpora, the situation is a bit complicated by the need to enter Hànzì characters. Although the plethora of input methods is a thing of the past and (in a non-professional setting) the prevalent input method is based on toneless Hànyǔ pīnyīn transcription, language model selecting (and ordering) the most probable Hànzì characters and the user picking up the appropriate character. While easy for native or fluent speakers, it can be challenging for students or less literate, less proficient non-native speakers. Also the specific tokenization matters – users have to be familiar with our chosen segmentation into words (词).

This is the basic motivation behind our *lemma* attribute – by default, the users can query the corpus by a single character (字) or a word (词); both of them can be written in Hànzì, in Hànyǔ pīnyīn with standard diacritics, in Hànyǔ pīnyīn with tones indicated by numbers, or in toneless Hànyǔ pīnyīn. Thus users with either technical obstacles preventing them typing Hànzì or diacritics, or users less proficient in written Chinese can still benefit from the corpus, by entering the search term in an intuitive way and still getting (a superset of) relevant results. This is obviously very important in teaching Chinese as foreign language.

In addition to searching for given words or characters, one of the nontrivial results we can obtain from huge corpora is the collocation analysis by various collocation measures. By default, the logDice measure is selected, empirically found to provide the best results for lexicographic purposes (and by extension, for almost any other purpose as well) (see Figure 4). The NoSketch Engine UI makes it very easy to search for collocation candidates in the corpus.



Attribute: word In the range from: -5 to: 5

Minimum frequency in corpus: 5

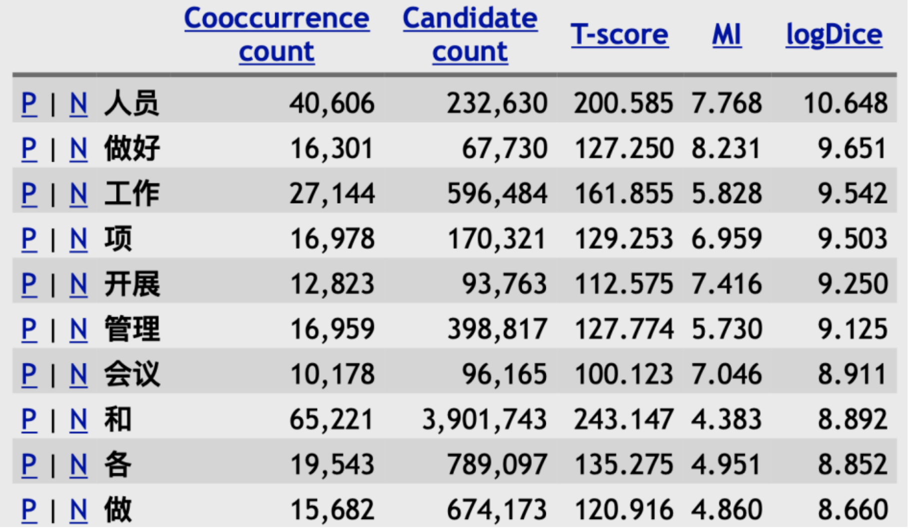
Minimum frequency in given range: 3

Show functions: T-score, MI, logDice

Sort by: logDice

Figure 4: The Collocations candidates parameters selection UI

The collocation candidates for e.g. the token *gōngzuò* 工作 (work) are presented in Figure 5.



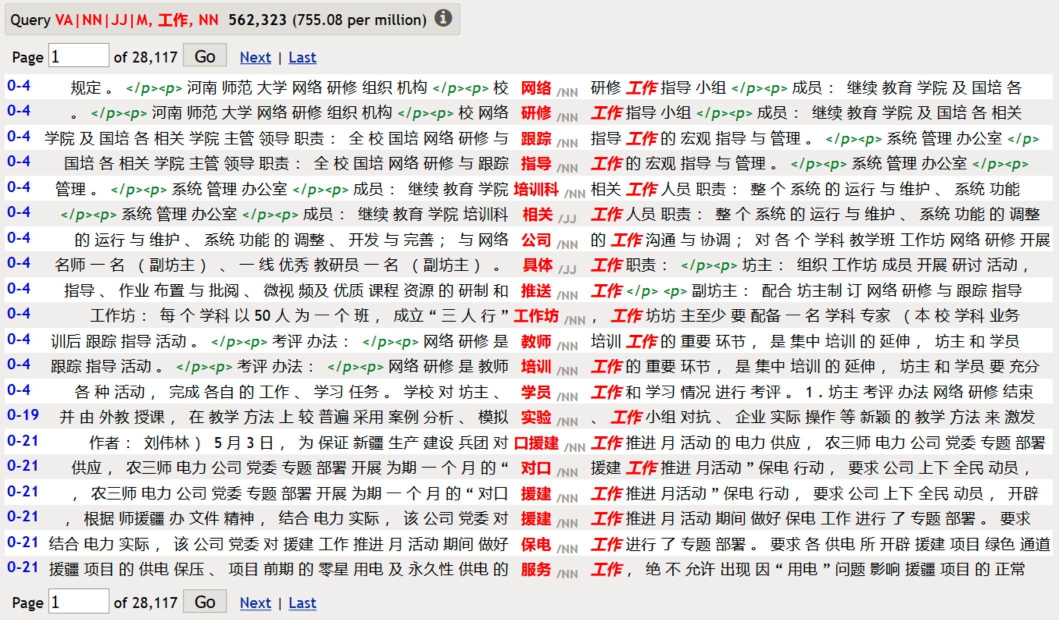
	Cooccurrence count	Candidate count	T-score	MI	logDice
P | N 人员	40,606	232,630	200.585	7.768	10.648
P | N 做好	16,301	67,730	127.250	8.231	9.651
P | N 工作	27,144	596,484	161.855	5.828	9.542
P | N 项	16,978	170,321	129.253	6.959	9.503
P | N 开展	12,823	93,763	112.575	7.416	9.250
P | N 管理	16,959	398,817	127.774	5.730	9.125
P | N 会议	10,178	96,165	100.123	7.046	8.911
P | N 和	65,221	3,901,743	243.147	4.383	8.892
P | N 各	19,543	789,097	135.275	4.951	8.852
P | N 做	15,682	674,173	120.916	4.860	8.660

Figure 5: Top 10 collocation candidates of a word *gōngzuò* 工作, word range -5 to 5 (i.e. up to five tokens in both directions)

Results of this simple query show that this word is frequently found in the phrases such as *gōngzuò rényuán* 工作人员 (staff member); it takes the verb *zuòhǎo* 做好 (to do/to finish), as in *zuòhǎo gōngzuò* 做好工作 (to do a job well), *zuòhǎo zìjǐ de gōngzuò* 做好自己的工作 (to do one's own work), *zuòhǎo yuǎnjiāo gōngzuò* 做好远教工作 (to do a distance teaching work); as a noun, it takes the classifier *xiàng* 项; it is expected to be used together with the conjunction hé 和 (with), e.g. to work with somebody, etc.

4.2 Advanced use

At the advanced level, it is possible to search for combinations of a few words conforming to a specified condition (e.g. usage of negation words (Gajdoš, 2019), concrete word order, part-of-speech tags, syntactic role, Boolean operators etc.) by using CQL expressions. In this example, we search for the most frequent attributives to the noun *gōngzuò* 工作 (work). The CQL query for this task would be (meet [tag="VA|NN|JJ|M"] 1:[word="工作" & tag="NN"]1 2). See Figure 6.

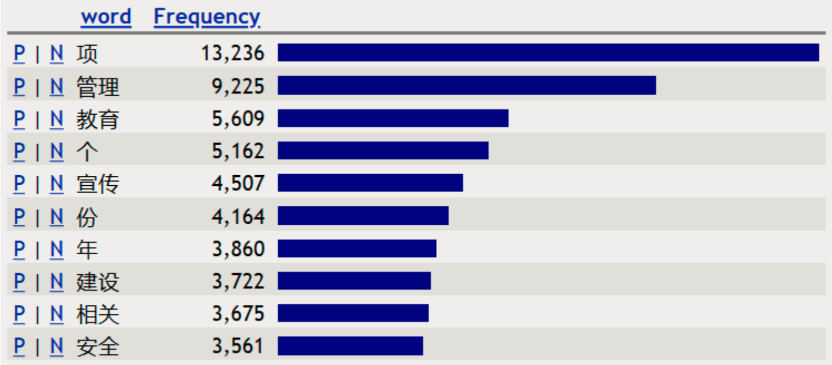


```
Query VA|NN|JJ|M, 工作, NN  562,323 (755.08 per million)

Page 1  of 28,117  Go  Next | Last
0-4    规定 。 </p><p> 河南 师范 大学 网络 研修 组织 机构 </p><p> 校  网络 /NN  研修 工作 指导 小组 </p><p> 成员 ： 继续 教育 学院 及 国培 各
0-4    。 </p><p> 河南 师范 大学 网络 研修 组织 机构 </p><p> 校 网络  研修 /NN  工作 指导 小组 </p><p> 成员 ： 继续 教育 学院 及 国培 各 相关
0-4    学院 及 国培 各 相关 学院 主管 领导 职责 ： 全 校 国培 网络 研修 与  跟踪 /NN  指导 工作 的 宏观 指导 与 管理 。 </p><p> 系统 管理 办公室 </p>
0-4    国培 各 相关 学院 主管 领导 职责 ： 全 校 国培 网络 研修 与 跟踪  指导 /NN  工作 的 宏观 指导 与 管理 。 </p><p> 系统 管理 办公室 </p><p>
0-4    管理 。 </p><p> 系统 管理 办公室 </p><p> 成员 ： 继续 教育 学院 培训科 /NN 相关 工作 人员 职责 ： 整 个 系统 的 运行 与 维护 、 系统 功能
0-4    </p><p> 系统 管理 办公室 </p><p> 成员 ： 继续 教育 学院 培训科  相关 /JJ  工作 人员 职责 ： 整 个 系统 的 运行 与 维护 、 系统 功能 的 调整
0-4    的 运行 与 维护 、 系统 功能 的 调整 、 开发 与 完善 ； 与 网络  公司 /NN  的 工作 沟通 与 协调 ； 对 各 个 学科 教学班 工作坊 网络 研修 开展
0-4    名师 一 名 （ 副坊主 ） 、 一线 优秀 教研员 一 名 （ 副坊主 ） 。  具体 /JJ  工作 职责 ： </p><p> 坊主 ： 组织 工作坊 成员 开展 研讨 活动 ，
0-4    指导 、 作业 布置 与 批阅 、 微视 频及 优质 课程 资源 的 研制 和  推送 /NN  工作 </p> <p> 副坊主 ： 配合 坊主制 订 网络 研修 与 跟踪 指导
0-4    工作坊 ： 每 个 学科 以 50 人 为 一 个 班 ， 成立 “ 三 人 行 ” 工作坊 /NN ， 工作 坊坊 主至少 要 配备 一 名 学科 专家 （ 本 校 学科 业务
0-4    训后 跟踪 指导 活动 。 </p><p> 考评 办法 ： </p><p> 网络 研修 是  教师 /NN  培训 工作 的 重要 环节 ， 是 集中 培训 的 延伸 ， 坊主 和 学员
0-4    跟踪 指导 活动 。 </p><p> 考评 办法 ： </p><p> 网络 研修 是 教师  培训 /NN  工作 的 重要 环节 ， 是 集中 培训 的 延伸 ， 坊主 和 学员 要 充分
0-4    各 种 活动 ， 完成 各自 的 工作 、 学习 任务 。 学校 对 坊主 、  学员 /NN  工作 和 学习 情况 进行 考评 。 1 . 坊主 考评 办法 网络 研修 结束
0-19   并 由 外教 授课 ， 在 教学 方法 上 较 普遍 采用 案例 分析 、 模拟  实验 /NN  、 工作 小组 对抗 、 企业 实际 操作 等 新颖 的 教学 方法 来 激发
0-21   作者 ： 刘伟林 ） 5 月 3 日 ， 为 保证 新疆 生产 建设 兵团 对 口援建 /NN 工作 推进 月 活动 的 电力 供应 ， 农三师 电力 公司 党委 专题 部署
0-21   供应 ， 农三师 电力 公司 党委 专题 部署 开展 为期 一 个 月 的 “  对口 /NN  援建 工作 推进 月活动 ” 保电 行动 ， 要求 公司 上下 全民 动员 ，
0-21   ， 农三师 电力 公司 党委 专题 部署 开展 为期 一 个 月 的 “ 对口  援建 /NN  工作 推进 月活动 ” 保电 行动 ， 要求 公司 上下 全民 动员 ， 开辟
0-21   ， 根据 师援疆 办 文件 精神 ， 结合 电力 实际 ， 该 公司 党委 对  援建 /NN  工作 推进 月 活动 期间 做好 保电 工作 进行 了 专题 部署 。 要求
0-21   结合 电力 实际 ， 该 公司 党委 对 援建 工作 推进 月 活动 期间 做好  保电 /NN  工作 进行 了 专题 部署 。 要求 各 供电 所 开辟 援建 项目 绿色 通道
0-21   援疆 项目 的 供电 保压 、 项目 前期 的 零星 用电 及 永久性 供电 的  服务 /NN  工作 ， 绝 不 允许 出现 因 “ 用电 ” 问题 影响 援疆 项目 的 正常
Page 1  of 28,117  Go  Next | Last
```

Figure 6: The results of the query
(meet [tag="VA|NN|JJ|M"] 1:[word="工作" & tag="NN"]1 2)

In the next step, results of the previous query may be ordered by Node forms in the Frequency menu, to get a list of the most frequent attributives (Figure 7).



```
           word    Frequency
P | N  项          13,236
P | N  管理         9,225
P | N  教育         5,609
P | N  个           5,162
P | N  宣传         4,507
P | N  份           4,164
P | N  年           3,860
P | N  建设         3,722
P | N  相关         3,675
P | N  安全         3,561
```

Figure 7: Top 10 most frequent attributives of the noun *gōngzuò* 工作 (work)

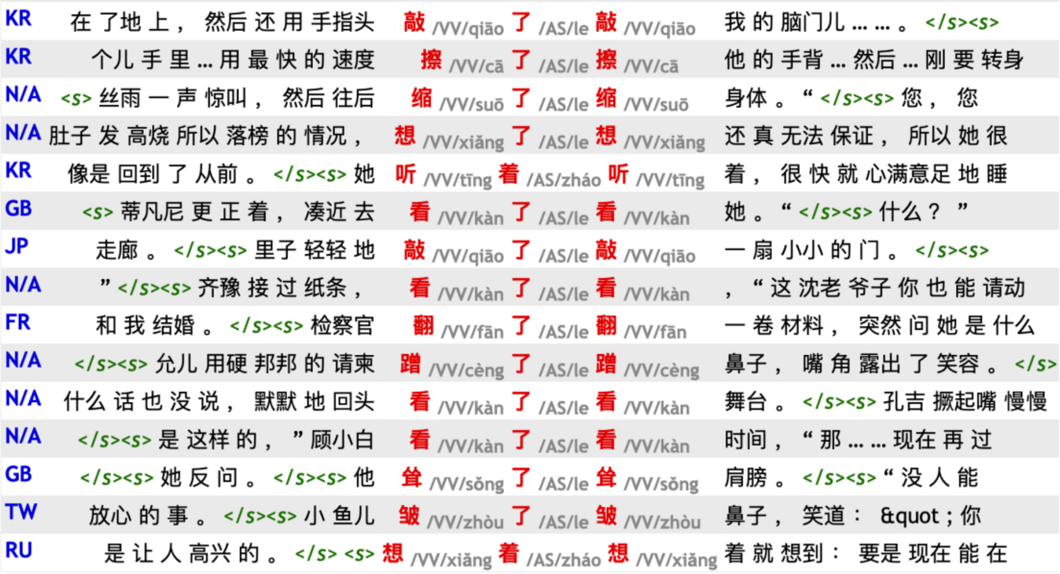
Results reveal that the most frequent noun phrases with the head noun *gōngzuò* 工作 (work) include *guǎnlǐ gōngzuò* 管理工作 (management work), *jiàoyù gōngzuò* 教育工作 (educational work), *xuānchuán gōngzuò* 宣传工作 (promotional work), *jiànshè gōngzuò* 建设工作 (construction work), and others. Its most frequent measure words are *xiàng* 项, *gè* 个 or *fèn* 份, etc.

In our opinion, this level of usage is suitable for most cases – language pedagogy as well as linguistics research.

4.3 Expert use

The expert level is an extension of the previous one and it is often used in linguistics research. The corpus manager offers an arbitrary combination of POS tags, word order, context filters (e.g. MEET, WITHIN), conditions for bibliographic annotation etc. For example, with bibliographic annotation in the Litchi corpus, it is possible to search for a concrete grammatical phenomenon in the works of one author (doc.author) or in the works of all female authors (doc.gender). The following figure demonstrates the possibility of conditions combination (search only in texts by authors *not* from Mainland China; find all "regular" verbs (VV) with an "aspect" marker (AS) followed again by the same verb).

CQL query:
(1:[tag="VV"] 2:[tag="AS"] 3:[tag="VV"] within <doc authors_origin!="CN"/>) & 1.word=3.word

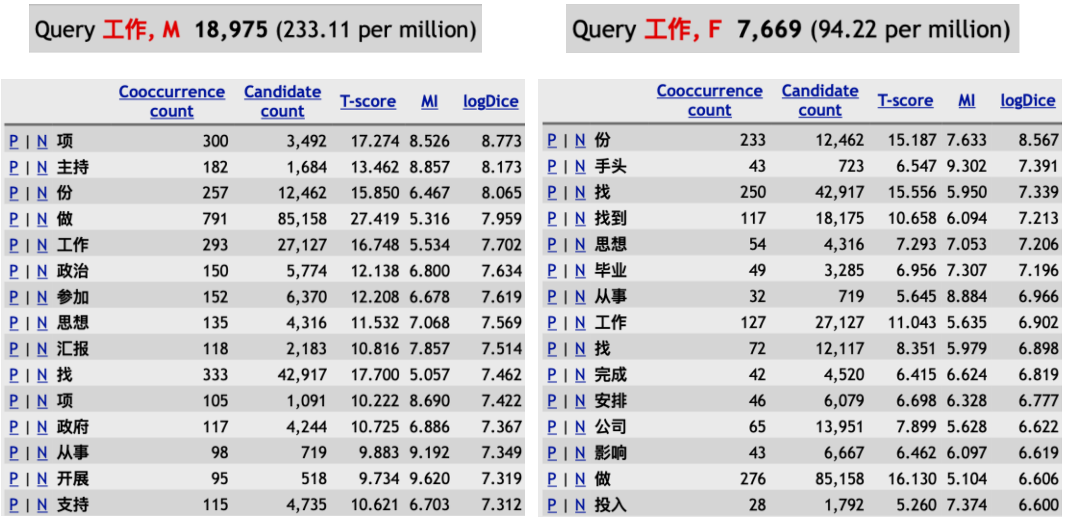


KR	在 了地 上， 然后 还 用 手指头	敲 /VV/qiāo 了 /AS/le 敲 /VV/qiāo	我 的 脑门儿 … … 。 </s><s>
KR	个儿 手 里 … 用 最 快 的 速度	擦 /VV/cā 了 /AS/le 擦 /VV/cā	他 的 手背 … 然后 … 刚 要 转身
N/A	<s> 丝雨 一 声 惊叫， 然后 往后	缩 /VV/suō 了 /AS/le 缩 /VV/suō	身体 。 " </s><s> 您， 您
N/A	肚子 发 高烧 所以 落榜 的 情况，	想 /VV/xiǎng 了 /AS/le 想 /VV/xiǎng	还 真 无法 保证， 所以 她 很
KR	像是 回到 了 从前 。 </s><s> 她	听 /VV/tīng 着 /AS/zháo 听 /VV/tīng	着， 很 快 就 心满意足 地 睡
GB	<s> 蒂凡尼 更 正 着， 凑近 去	看 /VV/kàn 了 /AS/le 看 /VV/kàn	她 。 " </s><s> 什么？ "
JP	走廊 。 </s><s> 里子 轻轻 地	敲 /VV/qiāo 了 /AS/le 敲 /VV/qiāo	一 扇 小小 的 门 。 </s><s>
N/A	" </s><s> 齐豫 接 过 纸条，	看 /VV/kàn 了 /AS/le 看 /VV/kàn	， " 这 沈老 爷子 你 也 能 请动
FR	和 我 结婚 。 </s><s> 检察官	翻 /VV/fān 了 /AS/le 翻 /VV/fān	一 卷 材料， 突然 问 她 是 什么
N/A	</s><s> 允儿 用硬 邦邦 的 请柬	蹭 /VV/cèng 了 /AS/le 蹭 /VV/cèng	鼻子， 嘴 角 露出 了 笑容 。 </s>
N/A	什么 话 也 没 说， 默默 地 回头	看 /VV/kàn 了 /AS/le 看 /VV/kàn	舞台 。 </s><s> 孔吉 撅起嘴 慢慢
N/A	</s><s> 是 这样 的， " 顾小白	看 /VV/kàn 了 /AS/le 看 /VV/kàn	时间， " 那 … … 现在 再 过
GB	</s><s> 她 反 问 。 </s><s> 他	耸 /VV/sǒng 了 /AS/le 耸 /VV/sǒng	肩膀 。 </s><s> " 没 人 能
TW	放心 的 事 。 </s><s> 小 鱼儿	皱 /VV/zhòu 了 /AS/le 皱 /VV/zhòu	鼻子， 笑道： &quot; 你
RU	是 让 人 高兴 的 。 </s> <s>	想 /VV/xiǎng 着 /AS/zháo 想 /VV/xiǎng	着 就 想到： 要是 现在 能 在

Figure 8: The combination of conditions in CQL

Or, to continue with an example using *gōngzuò* 工作 (work), it can be observed, that male authors tend to write more about work than female authors. Moreover, their focus seems to be on different aspects of work, as roughly indicated in the data. The most frequent collocation candidates in men works are *zhǔchí gōngzuò* 主持工作 (take charge of the work), *zhèngzhì gōngzuò* 政治工作 (political work) or *cānjiā gōngzuò* 参加工作 (participate in work); whereas the most frequent collocation candidates in works of female authors include *shǒutóu gōngzuò* 手头工作 (work at hand), *zhǎo gōngzuò* 找工作 (to look for a job), *zhǎodào gōngzuò* 找到工作 (to find a job).[11]

[11] For more relevant results, a thorough research should be conducted.



