Acta Linguistica Asiatica

TABLE OF CONTENTS

Foreword

The linguistic clumsiness of tourists and students might be the price we pay for the linguistic genius we displayed as babies, just as the decrepitude of age in the price we pay for the vigor of youth.

Steven Pinker

…, however, from the viewpoint of a linguist, it is definitely worth having it all.

The articles for the summer 2022 issue mainly involve topics either on second language learning and acquisition or historical language changes and the motives for them. They were carefully picked up from numerous proposals, and we are very grateful to every single contributor and also to the reviewers.

This issue opens with the article "Examples of Corpus Data Visualization: Collocations in Chinese" in which **Ľuboš GAJDOŠ** and **Elena GAJDOŠOVÁ** lightheartedly share a highly beneficial practical procedure that can be used in the visualization of language data, especially in language pedagogy.

In a very similar manner to the visualization, **DENG Qi** presents a tangible example from Japanese in the article "Choice Between the Synonymous Pairs of *Sutoppu* and *Teishi*: A Case Study on Synonyms of Western Loanwords and Sino-Japanese in Modern Japanese Based on Corpus", discussing their usage and functions.

Yet another article "The Roman Alphabet Within the Japanese Writing System: Patterns of Usages and Their Significance" by **Hironori NISHI** explores the usages of the Roman alphabet within the present writing system of Japanese, which is, as the author suggests, induced by more and more frequent horizontal writing and the ever-increasing international interaction.

The following article was written by **GUO Qingli** and **CHEW Fong Peng** and is entitled "*Liushu*-based Instruction and Its Effects on the Motivation and Intended Learning Efforts: The Case of Laos Learners of Standard Chinese". It introduces the *Liushu*-based instruction and examines its effects on the students' motivation and intended learning efforts.

LIU Sha wrote the article "Exceptions vs. Non-exceptions in Sound Changes: Morphological Condition and Frequency", in which the author tries a unique approach to locate factors that explain exceptions in the diphthongization of [i] to [ei] in Mandarin.

Last but not least, **Nina GOLOB** in her article "Word-Prosodic Typology: The Traps of Seemingly Similar Japanese and Slovene" offers a brief review of research trends on prosody, and by introducing the phonetic properties of the two languages and acquisition difficulties by Slovene speakers of Japanese questions the typological similarity between Japanese and Slovene.

Editors and Editorial board wish the regular and new readers of the ALA journal a pleasant read full of inspiration, and a rise of new research ideas inspired by these papers.

Editors

RESEARCH ARTICLES

Examples of Corpus Data Visualization: Collocations in Chinese

Ľuboš GAJDOŠ

Comenius University in Bratislava, Slovakia
lubos.gajdos@uniba.sk

Elena GAJDOŠOVÁ

Comenius University in Bratislava, Slovakia
gajdosova137@uniba.sk

Abstract

The article aims to show a practical procedure that can be used in the visualization of language data. The paper freely follows our previous articles about the visualization of language data in language pedagogy. We demonstrate how to retrieve language data – in our case from corpora, how to edit data in a spreadsheet program, and then in the last step, how to visualize it on the example of Legal Chinese and partly Legal German. The Javascript library Vis.js via Pyvis is chosen for the visualization of the language data.

Keywords: visualization, corpus, Chinese, Javascript library Vis.js, Python

Povzetek

Namen članka je predstaviti postopek vizualizacije jezikovnih podatkov. S tem se navezujemo na dosedanje prispevke o vizualizaciji jezikovnih podatkov pri poučevanju jezika. V članku najprej prikažemo, kako poteka pridobivanje jezikovnih podatkov, kar so v našem primeru korpusi besedil. Nato prikažemo, kako podatke urejamo z orodjem za delo s preglednicami, nazadnje pa se osredotočimo na vizualizacijo podatkov, za kar smo uporabili primer pravne kitajščine in deloma pravne nemščine. Za vizualizacijo jezikovnih podatkov smo uporabili dinamično knjižnico Vis.js in modul Pyvis v Pythonu.

Ključne besede: vizualizacija, korpus, kitajščina, dinamična knjižnica Vis.js, Python

1   Introduction

Digitized language data can be obtained from several sources. In this article, we chose a corpus as a source of data, though data can also be obtained relatively easily from other sources. A corpus or corpus linguistics brings some benefits when compared to other sources.[1] Let us mention a few major benefits: (1) language data are already preprocessed, which means that texts are cleaned, (2) very often texts are tokenized (divided into tokens), (3) texts or tokens are annotated, and (4) corpora are equipped with statistical tools.

2   Methods

As with the data sources, there are many ways to retrieve data from a corpus in general.[2] In this chapter, only a basic procedure is introduced. Followed by the preliminary stage of the visualization – data retrieving and data editing, the visualization of data via Vis.js[3] library[4] is also shown.

2.1   Retrieving data

The most conventional way to obtain data from a corpus is to use a built-in corpus manager and, for example, Corpus Query Language (hereafter CQL). As our goal is to find the most "common" collocations in a corpus, there are probably other, more appropriate, or perhaps easier ways to tackle this task.

As confirmed by our previous research on the identification of collocations, language data which are obtained directly from the built-in functionalities offered by the NoSketch Engine is advantageous. Both corpora – the subcorpus *Zh-law* of the Hanku[5] corpus and the corpus COLEGE,[6] which serve as the data source, use the NoSketch Engine. Because there are many queries for the identification of collocations (only some of the parameters change), it is convenient to use scripting languages. In

---

[1] Corpus linguistics is a well-established part of linguistics. Despite some existing methodology issues, results from corpus linguistics have already proven that some areas of linguistics and language pedagogy are unimaginable without it, such as lexicography or quantitative analyses, just to mention a few. For more details on using the corpus data in language pedagogy, see Petrovčič (2022, pp. 43–47).

[2] See Gajdoš (2022b) or Gajdoš (2020) for details.

[3] Vis.js is a dynamic, browser-based visualization (JavaScript) library (*Vis.js*, n.d.).

[4] As the data retrieving script is written in Python, Pyvis library should be used to produce Javascript code. Pyvis is designed as a wrapper around the popular Javascript Vis.js library (*Pyvis*, n.d.).

[5] The Hanku corpus is a corpus of the Chinese Language. The zh-law is a subcorpus of legal Chinese (Gajdoš et al., 2016).

[6] The COLEGE is a corpus of legal German. See Gajdošová & Gajdoš (2018) for details.

our case, by retrieving language data from corpora, programming languages like Python[7] and the JSON format[8] are used, however, the following procedure describes a "manual" way, and it may also be used.

1. Get keywords (hereafter KWIC)[9] from the corpus by using query, e.g. CQL:[10] [tag="VV|NN"], then use node forms functionality, and finally save as .txt format.[11]

2. Again, based on our previous research, it is advantageous to look for collocations on the right and left sides separately.[12] LogDice score is chosen as the basic criterion (the measure of association), and the span is set to 5 to the left and 5 to the right side from KWIC for Chinese, and 10 to the left and 10 to the right for German.[13] The results are sorted by node forms and frequency for Chinese, and then lemma for German. The results are saved as a .txt file. As spreadsheet programs may not work properly with .txt formats, it is advisable to simply change the .txt filename extension (a suffix to the name) to .csv.[14]

2.2   Data editing

In this step, the data will get equipped with parameters and will further be formed in the way to best suit the visualization. For our purposes, the Pandas[15] library was used to modify the data. The following steps show some manual data modifications.

3. In a spreadsheet program,[16] the results from both sides are merged into a single file based on the LogDice score maintaining the side designation.[17]

---

[7] Python (version 3) is a programming language, for more information see *Python* (n.d.).

[8] JSON (JavaScript Object Notation) is a lightweight data-interchange format, for more information see *JSON* (n.d.).

[9] Keywords can be selected based on one's own requirements, e.g. only verbs and nous, as it is in our case (tags VV or NN for Chinese).

[10] The CQL query means "searching for all verbs VV or (symbol |) nouns NN". In the next step, the functionality *Node forms* sorts tokens by frequency.

[11] The most frequent tokens in a corpus do not necessarily create the strongest collocations (measured by the Logdice score).

[12] As will be shown in the following chapters, this information is used in the visualization.

[13] See Gajdošová (2022) or Gajdoš (2022a) for more information.

[14] Abbreviation CSV stands for "Comma-Separated Values". See *CSV* (n.d.) for details.

[15] See *PANDAS* (n.d.) for more details.

[16] Here it is worth noting that not all spreadsheet programs work properly with the CSV format. Libreoffice Calc has proven to be a good solution in this regard. For more details, see https://www.libreoffice.org/discover/calc/.

[17] The side designation may be added manually after obtaining data from a corpus for each side separately.

4. The previous steps are repeated for all the keywords. The results are then combined into one single .csv file (sorted by the LogDice score). This can be considered as the basis for visualization.