Query 工作, M 18,975 (233.11 per million)

	Cooccurrence count	Candidate count	T-score	MI	logDice
P | N 项	300	3,492	17.274	8.526	8.773
P | N 主持	182	1,684	13.462	8.857	8.173
P | N 份	257	12,462	15.850	6.467	8.065
P | N 做	791	85,158	27.419	5.316	7.959
P | N 工作	293	27,127	16.748	5.534	7.702
P | N 政治	150	5,774	12.138	6.800	7.634
P | N 参加	152	6,370	12.208	6.678	7.619
P | N 思想	135	4,316	11.532	7.068	7.569
P | N 汇报	118	2,183	10.816	7.857	7.514
P | N 找	333	42,917	17.700	5.057	7.462
P | N 项	105	1,091	10.222	8.690	7.422
P | N 政府	117	4,244	10.725	6.886	7.367
P | N 从事	98	719	9.883	9.192	7.349
P | N 开展	95	518	9.734	9.620	7.319
P | N 支持	115	4,735	10.621	6.703	7.312

Query 工作, F 7,669 (94.22 per million)

	Cooccurrence count	Candidate count	T-score	MI	logDice
P | N 份	233	12,462	15.187	7.633	8.567
P | N 手头	43	723	6.547	9.302	7.391
P | N 找	250	42,917	15.556	5.950	7.339
P | N 找到	117	18,175	10.658	6.094	7.213
P | N 思想	54	4,316	7.293	7.053	7.206
P | N 毕业	49	3,285	6.956	7.307	7.196
P | N 从事	32	719	5.645	8.884	6.966
P | N 工作	127	27,127	11.043	5.635	6.902
P | N 找	72	12,117	8.351	5.979	6.898
P | N 完成	42	4,520	6.415	6.624	6.819
P | N 安排	46	6,079	6.698	6.328	6.777
P | N 公司	65	13,951	7.899	5.628	6.622
P | N 影响	43	6,667	6.462	6.097	6.619
P | N 做	276	85,158	16.130	5.104	6.606
P | N 投入	28	1,792	5.260	7.374	6.600

Figure 9: Comparison of frequency and collocation candidates for the word *gōngzuò* 工作 (work) in relation to authors' gender

Data also show that there are more works written by men, but this does not influence the relative frequency of the selected word.

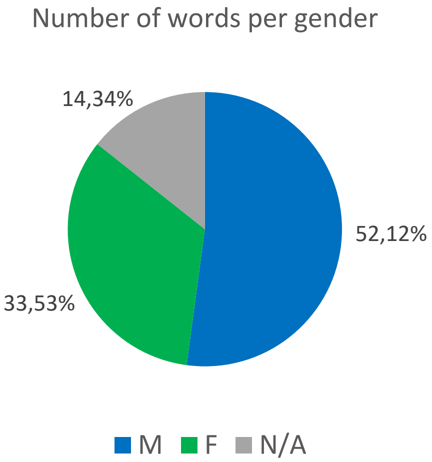


Figure 10: Distribution of the gender annotation value, as a percentage of the number of words (词) in the corpus

Following two tables demonstrate the use of the corpora to identify keywords that are more relevant in one corpus, as compared to the second (reference) corpus, using the Simple maths method (Kilgarriff 2009) – the words with their relative frequency much higher in one corpus. We focus on rare words.

Table 2: Comparison of most relevant keywords in the *zh-law* corpus, as compared against *zh-lit* as the reference corpus

	zh-law		zh-lit		
word	Freq	Freq/mill	Freq	Freq/mill	Score
行政	33949	4712.4	197	2.4	1378.1
国务院	11302	1568.8	27	0.3	1178.8
下列	9974	1384.5	15	0.2	1169.9
应当	57301	7953.8	530	6.5	1059.1
条例	6761	938.5	7	0.1	865.1
规定	50465	7004.9	583	7.2	858.3
生产	16099	2234.6	161	2.0	750.7
认证	5486	761.5	3	0.0	735.4
直辖市	5970	828.7	12	0.1	723.1
申请人	5643	783.3	11	0.1	690.9
法规	8449	1172.8	59	0.7	680.5
共和国	8772	1217.6	71	0.9	650.9
注册	8703	1208.0	83	1.0	598.6
自治区	6212	862.3	43	0.5	564.9
受理	4655	646.1	14	0.2	552.2

Table 3: Comparison of most relevant keywords in the *zh-lit* corpus, as compared against *zh-law* as the reference corpus

	zh-lit		zh-law		
word	Freq	Freq/mill	Freq	Freq/mill	Score
想	200,276	2460.4	4	0.6	1582.7
那	266,923	3279.2	16	2.2	1018.4
什么	219,236	2693.4	12	1.7	1010.8
你	720,026	8845.7	65	9.0	882.7
吧	80,800	992.6	1	0.1	872.5
那	149,508	1836.7	16	2.2	570.6
那么	48,539	596.3	2	0.3	467.5
却	124,025	1523.7	18	2.5	435.8
里	208,813	2565.3	39	5.4	400.1
句	41,029	504.0	2	0.3	395.3
呢	48,497	595.8	4	0.6	383.7
为什么	39,129	480.7	2	0.3	377.0
拿	32,804	403.0	1	0.1	354.8
东西	29,531	362.8	1	0.1	319.5
现在	69,307	851.5	13	1.8	304.0

Last but not least, the Litchi mainly reflects the language use of speakers born in recent decades, as shown in Figure 11. Therefore, this corpus is also appropriate for studies focusing on some specific features of the most recent language use.

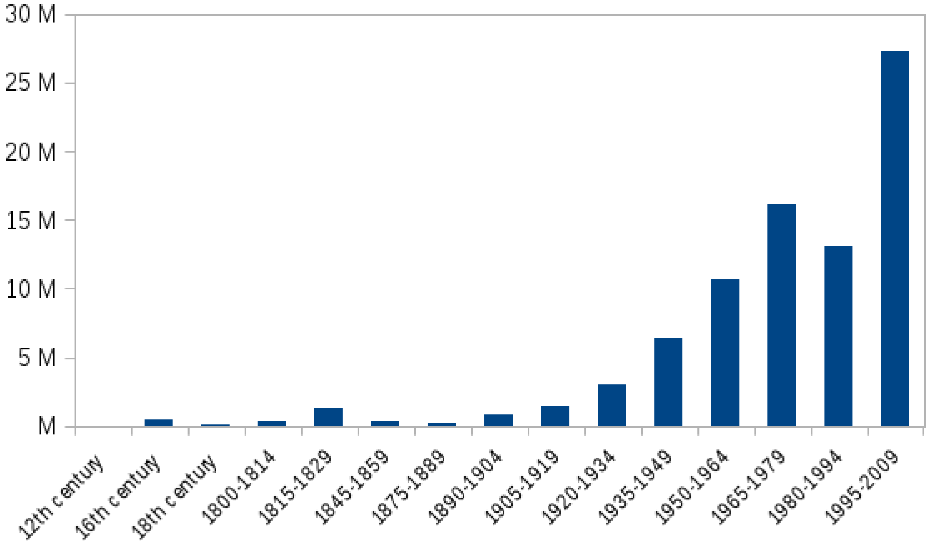


Figure 11: Distribution of author birth dates, by the number of words (词) in the corpus

5 Conclusion

The Litchi corpus is the third corpus of a family of Chinese language corpora used at the Comenius University. It adds a corpus of a different language variety and register to the existing corpora of Chinese (texts of laws, web corpus), while keeping compatible structure and annotations. The corpus manager offers the possibility of quantitative/qualitative analysis of various Chinese language registers - comparison of the three corpora, but it can also be used for comparison between. Chinese language usage in different situations or contexts (e.g. between translation and original texts; analysis of different expressions used by authors based on their gender, historical period etc.). The corpus is accessible through a web interface upon registration and aims to be a valuable resource for both teachers and students of Chinese as a foreign language, but also for linguistic research.

To conclude, the Litchi is a unique corpus in many respects. It provides a rich bibliographic, phonological, morphological and syntactic annotation and thus offers wide range of possibilities for linguistics research, e.g. lexicography/lexicology, morphology, syntax and to some extend also sociolinguistics.

Acknowledgments

This work was partially supported by the Slovenian Research Agency under research program Asian languages and cultures (P6-0243).

References

Council of Europe. (2011). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, U.K: Cambridge University Press. https://rm.coe.int/1680459f97

Gajdoš, Ľ. (2019). Retrieving Linguistic Information from a Corpus on the Example of Negation in Chinese. *Acta Linguistica Asiatica*, *9*(2), 103-115. https://doi.org/10.4312/ala.9.2.103-115

Gajdoš, Ľ., Garabík, R., & Benická, J. (2016). The New Chinese Webcorpus Hanku – Origin, Parameters, Usage. *Studia Orientalia Slovaca, 15*(1), 53–65.

Kilgarriff, A. et al. (2014). The Sketch Engine: Ten Years on. *Lexicography*, 1.1, 7-36.

Kilgarriff, A. (2009). Simple maths for keywords. In M. Mahlberg, V. González-Díaz & C. Smith (Eds.), *Proceedings of Corpus Linguistics Conference CL2009*. University of Liverpool, UK.

Lunde, K. (2009). CJKV Information Processing: Chinese, Japanese, Korean & Vietnamese Computing, 2nd edition. O'Reilly Media.

Rychlý, P. (2007). Manatee/Bonito-A Modular Corpus Manager. In P. Sojka & A. Horák (Eds.), *RASLAN 2007* (pp. 65-70). Brno: Masaryk University.

Zhang, Y., & Clark, S. (2011). Syntactic Processing Using the Generalized Perceptron and Beam Search. *Computational Linguistics*, *37*(1), 105-151.

THE MANY MEANINGS OF THE JAPANESE CAUSATIVE: WIDENING THE PRAGMATIC TAKE ON THE -*(SA)SERU* CAUSATIVE SENTENCE

Petra JAKLIN
University of Tsukuba, Japan
5rajaklin@gmail.com

Abstract

Causative sentences have long been a topic of research in Japanese linguistics due to the different meanings expressed by the use of the -*(s)aseru* inflection forming the causative verbal form. This paper presents a wider range of possible interpretations and meanings carried by Japanese causative sentences, the analysis of which was based on Fukada's (2010) paper. The aim of this paper is to present the Japanese causative in simple terms, with the focus placed on the causer and the causee, i.e. on how their relationship connects to the overall meaning of the sentence. Through the analysis of example sentences, the transfer and expression of different meanings will also be discussed on examples from English and Croatian. Since the meaning of causative sentences often depends on the context, interpretations presented in this paper can serve as guidelines to understanding the versatility of the Japanese causative, and help when expressing nuances of meaning in other languages.

Keywords: causative; meaning; Japanese; causer; causee

Povzetek

Vzročni (kavzalni) odvisniki so pogosto obravnavani v japonskem jezikoslovju, predvsem zaradi številnih različnih pomenov, ki jih lahko izraža vzročna pripona -*(s)aseru* in posledično vzročna glagolska oblika. Na osnovi klasifikacije, ki jo predlaga Fukada (2010), članek obravnava različne možne interpretacije in pomene, ki jih lahko izražajo vzročni odvisniki. Namen članka je čim bolj enostavno predstaviti vzročnost v japonskih povedih, pri čemer se osredotoča na povzročitelja in z dejanjem prizadetega, oz. na dejstvo, kakšno je njuno razmerje v povezavi s celotnim pomenom povedi. Na kratko je obravnavan tudi način, kako so isti pomeni povedi izraženi v angleškem in hrvaškem jeziku. Glede na to, da je pomen vzročnih stavkov mnogokrat odvisen od konteksta, lahko možne interpretacije povedi v tem članku služijo kot rdeča nit pri razumevanju vzročnih povedi in izražaju le-teh v drugih jezikih.

Ključne besede: vzročni odvisnik; pomen; japonščina; povzročitelj dejanja; prizadeti v dejanju

Acta Linguistica Asiatica, 10(2), 2020.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.10.2.83-103

1 Introduction

The Japanese causative verb and its uses in sentences have been a point of interest to many linguists and a stumbling block to Japanese language learners due to its inflexibility of form. In simple terms, the Japanese causative verbal form is easily identifiable by its *-(s)aseru* inflection. Although it might be easily distinguishable, the problem with having an unaltering verbal form is that its meaning is not easily understandable. As opposed to Japanese, both English and Croatian have a very small number of causative verbs that are formed using inflections, not to mention that there exists no designated causative verbal form; hence the Japanese causative has no counterpart in these two (or for that matter, many other) languages. Therefore, rendering the Japanese causative sentence and its meaning into other languages might prove a challenging task syntactically, semantically and pragmatically.

In this paper I analyze Japanese causative sentences that appear in the Haruki Murakami's novel *Noruwei no mori* (orig. ノルウェイの森, Engl. transl. *Norwegian wood*), a novel that has been known to a wider European readership, and discuss their meanings using Fukada's (2010) classification. I focus on the causer and the causee (animate/inanimate) within a sentence, and the variety of meanings of such sentences in general. Through this analysis I give an insight into the possible understanding and interpretations of causative sentences and clarify in which contexts they are to be used. Lastly, I touch upon how the meanings of such sentences were transferred into English and, if possible, suggest solutions for Croatian.[1] By doing so, different meanings of the Japanese causative are compared to a major European language, English, and contrasted with some possible ways of expressing causation in another, albeit minor, European language, showing the variety of causative expressions.

The general aim of this paper is to present causation in simple terms and make it more understandable for those who carry interest in linguistics and/or Japanese, and especially for learners of Japanese.

2 Defining the causative

2.1 The Japanese *-(s)aseru* causative

The causative expresses that "somebody made somebody else do something", such as "The lady made me call my boss". However, sentences such as "The lady made me angry." are also valid. To broaden the definition, the causative represents the influence

[1] There exists no direct translation of the aforementioned novel *Norwegian wood* in Croatian. The Croatian expressions mentioned come from the Croatian translation of the English version of the book as well as the author's own suggestions; the latter should be considered as possibilities of expression.

exerted by entity A on entity B (be that influence emotional, material or of some other nature; see Glumac (2015, p. 19), causing a change is entity B's emotional or physical state (see Baron (1987) for the temporal definition of the causative action). In order to exert influence, entity A – the causer, is usually animate and exercises some sort of power and/or willpower over entity B – the causee. In general, the causer uses their higher status to either make the causee perform an action, or to allow the causee to carry out an action. These are the two basic meanings of the causative scholars agree on – coercion and permission (Heycock, 1987; Glumac, 2015; Fukada, 2010; Hayatsu, 2016, etc.). These two meanings have a similar syntactical structure in Japanese, and their understanding is, in some cases, heavily dependent on the context they are used in (Glumac, 2015, p. 217).

(1a) 姉は弟を**家事させた**。
Ane wa otōto o kajisaseta.
'The older sister made the younger brother do the housework.'

(1b) 姉は弟に**歌わせた**。
Ane wa otōto ni utawaseta.
'The older sister let the younger brother sing.'

In (1a), the sister is making (coercing) her younger brother to do housework, i.e. the causer (the entity causing the action to happen)[2] is exerting her influence on an (un)willing causee (the entity doing the action)[3] of a "lower status". These kinds of simple sentences can be distinguished by the difference in the particle に (*ni*) and を (*wo*), namely the difference in case, but the following sentence might not be as easily understandable:

(2) JP: 先に先にと**行かせないで**立ちどまって**考えさせる**こと。
(ノルウェイの森(下), p. 14)
Saki ni saki ni to ikasenaide tatchidomatte kangaesaseru koto.
EN: 'And you don't let them rush ahead from one thing to the next: you make them stop and think.' (Norwegian wood, p. 202)
CR: 'I ne smiješ ih pustiti da jure s jedne stvari na drugu, nego ih natjerati da stanu i razmisle.' (Norveška šuma, p. 176)

[2] Hayatsu (2016).

[3] Hayatsu (2016).

Example (2) does have a coercive meaning, but unlike simple sentences preceding it, its meaning is dependent on the context. Furthermore, such sentences are predominant when it comes to the Japanese causative.

As already mentioned, the Japanese causative is easily distinguishable thanks to the usage of the *-(s)aseru* morpheme, which is agglutinated to the base of the verb. In simple sentences with an easily distinguishable structure and case pattern, the *ni* causative is permissive and the *wo* causative coercive.[4] The possible syntactical and semantical meanings of both causative verbs and causative sentences employing those verbs have been discussed by scholars; in addition, it is important to identify the causer and the causee in order to correctly interpret the Japanese causative sentence, which is the focus of this article.

2.2 Introduction of the English and Croatian causative

To reiterate, the causative construction has two basic meanings – that of "coercion" and that of "permission", which could easily be expressed by the verbs "make" and "let" in English, and "natjerati" and "dopustiti" in Croatian. Fukada (2010) used a variety of English terms for expressing different meanings of causative sentences he discussed; for the two basic usages, Fukada (2010) used expressions "to get [somebody] do [something]", "(Someone) made [somebody unwilling][do something]" (p. 21, 28) to express the coercive, and "To let [somebody] [do something]", "(someone) let [somebody] [do something]" (p. 21, 29) to convey the permissive meaning. The author notes that there are variations to these expressions which can be used in much the same manner, such as "to allow" instead of "to let", or "to order somebody to do something" instead of "to make somebody do something" (also see Baron (1974)). Similarly, downscaling is possible, such as "to ask somebody to do something" instead of "to make somebody do something". Such variations are dependant on different factors such as the setting and the relationship between the causer and causee[5].

Baron (1987) defines causation as a relationship between the state of affairs X and X' at the times T_1 and T_2, respectively, with conditions Z necessary for the action to happen (1974, p. 299). In her work Baron discusses the types of causation in English (morphological and syntactic) and the means of expression of both, as well as English causative verbs. Causation is possible and actually very commonly expressed in English through suppletion, e.g. believe-persuade (Baron, 1974, pp. 303–304), through lexical (same verb) causatives (e.g. bend), derivation from adjectives and nouns (e.g. just-

[4] See Heycock (1987).

[5] The relationship between the causer and causee needs to be paid attention to when translating from Japanese, as the Japanese language is especially sensitive to such factors, and expresses them in both written and spoken language.

justify), and syntactically[6] with periphrastic causative constructions formed by combining the verbs have/make/get/cause/let with a complement, e.g. make cry, or object of result, (e.g. The bomb demolished the wooden structure. (Baron, 1974, p. 309)) (Baron, 1974, pp. 302–310). When discussing causation as an underlying form (Baron, 1974, p. 312), Baron mentions Lakoff's derivation of causative verbs from adjectives through the inchoative rule, such as in "The metal hardened." from the adjective *hard* (Baron, 1974, p. 312). The author later writes about the inherent inchoative of a causative sentence when discussing the structure of X at T_1 (Baron, 1974, p. 317). She points out that the reason causative sentences are inchoative is because the state X at T_1 must be able to change into X' at T_2 due to Z, otherwise there can be no causation.

Similar to English, Croatian does not have a dominant morphological causative form like the Japanese *-(s)aseru*, and morphological causatives are not common[7] (Sinčić, 2018). In order to present a wider definition of causation, I will refer not only to works in Croatian, but also in two other Slavic languages closely related to Croatian, namely Slovene and Serbian.[8]

Glumac (2015) compares the Japanese *-(sa)seru* causative to its Serbian translations in her doctoral thesis, focusing on semantic and syntactic features. Glumac's work gives an overall picture of the problems faced when translating the Japanese causative and presents the issues in research of the Japanese causative expressions that may possibly be connected to Slavic languages. She also states that causation cannot be uniformly translated into Serbian since it is at times heavily dependent on the context of the sentence. Similarly, Shigemori-Bučar (2006) analysed and contrasted the Japanese causative to its Slovene verbal equivalents from a typological standpoint using a corpus analysis. In her work she compares causative expressions in Slovene and Japanese, giving an overview of the type and nature of verbs used to form the causative in these two languages, discusses the agency of the verbs and the roles of the causer and the causee, presenting the range of verbs used to translate the Japanese causative verbal form.

Little research is done on the causative in Croatian, and Sinčić (2018) gives a general overview of the Croatian causative in her master's thesis. She lists the forms of the causative and factitive expressions while comparing them to the French construction "faire+infinitive". In addition, Žagar-Szentesi (2011) writes about the grammaticalization of the Croatian *dati (se) + infinitive* construction, which, amongst others, carries a causative meaning of "having somebody do something for the subject

[6] Baron also mentions the quasi-causative (1974, p. 309).

[7] Sinčić also referred to V.A. Plungjan (2016, p. 73), writing that morphological expression of causation is not typical for Slavic languages.

[8] Since all three of these languages are similar (Croatian and Serbian especially so), the (basic) causative expressions are generally similar as well.

(causer)", in which case the causee (doer of the action) is in most cases unknown/not expressed (2011, pp. 305–306). Although "dati (se)+infinitive" is an expression carrying a causative meaning, it is not an expression which conveys a notable amount of pressure put on the causee as opposed to, e.g. the imperative or the verb "natjerati [nekoga] (da)", so the author believes there exists a possibility of using "dati (se) + infinitive" for expressing some of the Japanese *-temorau* sentences in causative use (as mentioned in Jaklin (2020)).[9]

The findings on causative in the three Slavic languages and English, and their further comparison to Japanese, brings us to the conclusion that transferring causation, the meaning of which in many cases is not apparent, does not have a set pattern. However, if the meanings of Japanese causative sentences get classified in as much detail as possible, useful guidelines on which expression suits a particular meaning could be formed.

3 Research methodology

3.1 Research question

A single causative verbal form used to express both coercive and permissive causative sentences might present a difficulty in interpreting complex sentences that do not strictly follow the formula X は Y に/を V させる:

(3) JP: それからその子にもう一度**弾かせる**の。(ノルウェイの森(下), p. 12)
Sorekara sono ko ni mō ichido hikaseru no.
EN: 'Then I'd have her play the piece again, and her performance would be ten times better than the first time though.' (Norwegian wood, p. 201)
CR: 'Onda bih joj rekla da je opet odsvira, i izvedba bi joj tada bila deset puta bolja nego prvi put.' (Norveška šuma, p. 175)

The above sentence can be seen as either coercive or permissive without context – either the teacher "makes" the child play or "lets" them play (the piano) one more time. In *Norwegian wood* the context is lightly coercive (the teacher has the child play the composition once again during their first class so that she can judge their ability), but without that background the sentence itself is ambiguous. The same cannot happen in English, where either the verbs "let" or "make (do)" would be used.

The same problem was addressed by Fukada (2010), who describes the morpheme *-(s)aseru* as ambiguous, and raises a question on how to understand which

[9] Jaklin (2020).

interpretation was intended (2010, p. 22). He suggests that pragmatics may give an answer to that question (2010, p. 22), and concludes that if the interpretations are indeed pragmatic inferences, the *-(s)ase* morpheme only means "to cause" in general (2010, p. 40). Therefore, Fukada (2010) gives several possible meanings a causative sentence may convey, and the analysis in this research bases on his classification.

3.2 Research scope and the data

The causative sentences analysed and cited in this paper were extracted from the original Japanese literary work *Noruwei no mori* (orig. ノルウェイの森) written by Haruki Murakami, and its direct English translation *Norwegian wood* (translated by Jay Rubin). For Croatian, examples were taken from the indirect translation into Croatian called "Norveška šuma" (translated by Maja Tančik), some of the examples are author's own translation variants. Causative sentences quoted from the original work *Noruwei no mori* are causative sentences with the *-(s)aseru* causative morpheme, i.e. sentences employing verbs in the causative form (行く(iku)→行かせる(ikaseru)); verbs 思わせる (omowaseru), 曇らせる (kumoraseru), 寝かせる (nekaseru) and 眠らせる (nemuraseru) are also included, since they carry causative meaning of "to make somebody believe, give the impression" for 思わせる(omowaseru), "to cloud, to make dim or dull" for 曇らせる(kumoraseru), "to put to bed" for 寝かせる(nekaseru) and "to put to sleep" 眠らせる (nemuraseru), as defined by Jisho.org. A total of 141 Japanese sentences were extracted and analyzed; any brackets, bold letters or underlining seen in the sentences cited in this paper are changes made by the author and, unless stated differently, are cited from *Noruwei no mori*, its English translation *Norwegian wood*, or its Croatian translation *Norveška šuma*.

The romanization system used throughout his work is the Hepburn system.

The sentences are analyzed according to the pragmatic analysis by Fukada (2010), and compared to the English and Croatian translations to investigate the expressions used, and the ways their meanings were transferred in translations.

4 Analysis

According to Fukada (2010), whose analysis discusses Shibatani's on the sociative meaning of the Japanese causative, there are several pragmatic meanings the causative can express: manipulative, coercive, permissive, hands-off, adversity, and sociative pragmatic meaning. In this paper, definitions for each of these meanings will be presented as given by Fukada (2010), and discussed through examples from Murakami's novel in an attempt to widen their possible definition.

Through the analysis, the following points are taken into consideration.

1. The causer and the causee (animate/inanimate)
2. The power relationship between the two (and possible external influences)
3. The overall meaning of the sentence and its usage

In this analysis, all animate agents are human entities {+human, +animate}, while inanimate agents are all non-human entities (i.e. natural phenomena, objects, etc.) and are marked as {-human, -animate}[10].

Definitions of Fukada's interpretations will be discussed through example sentences, and will be elaborated further in order to present possible meanings in as many details as possible; the meanings for which variations of Fukada's original definition were not found will be considered to be out of scope. In addition, English and Croatian expressions will be discussed according to each meaning.

It is here to be mentioned that the analyzed sentences were cited from a literary work, and although the meanings and the classification discussed are considered applicable to causative sentences in different registers (i.e. non-fiction writing or spoken language, etc.), the usage of the *-(sa)seru* causative itself might differ in non-literary Japanese.

4.1 Manipulation

Fukada's manipulative meaning is defined as follows – when the causee has no propensity over the caused action, such as in Fukada's example where the patient cannot take the medicine on their own and the causer (presumably doctor or caretaker) has to bring about the event, i.e. make the patient take the medicine (Fukada, 2010, p. 29).

There is a representative example in *Noruwei no mori*.

(4) JP: 緑は父親に水さしの水を少し**飲ませ**、果物かフルーツ・ゼリーを食べたくないかと訊いた。

Midori wa chichioya ni mizusashi no mizu o sukoshi nomase, kudamono ka furūtsu zerī o tabetakunai ka to kīta. (*Noruwei no mori, Part 2,* p. 73)

EN: 'Midori gave her father a drink of water and asked if he'd like a piece of fruit or some jellied fruit dessert.' (*Norwegian wood*, p. 241)

CR: 'Midori je ocu dala da popije vode i pitala bi li pojeo neko voće ili voćni žele.' (*Norveška šuma*, p. 209)

[10] Although no example sentences with animal agents are cited in this work, such would be classified as {-human, +animate}.

Midori {+human, +animate} comes to take care of her sick father {+human, +animate} and therefore has to provoke and carry out the action herself since he is unable to do it by himself.

The same interpretation can also be applied to the following example.

(5) JP: あなたが真剣に直子を**回復させたい**と望んでいるなら、そうしなさい。

Anata ga shinken ni Naoko o kaifukusasetai to nozondeirunara, sō shinasai. (*Noruwei no mori, Part 1,* p. 236)

EN: 'That's what you should do if you're serious about making Naoko well again.' (*Norwegian wood*, p. 153)

CR: 'Tako se moraš ponašati ako zbilja želiš da Naoko ozdravi. (*Norveška šuma*, p. 135)'

Although Naoko {+human, +animate} knows she is sick and is hospitalised, the people {+human, +animate} around her (such as Watanabe – the "あなた" in the sentence) can also help her get well. In that case, Naoko herself cannot do what she does not perceive, which would make her unable to perform that specific action of "making better" (回復する). Although in this instance there is no direct influence of the causer on the causee as in Fukada's example[11] (since a third person is the speaker), it still points to an action the causee has no propensity to perform on her own, and the causer (Watanabe) brings about the action, i.e. helping Naoko get better.

In the English example, "making Naoko well again" implies that Naoko herself is not actively participating in the action, i.e. somebody is doing it (making her better) instead of her. A nuance of Naoko's more active participation in her recovery could be expressed through paraphrasing, such as "if you're serious about helping Naoko get better/well again", although that nuance remains rather weak.

The Croatian sentence does not imply a strong nuance of the causer's participation, but rather only their desire to do so ("ako zbilja želiš da (…)" *if you really want to* (…)).

4.2 Hands-off or non-involvement

The hands-off interpretation is one that might arguably cause questions on why a causative verb would be used to express such a meaning and/or situation. Namely, hands-off is where a causative action is not shown to have necessarily been provoked by the causer, and involves an inanimate causee (Fukada, 2010, p. 29), and could therefore possibly be interpreted as a natural course of events such as in Fukada's

[11] "Kizetsu shiteiru kanja ni kusuri o nomaseta." (Fukada, 2010, p. 29).

example “Niku o kusaraseta.”. This example was translated as “(Someone) let the meat spoil.” and is presented as the most passive form of causation (Fukada, 2010, pp. 29–30).

In the original definition of a hands-off meaning the causer is not expressed (although presumably animate {+human, +animate}) while the causee is inanimate {-human, -animate}. However, let us look at the following example.

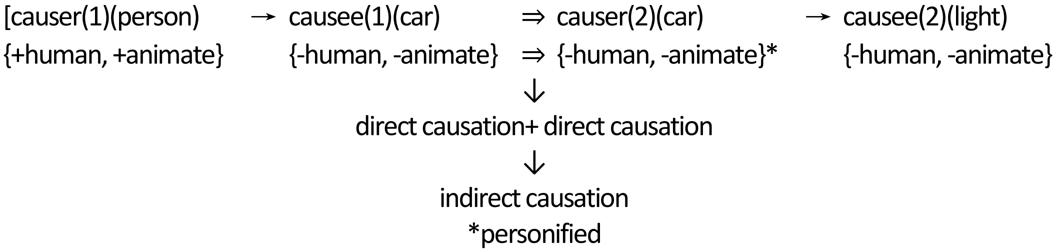
(6) JP: 交通を規制するパトカーが残って路上でライトをぐるぐると**回せていた**。

Kōtsū o kisēsuru patokā ga nokotte rojō de raito o guruguru to mawaseteita. (*Noruwei no mori, Part 1,* p. 163)

EN: ‘One police car remained to direct traffic, its rooftop light spinning.’ (*Norwegian wood*, p. 105)

CR: ‘Jedan policijski automobil ostao je regulirati promet, a na krovu mu se vrtjela plava svjetiljka.’ (*Norveška šuma*, p. 95)

Both the causer and the causee are inanimate, and the causer cannot be in a position of power in comparison to the causee. In example (6), it is logical to assume that a car as an inanimate entity cannot perform an action, so presumably there is another, animate causer “hidden” in the sentence, similar to Fukada’s “someone” {+human, +animate}. This describes indirect causation and is similar to the example presented by Neeleman and van de Koot (2012, p. 6). The only possibility of the action happening is if the car was operated by an animate being, such as a human driver, causing it to turn on and switching on the lights. Thus, it could be argued that the real causer is the person in the car, acting as a non-expressed causer, illustrated in the schematic below:



[causer(1)(person) → causee(1)(car) ⇒ causer(2)(car) → causee(2)(light)

{+human, +animate} {-human, -animate} ⇒ {-human, -animate}* {-human, -animate}

↓

direct causation+ direct causation

↓

indirect causation

*personified

Figure 1: The relationships of causer(s) and causee(s) in example (6)

The causative action is divided into two parts: a person {+human, +animate} turning the car {-human, -animate} on, and then the car ({-human, -animate} but personified) indirectly through the person making the lights {-human, -animate} spin. Although an inanimate object, a car is thought to be a causer in this sentence, as it is personified and as such the causer of the action. Each of these two actions separately are direct causative actions, but the whole sentence has indirect causation. Although not explicit, it is followed by a logical conclusion that there has to be another causer bringing up the event of making the car rotate the lights, making it an indirect causative event brought about by an animate causer(1) and causing a chain reaction.

In both English and Croatian, a car is the subject and the above interpretation is applicable, i.e. there has to be somebody (a person) not referred to in the sentences operating the car. The light is seen to be part of the car (expressed by “its” in English and “mu” in Croatian) and considered a whole (car+light). Sentences such as “One police car remained to direct traffic, making the rooftop light spin /spinning the rooftop light.” / “Jedan policijski automobil ostao je regulirati promet, a na krovu je vrtio plavu svjetiljku.” are not considered to be natural.

The other possible subdivision of hands-off meaning is similar to the abovementioned:

(7) JP: 夏の名残りの光が煙を余計にぼんやりと**曇らせていた**。
Natsu no nagori no hikari ga kemuri o yokē ni bonyari to kumoraseteita. (*Noruwei no mori, Part 1,* p. 125)
EN: ‘The fading summer light gave the smoke a soft and cloudy look.’ (*Norwegian wood*, p. 80)
CR: ‘U ljetnom sumraku dim je djelovao meko i mutno.’ (*Norveška šuma*, p. 73)

The animate causer is not mentioned nor involved in the event, i.e. has no connection with the action taking place since it is a natural phenomenon – the light (causer(1)) {-human, -animate} shone on the smoke (causee) {-human, -animate}, making it scatter. The only way an animate causer could have been involved in the action is to have been able to observe it; there is also no power relation between the inanimate causer and causee.

Similarly, in both the English and the Croatian sentences the animate causer is not referred to, with the only causer being the “fading summer light”and “ljetni sumrak” respectively.

A similar interpretation could be employed when an animate causer (1) has no direct influence or involvement in the causative action:

(8) JP: 突撃隊は世界中の人を**楽しい気持ちにさせる**ようだった。

Totsugekitai wa sekaijū no hito o tanoshii kimochinisaseru yō datta. (*Noruwei no mori, Part 1,* p. 133)

EN: 'Storm Trooper gave Midori an especially big laugh, as he seemed to do with all the world's people.' (*Norwegian wood*, p. 85)

CR: 'Smeđekošuljaš je posebno nasmijao Midori, kao, uostalom, i sve druge.' (*Norveška šuma*, p. 78)

In (8), the "Storm Trooper" a.k.a. Watanabe's roommate has no idea that Watanabe is using his peculiar habits as comic relief, and therefore has no control over the action. However, such an interpretation is heavily reliant on the context, since without it the sentence implies that "Storm Trooper" is, in fact, actively and knowingly participating in the action. The same conclusion can be drawn for the English and Croatian examples, i.e. they make "Storm Trooper" seem like an active participant in the action ("Storm Trooper gave Midori an especially big laugh (…)" / "Smeđekošuljaš je posebno nasmijao Midori"), and not just a character in his roommate's story.

In Japanese, oftentimes a sentence's causer and causee are one and the same, with the causee being either the causer's body part or the causer's feelings or mental state (see Glumac (2015, pp. 32–35); also mentioned in Jaklin (2020)). Such sentences will be classified as a type of hands-off sentence, and in the sentences in *Noruwei no mori*, the causer either has a degree of control over the causative action (9), or doesn't (10).

(9) JP: 「いつも自分を変えよう、向上させようとしていたけれど」と直子はソファーの上で脚を組みなおした。(ノルウェイの森(上), p. 261)

"Itsumo jibun o kaeyō, kōjousaseyō to shiteitakeredo" to Naoko wa sofa no ue de ashi o kuminaoshita.

*「[彼（きずき）は]いつも自分を変えよう、向上させようとしていたけれど」と直子はソファーの上で脚を組みなおした。

*"[kare (Kizuki) wa] itsumo jibun o kaeyō, kōjousaseyō to shiteitakeredo" to Naoko wa sofa no ue de ashi o kuminaoshita.

EN: 'He did keep trying to change himself, to improve himself, though.' (*Norwegian wood*, p. 169)

CR: 'Ali trudio se promijeniti, poboljšati.' (*Norveška šuma*, p. 148)

(10) JP: 「一緒に死んでくれるの？」と緑は目をかがやかせて言った。

“Isshoni shinde kureru no?” to Midori wa me o kagayakasete itta. (*Noruwei no mori, Part 1,* p. 155)

EN: ‘“You’ll die with me?” Midori asked with shining eyes.’ (*Norwegian wood*, p. 100)

CR: ‘Umrijet ćeš sa mnom? - pitala me Midori blistavih očiju.’ (*Norveška šuma*, p. 90)

In (10), Midori’s eyes shining is not her doing, but a physical reaction to her emotional state. Although it’s her own body, she has no control over it and is therefore classified as hands-off. In the English (“with shining eyes”) and Croatian (“blistavih očiju” *with sparkling eyes*), Midori’s non-involvement in her eyes’ condition is easily understandable, since the sentences describe the state of her eyes as “shiny”, and not as an action in which Midori is a participant.

On the other hand, in 9), Kizuki himself tried to consciously change who he is, putting pressure on himself[12], making him an active participant in an action that results in a change upon his (mental) person. Kizuki’s involvement in the action upon himself is translated into English and Croatian by the usage of the reflexive “himself” and the reflexive verb “se promijeniti, poboljšati” respectively.

Another hands-off, no-power, no-control interpretation is possible when an inanimate causer influences an animate causee.

(11) JP: そしてそのメロディーはいつものように僕を**混乱させた**。

Soshite sono merodī-wa itsumo no yōni boku o konransaseta. (*Noruwei no mori, Part 1,* p. 7)

EN: ‘The melody never failed to send a shiver through me, but this time it hit me harder than ever.’ (*Norwegian wood*, p. 3)

CR: ‘Od te bih melodije uvijek duboko uzdrhtao, ali ovoga puta potresla me jače nego ikad.’ (*Norveška šuma*, p. 9)

An inanimate entity cannot really exert power and “make” and animate, conscious one do its bidding, but it is influencing it, i.e. the melody {-human, -animate} is making a person feel different emotions.[13] The influence of the music is, in this case, out of the causee’s control.

[12] For a more detailed treatment of such sentences see Glumac (2015).

[13] Shigemori-Bučar notes that the non-animate/non-volitional causers can be assumed to be characteristic of the literary genre, continuing to state that in the analysed passages such causers were usually natural phenomena, while some others were personified (2006, p. 200).

The same can be understood from English and Croatian sentences.