Table 1 below shows the results for Legal Chinese. POS tags of KWIC (column POS_kwic) and collocators (POS_item) are added manually at the end of the entire search. According to the requirements, it is possible to add other information that can be obtained from the corpus, such as the author's gender, period of origin, and others.

Table 1: Identified collocations in the corpus *Zh-law*

KWIC	Side	Item	Logdice	Corpus	POS_kwic	POS_item
部门	LS	主管	12,921	Zh-law	NN	NN
人民	RS	政府	12,914	Zh-law	NN	NN
不	RS	得	12,856	Zh-law	AD	VV
人民	LS	中华	12,803	Zh-law	NN	NR
人民	RS	共和国	12,789	Zh-law	NN	NN
部门	LS	行政	12,483	Zh-law	NN	NN
行政	RS	部门	12,455	Zh-law	NN	NN
人民	RS	法院	12,183	Zh-law	NN	NN
规定	LS	本	12,116	Zh-law	NN	DT

2.3   Visualization

Data visualization can be done in different ways. We here demonstrate only one of the possible ways of using Vis.js – network. The network in the Vis.js library mainly consists of nodes and edges. In our case, the keywords (KWIC) and identified collocators (item) are chosen as the nodes. The following code shows how to create a simple network with three nodes and two connections (edges) as an arrow based on the side (LS).

```
// create an array with nodes
var nodes = new vis.DataSet([
    { id: "部门", label: "部门" , shape: "dot", value: 2, group: 1},
    { id: "主管", label: "主管" , shape: "dot", value: 1, group: 2},
    { id: "行政", label: "行政" , shape: "dot", value: 1, group: 3},
]);
// create an array with edges
var edges = new vis.DataSet([
    { from: "主管", to: "部门" },
    { from: "行政", to: "部门" },
]);
var edges = [{ from: "部门", to: "主管", arrows: "from" },
{ from: "部门", to: "行政", arrows: "from" },
];
```

As can be seen from the code above, there are many parameters that can be used for nodes and edges. In our visualization, the following parameters are used – shape (dot), value (based on the number of connections/edges to nodes), and group (based on POS tag). Nodes can be dragged via a left mouse click. Figure 1 shows the visualization of a given code.
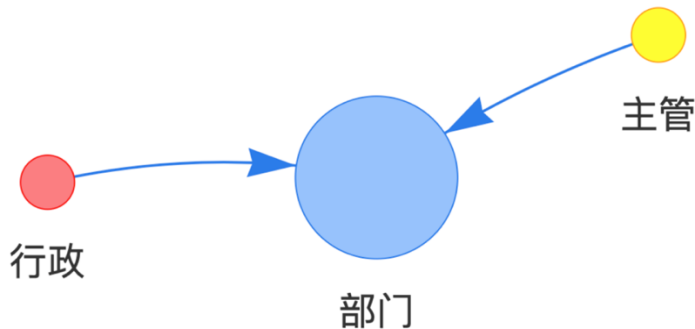
Figure 1: Example of creating a simple network in Vis.js

It is possible to create nodes and edges manually, yet with a large amount of data, this procedure is very laborious and time-consuming. For these reasons, it is very convenient to use one of the scripting languages, e.g. Python.[18]

---

[18] In the first step, the nodes are added within a for loop. In Python with Pandas are used, i.e. "for index, row in df.iterrows():". For more details, see https://pandas.pydata.org (*PANDAS*, n.d.).

The node ID is then the KWIC or item token. One KWIC or item (any token) is only one node. Edges are connected according to the side. The value of the node (the size) is based on the number of edges to other nodes. The result of the visualization is a .html file that may be displayed in a web browser such as Firefox, Chrome, or any other.

Because displaying a large amount of nodes (more than 1000) is computationally intensive, it is a good idea to choose a compromise number of nodes that are meaningful. It is appropriate to take this step when modifying/editing the data. Alternatively, select only some keywords (such as nouns and verbs) and then display them separately.

3   Practical use of visualization

Before showing a practical use of the visualization, is necessary to draw attention to the limits of the visualization:

- since the tagsets for corpora are different, one must expect a different proportion of POS tags in different corpora
- in the case of using the German corpus Colege, it is necessary to consider the relatively small size of the corpus, which may affect the result from the statistical point of view
- the visualization is only as relevant as the obtained data
- when comparing more languages, it is important to pay attention to genetic and typological differences between them.

3.1   Strongest collocations in Legal Chinese

The following example demonstrates possibilities of visualization in the Vis.js library. As already demonstrated, searching for the strongest collocation to one token (KWIC) in Legal Chinese, measured by the Logdice statistical measure, is quite simple. A recursive algorithm is used to search for the strongest collocations in the corpus.[19] The following POS tags are excluded from the search: punctuation (PU), time nouns (NT), numbers (CD, OD), sentence particles (SP) and all markers DE 的, 得, 地 (D.*).[20] Table 2 shows only a small portion of the result.[21] The whole table has 935 rows.

---

[19] The searching function is called itself until the Logdice value is below a certain value. For more details about recursion, see https://openbookproject.net/thinkcs/python/english3e/recursion.html

[20] Needless to say, these restrictions may be set arbitrary, and, in this case, these restrictions are applied to KWIC and collocators (Item).

[21] The very first result is particularly interesting. The formula for calculating Logdice shows that the maximum theoretical value is 14, yet this result is higher (Rýchly, 2008). The explanation is in this case quite simple – cooccurrence count (689) of tokens *yǒuqī* 有 期 (period) *túxíng* 徒 刑

Table 2: Strongest collocations in Legal Chinese

KWIC	Side	Item	Logdice	Corpus	POS_kwic	POS_item
徒刑	LS	有期	14,022	Zh-law	NN	JJ
中华	RS	共和国	13,979	Zh-law	NR	NN
共和国	LS	中华	13,979	Zh-law	NN	NR
自治区	RS	直辖市	13,895	Zh-law	NN	NN
直辖市	LS	自治区	13,895	Zh-law	NN	NN
徇私	RS	舞弊	13,894	Zh-law	VV	NN
舞弊	LS	徇私	13,894	Zh-law	NN	VV
款罪	LS	犯前	13,882	Zh-law	NN	VV
犯前	RS	款罪	13,882	Zh-law	VV	NN
有期	RS	徒刑	13,861	Zh-law	JJ	NN
自治区	LS	省	13,779	Zh-law	NN	NN
省	RS	自治区	13,779	Zh-law	NN	NN
省	RS	直辖市	13,733	Zh-law	NN	NN
直辖市	LS	省	13,733	Zh-law	NN	NN
交通管	RS	制员	13,601	Zh-law	NN	NN
刑事	LS	追究	13,596	Zh-law	NN	VV
追究	RS	刑事	13,596	Zh-law	VV	NN
撤职	LS	降级	13,484	Zh-law	VV	VV
降级	RS	撤职	13,484	Zh-law	VV	VV

(imprisonment; fixed-term imprisonment) is higher than candidate count (619) of 有期. This phenomenon may be caused by errors in tokenization.
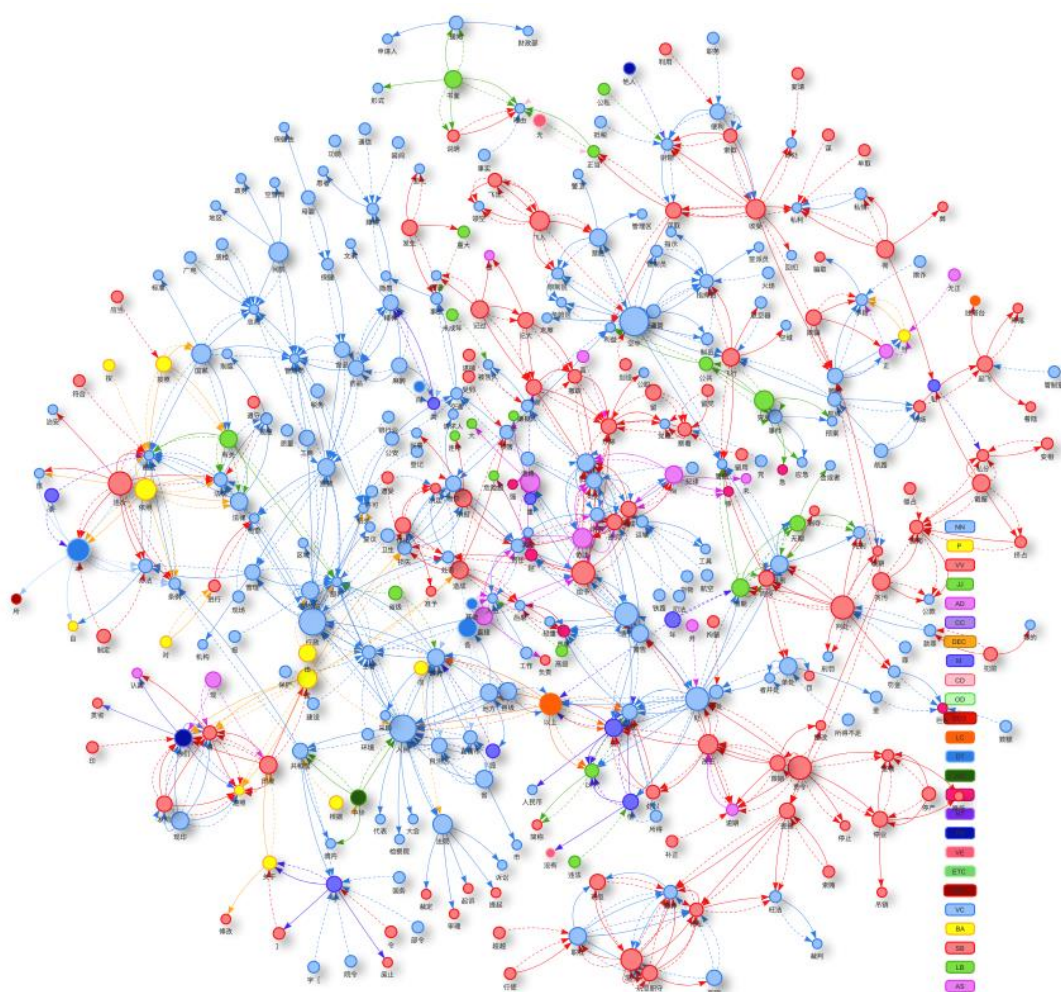
Figure 2: Network of 935 strongest collocations in Legal Chinese