All four instances can be argued to be a part of the hands-off category since an animate entity (a person) has no direct control or involvement in a causative action. Therefore, although hands-off may be considered an umbrella term, I would classify the latter examples employing inanimate causers {-human, -animate} as showing non-involvement. The difference, although slight, between indirect causation and non-involvement would be that while indirect causation covers any causative action in which there is no direct correlation between causer A and causee B, no matter whether they are animate or inanimate, non-involvement would entail only inanimate {-human, -animate} causers.

4.3 Permission and enabling

When expressing the "basic" causative meaning, permission presents the opposite of coercion, i.e. the causer allows or lets the causee perform the causative action instead of making them do it. Permission discussed here is much the same – the causee is assumed to be willing or desiring to carry out an action, meaning that the causative event will most likely occur (Fukada, 2010, p. 28). In the example below, Naoko (causee) seems to want to talk, so Watanabe (causer) lets her do so.

(12) JP: でも直子がそんなに夢中になって話すのははじめてだったし、僕は彼女にずっと**しゃべらせて**おいた。

Demo Naoko ga sonnani muchū ni natte hanasu no wa hajimete dattashi, boku wa kanojo ni zutto shaberaseteoita. (*Noruwei no mori, Part 1,* p. 82-83)

EN: 'I have never heard her speak with such intensity before, and so I did nothing to interrupt her.' (*Norwegian wood*, p. 52)

CR: 'Nikada je prije nisam čuo da govori s takvim žarom, pa je nisam prekidao.' (*Norveška šuma*, p. 50)

No major differences occur in the English and Croatian sentences, except for the nuance that more than "letting Naoko speak", i.e. giving direct permission, the causer "does not interrupt Naoko"/" (pa) je nisam prekidao", i.e. passively allows the action to take place by not interfering.

However, permission does not entail only giving, denying or asking for permission in the following sentence.

(13) JP: そしてそういう気持ちに**させて**くれたことだけで、私は彼に心から感謝したわ。

Soshite sōiu kimochi ni sasetekureta koto dake de, watashi wa kare ni kokoro kara kanshashita wa. (*Noruwei no mori, Part 1,* p. 245)

EN: 'If only for having made me feel that way, I was tremendously grateful to him.' (*Norwegian wood*, p. 159)

CR: 'Ako ni zbog čega drugoga, bila sam mu silno zahvalna što je u meni probudio te osjećaje.' (*Norveška šuma*, p. 140)

In (13), no permission is actually asked for or given, but the causee {+human, +animate} was able to feel something that made her grateful to the causer {+human, +animate}. In this case, the causative action represents what Žagar-Szentesi described as "enabling"[14] when discussing the causative meaning of the Croatian construction "dati (se)+infinitive" (2011, p. 303). In such cases, permission is not given but the causer helps, enables, or supports the causee (see also Hayatsu, (2016)[15]), exemplified by the following sentence.

(14) JP: 朝起きて家事して子どもの世話して、彼が帰ってきたらごはん**食べさせ**て…毎日毎日がそのくりかえし。

Asa okite kajishite kodomo no sewashite, kare ga kaettekitara gohan tabesasete… mainichi mainichi ga sono kurikaeshi. (*Noruwei no mori, Part 1,* p. 247)

EN: 'I'd get up in the morning and do the housework and take care of the baby and feed my husband when he came home from work. It was the same thing day after day, but I was happy.' (*Norwegian wood*, p. 160)

CR: 'Ujutro bih ustala i obavila kućanske poslove, te se brinula o djetetu i kuhala mužu ručak kad bi došao s posla. Svaki dan isto, ali bila sam sretna.' (*Norveška šuma*, p. 141)

In the above example, the meaning intended was not that Reiko (causer) spoon-fed her husband (causee) as though he were unable to do so himself, but rather that by cooking food she enabled and made it possible for him to have dinner. Neither did she "allow" him to eat, as they were a married couple on equal grounds, and she had no reason or power to stop him from eating the lunch she made for him.

[14] „Dopuštanje/omogućavanje drugoj osobi da nešto radi (permission/enabling another person to do something; author's translation) (Žagar-Szentesi 2011, p. 303).

[15] Hayatsu (2016) discusses みちびき (*michibiki*) in Japanese causative sentences, where the result of the action is favourable for the causee, with the causer providing "guidance".

Therefore, in cases where permission does not fit the mould either because it was not asked for, the causer is in no position to give it, or the situation does not call for it, i.e. there is no power play between the causer and the causee, the meaning expressed is that of enabling the causee to do the specified causative action.

4.4 Sociative causative

Sociative causative sentences involve participation of a causer, but unlike the manipulative interpretation in which the causee is passive, both the causer and the causee are animate {+human, +animate} and actively (willingly) participate in the action. As Fukada (2010) exemplifies, the pragmatic inference associated with this interpretation might be assistive, e.g. a mother assisting her young child in the toilet (Fukada, 2010, p. 31).

In *Noruwei no mori*, sociative causation can be seen as used for general social norms, where there is no pressure expressed or intended by the causer, and the causee does the action willingly, almost unconsciously following a pattern of behaviour in social occasions, such as being encouraged to sit in the following situation:

(15) JP: 「あなたはここの冬を知らないからそう言うのよ」とレイコさんは僕の背中を叩いてソファーに**座らせ**、自分もそのとなりに座った。

“Anata wa kono fuyu o shiranai kara sō iu no yo” to Reiko-san wa boku no senaka o tataite sofā ni suwarase, jibun mo sono tonari ni suwatta. (*Noruwei no mori, Part 1,* p. 211)

EN: ‘“Ah but you haven’t seen the winters here,” said Reiko, touching my back to guide me to the sofa and sitting down next to me.’ (*Norwegian wood*, p. 137)

CR: ‘- E, ali niste vidjeli kakve su ovdje zime - rekla je Reiko, lagano me gurnuvši prema kauču i sjedajući kraj mene.’ (*Norveška šuma*, p. 121)

Although the sentence looks as though it were coercive, Reiko (causer) does not pressure Watanabe (causee) to do anything - she suggests he sit down by touching him on the back, and then sitting herself. Therefore, as there is no explicit pressure from the causer, and the causee accepts and does the action willingly, the action can be classified as sociative – it fulfils a social norm of a guest (Watanabe, causee) taking the seat offered by the host (Reiko, causer), making both the causer and the causee participate in the action.

Another example is the case of Naoko’s sister asking Naoko about her day and listening to her talk; the older sister encourages her little sister and takes interest in the events of her little sister’s day:

(16) JP: 私が学校から戻ると部屋に呼んで、隣りに**座らせて**、私のその日いちにちのことを聞くの。

Watashi ga gakkō kara modoru to heya ni yonde, tonari ni suwarasete, watashi no sono hinichi no koto o kiku no. (*Noruwei no mori, Part 1,* p. 296)

EN: ‘When I came home from school, she’d call me into her room and sit me down next to her and ask me about my day.’ (Norwegian wood, p. 193)

CR: ‘Kad bih se vratila iz škole, pozvala bi me u svoju sobu, posjela kraj sebe i pitala kako sam provela dan.’ (Norveška šuma, p. 168)

As the little sister, Naoko is in a position of less power compared to her sister, but her older sister does not take advantage of her own higher status. The action is consensual and welcomed by both the causer and the causee, with no amount of pressure exerted by the causer.

In neither of the English or Croatian sentences presented in this section is there any nuance of “force” or pressure to do the action that was expressed through linguistic means.

5 The meanings and expressions in English and Croatian

There is no uniform way of expressing each causative meaning presented, but there are possibilities of expression that can serve as guidelines when transferring the causative meanings in other languages, in this case English and Croatian.

For some meanings causative expressions might not differ from the expressions of coercion or permission in general, and the intention behind the sentence needs to be understood through the context – manipulation is one of such meanings. Manipulative sentences do not differ significantly from the possible coercive interpretation and the meaning is contextual; some sentences tend to use milder expressions of coercion, for ex. “feeding him/hraniti ga” instead of “make him eat/tjerati ga da jede”, which also carry the meaning of the causer physically doing the action (giving food to somebody and them eating it). The nuance of “helping” or “being of assistance” was expressed as well with the expression such as “give (him) water/daš (mu) vode”. There appeared various constructions in the English translation (*Norwegian wood*), but generally the Japanese manipulative meaning did not noticeably differ from the coercive. Coercion was expressed more strongly in some instances, such as “make (a girl) say/prisiljavaš da kaže“, as a causer’s wish (“I wanted to finish making my point/htio sam istjerati svoje”), an obligation (“we’ve got to put a stop to (it)/moramo to zaustaviti”) or internal obligation of the causer (“he did keep trying (…) to improve (himself)/trudio se (...) poboljšati”), or through paraphrasing.

In both English and Croatian, the inanimate causer was the subject of the sentence when expressing hands-off and non-involvement causation. Hands-off and non-involvement sentences used a plethora of different linguistic expressions, meaning that a uniform way of expressing these meanings was not observed[16]. As was the case with the inanimate Japanese causers (e.g. grass, sunlight, car, etc.), the English and Croatian causers were personified (as Shigemori-Bučar (2006) noted), not hinting to a person's involvement in the causative action, despite the logical conclusion that a person had to have been involved.

Permission and enabling did not always significantly differ in the expressions used, but in Croatian the "dati (se) + infinitive" (Žagar-Szentesi, 2011) expression could be an appropriate choice when transferring the nuance of enabling when both the causer and causee are animate. Although is not always possible or observed, the distinction between permission and enabling is clearer than between manipulation and coercion. The reason for this is that sentences analysed as permissive sometimes used clearer expressions of permission than the coercive sentences did for duress, such as the verb "let/pustiti" (e.g. "let me sleep, let her talk herself out, etc. /pusti me da spavam, [...] je pustim da sama prestane pričati"), since coercive sentences did not tend to use verbs with a strong coercive meaning, e.g. to make, to force/natjerati, tjerati, prisiliti. Enabling did not express actual permitting, nor was it suggested in the context of the sentences.

The sociative causative does not use any specific causative expressions in English or Croatian, although there is an overall use of polite language in different social occasions. What distinguishes these sentences is also the lack of any kind of pressure put on the causee by the causer (similar to enabling), e.g. "(she) got a laugh from them/nasmijala ih je". The situations in which the sociative causative is used could have, similar to some variations of hands-off, been expressed without the use of the causative.

6 Summary

In this paper the possible interpretations of Japanese causative sentences have been discussed, and the meanings they can convey according to both the context in which they are used and according to the expression of the causer and the causee. The analysis touched upon the sentences' transfer of meaning into English and Croatian as well, but the main aim of the paper was the discussion of Japanese causative sentences' nuances of meaning. Rather than translation, which is out of scope of this paper, the

[16] Croatian reflexive verbs with the *se* pronoun such as "lelujale se (set swaying)", "[mozak mi] se pokrenuo (I got my head working) were used (amongst others) where the causee is the subject of the sentence and the action is performed by, and on, that subject alone.

possibilities of expression in the two languages mentioned serve as indicators of meaning and a point for understanding the causative meaning. Through the analysis of example sentences, I focused on the possible meaning a sentence can convey, since classifying it simply as "coercive" or "permissive" poses an obstacle for understanding causation in the Japanese language.

Although the analysis did not show major variations from those in pervious research (as summarised in Table1. below), it did illuminate some peculiarities, and can serve as a guideline for the ways of expression in English and Croatian; however, more often than not the way a causative sentence will be understood and expressed in another language heavily depends on the context and the placement of emphasis. One such example are manipulative causative sentences – although for expressing manipulation milder expressions of coercion and pressure can be used, no major difference in expressing a coercive and a manipulative meaning has been observed. On the other hand, permissive sentences were clearer in their meaning, and a subdivision of "enabling" has been discussed.

The analysis and the points addressed so far have been summarised in Table 1 below:

Table1: Causative sentence meanings and points

Meaning	Characteristics (Fukada, 2010)	Variants (author)
manipulative	"[…] the causee has no propensity to do the caused action", so the causer carries out the action (Fukada, 2010, p. 27)	According to Fukada's example, it was concluded that the meaning referred to the physical inability of the causee to carry out an action; the same interpretation can be used when the action in question is mental/not physical.
[coercive]	the causee does not wish to carry out a causative action (2010, p. 28) the nuance of "pressure" is evident	no variants observed There was not enough evidence to claim that there is a clear distinction in the expressions used for manipulative and coercive meanings, although when expressing manipulation, milder expressions might be preferable.
permissive /enabling	the causee is assumingly voluntarily carrying out a causative action (2010, p. 28)	enabling*: a variation that was discussed as falling under the scope of "permissive" although no permission is granted; for "enabling", both the causer and the causee are animate (human) and of equal standing, which means that the basic condition for a permissive meaning – the higher position of the person giving permission, is not met and therefore non-existent *based on Žagar-Szentesi (2011)

hands-off	involves an inanimate causee and the inference is that of not preventing (2010, p. 29–30) [author: the causer is not explicitly referred to, although it is logically assumed to be animate {+human, +animate}]	inanimate causer and causee: it is proposed that in such cases there has to be an animate causer, however unknown as per original definition inanimate causer and animate causee: influence that cannot be controlled (ex. invoking memories) animate causer and causee: the causee is not directly involved in the action and/or is unaware that the action is taking place: the causee is the causer's body part/psyche
[adversity]	similar to the above "hands-off", however the action has unwanted consequences on the causer, i.e. the inference is that of undesirability for the causer (2010, p. 30–31)	No variants observed.
sociative	(Shibatani and Chung (2002)) sociative causation means that the causer is involved in the action "[...] in various capacities." (Fukada, 2010, p. 23) "[...] inference not related to the propensity of the cause to perform the event is crucially involved [...]", which concerns the causer's involvement in said event (2010, p. 31) not linguistically encoded (2010, p. 32).	Since both the causer and the causee participate in a given action/event, the nuance of "pressure" usually inherently present in causative sentences is not there; rather, the action/event is assumed to be participated in willingly by both parties. Additionally, strengthened by the use of polite language, this meaning is connected to social norms/polite behaviour.

7 Conclusion

This paper discussed possible meanings a causative sentence can be used to express based on the pragmatic meanings of causative sentences presented by Fukada (2010), as just labelling them "coercive" or "permissive" was felt to be inadequate. Rather than seen as equivalent to coercion, the causative can be a versatile verbal form with many possibilities of use. As Glumac (2015) stated, the meaning of a causative sentence is also dependent on the context; however, understanding the possibilities of expression helps choose the appropriate expression in another language – milder coercion used for the manipulative meaning and conveying the sense of helping another, enabling being different from permission and sociative causation used in the context of social conventions and politeness.

Although the *-(sa)seru* causative sentences dealt with in this paper are one of, if not the most common way of expressing the causative in Japanese language, causation can also be expressed by using lexical causative verbs or the *-temorau* form. The latter might be considered as a topic of further study, both for the possible variation in causative meanings it is used to express, and the possible difficulty and variation in its translations.

Acknowledgement

I would like to thank Prof. Masaki Ono from the University of Tsukuba and Prof. Nagisa Moritoki from the University of Ljubljana for their useful comments, and Vishal Gor for the native English check.

References

Baron, N. S. (1974). The structure of English causatives. *Lingua, 33*(4), 299-342.

Hayatsu, E. 早津恵美子 (2016). *Gendai Nihongo no shieki* 『現代日本語の使役』. Hitsuji kenkyūsōsyo ひつじ研究叢書.

Heycock, C. (1987). *The structure of the Japanese causative*. The University of Edinburgh. Edinburgh Research Explorer.

Fukada, A. (2010). The Japanese Causative Controversy: A Pragmatic Perspective. *Japanese Language and Literature* 44, 21-43.

Glumac, D. (2015). Kauzativ i glagolski rod u japanskom jeziku u poređenju sa srpskim. Faculty of Phililogy, Belgrade. (Doctoral thesis).

Jaklin, P. ヤクリン・ペトラ (2020). Nihongo no shieki to sono Kuroachiago deno hyōgenhōhō: imikaisyaku no kanten kara atsukatta Akutagawa Ryūnosuke no Kappa ni okeru shiekibun 「日本語の使役とそのクロアチア語での表現方法―意味解釈の観点から扱った芥川竜之介の『河童』における使役文―. Nihongo komyūnikēsyon kenkyūronsyū 『日本語コミュニケーション研究論集』(9), 46-57.

Neeleman, A., & Van de Koot, H. (2012). The Linguistic Expression of Causation, pp. 1-27. https://www.researchgate.net/publication/263235725_The_Linguistic_Expression_of_Causation

Shigemori Bučar, C. (2006). Causative constructions in Japanese and Slovene. *Linguistica, 46*(1), 191-202.

Sinčić, B. (2018). Francuska konstrukcija “faire“+ infinitiv i njezini prijevodni ekvivalenti u hrvatskom jeziku: izricanje kauzativnosti u hrvatskom jeziku. Hrvatska znanstvena bibiografija (bib.irb.hr). Sveučilište u Zagrebu, Zagreb. (Master’s thesis)

Žagar Szentesi, O. (2011). Funkcionalne varijante konstrukcije dati se + infinitiv u hrvatskom jeziku – u okviru gramatikalizacije. *Suvremena lingvistika*, *37*(72), 295-318.

Murakami, H. 村上春樹 (2014, 2018). *Noruwei no mori (ka/ge)*)『ノルウェイの森(上/下)』. Kōdansya bunko 講談社文庫.

Murakami, H. (2011). *Norwegian wood.* New York: Vintage International, Vintage books.

Murakami, H. (2015). *Norveška šuma.* Zagreb: Vuković & Runjić.

Normative forms and synthetic structure of Japanese in the incubation period of L2: subject to sentence-final forms in longitudinal discourse data of Korean returnee sisters' Japanese

HWANG Young Hee
Hanyang Cyber University, Korea
hwang@hycu.ac.kr

Abstract

In this paper, I examine the change mechanism of Japanese sentence-final forms (SFF) maintained by two Korean returnee sisters for over 10 years after the cessation of L2 contact, and focus on the negative formal style of verb sentences and its deviation from the actual use of norms (analysis form) and non-norms (synthetic form). Findings are based on a comparison of two Korean sisters' Japanese with that of thirteen Korean adults' colonial Japanese maintained for over 60 years, which is also in the incubation phase. In the sisters' Japanese sentence-final forms that were incubating as their L2, they rarely used the non-norms, while the norms were stably maintained, and the retention of the synthetic structure of their returnee Japanese correlated with the duration of the language acquisition period and with the elapsed time of contact cessation. That is, the sisters used more of the norms in the Korean colonial Japanese SFF than they did in their Japanese; I attribute this to the sisters' 10-year incubation period. Specifically, the Korean returnee sisters' speech included interventions of the explanatory *[N]* in the past affirmation of *-ta desu*, heavy use of non-norms in adjective and noun sentences, and connecting sentence-final particles to further grammatical structures. However, there were fixed and conventional Norms in the Korean returnee sisters' Japanese, and once these are acquired, *masu* forms are retained for long periods in mirror image, especially on elder *A*. To summarize, in terms of the format of the returnee Japanese SFF, the two Korean returnee sisters were slower to shift from norms (*masu*) to non-norms (*desu*) than were adult Korean speakers of colonial Japanese. The same shift is observed with synthetic structure even after cessation of the language contact.

Keywords: normative forms; incubation period; sentence-final forms of Japanese; synthetic structure; analysis; Korean returnee sisters

Povzetek

V tej študiji preučujem mehanizem sprememb v končniških oblikah povedi v japonščini, ki sta jih dve korejski sestri povratnici uporabljali v obdobju 10 let po prenehanju neposrednega stika z L2. Predvsem se osredotočam na negativni formalni slog glagolskih stavkov ter odstopanj v dejanski uporabi norme (analitična oblika) in nenorme (sintetična oblika). Ugotovitve temeljijo

Acta Linguistica Asiatica, 10(2), 2020.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.10.2.105-125

na primerjavi japonščine omenjenih sester ter japonščine trinajstih odraslih Korejcev kolonialne japonščine, ki se je ohranila več kot 60 let. Obe vrsti japonščine sta ocenjeni kot jezika v fazi inkubacije. V oblikah korejskih sester so le redko opazne nenorme, medtem ko so bile norme stabilno vzdrževane. Njuno ohranjanje sintetičnih oblik je sorazmerno tako s trajanjem usvajanja japonskega jezika kot tudi s pretečenim časom od prekinitve stika z jezikom. Poleg tega je razvidno, da sta po povratku v domovino uporabljali več analitičnih oblik kot v času bivanja na Japonskem, kar pripisujem desetletnemu obdobju inkubacije. Njun govor je na primer vključeval razlagalni *[N],* ki se je pogosto vmešal v pozitivno preteklo obliko *ta desu*, prekomerno rabo sintetičnih oblik v pridevniških in samostalniških stavkih ter nadaljno navezavo na končniški povedne členke. Hkrati je bilo v njunem govoru opaziti precej ustaljenih in običajnih analitičnih oblik, ki so se, potem ko so bile usvojene, v zrcalnih slikah obdržale zelo dolgo. Tako se je -masu oblika še posebno dolgo obdržala pri starejši od sester. V splošnem lahko rečem, da se je japonski govor korejskih sester povratnic počasneje preusmeril iz norme (*-masu*) v nenormo (*desu*) v primerjavi z govorom korejskih govorcev kolonialne Japonske. Podoben premik je opazen tudi pri sintetični strukturi navkljub prenehanju jezikovnega stika.

Ključne besede: normativne oblike; inkubacijska doba; končniške oblike povedi v japonščini; sintetična struktura; analiza; korejski sestri povratnici

1 Introduction

The field of applied linguistics has had a recent focus on "L2 maintenance" or "L2 attrition". In that focus, the mechanism of language change is either maintained, or else declines after L2 is acquired, and it is important to provide feedback and savings effect to the relearning of L2 learner (Tomiyama 2004). This includes cases of Japanese learners who learned in high school, expatriates who had returned to Korea after staying in Japan, and international students who wanted to maintain a certain level of Japanese. I refer to the language of such Japanese speakers as the "Japanese of the incubation period."

In addition to studies on L2 acquisition, L2 attrition researchers such as Weltens and Cohen (1989), Ono (1994), Shibuya (1995), Hansen (1999), Tomiyama (2000), Kiguchi (2004), Kim (2010), and Matsumoto (2013) have conducted studies on Japanese learners to examine subtractive changes over time. However, very few researchers have explored the changes in sentence-final forms (*desu* and *masu* style) that occur over time among learners of Japanese as L2.

In this study, I investigate the latent Japanese that remains in two Korean children who returned to Korea after six years in Japan. By examining characteristics of sentence patterns in different stages of L2 acquisition, namely maintenance, attrition, and decline, I aim to disprove the hypothesis of "analysis" that Dorian (1977), Shibuya (1995), Kiguchi (2004), and Hwang (2013) have asserted to be normal in language retention.

Furthermore, some researchers, such as Noda et al. (2001), have proposed the unstable interlanguage states of the sentence-final forms (SFF) (*desu* and *masu* style) with Japanese learners. However, because learners involved in earlier studies had ceased learning Japanese during their stays in Japan, including the current Japanese learner, their Japanese had been latent for more than 10 years. Further research is needed on the maintenance status of the Verb, *i*-adjective, Noun and *na*-adjective SFF to elaborate on results by Kiguchi (2009), Sanada (2009), and Hwang (2010).

I aim to clarify the processes of maintaining and attrition, which relate to the processes of learning SFF and mirror relationships, in order to elucidate the holistic changes in learners' Japanese.

2 Previous research

2.1 The L2 life cycle

Among L2 researches on language changes after active learning stops that were conducted around 1980s, Weltens and Cohen (1989) on L2 attrition identified the early stage of experimental verification as "order of attrition, relations with literacy, age, and supplemental strategies" from a symmetric perspective on language acquisition. Education for returnee from overseas began to be recognized as an issue in Japan already in the late 1960s (Kubo 1994). In the 1990s, language problems of Japanese returnees attracted attention, with the main theme being the changes in young people's English ability.

For example, Ono (1994) pointed out that Japanese returnees' English vocabulary skills had rapidly declined regardless of how long they had stayed abroad, their ages at return, or their English vocabulary skills immediately after their return. Tomiyama (2000) surveyed Japanese returnees (all eight years old) from the United States for four years after they had returned to Japan, and found that signs of decline appeared after six months though the speakers had maintained the capacity to accept and produce English after four years. Tomiyama (2008) investigated the differences in age and proficiency in L2 attrition between two siblings and suggested maturational factors. Meanwhile, Kim (2010) discussed attrition-related changes in fluency, vocabulary, and postpositional particles in Japanese as L2 according to the passage of time in Korean child returnees. The author also found out that fluency degraded over time.

In the present study, I compare children's L2 Japanese in the incubation period with that of Korean colonial adults who had had little contact with Japanese after liberation in 1945. Korean colonial Japanese has been maintained for over 60 years since its acquisition and can therefore be defined as an attrition language similar to Korean returnee Japanese for over 10 years after the cessation of L2 contact. There are very few longitudinal studies on the maintenance status of Koreans' L2 Japanese after language contact. For example, "Japanese Learners of Chinese and Korean Native Language Longitudinal Utterance Corpus (C-JAS [http://c-jas.ninjal.ac.jp/index.py])," which was released in Japan, includes a three-year longitudinal discourse data of six Japanese learners, in which authors focused mainly on language acquisition, and less on Japanese maintenance. Therefore, I consider findings of Korean colonial adults' Japanese to be optimal as a contrast data.

Researchers have recently investigated Japanese attrition with Japanese L2 learners who speak English (Hansen ed. 1999, Hayashi 1999, Tomiyama 2000), and with colonial Japanese speakers who remained in Taiwan and the South Sea Islands (Kan 2004, Shibuya 2001, Matsumoto 2013). Their works are cited examples of colonial Japanese in which L2 has been restructured without suffer decline for over 50 years. Kiguchi (2004) and Hwang (2009) have also examined colonial Japanese among Koreans who maintain Japanese as their L2. On the other hand, Kiguchi (2004) worked on the context of language attrition, and Hwang (2009) elucidated the relevance of variations between speakers and their language acquisition environment at the point of language maintenance. There has also been little research on individual circumstances after L2 active learning stops and L2 cessation because of the difficulties associated with continuing the longitudinal research.

2.2 General characteristics of L2 in the L2 life cycle

Dorian (1981) examined the dying language of a Scottish Gaelic dialect and pointed out the following linguistic features (instances of Japanese usage being regressed are excerpts from Kiguchi 2004).

(a) No possibility of selecting formal versus informal style

(b) Replacing a synthetic structure with an analytic structure (e.g., *mashi-ta* [FML-Past] and *ta-desu* [Past-FML])

(c) Occurrence of analogical leveling (e.g., *shabereru* [POT] → *shaberareru* [POT])

In addition, Shibuya (1995) reported on the declining status of Japanese remaining in Palau in the South Sea Islands in terms of the following five features of speech structure.

(1) simplified formal style usage,

(2) simplified discourse management with reduced use of sentence-final particles,

(3) analysis of using *nai-desu* [NEG-COP] rather than *masen* [FML-NEG] or using *kitto* 'surely' to indicate likelihood,

(4) use of chunks such as sentence-final particles *yone* [Compound SFP] and negatives like *masen* and *wakara-nai-ne* 'understand' [NEG-SFP],

(5) avoiding complex phrases such as passive and causative sentences.

In the above examples, characteristics observed in the L2 in its incubation state are considered as “simplification, analysis, or transfer of the mother tongue”, and are the factors driving major language change.

2.3 Characteristics of Japanese sentence-final forms (SFF) appearing in the L2 life cycle

Noda et al. (2001) found that interlanguage speakers construct formal sentences with past-tense verbs using nonstandard Non-norms series (past + formal) much more often than using standard Norms series (formal + past). Table 1 presents the two series of formal SFF.

Table 1: The two series of Japanese SFF [FML]

		Past		Non-past	
Sentence		Norms	Non-norms	Norms	Non-norms
Verb	Positive	*V-masita*	*V-tadesu**	*V-masu*	*V-desu**
	Negative	*V-masendesita*	*V-nakattadesu*	*V-masen*	*V-naidesu*
i-adj.	Positive	*i-desita**	*i-tadesu*	*i-desu*	
	Negative	*i-arimasendesita*	*i-nakattadesu*	*i-arimasen*	*i-naidesu*
Noun	Positive	*N-desita*	*N-dattadesu*	*N-desu*	
	Negative	*N-dewaarimasendesita*	*N-dewanakattadesu*	*N-dewaarimasen*	*N-dewanaidesu*
na-adj.	Positive	*na-desita*	*na-dattadesu*	*na-desu*	
	Negative	*na-dewaarimasendesita*	*na-dewanakattadesu*	*na-dewaarimasen*	*na-dewanaidesu*

* non-standard type

This table is a modified version of the SFF in Noda et al. (2001), with additional non-past forms. Norms conjugation indicates synthetic forms as chunks or units (following synthetic structure by Dorian 1981), and Non-norms conjugation indicates analytic forms (following analytic structure by Dorian 1981). For example, if Japanese learners recognize and use expressions such as *-tarayokatta* and *-naitodame* as chunks when making conditional statements then they are applying the synthetic forms, whereas if these chunks are recognized and used separately, such as *-tara+yokkatta* or *-nai+to+dame(ikenai)*, they are applying the analytic forms. Actually, in colonial Japanese, an analytic form is a case where a pause (,) is inserted into the +part such as *Netara, yokkatta* ‘I wish, I had gotten to sleep.’, *Sinakereba, dameda* ‘You have to, do it.’. On the other hand, in returnees’ Japanese there is no pause in the +part and it is spoken as a chunk, like *Okitarayokkatta* ‘I wish I woke up.’, *Ittaradameda* ‘Don't go’, which corresponds to the synthetic form.

Kiguchi (2004), who described the colonial Japanese of Korean elders, observed that Non-norms such as *ta-desu* [Past-COP], *nai-desu* [NEG-COP], and Verb+*desu* [COP] tend to be productive, and also further detailed the declining tendency at the stage of language attrition.

In addition, Hwang (2013) examined Chinese-Korean colonial Japanese and found universal simplification and analysis in the Japanese learners’ interlanguage acquisition process in case of non-norms. Besides, Hwang also presented various potential interlanguage factors, such as the possibility of influence from Korean, the correlation with verbal internal factors, predominantly using the norms of speakers with high academic ability, and discrimination based on discourse functions in sentences.

In this study, I will discuss similarities and differences in the usage of verb SFF with a particular focus on speaker attributes, age of contact cessation, types of verbs, collocation relationships with sentence-final particles, and insertions of *[N]*. In particular, in relation to the age of contact cessation, I aim to present how both the synthetic and the analytic predicate structures have been maintained in both the Japanese of Korean returnee children and colonial Japanese of Korean adults.

3 Survey overview

I have collected natural discourses of two Korean-Japanese bilingual Korean sisters for over 4 years, 11 months in order to examine the mechanism of acquisition, maintenance, attrition, and regression of Japanese as L2 that was acquired approximately 15 years earlier.

The discourse data I use in this study consist of language data that were collected roughly 10 years before and after Japanese contact cessation. Details of the two sisters Japanese L2 acquisition life cycle will be given shortly in Table 2 below.

3.1 Informants

The informants I studied are two Korean returnee sisters: *A*, born in 1999, and *B*, born in 2001. They lived in Tokyo and Osaka for six years from 2001 to 2008, with two short returns to Korea, and received just over 4 years of local kindergarten and elementary education (1st and 2nd stage; acquisition period). Ten years after they had returned permanently to Korea in 2008, they were experiencing regression of their Japanese proficiency.

As of July 2008 the sisters (elder sister *A* was in the 3rd grade and younger sister *B* was in the 2nd grade) were gradually deteriorating in their Japanese because they had no contact with the language. At that time they attended the "returned children's class" established in Korea's elementary school until February 2009 (3rd stage; retention period). After that, until August 2011, they gradually became monolingual speakers; during that time, they had no contact with Japanese (4th stage; incubation and attrition period). *A* entered high school in March 2015 and *B* in March 2017, and at that point they both began studying Japanese again (5th stage; re-contact period).

In short, *A* and *B* learned Japanese in the 1st and 2nd stages of the L2 acquisition cycle and often used to speak Japanese on a regular basis instead of Korean. In the 3rd stage, Japanese was maintained as their basic (default) language, but in the first half of the 4th stage, their basic language changed from Japanese to Korean. In the second half of the 4th stage, the sisters ceased using Japanese.

As noted above, I classify the data on the two informants across the stages of the L2 acquisition life cycle: 1st and 2nd stage (acquisition period), 3rd stage (retention period), 4th stage (incubation and attrition period), and 5th stage (re-contact period).

Table 2: Information of informants and Japanese contact

Stage	Residential area and period	Education Information
1st stage (acquisition 1)	① Seoul (*A,* 0:0-2:0; *B*, 0:0-0:7)* ② Tokyo & Yokohama (*A*, 2:1-4:6; *B*, 0:8-3:1) ③ Daegu (*A*, 4:7-4:10; *B*, 3:2-3:5) ④ Osaka (*A*, 4:11-5:1; *B*, 3:6-3:8) ⑤ Daegu (*A*, 5:2-5:6; *B*, 3:9-4:1)	① Resident in Korea ② Resident in Japan ③ Kindergarten in Korea ④ Kindergarten in Japan ⑤ Child-care institutions in Korea
2nd stage (acquisition 2)	⑥ Osaka (*A*, 5:7-8:9; *B*, 4:2-7:4)	⑥ Kindergarten and Elementary school in Japan
3rd stage (retention)	⑦ Seoul (*A*, 8:10-9:4; *B*, 7:5-7:11)	⑦ Elementary school in Korea
4th stage (incubation and attrition)	⑧ Seoul (*A*, 9:5-15:3; *B*, 7:12-13:10)	⑧ Elementary and junior high school in Korea
5th stage (re-contact)**	⑨ Seoul (*A*, 15:4-present; *B*, 13:11-present)	⑨ High school in Korea

* In column 2, (year: month) indicate the elapsed time since the births of informants *A* and *B*.

** Results are taken from the surveys conducted in November 2016, March 2017, and March 2020.

3.2 Discourse data information

I began recording the linguistic data in September 2006 (*A*, 7:11; *B*, 5:6) and recorded informants' conversations until August 2011 (*A*, 11:10; *B*, 10:5). This includes having the girls look at a picture book in Japanese (Kim 2010) four times, from just before they returned to Korea in July 2008 to December 2009. I recorded the discourse materials at the sisters' home in stable scenes. I recorded conversations when the sisters or their family members were playing together or talking to each other before sleeping. At that time, the parents did not use Japanese, as they communicated mostly in Korean.

Six months after the girls returned to Korea, after the January 2009 survey, most of their natural conversation occurred in Korean, and in the March 2009 survey that was conducted eight months after their return, I observed examples of misuse of Japanese and mixed usage of Korean and Japanese. Then, in the January 2010 survey, which was conducted 17 months after the return to their home country, they were no longer using Japanese.

After the 5th period when Japanese was no longer used, a translation survey was conducted on the use of words. I expected some variants to be retained as latent knowledge, however, they were absent from the discourse data. To clarify the latent linguistic ability, an additional survey was conducted in November 2016, March 2017 and March 2020 when voluntary use of Japanese disappeared completely to obtain data on translations of Korean sentences.

Previous researchers mainly gathered data for their studies through artificially manipulated speech, reading, and survey responses. However, the discourse in this study contains natural Japanese used in conversations involving only two sisters and their family members. The girls mostly spoke casually with each other, but role plays in formal style conversation have also been detected. The proportions of formal style were 12.4% for *A* and 15.5% for *B* (see Table 3).

Moreover, the recorded data I have collected for five years are presented in Table 2 [④~⑧] account for 1,149 minutes (~19 hours) in total. Considering the entire duration of the investigation, this total is an equivalent to one-hour conversations recorded every three months. In addition, considering the recording time, the number of characters, and the number of times, the ratio of the volume of discourse from the 1st to the 5th stages is 4:4:3:3:4 with rounding to the nearest integer. This does not include the time I presented Kim's (2010) materials and collected those data.

3.3 Contrast data information

To clarify the characteristics of the SFF in the Japanese spoken by the two returnee-sisters to Korea, I contrast their data with the discourse data of the colonial Japanese spoken by Korean and Chinese adults that have undergone similar regressions. Although the two samples are very different in size, they are appropriate for comparison because the conversations with the Japanese native speakers and the aspects of discourse development are similar.

I furthermore include two sets of colonial Japanese data for Korean elders, the first of which was recorded in Daegu and Seoul in South Korea in 2006. The informants were eight Korean speakers over 80 years old (3 male and 5 female) who had been screened for having little experience with relearning Japanese after liberation in 1945. Their 61 years since Japanese language contact cessation was significantly longer than the equivalent period for the Korean returnee sisters. The total recording time was 530 minutes at an average of 44 minutes per person.

The second data set, discourse among five Korean elders (1 male and 4 female) and two Japanese native speakers, was recorded in 2010 in the Yanbian area (Yanji city and Tumen city) of northeastern China, an average of 42 minutes per person. These speakers were over age 80 and had had little contact with Japanese since 1945. For more information on the survey subjects, refer to Hwang (2013).

Further, these thirteen elderly were about the same age as elementary school students, and were in contact with Japanese students and Japanese people around them. As such, in terms of the degree of Japanese contact, the duration of Japanese acquisition period, and Japanese learning method similar to current immersion programs, these thirteen elderly had a similar background of Japanese acquisition that was similar to that of the Korean returnee sisters.

4 Results and discussion

4.1 Informants' actual use of sentence-final forms (SFF)

Before proceeding to the main text, I examine the overall outline and characteristics of SFF through the frequency of use of formal and informal verb sentences in Japanese in incubation. Table 3 shows the informants' rates of using the whole SFF, including both noun and adjective as well as verb sentences.

Table 3: Informants' usage of SFF

SFF	Returnee *A*	Returnee *B*	Koreans	Chinese Koreans
❶ Total	2397	1789	1626	1112
❷ Informal style	2100	1512	603	225
❸ Formal style	297	277	1023	887
❹ Percentage of formal style	12.4%	15.5%	62.9%	79.8%

Note: ❹=❸/❶

As shown in Table 3, the two Korean returnee sisters used substantially less formal style Japanese in the incubation period than did the Korean and Chinese-Korean elders. This is attributed to the differences in the attributes of the conversation partners, i.e., Korean returnee children versus Japanese native adults.

4.2 Use of formal verb sentences

Table 4 shows the number of appearances in the L2 life cycle by period, and the data is divided into positive, negative, informants (vertical axis) and tense, series, and period (horizontal axis). Furthermore, regarding L2 incubation, I focus on the time after acquisition, namely the 3rd or the retention period.

Table 4: Usage of formal verb sentences

Sentence		Past										Non-Past										Total
		Norms (*masi-ta*)					Non-norms (*ta-desu*)					Norms (*masu*)					Non-norms (*ru-desu*)					
		I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V	
Positive	*A*	15	12	-	-	-	8	2	-	-	-	44	45	7	1	18	3	5	1	-	-	160
		27			-		10(4)			-		89			26		8(5)[*]			1		
	B	2	2	-	112	-	4		-	-	-	14	33	2	-	13	4	2	-	-	-	188
		4			112		4(2)			-		47			15		6(3)			-		
Sentence		Past										Non-Past										Total
		Norms (*masen-desi-ta*)					Non-norms (*nakat-ta-desu*)					Norms (*masen*)					Non-norms (*nai-desu*)					
		I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V	
Negative	*A*	-	-	-	-	4	1	-	-	-	1	8	5	1	-	14	1	-	-	-	-	35
		-			4		1			1		13			15		1			-		
	B	1	-	-	1	2	-	-	-	-	7	-	5	4	1	6	3	-	1	-	1	32
		1			3		-			7		5			11		3(2)			2		

* Examples with *ta-desu, nakatta-desu, ru-desu* and *nai-desu* followed by *[N]* are indicated in parentheses.

Table 4 shows the following:

(a) In Japanese SFF in the incubation period, speakers use Norms more than Non-norms in all areas except for the past negative (*nakat-ta-desu*) used by *B*.

(b) *A* shows use of more appropriate SFF than *B* does (Examples 1-5), indicating that *A* better distinguished the use of the two series. In addition, fixed and idiomatic Norms were prominent in the speech of *B*, who had less contact with Japanese (Example 6).

(1) *A: Meeru-janakute koko-ni atebi-o* ***kaku-n-desu-yo***. (II 200711)[1]
e-mail-COP-NEG here-LOC date-OBJ write-*[N]*-COP-SFP
'You must write the date here, not on the e-mail.'