As the printed version of the visualization is fairly limited, possibilities of the visualization are demonstrated in the multiple figures. Figure 3 shows the following:

- there are nodes (tokens) that make more connections than others, the size of a node reflects this information (mostly blue are nouns and red nodes are verbs)
- some parts of speech are more common than others, e.g. blue nodes (nouns)[22]
- clusters of blue nodes (nouns) and red (verbs) are the most common
- arrows point to the collocators (dashed edges to the collocators on the right side, solid-line edges to the collocators on the right side) in the word order.

---

[22] It is worth noting that due to polysemy or conversion (zero derivation), POS tags are not always adequate in a certain collocation in Chinese.

Though the above information is also available in the .csv file, we believe that their visualization is more easily readable to a student or in the field of second language acquisition in general. This is also because some information, as shown in Figure 3, can only be retrieved from the table via a search, while in the figure, this information is available by clicking on a node. Also, the figure can be zoomed in or out using the mouse.
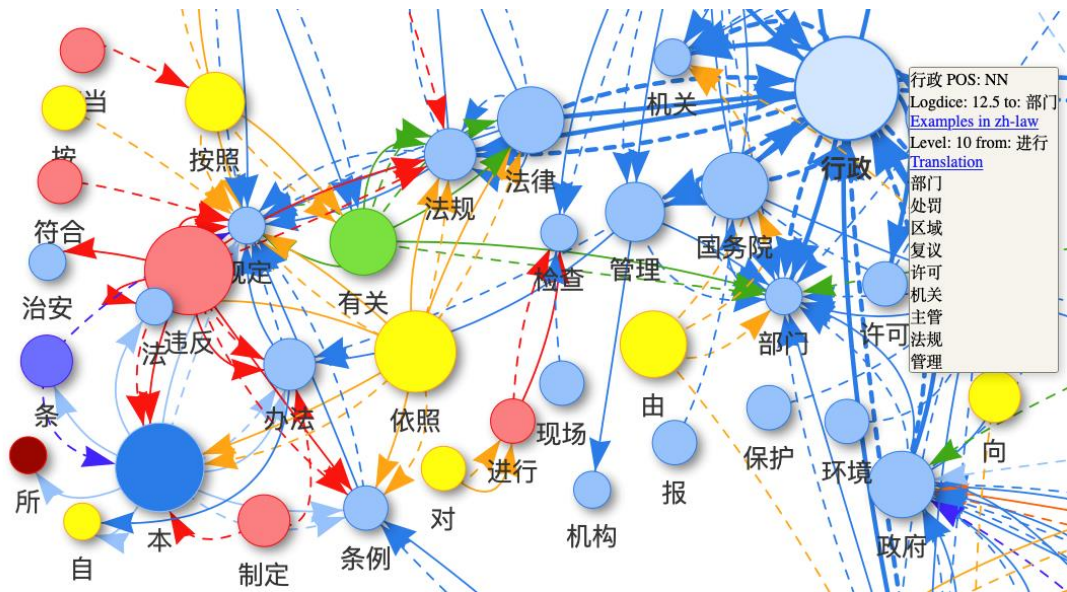
Figure 3: Zoom in and viewing connection to other tokens

After clicking on a node, the pop-up menu would show that the node *xíngzhèng* 行政 (administrative) is a noun (NN), which makes collocations with the following tokens in the column: *bùmén* 部门 (department), *chǔfá* 处罚 (punish), and others. At the same time, the edges are also highlighted. Furthermore, by clicking on the connected node, one can obtain information on the collocators to the connected node as well. The pop-up menu in Vis.js also offers other options, such as a hyperlink directly to examples in the corpus, to the web translator, statistical data, and others.

3.2   Comparison of synonyms

There are synonyms in every natural language and one of the difficulties in translating them or in the L2 acquisition is to find their equivalences. Our empirical experiences show that it is appropriate to seek an equivalence, not at the level of words (tokens) but at least at the level of collocations (bigrams, n-grams), as has already been shown by Benická (2017), and others.[23]

---

[23] For more details about translation issues, see Benická (2017).

Let us start with an example of the following three Chinese synonyms: "according to" *ànzhào* 按照，*gēnjù* 根据，*yīzhào* 依照. Table 3 shows the first 10 rows. The whole table consists of 193 rows.

Table 3: Collocations to given prepositions

KWIC	Side	Item	Logdice	Corpus	POS_kwic	POS_item
按照	RS	规定	11,477	Zh-law	P	NN
依照	RS	法	11,02	Zh-law	P	NN
根据	RS	需要	11,007	Zh-law	P	VV
依照	RS	法律	10,958	Zh-law	P	NN
按照	RS	有关	10,86	Zh-law	P	JJ
按照	LS	应当	10,809	Zh-law	P	VV
根据	RS	共和国	10,781	Zh-law	P	NN
根据	RS	中华	10,78	Zh-law	P	NR
依照	RS	规定	10,769	Zh-law	P	NN
依照	RS	条例	10,685	Zh-law	P	NN
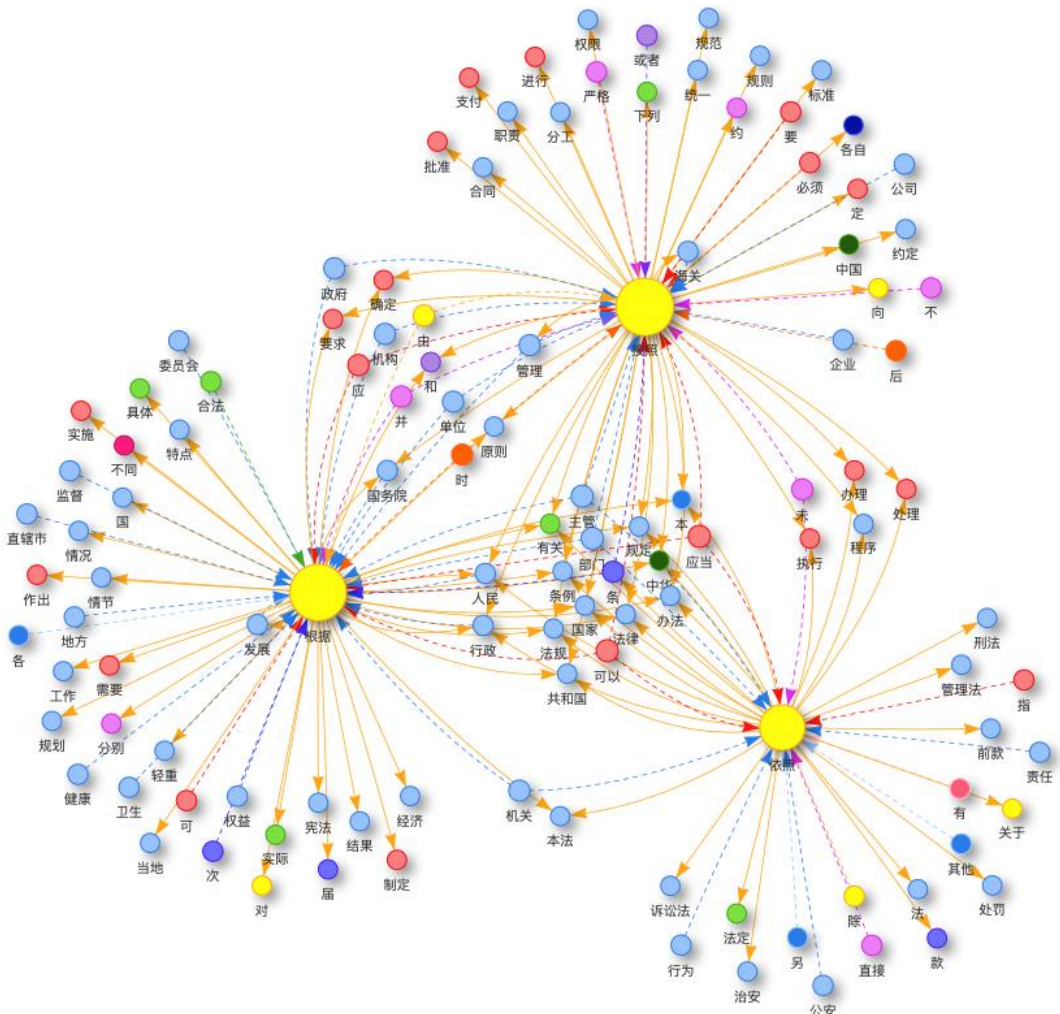
Figure 4: Collocators to prepositions *ànzhào* 按照, *gēnjù* 根据 and *yīzhào* 依照