(2) *A: Mousikomi anata-ga* ***si-ta-n-desu-ka.*** (II 200807)
Application you-SUB do-Past-*[N]*-COP-INT
'Did you apply for it?'

(3) *A: Minna* ***wakari-masi-ta-ka.*** (III 200810)
everyone understand-FML-Past-INT
'Did everyone understand?'

(4) *A: Kochosensei-ga kimeru-tte okaasan-ga* ***ii-masi-ta.*** (III 200810)
principal-SUB decide-Quotation mother-SUB say-FML-Past
'The mother said that the principal decides.'

(5) *A: Nani-****degozai-mashoo-ka.*** (IV 200903)
what-COP-Extraordinary polite-FML-SPEC-INT
'What is going on?'

(6) *B: CD-wa* ***mot-teori-masu-ka***. ***Mot-teori-masen-ka***. (III 200812)
CD-TOP have-ASP-FML-INT have-ASP-FML-NEG-INT
'Do you have the CD? Or don't you?'

(c) There are characteristic differences by period, for instance that *B*'s use of Non-norms in the negative past is more prominent toward the 5th period, the period of attrition. Specifically, *A* generally used Norms whereas *B* used both Norms and Non-norms and transitioned to Non-norms. *B*'s usage was superior to *A*'s method in terms of her ability to convert the sentence styles in proportion to Japanese contact.

(7) *B: Isogasiku-te tegami-o* ***kak-e-nakat-ta-desu***. (V 201611)
busy-Cause letter-OBJ write-POT-NEG-Past-COP
'I was busy and could not write a letter.'

[1] The parentheses indicate the period of the utterance as the year and month.

(8) *B: Moo [gamansuru]*[2] ***gaman-deki-nakat-ta-desu***. (V 201611)

anymore [endure] endurance-can-NEG-Past-COP

'I could not endure it anymore.'

Table 5 summarizes the percentage of use per person of colonial Korean Japanese and colonial Chinese-Korean Japanese of the incubation period in Table 4 (Decimal point, rounding off). The use of *A* and *B* were only for the 3rd to 5th periods, the incubation period of L2, and in the same way, I regarded colonial Japanese as Japanese of the incubation period from the 3rd to 5th period.

Table 5: Usage of formal verb sentences in the L2 incubation period

Sentence	Informants	Past		Non-past		Total
		Norms	Non-norms	Norms	Non-norms	
Positive	Returnee *A*	-	-	26	1	27
	Returnee *B*	112	-	2	-	114
	Koreans	22	39	69	32	161
	Chinese-Koreans	52	79	90	61	282
Negative	Returnee *A*	4	1	15	-	20
	Returnee *B*	3	7	11	2	23
	Koreans	-	2	4	30	36
	Chinese-Koreans	4	10	13	17	43

The distribution of usage of formal verb sentences in Table 5 can be generally understood in terms of the ability to use sentence forms according to linguistic functions in sentences, as well as the differences in the different informants' levels of Japanese contact. In other words, there are many Norms in the order of *A* > *B*, but *B* uses Non-norms prominently in the Japanese negative past tense; this is also a common phenomenon with the colonial Japanese speakers in this study (Hwang 2010, 2013). Hwang's conclusions on Japanese acquisition, maintenance, attrition, and regression for Korean learners were tentatively a mirror image as follows.

(a) Non-past negative forms of formal verb sentences

(i) *masen* > (ii) *nai[N]desu* > (iii) *masen* > (iv) *nai[N]desu* > (v) *naidesu*

[2] [ ] is the interviewer's utterance.

(b) Past negative forms of formal verb sentences

(i) *masendesita* > (ii) *nakatta[N]desu* > (iii) *masendesita* > (iv) *nakatta[N]desu* > (v) *nakattadesu*

In the incubation period, the normative and synthetic *masu* conjugation forms in formal verb sentences are in unmarked positions in terms of the numbers used (Examples 10 and 11), but in colonial Japanese, non-normative and analytical *desu* conjugation forms are in the unmarked position.

(9) *K: (Chuugakkoo-wa)* ***iki-masen-desi-ta****.*
junior high school-TOP go-FML-NEG-COP-Past
'I did not go to (Junior High school).'

(10) *A:* ***Ki-te-masen-desi-ta-ka****.* (II 200807)
come-ASP-FML-NEG-COP-Past-INT
'Did not you come?'

(11) *B: Urusaku-te* ***nemure-masen-desi-ta****.* (II 200701)
noisy-Cause sleep-FML-NEG-COP-Past
'I could not sleep because it was noisy.'

In addition, Non-norms (*ta-desu*) of positive and past only appears in the acquisition period (1st and 2nd stage), which is characterized by being used in the typical form *ta-deshoo* with the confirmation and requirement functions (Examples 12-16).

(12) *A: Oneechan-ga* ***mot-teki-ta-deshoo****.* (II 200607)
sister-SUB bring-AUX-Past-COP-SPEC
'Sister brought it, did not you?'

(13) *A: Jeojjok-e hwajangsil.* ***Mie-nakat-ta-deshoo****.* (II 200612)
there-LOC toilet. see-NEG-Past-COP-SPEC
'There's a toilet over there, did not you see?'

(14) *A: Oneechan-ga* ***tot-teta-deshoo****.* ***Kime-ta-deshoo****.* *Hutari-de.* (II 200701)
sister-SUB take-ASP-Past-COP-SPEC. decide-Past-COP-SPEC. two people-with
'Did sister take it? We decided with two people?'

(15) *A: Kore junbisuru-kara doko-de* ***it-ta-deshoo****.* *Osoku dasi-te-nai-tte.* (II 200712)
This prepare-Cause where-LOC say-Past-COP-SPEC. late send-ASP-NEG-Quotation
'Because I'll prepare this. Somewhere did I say that 'Do not send late'?'

(16) *B: Koo ioo-to* ***omo-tta-deshoo****.* (II 200807)
like.this say-Will-Quotation think-Past-COP-SPEC
'You thought to say like this?'

In the above examples, the returnee sisters use non-norms in the verb SFF of Japanese in the incubation period substantially less than do the colonial speakers of Japanese. This indicates that the girls have well maintained norms inputted as chunks in the background rather than following the strategy of "rationalization (analysis) to *desu*" as a universal strategy that appears mainly in adult Koreans learning Japanese.

4.3 Use of *i*-adjective and *na*-adjective (noun) sentences

Next, I examined the interrelationships between the SFF used in *i*-adj. and *na*-adj. (noun) sentences, as well as the forms used in the verb sentences described in previous sections. Table 6 summarizes each informants' actual use of the formal forms of *i*-adj. sentences. After the acquisition period, the use of formal *i*-adj., *na*-adj. and Noun sentences decreased dramatically for the returnee sisters. The use of informal styles is indicated with brackets. Even though few in number, the translation surveys clearly show the tendency of decreasing Norms. I examined the results of the translational survey described in §3.2, and its use in a discourse.

Table 6: Usage of the formal forms in *i*-adjective sentences

Sentence style		Returnee *A*	Returnee *B*	Koreans	Chi.-Kor.	Total
Positive	Past	- (4)	- (-)	10	39	49
	Non-Past	- (25)	1 (13)	97	59	157
Negative	Past	5 (2)	4 (2)	4	3	16
	Non-Past	- (2)	- (-)	37	4	41

Table 6 shows that *A* used *masu* forms such as *atsuku arimasendesita* 'It was not hot' (V 202003) in *i*-adj., whereas *B* only used *desu* forms: *atsuku nakattadesu* 'It was not hot' (V 202003). This suggests that in the *i*-adj., the learner's analysis strategy tends to commence after the learner reaches the 5th period of re-contact and attempts to unify SFF into the *desu* forms.

Table 7: Usage of formal *na*-adjective (Noun) sentences

Sentence style		Returnee *A*	Returnee *B*	Koreans	Chi.-Kor.	Total
Positive	Past	1 (3)	- (-)	7	24	32
	Non-Past	2 (87)	2 (46)	282	140	426
Negative	Past	6 (4)	5 (4)	-	3	14
	Non-Past	- (1)	- (3)	17	13	30

Table 7 summarizes the actual usage of formal noun sentences from the survey in the 5th period; the table shows that although *A* used both Norms and Non-norms (Examples 17ab, 18ab), *B* only used Non-norms (Examples 19 and 20).

(17a) *A: Kankokujin-**janaka-tta-desu**.* (V 201611 and V 202003)
Korean-NEG-AUX-Past-COP
'He was not a Korean.'

(17b) *A: Chuugokujin-**dewaarimasen-desi-ta***. (V 201703)
Chinese-AUX-NEG-FML-COP-Past
'He was not a Chinese.'

(18a) *A: Sizuka-**janaka-tta-desu.*** (V 201611 and V 202003)
quiet-AUX-NEG-Past-COP
'It was not quiet.'

(18b) *A: Sizuka-**dewaarimasen-desi-ta***. (V 201703 and V 202003)
quiet-AUX-NEG-FML-COP-Past
'It was not quiet.'

(19) *B: Chuugokujin-**janaka-tta-desu***. (V 201611, V 201703 and V 202003)
Chinese-AUX-NEG-Past-COP
'He was not a Chinese.'

(20) *B: Sizuka-**janaka-tta-desu**.* (V 201611, V 201703 and V 202003)
quiet-AUX-NEG-Past-COP
'It was not quiet.'

Thus, considering that *A* had remembered Norms as chunks in the early learning stage and that the sentences I investigated were formal rather than from discourse data, I consider the phenomenon to represent a style shift for *A* that reflected her higher Japanese ability. These results were the same for verb sentences (Examples 21, 22).

(21) *A: Isogasiku-te terebi-o* ***mi-masen-desi-ta***. (V 201703)
busy-Cause TV-OBJ watch-FML-NEG-COP-Past
'I did not watch TV because I was busy.'

(22) *B: Sono-hito-wa* ***kekkonsi-naka-tta-desu***. (V 201703)
That.person-TOP marry-NEG-Past-COP
'He did not get married.'

4.4 Linguistic factors related to the use of commentary predicate *[N]*

Next, I examined the characteristics of the informants' *[N]* through their discourse functions. As was the case in Examples 23-25, *A* distinguished *[N]* according to its discourse function in the sentence.

(23) *A:* ***Ayamat-ta-n-da-ne.*** (III 200812)
apology-Past-*[N]*-COP-SFP
'I apologized.'

(24) *A:* ***Si-teru-n-da-yo.*** (IV 200903)
do-ASP-*[N]*-COP-SFP
'I'm doing it.'

(25) *A:* ***Watara-nakyaikenai-n-dat-ta.*** (IV 200908)
cross.over-Obligation AUX-*[N]*-COP-Past
'I had to cross over it.'

In the context of politely conveying simple facts, *A* was distinguishing the Norms and the proper *ru-[N]* and *ta-[N]* as she explained information that was unknown to her speaking partner.

One case of using *[N]-da* to connect the sentence-final particle *yo* with the information presentation function and one case of connecting other sentence-final particles behind can be understood as examples of proper use according to the discourse function.

However, *B*, who did show the use of Norms, also showed non-normative variations; her mixture of variations containing *[N]* is remarkable (Examples 26-28). There were five cases with sentence-final particle *yo*, two cases with other sentence-final particles, and eight cases without these particles. That is, in the discourse in the L2 attrition period, *A* had only two cases of no discourse function *[N]* of explanation, confirmation, or assertion, which was much fewer than *B*'s eight cases. I attribute this to *A*'s lower usage of Non-norms.

(26) *B:* ***Suru-n-da-yo.*** (III 200812)
do-*[N]*-COP-SFP
'I will do it.'

(27) *B:* ***Nac-chat-ta-n-da-yone.*** (IV 200905)
happen-ASP-Past-*[N]*-COP-SFP
'That's happened.'

(28) *B:* ***Nat-ta-n-daroo.*** (IV 200908)
happen-Past-*[N]*-COP-SPEC
'Is it happened?'

4.5 Collocation relationships with sentence-final particles (SFP)

Finally, as shown in Table 8, I observed a continuous pattern of usage classification in the collocation relationships with sentence-final particles in the speech of *A* and *B*.

Table 8: Korean returnee sisters' usage of sentence-final particles

			A					*B*		
Sentence styles	I	II	III	IV	V	I	II	III	IV	V
Informal styles + Formal styles	1955	251	92	-	92	1419	144	143	1	82
yo	276	8	4	-	2	189	15	4	-	-
ne	132	2	-	-	-	76	2	5	-	-
na	54	6	1	-	-	34	4	2	-	-
The ratio of sentence-final particles	24%	6%	5%	0%	2%	21%	15%	8%	0%	0%

Notes: Only verb sentences are included. In addition, *A* had 140 cases (*no*, 72 cases) and *B* had 44 of sentence-final particles *kai, ka, sa, ze, zo, no, mono, ya,* and *wa*. This is considered a feature of child L2 in contrast with adult L2.

For both *A* and *B*, the total usage of sentence-final particles decreased drastically beginning in the 3rd period. However, *A* showed more relevant sentence-final particles based on the degree of Japanese contact. *A* distinguished between Norms and Non-norms according to the discourse function in sentences. By contrast, *B* misused Non-norms that were inappropriate for the scene. She then began to use Norms incorrectly, and there was no use of *[N]* with explanation, confirmation, and assertion function where there should have been (Example 29).

(29) *B: Kak-e-nai,* ***dat-ta-desu***. (V 201611)
write-POT-NEG, COP-Past-COP
'I can write, it was.'

However, *A* and *B* simultaneously used fixed Norms more than colonial Japanese. From this reason I conclude that the verb SFF in L2 in its incubation period maintained sequentially according to the elapsed time after Japanese contact cessation along with the linguistic factors such as the usage according to the discourse function in the sentence.

5 Conclusions

In this paper, I have described actual usages of maintenance, attrition, and regression of SFF in the L2 Japanese of two Korean returnee children compared with that in the colonial Japanese of thirteen elderly. Based on these observations which focused on the verb sentences of Japanese in the incubation period, I present the following study conclusions.

(1) The lower the level of Japanese contact in the L2 incubation period, the more likely that the conjugation forms that predominantly use sentence Non-norms will mainly appear in analytical forms in the 5th regression period. *[N]* begins to intervene irrelevantly for a discourse function in the past and affirmative *ta-desu.* However, the usage of fixed and idiomatic Norms is maintained. These tendencies are stronger in the Japanese of two Korean child returnees than in colonial Japanese. There is also a higher ratio of formal Norms observed.

(2) The informants all tended to use Non-norms in not only verb, but also adjective and noun, sentences.

(3) The use of sentence-final particles and the use of Non-norms varied according to the sentence discourse function, which affected the analysis of SFF, correlated with the Japanese contact level as well as with the elapsed time (=age) of contact cessation.

(4) In terms of the forms of Japanese sentences at the learning stage and at the cessation of contact with Japanese, it is possible to deduce the process of creating verb sentences across the life cycle by partly modifying Hwang's (2013) conclusions that Japanese acquisition, maintenance, attrition, and regression for Korean learners is tentatively a mirror image.

To conclude, I propose that the SFFs of Japanese in the incubation period are recursively reconstructed and treated as the interlanguage in which the ongoing analysis, which is often observed in adult Korean learners of Japanese, takes place in terms of conjugation forms.

Acknowledgments

This study was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2019S1A5A2A01035182).

References

Anderson, R. W. (1982). Determining the linguistic attributes of language attrition. Lambert, R.D. & Freed, B.F. (eds.) *The Loss of Language Skills,* 83-118, Rowley MA: Newbury House.

Dorian, N. C. (1981). *Language death*: *The life cycle of a Scottish Gaelic dialect*. Philadelphia: University of Pennsylvania Press.

Han, K. (2000). A Study on the English Loss of Returnees; focused on elementary school students (귀국학생의 영어손실에 관한 연구: 초등학생을 중심으로). Yonsei University Graduate School of Education Master's Thesis, Seoul: Yonsei University Press.

Hansen, L. (1999). Investigating second language attrition: An Introduction, In Hansen, L. (ed.) *Second language attrition in Japanese contexts*; Tokyo, 3-18. Oxford University Press

Hansen, L. (ed.) (1999). *Second language attrition in Japanese contexts.* New York; Tokyo: Oxford University Press.

Hayashi, B. (1999). Testing the regression hypotheses, the remains of the Japanese negation system in Micronesia. In Hansen, L. (ed.) *Second language attrition in Japanese contexts.* 154-168. Oxford University Press.

Hwang, Y. (2009). Pragmatic approach to interlanguage in the second language retention period; focused on the sentence-final particle *Yo* of Korean elder Japanese (第二言語保持期の中間言語への誤用論的アプローチ: 韓国高年層日本語の文末詞ヨを対象に), *Japanese Language Studies*, 24, 225-245. The Japanese Language Association Of Korea.

Hwang, Y. (2010). Preservation of sentence forms in Japanese learners during Japanese colony; focused on verb formal sentences (일제강점기 일본어학습자에 있어서 문말형식의 보존: 동사문 정중체를 중심으로). *Japanese Language Studies*. 28, 286-300. The Japanese Language Association of Korea.

Hwang, Y. (2013). Preservation of *desu* and *masu* in Japanese of incubation period in Korean and China (韓国と中国の潜伏期日本語におけるデス形とマス形の保持), *Japanese language education research*, 25, 95-110. Korean Academic Society of Japanese Education.

Hwang, Y. (2015). Second Language Retention Mechanism of Korean Japanese Learners; Variations according to Japanese contact type (韓国人日本語学習者の第二言語保持メカニズム: 日本語接触の類型による変異形), *Japanese language education research*, 31, 273-291. Korean Academic Society of Japanese Education.

Kan, K. (2004). Remaining Japanese in Taiwan (台湾に残存する日本語の実態). Osaka University Graduate School of literature Doctoral Thesis, Osaka: Osaka University Press.

Kanazawa, H. (2008). The Japanese of foreign students is Japanese in the future; Dynamism of Japanese change (留学生の日本語は、未来の日本語: 日本語の変化のダイナミズム), Hitsuji Shobo.

Kiguchi, M. (2004). Studies on the attrition of residual Japanese in Korea; Through the analysis of the Korean elder corpus (韓国における残存日本語の摩滅に関する研究: 韓国人高齢者コーパスの分析を通して), Chung-ang University Doctoral thesis, Seoul: Chung-ang University Press.

Kiguchi, M. (2009). A Study on the Change of Remaining Japanese in Korea (韓国における残存日本語の変容をめぐって). *Japanese Studies,* 24, Seoul: Japan Research Institute of Chung-ang University.

Kim, C. (2002). The need for bilingual education for returnees (귀국학생을 위한 이중언어교육의 필요성). *Bilingualism*, 21, 122-140. Institution of Bilingualism.

Kim, M. (2010). Maintenance and attrition of the Japanese in child returnees (帰国子女の日本語の維持と摩滅), *Japanese Studies*, 29(14), 183-195. Meiji Shoin.

Kubo, M. (1994). Current Status of Child Returnees Education (帰国子女教育の現状), *Japanese language*, 13(3), 4-12. Meiji Shoin.

Matsumoto, K. (2013). Pragmatic Variations and Changes in Palau Japanese (パラオ日本語の語用論的変異と変化), Okamura, T. (ed.) *The linguistic world of Oceania*, 220-262. Keisuisha.

Noda, S., Sakoda, K., Shibuya, K., Kobayashi, N. (2001). *Grammar acquisition of Japanese learners* (日本語学習者の文法習得), Taishukan Shoten.

Noda, H. (2004). Factors related to the use of negative and non-negative statements; Based on the case study and the younger year anchor survey (否定ていねい形「ません」と「ないです」の使用に関わる要因: 用例調査と若年層アンケート調査に基づいて), *Quantitative Japanese Linguistics*, 24(5), 228-244.

Ono, H. (1994). Retention of the ability of the returning child (帰国子女のバイリンガルの能力の保持). *Japanese language*, 13(3), 4-12. Meiji Shoin.

Park, A. (2014). Qualitative Research on the Reminiscence of Middle and High School Students in Korea; Focusing on the case of living in Japan with parents (귀국학생의 중·고등학교생활 회고담에 관한 질적 연구: 부모와 동행하여 일본 생활을 한 경우를 중심으로). *Educational Anthropology Research*, 17(2), 231-286

Sanada, S. (2009). Dynamics of Eastern Asia Residual Japanese (東アジア残留日本語のダイナミックス), *Monthly language*, Vol.38. No.1, 82-89.