As can be seen from Figure 4, these prepositions have mutual collocators, of which some are typical for two prepositions and some even for one preposition only. Let us zoom in to see more details.

On the left side of Figure 5, there is a group of *gēnjù* 根据 and *ànzhào* 按照 collocators. The group in the middle are collocators of all three prepositions. The group on the right side are the mutual collocators to *ànzhào* 按照 and *yīzhào* 依照.
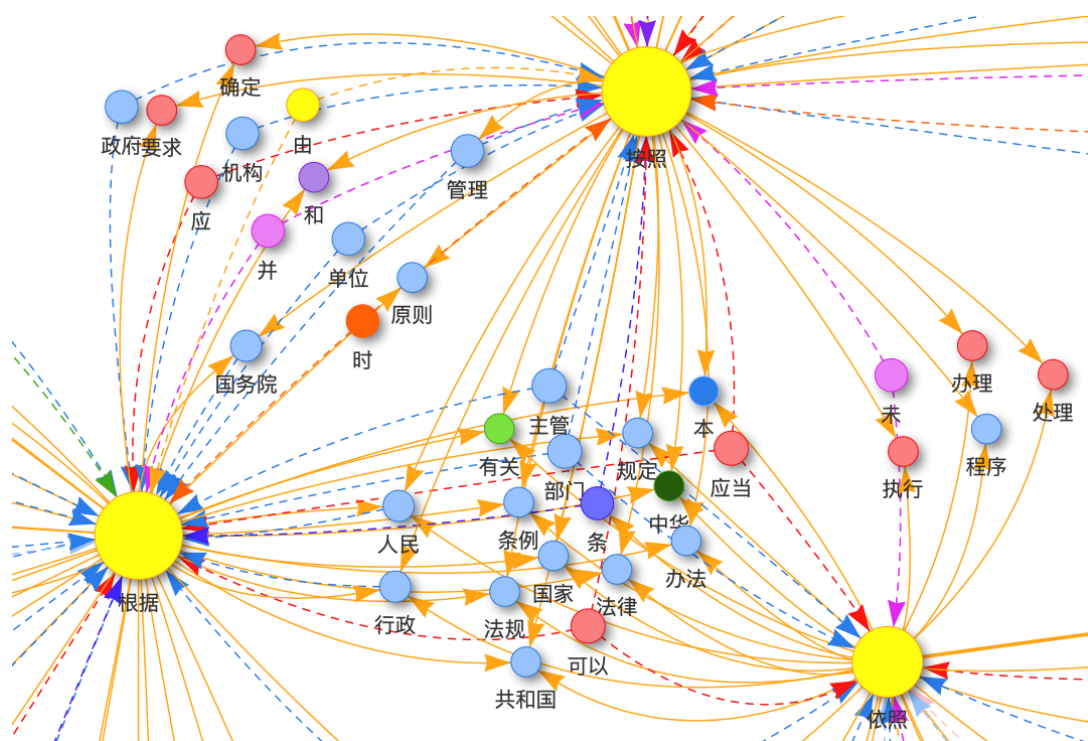
Figure 5: Mutual collocators of the prepositions

3.3   Comparison of modal verbs in legal texts

In the case of translating legal texts, for example, it is very important to find an equivalence in translated languages, in our case for modal verbs. Due to the ongoing research,[24] the language pair Chinese – German is chosen as an example.

There are no specific tags for modal verbs[25] in the Chinese corpus, so modal verbs must be selected manually as follows by the CQL query:

[word="要|能|会|可以|可|想|能够|需要|应|可能|必须|得|应该|需|应当|该|必|想要|愿意|能否|愿|须|得以" & tag="VV"]

Table 4 below shows the first 10 collocations only. The whole table has 984 rows.

---

[24] For more details, see Gajdošová (2022).

[25] Identifying modal verbs is not a trivial operation, since many modal verbs in Chinese may also be e.g. transitive verbs.

Table 4: Collocations of modal verbs

KWIC	Side	Item	Logdice	Corpus	POS_kwic	POS_item
得	LS	不	12,846	Zh-law	VV	AD
会	LS	保监	11,961	Zh-law	VV	VV
会	RS	同	11,601	Zh-law	VV	DT
会	LS	银监	11,55	Zh-law	VV	NN
必	RS	修课	11,415	Zh-law	VV	VV
可	LS	行政许	11,366	Zh-law	VV	NN
需	LS	确	11,342	Zh-law	VV	AD
可能	RS	影响	11,134	Zh-law	VV	NN
能	LS	只	11,068	Zh-law	VV	AD
可以	LS	也	11,018	Zh-law	VV	AD

As can be seen from Figure 6 below, some modal verbs collocate more often than others.
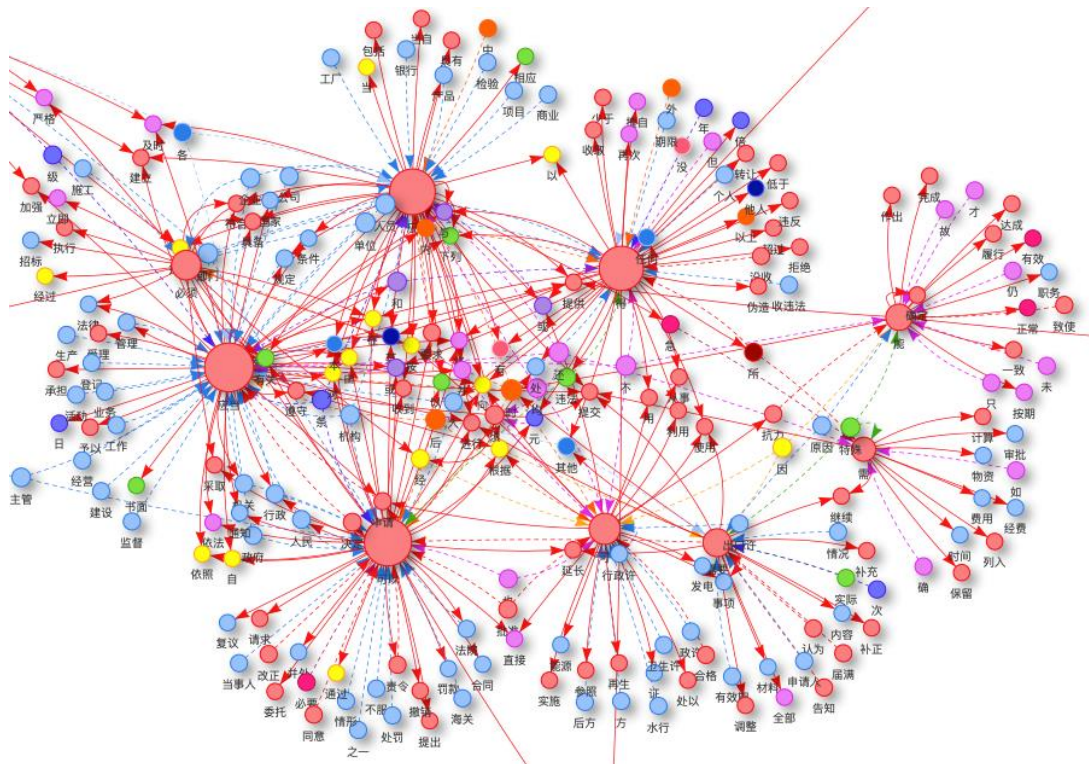
Figure 6: Collocability of modal verbs in Chinese

From the list of modal verbs, the token that collocates the most is the verb *yīngdāng* 应当 (should; must), *yīng* 应 (should), *kěyǐ* 可以 (may), *dé* 得 (must), and may express deontic modality. The collocability of other verbs is rather limited or they do not collocate at all. Such are *yuànyì* 愿意 (willing to), *yuàn* 愿 (willing to), *yīnggāi* 应该 (should), and others.

Let us zoom in. It is obvious that the modal verbs *yīngdāng* 应当 (should; must), *yīng* 应 (should), and *dé* 得 (must) have mutual collocators. On the other hand, some modal verbs such as *yīnggāi* 应该 (should) and *gāi* 该 (should) rarely collocate in Legal Chinese, if at all.
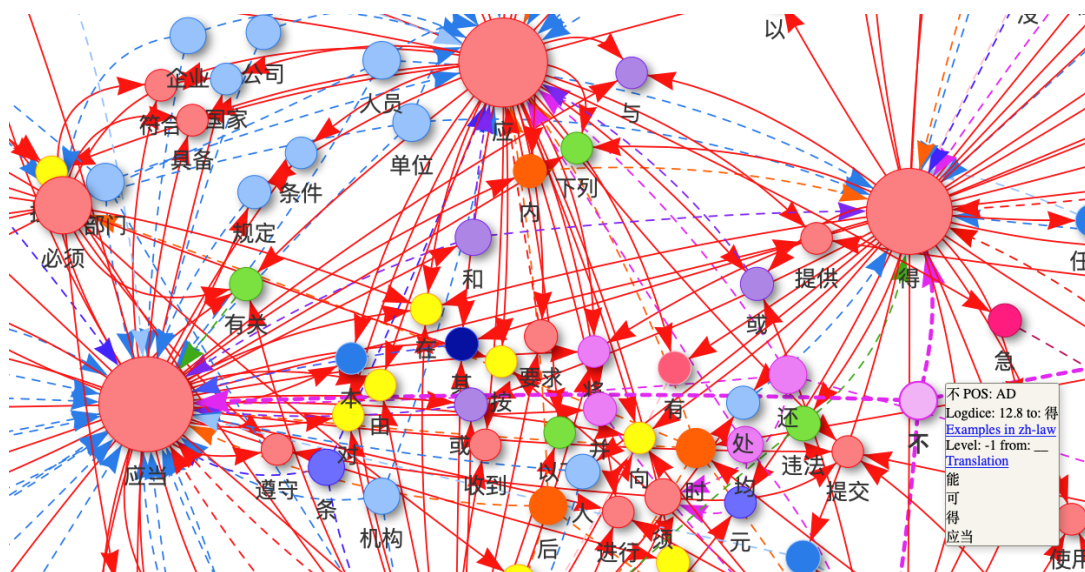
Figure 7: Collocability of verbs (from left) *yīngdāng* 应当, *yīng* 应 and *dé* 得

Now let us look at a practical example of using the network in L2 acquisition and select the node *bùmén* 部门 (department; division). This node points to the verb *yīngdāng* 应当 (must) and this verb, among others, points to the preposition *gēnjù* 根据 (according to), which further points to the verb *xūyào* 需要 (a need; to need). Searching for the above tokens in the Zh-law subcorpus brings the results presented in Figure 8.

Figure 8: Example of the n-gram in Legal Chinese

The visualization may not appear clear enough at first sight. If so, it is advantageous to reduce the number of collocators or modify the parameters of the physical model, e.g. change the value of constants (gravitationalConstant), extend the length of the edges (springLength), etc. though these are rather limitations resulting from the printed version. The .html version solves these problems partially or completely.

The downside of the visualization, as the parameters are set in this example (node size based on the number of collocators), is also the inability to display collocation strength based on the Logdice score. An example in this case is the negation *bù* 不 (not) which makes the strongest collocation of all modal verbs, in this case with *bù dé* 不得 (must not). Besides, the negation *bù* 不 can also collocate with other modal verbs in Legal Chinese.

Now, let us compare the above situation with the one in Legal German. Taking the same example, we would like to illustrate a different approach to the same problem in two different corpora and languages. This is not an exhaustive analysis of the issue of equivalence in these languages but just a sample of the possibilities.

As for the modal verbs in German, there is a dedicated tag for modal verbs in the Colege corpus and it may be identified very easily by the CQL:[26] [tag="VMFIN| VMINF"]

Table 5: Modal verbs in Legal German

lemma	Frequency	Items: 6 || Total frequency: 10,949
P | N können	7,034	
P | N dürfen	1,458	
P | N müssen	1,231	
P | N sollen	1,083	
P | N wollen	138	
P | N mögen	5	

---

[26] The VMFIN tag means a finite form of a modal verb, the symbol "|" OR and VMINF means an infinitive form of a modal verb.
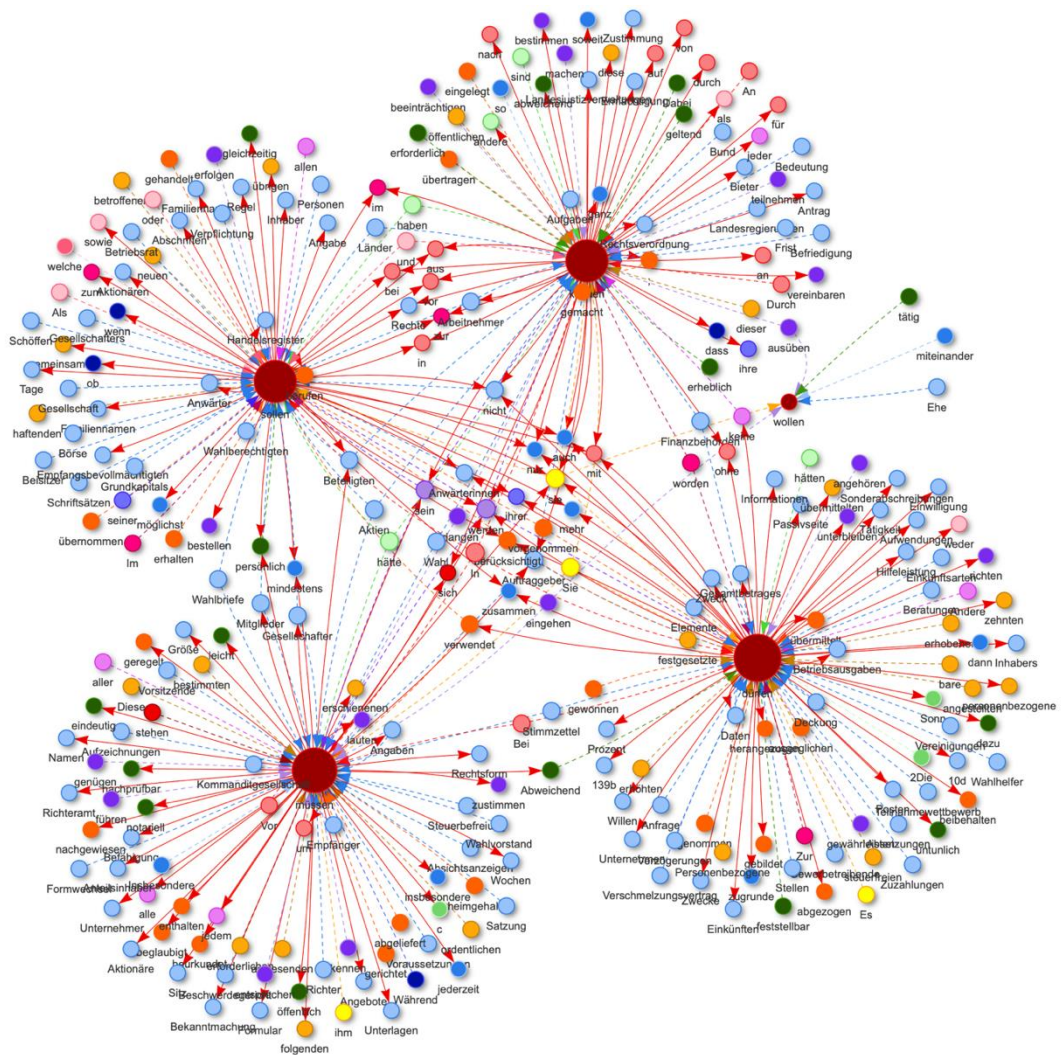
Figure 9: Modal verbs in Legal German
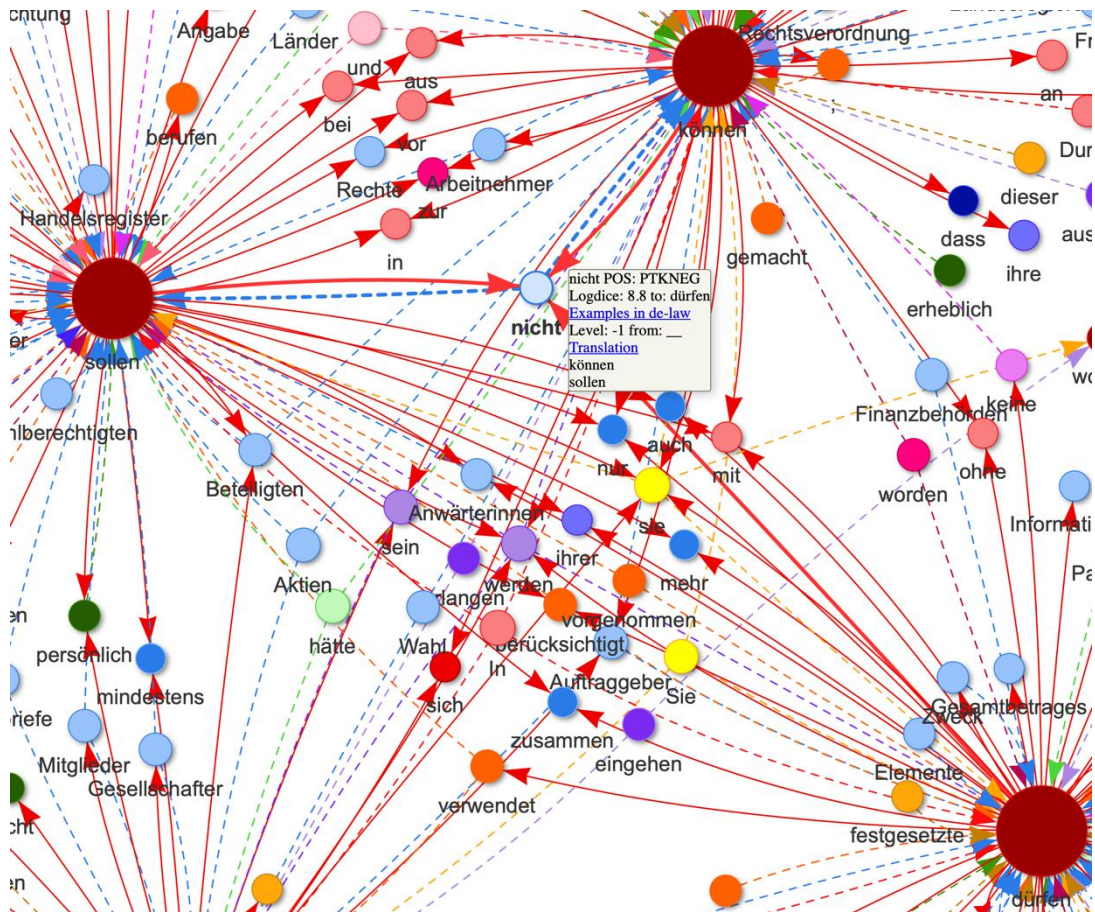
Figure 10: Modal verbs with negations in Legal German