Sibuya, K. (1995). Possible expressions of Japanese remaining in the South Sea Islands (旧南洋群島に残存する日本語の可能表現), *Musha*, 2, 81-96.

Sibuya, K. (2001). The Status of Japanese Remaining in the Palau; Report: Preface (パラオにおける日本語残存の実態 報告書: 序章), Sanada, S. (eds.) *Investigations on dialects that have been disappearing in Japanese*, Grant-in-Aid for Scientific Research (A), Emergency Surveys on the 'Endangered Languages' of the Pacific Rim, Achievement report, 285-302.

Tanomura, T. (1994). A Quantitative Survey on the Choice of Negative Forms of Politeness; *masen and naidesu* (丁寧体の述語否定形の選択に関する計量的調査: 「～ません」と「～ないです」), Osaka University of Foreign Studies, 11, 51-66.

Tomiyama, M. (2000). Child second language attrition; A longitudinal case study. *Applied Linguistics 21(3),* 304-332.

Tomiyama, M. (2004). Loss and maintenance of L2 (第二言語の喪失と維持), Terauchi, M. Kinoshita, K & Narita, M (eds.) *The present of second language acquisition research,* 239-256. Taishukan Shoten.

Tomiyama, M. (2008). Age and proficiency in L2 attrition: Data from two siblings, *Applied Linguistics 30(2),* 253-275.

Weltens, B. & Cohen, A. D. (1989). Language attrition research; An introduction. *Studies in Second Language Acquisition, 11*, 127-133.

Winford, D. (2003). *An introduction to contact linguistics*. Oxford: Blackwell.

Abbreviations and symbols

A.	= Elder Korean returnee
ASP.	= Aspect
AUX.	= Auxiliary
B.	= Younger Korean returnee
COP.	= Copula, Be verb (e.g. *da, desu*)
FML.	= Formal
i-adj.	= *i*-adjective sentences (e.g. *takai*, *yasui*)
INF.	= Informal
INT.	= Interrogative
LOC.	= Locative
[N].	= Nominalizer, Commentary predicate *n(o)da*
na-adj.	= *na*-adjective sentences (e.g. *genkida, sizukada*)
Norms.	= Normative sentences-final forms
Non-norms.	= Non-normative sentences-final forms
NEG.	= Negative
OBJ.	= Object
POT.	= Potential
SFF.	= Sentence-final forms
SPEC.	= Speculation
SUB.	= Subject
TOP.	= Topic

EXAMINING THE PART-OF-SPEECH FEATURES IN ASSESSING THE READABILITY OF VIETNAMESE TEXTS

An-Vinh LUONG
Computational Linguistics Center, University of Science, Ho Chi Minh City, Vietnam
anvinhluong@gmail.com

Diep NGUYEN
Department of Linguistics, University of Social Sciences & Humanities, Ho Chi Minh City, Vietnam
nhudiep2004@gmail.com

Dien DINH
Computational Linguistics Center, University of Science, Ho Chi Minh City, Vietnam
ddien@fit.hcmus.edu.vn

Abstract

The readability of the text plays a very important role in selecting appropriate materials for the level of the reader. Text readability in Vietnamese language has received a lot of attention in recent years, however, studies have mainly been limited to simple statistics at the level of a sentence length, word length, *etc.* In this article, we investigate the role of word-level grammatical characteristics in assessing the difficulty of texts in Vietnamese textbooks. We have used machine learning models (for instance, Decision Tree, K-nearest neighbor, Support Vector Machines, *etc.*) to evaluate the accuracy of classifying texts according to readability, using grammatical features in word level along with other statistical characteristics. Empirical results show that the presence of POS-level characteristics increases the accuracy of the classification by 2-4%.

Keywords: text readability; text difficulty; Vietnamese text readability; text classification; school textbooks

Povzetek

Berljivost besedila ima zelo pomembno vlogo pri izbiri ustreznih gradiv za raven bralca. Berljivost besedil v vietnamskem jeziku pridobiva pozornost šele v zadnjih letih in dosedanje študije so omejene na preproste ocene na osnovi statističnih podatkov za dolžino stavka, dolžino besed in podobnih značilnosti. V tem članku raziskujemo vlogo slovničnih značilnosti na besedni ravni pri ocenjevanju težavnosti besedil v vietnamskih učbenikih. Za oceno natančnosti razvrščanja besedila glede na berljivost smo uporabili modele strojnega učenja (na primer drevo odločitve, K-najbližji sosed, podporni vektorski stroji itd.) Empirični rezultati kažejo, da upoštevanje različnih značilnosti na nivoju besednih vrst poveča natančnost klasifikacije za 2-4%.

Ključne besede: berljivost besedila; raven enostavnosti; berljivost vietnamskih tekstov; klasifikacija tekstov; šolski učbeniki

Acta Linguistica Asiatica, 10(2), 2020.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.10.2.127-142

1 Introduction

In today's era of information explosion, thousands of documents with different contents and in different languages get released every second. Such documents have different levels of readability; some are easy to read and understand while others are more difficult and demand larger amount of time and knowledge to get through. It is generally known that the best way to assess whether a text is easy or difficult is to ask readers to read or skim that text, though this can be time-consuming for the readers. Therefore, we suppose that there exists some kind of a method that assists a reader to determine the readability of the text, upon which they make further decision on whether they would continue reading or not. Recently gaining a lot of focus is readability, which is one of such methods.

Brown and colleagues state that

> readability is a concept that describes the degree to which a text is easy or difficult to read. A readability index is a numerical scale that estimates the readability or degree reading difficulty that native speakers are likely to have in reading a particular text. (Brown et al., 2012).

Determining the readability index of a document is to determine how difficult the text is, which gives a reader information on whether the document is suitable for them to read to understand it in a reasonable amount of time. Information on readability is useful in many different fields of science as well as in everyday life. It can be used when assisting scientists at publishing articles, helping text editors (writers, journalists, *etc.*) to create documents suitable specific audience, or else for manufacturers to produce readable manuals. Above all, information on readability is of most importance in education, especially in second language education. It is used when textbooks are compiled, or for educators to make decisions on appropriate texts are made.

Research on the difficulty of texts originates back to the late 19th century when Lucius Adelno Sherman wrote that "the average length of sentences has been decreasing over time" (Sherman, 1893). Many books on readability have been published since, however, they mostly applied for English, such the work of Dale and Chall (1948), Si and Callan (2001), Schwarm and Ostendorf (2005), Chall and Dale (1995), Chen and Meurers (2018), etc., and some languages that were treated as lingua franca at some point in the history or some part of the world. Thus we find works on French (François (2014), François & Fairon (2012), *etc.*), Chinese (Chen et al. (2013); Jiang et al. (2018); Sun et al. (2014), *etc.*), Spanish (Coco et al. (2017); I. Parkeret al. (2001); Spaulding (1956), etc.), Arabic (Al-Ajlan et al. (2008); Al-Tamimi et al. (2014); Al Khalil et al. (2018); Saddiki et al. (2015); Saddiki et al. (2018), *etc.*), and other languages.

For less-resource languages, studies on the readability of texts are still limited, and Vietnamese is one of such languages. In Vietnamese, there some publications that date back to 1980ies (Nguyen & Henkin 1982, 1985), and recent studies of Luong et al. (2017, 2018a, 2018b), Điệp (2019) and Luong & Tran (2019). These studies have shown some valuable features for assessing text readability in Vietnamese, but the results are limited and further research on the topic is necessary.

In this research, we examine features on the level of parts of speech (POS-level) and assess the readability of literary texts based on them. The texts were taken from literary textbooks for school students from grade 2 to grade 12, which corresponds to the students' age 7 to age 17. This method inherits the results of the Luong et al. (2017, 2018b) with the addition of a number of grammatical features at word-level to build a text-based classifier based on readability through some machine learning methods like Decision Tree, K-nearest neighbor, Support Vector Machines, *etc.*

The article is thus organized as follows. Section 2 presents some ground works on text readability and previous literature on Vietnamese text readability. It also introduces the features that we surveyed and used in this study to develop models for assessing the readability of Vietnamese texts, using some classification algorithms along with experimental results. Section 0 presents the results of the study and discusses them, and the final Section 5 offers an overall conclusion to the topic.

2 Related works

2.1 Different approaches in previous studies

Previous studies of text readability can be grouped into two groups based on either they undertake traditional approach or corpus-based approach.

Traditional approach uses conventional statistical methods on the documents to select high correlation factors with the readability of texts and then use regression analysis to create formulas for measuring the readability. The factors examined are typically shallow features, also called easy-to-extract features, such as average sentence length, average word length, percentage of difficult words in the documents. Representative researches with this approach produces the Dale-Chall formula (Dale & Chall, 1948), the Gunning Fog Index (Robert, 1952), the SMOG formula (Mc Laughlin, 1969), the Flesch-Kincaid grade level readability (Kincaid, Fishburne, Rogers, & Chissom, 1975), the new Dale-Chall formula (Chall & Dale, 1995), and others.

On the other hand, corpus-based approach approach has been developed in recent years due to the fast development computer science and machine learning algorithms. Studies in this approach see the problem of assessing the readability of text as a classification problem, and use machine learning models to classify text by layer of readability based on extracted features. Representative works are those of Si & Callan (2001); Collins-Thompson & Callan (2005); Schwarm & Ostendorf (2005); Heilman et al.

(2007); Pitler & Nenkova (2008); Feng et al. (2010); Vajjala & Meurers (2012); Jiang et al. (2015); Wang & Andersen (2016); Chen & Meurers (2018), and others.

2.2 Studies on text readability in Vietnamese

The research on the readability of the text in Vietnamese is still quite small and their results are limited. Nguyen et al., have introduced two formulas to measure the readability of Vietnamese texts (Nguyen & Henkin, 1982, 1985). These two formulas base on features such as the average length of sentences or words, and the ratio of difficult words in texts. The weak point in these works is that the two formulas were surveyed and evaluated on a relatively small amount of data; on 20 documents in Nguyen & Henkin, 1982 and 54 documents in Nguyen & Henkin, 1985.

Luong et al. (2017) conducted a survey of texts extracted from literary textbooks for Vietnamese high school students and suggested to use the feature of text length to classify texts according to readability. Experimental results show that the length of texts has a great influence on the classification results, and is to be used to evaluate texts in Vietnamese textbooks.

Luong et al. (2018a) introduced a new formula for measuring the readability of Vietnamese texts. This formula is based on a survey of 1,200 documents classified into 3 levels of difficulty (easy, medium and difficult). The features of the average length of the sentence, of the word and the ratio of difficult words in the text have been chosen formulas criteria.

In addition, Luong et al. (2018b) published another study on the readability of texts using the proportional features of proper nouns and Vietnamese specific characteristics such as Sino-Vietnamese words ratio, borrowed words ratio, and dialect words ratio within documents. Experimental results show the contribution of these features in improving the accuracy of classification processes.

Điệp et al. (2019) presented the statistical analyses on the frequency of POS tags in Vietnamese texts. They conducted a survey of 209 texts extracted from Vietnamese textbooks from grade 2 to grade 5 (corresponding to age 7 to 10) for primary students according to the general curriculum in Vietnam. Their results showed that words such as common nouns and volatile verbs were common in the examined documents. In addition, through the correlation analysis between the frequencies of POS tags and the readability of the surveyed documents, they also proved a high correlation between the ratio of common nouns and the ratio of prepositions with the readability level in the examined texts.

Furthermore, Luong & Tran (2019) introduced a method of evaluating the readability of documents by comparing difficulty correlation between different documents. They built a set of 30 texts – which were graded the readability level – as the standard.

The documents in this research will be compared to the standard texts proposed by Luong & Tran (2019) to determine the readability level.

2.3 Features

In this section, we will introduce features that used to classify texts with readability level. These features include the so-called traditional or grammatical features at the word level (features (1) – (4)), and features (5) – (10) proposed by Luong et al. (2017, 2018b) that have been defined relevant in Vietnamese literary texts and are the focus of this research.

(1) Average sentence length. The average length of sentences is a common factor in most studies of text readability. The length of sentences is very important in the process of documenting and reading texts. If a text has too many long sentences, it may make it difficult for the reader to fully understand meanings of its long sentences. On the other hand, using only short sentences may make the text discrete and incoherent, which could make the reader experience difficulties. Therefore, the length of sentences is a very important factor in assessing the readability of text. The average sentence length features commonly used are average sentence length in words (ASLW), in syllables (ASLS), and in characters (ASLC).

(2) Average word length. Carver (1976) showed a linear correlation between the length of words and the readability of text, which is a commonly used factor in studies on text readability. The average word length in syllables (AWLS) and in characters (AWLC) are commonly used with length features.

(3) Percentage of difficult words. In many studies, the percentage of difficult words is one of the most valuable factors when assessing the readability of texts. These studies often use a list of easy vs. difficult words in a language as the base for calculation. However, building such a list takes a lot of effort, and therefore many studies used statistical lists of words according to their frequency of use instead; they chose most commonly used words in a particular language and treated them as a list of easy words.

In this study, we use 3,000 most popular words extracted from the statistical list of words in the Vietnamese texts of the researching group of Dinh et al. (Dinh, Nguyen, & Ho, 2018) as the basis for calculating difficult word rate. The features we surveyed are Percentage of Difficult Words (PDW), Percentage of Unique Difficult Words (PDDW).

(4) Percentage of difficult syllables. Vietnamese writing is monosyllabic in nature. Every “syllable” is written as though it were a separate dictation-unit with a space before and after. Such a unit is called morphosyllable or “tiếng” in Vietnamese. Each morphosyllable tends to have its own meaning and consequently a strong identity. However, these morphosyllables are not automatically combined into ‘words’ as the linguistic notion of word commonly applies for European languages (Tran et el., 2007), which leads to difficulties for readers, especially those with low reading skills, to distinguish the boundaries between words.

For this reason, we consider syllables as an important language unit of Vietnamese to make statistics and use as a characteristic for examination. In this work, we use 3,000 most popular syllables of Dinh et al. (2018) to extract the following two features: percentage of difficult syllables (PDS) and percentage of distinct difficult syllables (PDDS).

(5) Text length features. In the study by Luong et al. (2017), results showed the essential role of the text length features in assessing readability. The features that Luong et al. surveyed and are relevant to this study are as follows. The total number of sentences (NSen), the total number of words (NWo), the total number of syllables (NSyl), the total number of characters (NCha), the total number of distinct words (NDWo), and finally the total number of distinct syllables (NDSyl).

In the article published in 2018, Luong et al. introduced some additional features for assessing readability of Vietnamese texts:

(6) Percentage of Sino-Vietnamese words. Vietnam has spent more than 1,000 years of domination by the Chinese feudal dynasties (111BC - 905AC). During that period, the Vietnamese language was strongly influenced by Chinese culture and language, and those influences continue to this day. Vocabulary of the Vietnamese language consists of more than 60% of words of Chinese-origin, the called as Sino-Vietnamese words (DeFrancis, 1977). These Sino-Vietnamese words are often used in the official and ceremonial language and are therefore considered as more difficult compared to originally Vietnamese words of the same meaning.

In this study, we examined features such as percentage of Sino-Vietnamese Words (PSVW), percentage of distinct Sino-Vietnamese words (PDSVW), and the proportion of distinct Sino-Vietnamese within all distinct words (DSVW/DW).

(7) Percentage of borrowed words. Similar to Sino-Vietnamese words, many words from other languages entered Vietnamese. This foreign influence was especially strong during the French invasion of Vietnam in the middle of the 19th century. Such words of French, English and other origin undertook Vietnamese phonetic transcriptions (Alves, 2009) and are nowadays used in the official and scientific language. It is estimated that they influence the readability and are therefore taken into account as the percentage of borrowed words (PBW), the percentage of distinct borrowed words (PDBW), and the proportion of distinct borrowed words within all distinct words (DBW/DW).

(8) Percentage of dialectal words. Vietnamese territory has many different regions with different cultural and linguistic characteristics. Regions tend to localize general Vietnamese language and use it as their own regional language. Consequently, such dialectal vocabulary, which is not available in the standard Vietnamese language, is thought to be difficult for readers. In assessing text readability the feature appears either as the percentage of dialect words (PDiaW), the percentage of distinct dialect words (PDDiaW), or the proportion of distinct dialect words within all distinct words (DDiaW/DW).

(9) Percentage of proper nouns. According to Luong et al. (2018b), the more proper nouns in the text, the more effort the reader will have to memorize those objects, and therefore the text is considered more difficult. For this reason we have decided to take into account the characteristics of proper nouns in this experiment. The features are defines in the following way. Nr/Sen is the abbreviation for the proportion of proper nouns within sentences. Nr/W is the abbreviation for the number of proper nouns in comparison to all words. Nr/DW stands for the number of proper nouns that is divided by the number of distinct words. DNr/Sen points to the proportion of distinct proper nouns within the overall number of sentences. DNr/W stands for the number of distinct proper nouns divided by the number of words. Finally, DNr/DW is the abbreviation for the number of distinct proper nouns divided by the number of distinct words.

(10) Other parts of speech and their elements. In this study, we also used other POS tags such as countable noun, directional verb, parallel association, *etc.* to experiment with the model. Table 1 is a list of tags used. These POS tags are derived from the CLC_VN_Toolkit tool, which has been developed by the Computational Linguistics Center, Ho Chi Minh City University of Science[1]. This is a tool for pre-processing, sentences segmentation, words segmentation, part-of-speech tagging (POS), named entity labeling, *etc.* for Vietnamese texts. Similar to proper nouns, we use features with abbreviated symbols for each POS element. The number POS divided by the number of sentences is POS/Sen. The number of POS divided by the number of words is POS/Wo. The number of POS divided by the number of distinct words is POS/DWo. The number of distinct POS divided by the number of sentences is DPOS/Sen. The number of Distinct POS divided by the number of words is DPOS/Wo. The number of Distinct POS divided by the number of distinct words is DPOS/DWo. The abbreviation 'POS' is generally replaced by POS tags as shown in Table 1 except for proper nouns (Nr) already presented in the work of Luong et al. (2018b).

Table 1: List of Vietnamese POS tags used in CLC_VN_Toolkit

POS	**Tag**	**POS**	**Tag**
Countable nouns	Nc	Quality adjectives	Aa
Concrete nouns	Nu	Demonstrative pronouns	Pd
Temporal nouns	Nt	Personal pronouns	Pp

[1] CLC website: http://clc.hcmus.edu.vn

POS	Tag	POS	Tag
Numerals	Nq	Adverbs	R
Common nouns	Nn	Prepositions	Cm
Proper nouns	Nr	Parallel conjunctions	Cp
Directional verbs	Vd	Subordinating conjunctions	Cs
State verbs	Ve	Modifiers	M
Comparative verbs	Vc	Emotion words	E
Volatile verbs	Vv	Foreign words	FW
Directional co-verb	D	Onomatopoeia	ON
Quantity Adjectives	An	Idioms	ID

3 Experiment

In this study, we used the corpus of 371 literary texts of Luong et al. (2018b) for experimentation. These documents were taken from Vietnamese textbooks for primary, middle and high school students in Vietnam. We divided the texts into groups based on:

(1) grade level (from grade 2 to grade 12);
(2) level of education (Primary, Middle and High school).

Table 2 presents the basic statistics of the corpus. The features mentioned in Section 2.3 are used to build the classification models for text readability.