As can be seen from the visualization, the modal verbs *sollen* (should, shall), *dürfen* (may), *können* (can) can be associated directly with the negation *nicht* (not) (similar to *keine* none), but the verb *müssen* (must) does not take a negation. Therefore, when comparing the situation in Legal German vs. Legal Chinese, it is clear that the collocational preferences at the negation are different.

4   Conclusion

In this article, we have briefly demonstrated the whole process of a language data visualization – from retrieving, via data editing to the visualization. We have shown some examples of the visualization.

In the end, there is still one question to answer: is this kind of visualization useful? We think that this type of visualization is primarily intended for the area of language acquisition. We believe that visualization can help get a basic overview of the examined

data – whether it is a view of the language register (as in our case), or to a literary work of one author, or else just to compare the collocability of individual words (tokens). With appropriately selected parameters and the number of nodes, it is possible to very clearly show typical features of a text, register, or the relation between words. To conclude, the visualization of linguistic data has its place in language data processing and sometimes provides a clearer insight into the issue.

References

Benická, J. (2017). Archaizujúci jazyk v čínštine a jeho prekladanie do slovenčiny. (Historicizing Language in Chinese and its Translation into Slovak). In D. Veverková, I. Kolečáni Lenčová, M. Ľupták & Z. Danihelová (Eds.), *Aplikované jazyky v univerzitnom kontexte IV* (pp. 58-68). Technická univerzita.

CSV. (n.d.). Retrieved March 15, 2022, from https://www.w3.org/TR/tabular-data-primer/#tabular-data

Gajdoš, Ľ., Garabík, R., & Benická, J. (2016). The New Chinese Webcorpus Hanku—Origin, Parameters, Usage. *Studia Orientalia Slovaca, 15*(2), 21-33.

Gajdoš, Ľ. (2020). Verb Collocations in Chinese – Retrieving, Visualization and Analysis of Corpus Data. *Studia Orientalia Slovaca, 19*(1), 121-138.

Gajdoš, Ľ. (2022a). Vizualizácia jazykových dát ako didaktická pomôcka na príklade korpusu čínskych právnych textov (Visualisation of Linguistic Data as a Didactic Tool on the Example of the Corpus of Legal Chinese). In *Kontexty súdneho prekladu* (pp. 7-22).

Gajdoš, Ľ. (2022b). *Praktická korpusová lingvistika – čínština* (Practical Corpus Linguistics – Chinese Language). Univerzita Komenského.

Gajdošová, E., & Gajdoš, Ľ. (2018). Korpus nemeckého právneho textu COLEGE (Corpus of Legal German). In *Kontexty súdneho prekladu a tlmočenia*, 7, 40-47.

Gajdošová, E. (2022). Korpusbasierte Analyse von Rechtstexten in slowakischer und deutscher Sprache mit besonderem Augenmerk auf Verb-Nomen-Kollokationen [Unpublished doctoral dissertation]. Univerzita Komenského.

JSON. (n.d.). Retrieved March 15, 2022, from https://www.json.org/json-en.html

PANDAS. (n.d.). Retrieved March 14, 2022, from https://pandas.pydata.org

Petrovčič, M. (2022). Chinese Idioms: Stepping Into L2 Student's Shoes. *Acta Linguistica Asiatica*, *12*(1), 37-58. https://doi.org/10.4312/ala.12.1.37-58

Python. (n.d.). Retrieved March 15, 2022, from https://www.python.org

Pyvis. (n.d.). Retrieved March 16, 2022, from https://pyvis.readthedocs.io/en/latest/#

Rýchly, P. (2008). A Lexicographer-Friendly Association Score. In P. Sojka & A. Horák (Eds.), *Recent Advances in Slavonic Natural Language Processing* (pp. 6-9). Masaryk University.

VIS.JS. (n.d.). Retrieved March 16, 2022, from https://almende.github.io/vis/docs/network/

Choice Between the Synonymous Pairs of *Sutoppu* and *Teishi*: A Case Study on Synonyms of Western Loanwords and Sino-Japanese in Modern Japanese Based on Corpus

DENG Qi

School of Foreign Studies, Northeastern University, China
dengqixq123@gmail.com

Abstract

This paper discusses the results of a corpus-based study on the usage and functions of the western loanword *sutoppu* and its synonymous Sino-Japanese, *teishi*. Our analyses focus on the following four perspectives: (1) frequency, (2) conjugation types, (3) characteristics of genres used, and (4) collocations. The results show that *sutoppu* is used mostly in a causative form, implying something compulsory or intentional, whereas *teishi* is mostly used in its passive form to imply inevitability. In addition, *sutoppu* emphasizes instantaneity and has the meaning of intentionally stopping something with great resistance. Whereas *teishi* permits a certain duration of time and describes the state of being stopped.

Keywords: western loanwords, Sino-Japanese words, *sutoppu*, *teishi*, synonyms

Povzetek

Članek obravnava rezultate korpusne študije o uporabi in vlogah novejše prevzete besede *sutoppu* in njene sino-japonske sopomenke *teishi*. Analize se osredotoča na naslednje štiri vidike: (1) pogostnost, (2) vrste spreganja, (3) značilnosti uporabljenih uporabljeni žanrov in (4) kolokacije. Rezultati kažejo, da se *sutoppu* večinoma uporablja v vzročni obliki, kar nakazuje na obvezno ali namerno dejanje, medtem ko se *teishi* večinoma uporablja v pasivni obliki in s tem izraža neizogibnost. Poleg tega *teishi* dovoljuje določen čas izvedbe oziroma opisuje stanje ustavljanja, *sutoppu* pa poudarja trenutnost in ima pomen namernega ustavljanja nečesa z velikim odporom.

Ključne besede: novejše prevzete besede, sino-japonske besede, *sutoppu*, *teishi*, sinonimi

1   Introduction

Japanese vocabulary can be classified into four lexicon strata according to its origin: native (or Yamato), western loanwords, Sino-Japanese words, and hybrid vocabulary (Sugimoto & Iwabuchi, 1994; Ito & Mester, 1999; Nihongogakkai, 2018). Among the aforementioned four strata of Japanese, western loanwords and Sino-Japanese can function as verbs by adding *suru* to the noun. It should be noted that there are many synonymous pairs between *suru*-verbs in western loanwords and Sino-Japanese words, and their proper usage is one of the major difficulties for learners of Japanese.