Table 2: The statistics of the corpus of 371 literary documents of Luong et al. (2018b)

Grade	2	3	4	5	6	7	8	9	10	11	12
Number of texts	67	62	40	40	28	13	17	21	15	19	49
Average number of sentences	18.34	19.63	21.53	21.43	54.75	46.38	65.76	107.33	60.67	105.16	111.65
Average number of words	158.06	192.31	231.28	244.4	679.54	676.92	969.24	1447.4	861.73	1359.9	1710.3
Average number of distinct words	100.63	125.58	144.3	152.78	304.86	329.69	394.29	526.29	368.4	510	576
Average number of syllables	178.48	221.98	276.1	288	784.11	820.85	1131.5	1709.7	1006.5	1579.1	2179.4
Average number of distinct syllables	111.36	141.53	164.78	173.35	327.54	372.46	428.35	555.52	390.07	534.95	594.2
Average number of characters	826.8	1065.4	1335	1395.9	3709	3942.3	5401.9	8160	4860	7535.1	10761

Grade	2	3	4	5	6	7	8	9	10	11	12
Average sentence length in words	9.14	10.61	11.59	12.69	14.01	17.99	17.78	18.23	15.04	15.67	16.68
Average sentence length in syllables	10.36	12.34	14.08	15.21	16.14	22.3	21.34	22.07	17.72	18.72	22.17
Average sentence length in characters	48.3	59.57	68.61	74.37	76.67	108.69	103.38	106.62	85.79	90.66	111.21
Average word length in syllables	1.13	1.16	1.2	1.19	1.15	1.23	1.19	1.2	1.17	1.18	1.32
Average word length in characters	5.25	5.61	5.84	5.77	5.43	5.98	5.74	5.78	5.67	5.7	6.59

We conducted experiments by using several classification algorithms such as Decision Tree (denoted as D-TREE), K-nearest neighbor (K-NN), Multi-layer Perceptron (MLP), Random Forest (RND-FRST), and Support Vector Machines (SVM). In this study, we used Scikit-learn, a machine-learning library for the python programming language for the experiments. With D-TREE and RND-FRST, we used two common impurity measures: Entropy and Gini index. In order to avoid overfitting we used k-fold cross validation during training and testing; randomly dividing the corpus into 5 parts (4 parts for training and 1 for testing). The best features combinations of Luong et al. (2017) and Luong et al. (2018b) are used as the baselines for the experimental process. Tables 3 and 4 show the best practices on 4 metrics: accuracy (Acc), precision (P), recall (R), and F1-score (F1).

Table 3: Classification results performed on grade-level documents

Feature set	Acc	P	R	F1
	D-TREE (ENTROPY)			
Luong2017	0.3828	0.2893	0.2772	0.2728
Luong2018	0.4206	0.3488	0.3276	0.3142
Luong2017 + PSVW, Nq/Sen, DMSen	**0.4449**	**0.3749**	**0.364**	**0.3552**
Luong2017 + PSVW, Aa/Sen, DCm/Wo	0.4341	0.3714	0.3523	0.3492
Luong2017 + PSVW, DNr/DWo, Cp/DWo	0.4204	0.3709	0.3467	0.3359
	D-TREE (GINI)			
Luong2017	0.3855	0.3049	0.2973	0.289
Luong2018	0.3909	0.3038	0.2959	0.2888
Luong2017 + PSVW, Aa/Sen, Cm/Wo	0.4448	0.3506	**0.3829**	**0.3538**
Luong2017 + PSVW, DNr/DWo, Nq/Wo, DAa/Sen, Nq/DWo, Cm/Wo	0.4368	**0.3849**	0.3472	0.3409
Luong2017 + PSVW, Nq/Wo, DCm/Sen	**0.4502**	0.3649	0.3473	0.3375

Feature set	Acc	P	R	F1
	KNN			
Luong2017	0.4556	0.2996	0.3097	0.2928
Luong2018	0.4475	0.2929	0.3038	0.2877
Luong2017 + PSVW, Aa/Sen, Cm/Sen	**0.4609**	**0.3236**	**0.3283**	**0.3069**
Luong2017 + PSVW, Nr/Sen, Cm/Sen	0.4584	0.3075	0.3232	0.3035
Luong2017 + PSVW, Cp/Sen, Cm/DWo	0.4476	0.3134	0.3182	0.3025
	MLP			
Luong2017	0.3882	0.2916	0.3034	0.2696
Luong2018	0.3883	0.3246	0.2845	0.2724
Luong2017 + PSVW, DNr/DWo, DAa/Sen, Cm/Wo, DNr/Sen	0.4286	0.3258	0.3404	**0.3058**
Luong2017 + PSVW, DNr/DWo, DCp/Sen	**0.4421**	0.3128	**0.3447**	0.2993
Luong2017 + PSVW, Aa/Sen, DD/Sen	0.38	**0.3488**	0.3098	0.2955
	RND-FRST (ENTROPY)			
Luong2017	0.4529	0.3569	0.3577	0.3403
Luong2018	0.4689	0.3952	0.3503	0.3477
Luong2017 + PSVW, DNr/DWo, Aa/DWo	**0.5041**	0.4291	**0.4029**	**0.3897**
Luong2017 + PSVW, Nr/Sen, DM/Wo	0.4772	**0.4629**	0.382	0.3889
Luong2017 + PSVW, Nn/Sen, Aa/DWo	0.4826	0.4178	0.4011	0.3811
	RND-FRST (GINI)			
Luong2017	0.4392	0.3191	0.3206	0.3071
Luong2018	0.4636	0.345	0.3365	0.3195
Luong2017 + PSVW, Nr/Sen, Cp/Wo	**0.523**	**0.4256**	**0.4089**	**0.4051**
Luong2017 + PSVW, DNr/DWo, DPp/Sen	0.4989	0.402	0.3883	0.3766
Luong2017 + PSVW, Nq/Sen, Nn/Sen	0.4852	0.4078	0.3809	0.371
	SVM (LINEAR)			
Luong2017	0.4446	0.3402	0.3257	0.3177
Luong2018	0.477	0.3892	0.3611	0.3538
Luong2017 + PSVW, DNr/DWo, Nq/Wo, DAa/Sen, Cm/Wo	**0.5148**	**0.4657**	**0.4219**	**0.418**
Luong2017 + PSVW, DNr/DWo, Nq/Wo, Nq/DWo, Cm/Wo	0.5068	0.4479	0.4099	0.4134
Luong2017 + PSVW, DNr/DWo, Nq/Wo	0.5069	0.4464	0.4132	0.408

Table 4: Classification results performed on school-level documents

Feature set	Acc	P	R	F1
	D-TREE ENTROPY			
Luong2017	0.7845	0.7268	0.7006	0.7021
Luong2018	0.8167	0.7594	0.7489	0.7511
Luong2018 + Nc/DWo, DPp/DWo	**0.8329**	**0.7881**	**0.7729**	**0.7761**
Luong2018 + DVc/Sen, DID/Wo	0.8221	0.7792	0.7584	0.7623
Luong2018 + FW/Wo, DNn/Sen	0.8221	0.7739	0.7569	0.7607
	D-TREE GINI			
Luong2017	0.7925	0.7234	0.6985	0.7008
Luong2018	0.7925	0.7174	0.7033	0.7049
Luong2018 + Nu/Sen, D/Wo	**0.8169**	0.7531	**0.7429**	**0.743**
Luong2018 + D/Sen, DNn/Sen	0.8087	0.7568	0.7341	0.7322
Luong2018 + Nc/Sen, DCp/DWo	0.8114	0.7441	0.7338	0.7316
	KNN			
Luong2017	0.7708	0.6687	0.656	0.6594
Luong2018	0.7708	0.6687	0.656	0.6594
Luong2018 + Vv/Wo, DVv/Sen	**0.7815**	**0.6846**	**0.6688**	**0.6746**
Luong2018 + Aa/Sen, DNu/Wo	0.7762	0.6759	0.6612	0.6655
Luong2018 + Aa/Sen	0.7762	0.6759	0.6612	0.6655
	MLP			
Luong2017	0.6589	0.4973	0.5855	0.5169
Luong2018	0.6846	0.5701	0.631	0.5666
Luong2018 + Nr/Wo, DPp/Wo	**0.7954**	0.7124	**0.7029**	**0.6723**
Luong2018 + Aa/Sen, FW/Wo	0.7707	**0.7555**	0.7005	0.6652
	RND-FRST ENTROPY			
Luong2017	0.8221	0.757	0.7368	0.7367
Luong2018	0.8355	0.7743	0.7547	0.7596
Luong2018 + M/Wo, DNq/Sen	**0.8599**	0.8138	**0.7956**	**0.802**
Luong2018 + Nc/Sen, Nq/Sen	0.8544	0.8126	0.789	0.7939
Luong2018 + Nq/Sen, Aa/Sen	0.8571	**0.8169**	0.7824	0.7903
	RND-FRST GINI			
Luong2017	0.8222	0.7569	0.7411	0.7439
Luong2018	0.8302	0.7735	0.7528	0.7573
Luong2018 + Nq/Wo, DON/DWo	**0.8653**	**0.8182**	**0.8004**	**0.8062**
Luong2018 + Nc/Sen, Nq/Sen	0.8652	0.8173	0.7973	0.8031
Luong2018 + Nq/Wo	0.8491	0.7978	0.7837	0.7879

Feature set	Acc	P	R	F1
	SVM LINEAR			
Luong2017	0.8274	0.785	0.7626	0.7644
Luong2018	0.8517	0.8107	0.7842	0.7903
Luong2018 + Aa/Sen, DPp/Wo	**0.8787**	**0.8462**	**0.8206**	**0.8231**
Luong2018 + Aa/Sen, DFW/Wo	0.8733	0.8326	0.8153	0.8182
Luong2018 + D/Sen, DNn/Sen	0.8706	0.833	0.8163	0.8162

From the results presented in Table 3 and Table 4 we can see that, when adding POS features, some features have helped improve the performance of the classification model.

With the experiments in grade-level grouping, accuracy increased from the value 0.4770 of the work of Luong2017 to the value 0.5148 when adding the features PSVW, DNr/DWo, Nq/Wo, DAa/Sen, Cm/Wo with the SVM classifier. Similarly, precision, recall and F1-score also increased from 0.3892, 0.361, and 0.3538 respectively in Luong2017 to 0.4657, 0.4219, and 0.4180 respectively with the SVM classifier. In experimental results, the most accurate features combination is the combination (Luong2017 + PSVW, Nr/Sen, Cp/Wo), implemented on the Random Forest classifier (Gini index). However, the combination that yield the highest precision and F1-score is the combination (Luong2017 + PSVW, DNr/DWo, Nq/Wo, DAa/Sen, Cm/Wo). Among the POS features surveyed, the feature DNr/Dwo (Number of Distinct Proper Nouns divided by number of Distinct Words) feature appears the most in high performing experiments (appears 9 times in Table 3). This shows that the DNr/Dwo feature is a good feature for evaluating the readability of Vietnamese texts Besides, some other POS features also appear several times in the Table 3, such as Cm/Wo (5 times), Nq/Wo (5 times), Aa/Sen (4 times), etc. These POS features are also valuable for classifying Vietnamese texts according to difficulty level.

With school-level grouping, the highest experimental results belong to the feature combination (Luong2018 + Aa/Sen, DPp/Wo), implemented on the SVM classifier: the Accuracy, Precision, Recall and F1-score increased from 0.8517 (Luong2018) to 0.8787; from 0.8107 (Luong2018) to 0.8462; from 0.7842 (Luong2018) to 0.8206; and from 0.7903 (Luong2018) to 0.8231 respectively. The feature Aa/Sen (Number of Quality Adjectives divided by number of Sentences) appears the most (6 times) in the Table 4, therefore, this is a valuable feature for assessing the readability of Vietnamese texts. Similarly, features like DNn/Sen (appears 3 times), Nc/Sen (appears 3 times) or Nq/Sen (appears 3 times) are also good features for automatic classification of Vietnamese texts according to the difficulty level.

Experimental results also show that SVM classifier performs best on overall Accuracy, Precision, Recall and F1-score for most feature sets on both school and grade-level. The Random Forest classifier (Gini impurity) archives the best accuracy in grade-level with the feature set of (Luong2017 + PSVW, Nr/Sen, Cp/Wo). The other classifiers do not seem suitable for the problem of evaluating the readability of Vietnamese text.

4 Discussion and conclusion

Text readability is an important factor affecting the selection and understanding of documents. Numerous studies on text readability have been conducted for English and some other resource-rich languages, while for Vietnamese research results are rare and limited. In this study, we investigated the role of word-level grammatical characteristics in assessing the difficulty of texts in Vietnamese textbooks. We conducted empirical assessments of text readability in 371 literary texts extracted from Vietnamese textbooks primary school students and the literary textbooks for middle and high school students in Vietnam. Some machine learning algorithms for automatic text classification like Decision Tree, K-nearest neighbor, Support Vector Machines, *etc.* were used to classify the texts.

The experimental results presented in Table 3 and Table 4 show that some POS features such as DNr/Dwo, Cm/Wo, Nq/Wo, or Aa/Sen also contribute to the efficiency of classification. Comparing the results to the Luong 2017 results we can conclude that, he feature set (DNr/DWo, Nq/Wo, DAa/Sen, Cm/Wo), and the feature PSVW help increase precision value with SVM classifier in case of the group-by-grade-level corpus. On the other hand, the case of the group-by-school-level corpus, the feature set (Aa/Sen, DPp/Wo) helped the classification process to achieve the highest results for all measurements.

Experiments in this study only used those machine learning classification algorithms that assess whether a feature is valuable for the classification or not. For that reason it is not possible to discuss the potential influence that increasing or decreasing the use of a certain POS would have on the difficulty of the text. Such studies on the correlation of the extracted features with the text readability level are planned to be conducted in the upcoming investigations.

For the future works, we will proceed to collect additional corpora on different domains to look for features that could be useful for evaluating the readability of texts in the responding domains. Deeper features such as sentence-level grammar (syntax, coherence, cohesion, and others) should also be surveyed to find a better combination of features for assessing the readability of Vietnamese texts.

References

Al-Tamimi, A. K., Jaradat, M., Aljarrah, N., & Ghanim, S. (2014). AARI: Automatic Arabic readability index. *International Arab Journal of Information Technology, 11*(4), 370-378.

Alves, M. J. (2009). Loanwords in Vietnamese. *Loanwords in the world's language: A Comparative Handbook*, 617-637.

Brown, J. D., Janssen, G., Trace, J., & Kozhevnikova, L. (2012). A preliminary study of cloze procedure as a tool for estimating English readability for Russian students. In *Second Language Studies Paper* (pp. 1-22): University of Hawai'i at Manoa.

Carver, R. P. (1976). Word Length, Prose Difficulty, and Reading Rate. *Journal of Reading Behavior, 8*(2), 193-203.

Chall, J. S., & Dale, E. (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Northampton, Massachusetts: Brookline Books.

Chen, X., & Meurers, D. (2018). Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading, 41*(3), 486-510.

Chen, Y.-T., Chen, Y.-H., & Cheng, Y.-C. (2013). Assessing Chinese Readability using Term Frequency and Lexical Chain. *IJCLCLP, 18*(2), 1-18.

Coco, L., Colina, S., Atcherson, S. R., & Marrone, N. (2017). Readability Level of Spanish-Language Patient-Reported Outcome Measures in Audiology and Otolaryngology. *American journal of audiology, 26*(3), 309-317. doi:10.1044/2017_AJA-17-0018

Collins-Thompson, K., & Callan, J. (2005). Predicting Reading Difficulty with Statistical Language Models. *J. Am. Soc. Inf. Sci. Technol., 56*(13), 1448-1462.

Dale, E., & Chall, J. S. (1948). A Formula for Predicting Readability. *Educational Research Bulletin*, 11-28.

DeFrancis, J. (1977). Colonialism and language policy in Viet Nam. The Hague: Mouton.

Dinh, D., Nguyen, T. N., & Ho, H. T. (2018). Building a corpus-based frequency dictionary of Vietnamese. In (pp. 72-98).

Nguyễn Điệp T. N., Lươnga.-V., & Đinh Điền. (2019). Affection of the part of speech elements in Vietnamese text readability. *Acta Linguistica Asiatica*, *9*(1), 105-118. https://doi.org/10.4312/ala.9.1.105-118

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. e., mie. (2010). *A Comparison of Features for Automatic Readability Assessment.* Paper presented at the Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Stroudsburg, PA, USA.

François, T., & Fairon, C. (2012). *An AI readability formula for French as a foreign language.* Paper presented at the Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

Gunning, R. (1952). *The technique of clear writing*. New York: McGraw-Hill Book Co.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007, April). *Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts.* Paper presented at the Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, Rochester, New York.

Parker, R. I., Hasbrouck, J. E., & Weaver, L. R. (2001). Spanish readability formulas for elementary-level texts: A validation study. *Reading & Writing Quarterly*, 17(4), 307-322. doi:10.1080/105735601317095052

Jiang, Z., Sun, G., Gu, Q., Yu, L., & Chen, D. (2015). *An Extended Graph-Based Label Propagation Method for Readability Assessment.* Paper presented at the Web Technologies and Applications, Cham.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Technical Training, Research B*(February), 49.

Luong, A.-V., Nguyen, D., & Dinh, D. (2018a, November). *A New Formula for Vietnamese Text Readability Assessment.* Paper presented at the 2018 10th International Conference on Knowledge and Systems Engineering (KSE).

Luong, A.-V., Nguyen, D., & Dinh, D. (2018b, November). *Assessing the Readability of Literary Texts in Vietnamese Textbooks.* Paper presented at the 2018 5th NAFOSTED Conference on Information and Computer Science (NICS).

Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of Reading, 12*(8), 639-646.

Nguyen, L. T., & Henkin, A. B. (1982). A Readability Formula for Vietnamese. *Journal of Reading, 26*(3), 243-251.

Nguyen, L. T., & Henkin, A. B. (1985). A Second Generation Readability Formula for Vietnamese. *Journal of Reading, 29*(3), 219-225.

Pitler, E., & Nenkova, A. (2008). *Revisiting readability: A unified framework for predicting text quality.* Paper presented at the Proceedings of the conference on empirical methods in natural language processing.

Saddiki, H., Bouzoubaa, K., & Cavalli-Sforza, V. (2015). *Text readability for Arabic as a foreign language.* Paper presented at the Computer Systems and Applications (AICCSA), 2015 IEEE/ACS 12th International Conference of.

Saddiki, H., Habash, N., Cavalli-Sforza, V., & Al Khalil, M. (2018, July). *Feature Optimization for Predicting Readability of Arabic L1 and L2.* Paper presented at the Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, Melbourne, Australia.

Schwarm, S. E., & Ostendorf, M. (2005). *Reading Level Assessment Using Support Vector Machines and Statistical Language Models.* Paper presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA.

Sherman, L. A. (1893). *Analytics of literature: a manual for the objective study of English prose and poetry*. Boston, England: Ginn.

Si, L., & Callan, J. (2001). *A Statistical Model for Scientific Readability.* Paper presented at the Proceedings of the Tenth International Conference on Information and Knowledge Management, New York, NY, USA.

Spaulding, S. (1956). A Spanish Readability Formula. *The Modern Language Journal, 40*(8), 433-441. doi:10.1111/j.1540-4781.1956.tb02145.x

Tran, T., Pham, T., Ngo, H., Dien, D., & Collier, N. (2007). Named Entity Recognition in Vietnamese documents. *Progress in Informatics*, 5-13. doi:10.2201/NiiPi.2007.4.2

Al-Ajlan, A. A., Al-Khalifa, H. S., & Al-Salman, A. S. (2008, November). *Towards the development of an automatic readability measurements for arabic language.* Paper presented at the 2008 Third International Conference on Digital Information Management, University of East London, London, UK.

Vajjala, S., & Meurers, D. (2012, June). *On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition.* Paper presented at the Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, Montréal, Canada.

Sun, G., Jiang, Z., Gu, Q., & Chen, D. (2014, September). *Linear model incorporating feature ranking for Chinese documents readability.* Paper presented at the The 9th International Symposium on Chinese Spoken Language Processing, Singapore.

François, T. (2014, November). *An analysis of a French as a Foreign Language Corpus for Readability Assessment.* Paper presented at the Proceedings of the third workshop on NLP for computer-assisted language learning, Uppsala, Sweden.

Wang, S., & Andersen, E. (2016, December). *Grammatical Templates: Improving Text Difficulty Evaluation for Language Learners.* Paper presented at the Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan.

Luong, A.-V., Nguyen, D., & Dinh, D. (2017, October). *Examining the text-length factor in evaluating the readability of literary texts in Vietnamese textbooks.* Paper presented at the 2017 9th International Conference on Knowledge and Systems Engineering (KSE).

Al Khalil, M., Saddiki, H., Habash, N., & Alfalasi, L. (2018, May). *A Leveled Reading Corpus of Modern Standard Arabic.* Paper presented at the Proceedings of the 11th Language Resources and Evaluation Conference, Miyazaki, Japan.

Jiang, Z., Gu, Q., Yin, Y., & Chen, D. (2018, August). *Enriching Word Embeddings with Domain Knowledge for Readability Assessment.* Paper presented at the Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA.

Luong, A.-V., & Tran, P. (2019, November). *Assessing the Readability of Vietnamese Texts Through Comparison.* Paper presented at the Computational Data and Social Networks, Cham.