The reason why learners feel it difficult to learn the distinctive usage of the synonymous *suru*-verb in western loanwords and Sino-Japanese can be summarized as follows: (1) western loanwords are one of the most difficult strata to learn, and the acquisition of western loanwords is accompanied by various difficulties (Jinnai, 2008, Yamasita et al., 2018). (2) Moreover, it has been pointed out that there is no description in dictionaries on how to use them correctly (Yamashita et al., 2018). Although there have been some case studies on the criteria for the use of synonymous *suru*-verb pairs in western loanwords and Sino-Japanese, it is clear that there is still a lack of basic information to help learners distinguish the difference in the usage (Mogi, 2015).

To address the above-mentioned problems, in this study, we investigate the semantic differences between *suru*-verb western loanwords and Sino-Japanese, by taking the pair of *stoppu* and *teishi* as an example. Regarding the selection of the target words in this study, Yamashita et al. (2018) took up 9 pairs of synonymous *suru*-verb in western loanwords and Sino-Japanese and conducted an awareness survey among 110 native Japanese speakers. Yamashita's study revealed that 7 of the 9 pairs can be classified into three major types, while the difference between 2 pairs (ストップ *stoppu* and 停止 *teishi*, テスト *tesuto* and 試験 *shiken*) cannot be well explained. However, Yamashita (2018) focused his investigation on the introspection of native speakers. Since a corpus-based study is considered valid when the introspection of native speakers does not work (Ishikawa, 2012), this study will attempt to elucidate the usage of ストップ *sutoppu* (hereinafter referred to as *sutoppu*) and 停止 *teishi* (hereinafter referred to as *teishi*) as an example, to cover the first limitation.

2   Literature review

The study of synonyms has been a fruitful area of corpus linguistics (Gries & Otani, 2010), which is discussed in many introductory books on corpus linguistics, such as Kennedy (1998), McEnery & Wilson (2001), etc. Corpus data can not only provide insights into naturally occurring language (Sinclair, 1991) but are also regarded as an effective tool to distinguish the differences between synonyms (Biber et al., 1998; Hunston, 2002; Moon, 2010). Besides the works which have investigated the

differences between synonyms using corpora in English (Biber et.al, 1996; Liu, 2010; Chuang, 2011; etc.), many studies have compared the semantic functions of synonyms in Japanese (Sugimoto, 2009; Shinya, 2010; Zhao, 2013; etc.).

Regarding the works that have compared the semantic functions of synonymous western loanwords and Sino-Japanese, for example, Miyata & Tanaka (2006) took up the western loanword リスク *risuku* and compared it with its synonyms 危険 *kiken*, 危険性 *kikensei* (all of which mean "a risk") by using a newspaper database (*Asahi*, *Mainichi*, and *Yomiuri* newspaper from 2003 and 2004). Miyata (2007) compared メリット *meritto* and its synonym 利点 *riten* (both of which mean "a merit") also by using a newspaper database (*Asahi*, *Mainichi*, and *Yomiuri* newspapers from 2003, 2004, and 2005). Sato (2013) compared ムード *mūdo* with 雰囲気 *fun'iki* (both of which mean "atmosphere/mood") using a newspaper database (*Asahi* newspaper) and clarified the semantic features of each.

However, there are not so many studies that have compared synonymous *suru*-verb western loanwords and Sino-Japanese. The following is an overview of studies on synonymous pairs between *suru*-verb western loanwords and Sino-Japanese from the following two perspectives: the qualitative survey on the awareness of native Japanese speakers, and quantitative surveys of the corpus.

Regarding the awareness survey, Yamashita et al. (2018) conducted a survey among 110 native Japanese speakers (teachers and students) to clarify the differences between synonymous *suru*-verbs as either western loanwords or Sino-Japanese. The task was two-fold: the first was sentence production, in which the participants were asked to write down sentences they could think of, using each of the synonyms, and the second was freewriting about their findings on the differences in the usages of the same synonyms. Results revealed that 7 of the 9 pairs are classified into three major types: (1) differences are seen in terms of the semantic nuance (e.g. 開始する *kaishi-suru*/ スタートする *sutāto-suru* [begin]), B(2) differences are seen in terms of the number of senses (e.g. 点検する *tenken-suru*/ チェックする *chekku-suru* [investigate, check]), and (3) differences are seen in terms of the range of use (e.g. 練習する *renshū-suru*/ トレーニングする *toreningu-suru* [train, have a physical practice]). However, the difference between the two pairs (ストップ *stoppu* and 停止 *teishi*, テスト *tesuto* and 試験 *shiken*) could not be well explained.

About the quantitative surveys, Chen (2014) used a newspaper database to clarify the semantic frames of western loanwords, ケア *kea*, and Sino-Japanese, 介護 *kaigo* (both of which mean "care") by classifying the co-occurrences of these two words. Mogi (2015) conducted a survey using the *Balanced Corpus of Contemporary Written Japanese* to compare the similarities and differences between マークする *māku-suru* and 記録する *kiroku-suru* (all of which mean "to mark") from the perspective of feature genres and co-occurring objects.

Although these studies have contributed a lot to a better description of the differences between synonymous western loanwords and Sino-Japanese, there is still room for improvement both in the database used as well as in the perspectives investigated. Regarding the database, many studies use relatively small databases, especially newspaper databases (Miyata et al., 2006; Miyata, 2007; Sato, 2013; Chen, 2014). As for the perspectives, there are many studies that focus on the frequency of appearances and collocations (Miyata et al., 2006; Miyata, 2007; Chen, 2014; Mogi, 2015), while only a few studies focus on the parts-of-speech (conjugation types) and characteristically used genres. Frequency and collocations may be considered a useful measure in identifying the differences between synonyms (Biber et.al, 1998; Evison, 2010; Aroonmanakun, 2015), and the preferable register and/or part-of-speech in which the words appear may also help students to understand the differences (Shaw, 2011; Phoocharoensil, 2020). Therefore, this study takes up *stoppu* and *teishi* as an example to clarify the differences in semantic properties from the following four perspectives: (1) frequency of appearance, (2) parts of speech and conjugation types, (3) characteristically used genres, and (4) collocations.

3   Aims and methodology

This section outlines the research questions, data, and methodology used in this study. We give an overview of the definitions of the two words in several dictionaries in Section 3.1 and set out research questions in Section 3.2. The compilation of the corpora is delineated in Section 3.3. The methodology employed to address the research questions is described in Section 3.4.

3.1   Dictionary descriptions of *sutoppu* and *teishi*

Prior to conducting the survey, we first refer to the dictionary definitions of *sutoppu* and *teishi* in four kinds of Japanese-Japanese (JJ) and Japanese-English (JE) dictionaries. The definitions in JJ are translated into English by the author. Definitions related to technical terms are excluded.

In these definitions, both *sutoppu* and *teishi* mean to stop and to desist, and there is no specific information on the conjugation types or genres in which words are likely to be used. Therefore it is considered to be highly difficult for learners to understand the specifics of usages of the two words by looking up a dictionary. To better distinguish these two words according to their actual usages, further investigation on how these two words are used is needed.

Table 1: Definition of *sutoppu* and *teishi* in dictionaries

Source	*stoppu*	*teishi*
JJ: 大辞林 *Daijirin*	（名）する。(1)とまること。やめること。(2)とまれの信号。 *(mei)suru.* (1) *Tomaru koto. Yameru koto.* (2) *[tomare] no shingo.* (1) to stop. (2) stop signal.	（名）する。(1)動いていたものがとまること。また、とめること。(2)していたことをやめること。また、やめさせること。 *(mei) suru.* (1) *Ugoiteita mono ga tomaru koto. Mata, tomeru koto.* (2) *Shiteita koto wo yameru koto. Mata, yame saseru koto.* (1) To stop something that was moving. (2) To stop doing what you were doing, or to make someone stop.
JJ: 新明解 *Shin-Meikai*	(1)［ーする］止まること。(2)止まれ（のしるし） (1) [*-suru*] *tomaru koto.* (2) *tomare* (*no shirushi*). (1) to stop. (2) a signal sign to stop.)	［ーする］(1)［移動しているものが］中途で止まること。(2)活動をやめ（させ）ること。 [*-suru*] (1) [*Idou shiteiru mono ga*] *chuuto de tomaru koto.* (2) *Katsudou wo yame (sase) ru koto.* (1) To stop a moving object. (2) To stop an activity.
JE: *Geneus*	n. stop. v.stop, halt.	n. (1) a stop; (2) suspension. v. (1) stop; (2) come to a stop (halt); (3) suspend.
JE: *Wisdom*	n. a stop. v. stop.	n. (1) (a) stoppage, a stop; (2) (a) suspension. v. (1) stop, pause; (2) cease; (3) suspend.

3.2   Aims and a research question

This study will elucidate the use of *suru*-noun western loanwords and Sino-Japanese, taking *sutoppu* and *teishi* as an example, for which the differences in the meaning are sometimes unknown even to Japanese native speakers. Although there are many linguistic perspectives regarding their semantic functions, this study elucidates the distinction between *sutoppu* and *teishi* from four perspectives: (1) frequency, (2) conjugation types, (3) characteristically used genres, (4) collocations. Specifically, we set up the following 4 research questions (RQs).

1.   What are the differences in the frequency of use?
2.   What are the differences in conjugation types?
3.   What are the differences in the genres used?
4.   What are the differences in collocations?

To verify these research questions based on corpus data, we further developed hypotheses for each question to make the verification process more systematic.

1.   (H1) According to some previous studies, the frequency of western loanwords is higher than that of Sino-Japanese (e.g., Miyata (2007) (メリット *meritto* vs. 利点 *riten*: 7898 vs. 5125)), while others found the frequency of Sino-Japanese to be higher than western loanwords (e.g., Miyata (2006) (リスク *risuku* vs.危険 *kiken*: 4671 vs. 8091), Chen (2014) (ケア *kea* vs. 介護 *kaigo*: 1097 vs. 3766)). Since *sutoppu* and *teishi* are expected to be used more frequently in socio-economic contexts, such as ストップ高 *sutoppu daka* for stock prices and 緊急停止 *kinkyu teishi* in issues regarding nuclear power plants, they are closer to リスク *risuku*/ 危険 *kiken* and ケア *kea*/ 介護 *kaigo* described in the studies above, and because of this, we expect the frequency of Sino-Japanese to be higher.

2.   (H2) Regarding the conjugation types, in lexical descriptions, there was no clear difference in verbosity and intransitivity, although details of the glosses were slightly different. On this basis, it is expected that the part-of-speech and conjugation types of the two words will be equivalent.

3.   (H3) Regarding the characteristic genres in which they are used, it was pointed out that western loanwords are more likely to be used in a more casual style and in everyday conversation, while Sino-Japanese words are relatively more likely to be used in a more formal style (e.g. Zhou (2014), Baba (2018)). Consequently, it is expected that *sutoppu* will be used more frequently in less formal genres like 'blogs' and 'magazines', while *teishi* will be used more frequently in more formal genres like 'white papers' and 'laws'.

4.   (H4) Regarding the collocations, (1) since western loanwords are assumed to be less restrictive in terms of word types that they co-occur with (e.g. Chen (2018) stated that ケア *kea* co-occurs with native Japanese, Sino-Japanese and western loanwords, but 介護 *kaigo* mainly co-occurs with native Japanese and Sino-Japanese), it is expected that *sutoppu* co-occurs with native Japanese, Sino-Japanese and western loanwords, but *teishi* mainly co-occurs with native Japanese and Sino-Japanese. (2) In addition, as mentioned in Miyata et al. (2006), when there is an existing word with a similar meaning, the reason for the existence of the western loanwords is that there is a certain separation between it and the existing word. So we expect that there will also be a difference in the co-occurrence of *sutoppu* and *teishi*.

3.3   Corpora

In this study, we use the *Balanced Corpus of Contemporary Written Japanese* (現代日本語書き言葉均衡コーパス *Gendai nihongo kakikotoba kinkō kōpasu*, henceforth BCCWJ), which is the first large-scale balanced corpus on the Japanese language developed mainly by the National Institute for Japanese Language and Linguistics (NINJAL). To capture the diverse reality of the written language, BCCWJ consists of

three subcorpora (publication subcorpus, library subcorpus, and special-purpose subcorpus) and covers a wide range of text registers including 'books in general', 'magazines', 'newspapers', 'governmental white papers', 'best-selling books', 'internet bulletin-board', 'blogs', 'school textbooks', 'minutes of the national diet', 'publicity newsletters of local governments', 'laws', and 'poetry verses'. The amount of data is 100 million words, which is comparable to BNC (Ishikawa, 2012, Maekawa et al., 2014).

3.4   Methodology

Regarding the RQ1 (frequency of use), we investigate the frequency of *sutoppu* and *teishi* in the BCCWJ.

For RQ2 (conjugation types), we first investigate the frequency with which the two words *sutoppu* and *teishi* are used as nouns and verbs respectively. We follow the conjugation patterns of verbs in BCCWJ (11 categories including negative form, continuous form, hypothetical form, imperative form, and others). The specific conjugation types are shown in Table 2. English translations are taken from *A Handbook of Japanese Grammar Patterns for Teachers and Learners* (Jammassy, 2015).

Table 2: The conjugation types of verbs

Conjugation patterns	Form	Example
連用形 *renyōu-kei*	conjunctive form	し *shi-*
	polite form	します *shi-masu*
	te-form	して *shi-te/* しまして *shi-mashite*
	ta-form	した *shi-ta/* しました *shi-mashita*
連体形 *rentai-kei*	dictionary form	する *suru*
終止形 *shuushi-kei*	dictionary form	する *suru*
仮定形 *katei-kei*	conditionals (*ba*-form)	すれば *sure-ba*
意志推量形 *ishisuiryō-kei*	volitional form	しよう *shi-yō*
可能態 *kanō-tai*	potential form	できる *dekiru/* できます *dekimasu*
命令形 *meirei-kei*	command form	しろ *shiro/* せよ *seyo*
未然形 *mizen-kei* (される *sareru*)	passive form	される *sareru/* されます *saremasu*
未然形 *mizen-kei* (させる *saseru*)	causative form	させる *saseru/* させます *sase-masu*
未然形 *mizen-kei* (せ *se*)	*se*-form	せ *se-*
未然形 *mizen-kei* (一般 *ippan*)	negative form	しない *shi-nai*

Regarding the RQ3 (genres), we investigate the frequency of *sutoppu* and *teishi* in 12 genres of BCCWJ (excluding prosody). In order to make appropriate comparisons between the genres, we adjust the raw frequency into frequency per million words.

Regarding the RQ4 (collocations), we conduct correspondence analysis to compare the co-occurences of the two words.

Correspondence analysis (Benzécri, 1973; Greenacre, 1984, 2017; etc.) has recently been adopted in many corpus studies. It is a method of data visualization by translating two-way and multi-way tables into more readable graphical forms (Greenacre, 2017; Beh & Lombardo, 2021). Correspondence analysis simultaneously classifies cases and variables, which are both called categories, and displays the internal structure existing in a set of item-category data in a simple two-dimensional scatter plot, which enables us to intuitively examine how the categories or items are mutually interrelated and grouped (Ishikawa, 2016).

Specifically, we separate both *sutoppu* and *teishi* into two groups: when either used as nouns or as verbs.

(1) When the two words are used as nouns, we extract the first word and the second word on the left of *sutoppu* or *teishi*, and the first word on the right of *sutoppu* or *teishi*. In addition, we sort the collocations and extract the top 15 words (if there was a word whose frequency matched that of the 15th word, all words with the same frequency were included). Furthermore, we make a frequency table with each genre of *sutoppu* and *teishi* as the first item (18 genres excluding the genre in which the frequency of the top word is less than 20), and the top words of co-occurrence as the second item (84 words excluding redundancy), and then conduct a correspondence analysis.

(2) When the two words are used as verbs, we first manually extract the objects, and then extract the top 30 words (if there is a word whose frequency agrees with the 30th word, we include all words whose frequency are the same). Next, we make a frequency table in which each genre of *sutoppu* and *teishi* is the first item (10 genres excluding the genre in which the frequency of the top word is less than 20), and the top words of the objects are the second item (64 words excluding duplicates) and conduct a correspondence analysis.

4   Results and discussion

This section presents the results and discussions on each of the four perspectives described above: frequency (RQ1) in Section 4.1, conjugation types (RQ2) in Section 4.2, characteristically used genres (RQ3) in Section 4.3 and collocations (RQ4) in Section 4.4.

4.1 The comparison of frequency

First, the frequency of *sutoppu* and *teishi* in BCCWJ are shown in Figure 1 below.
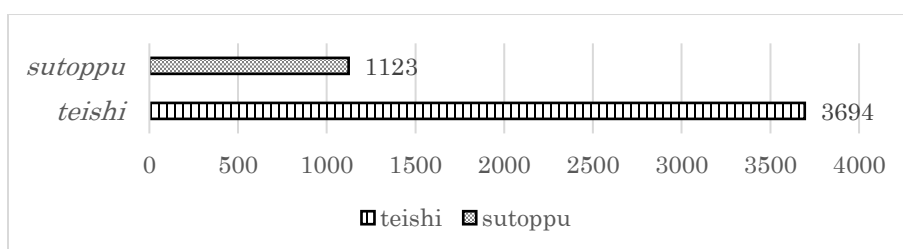

Figure 1: Frequency of *sutoppu* and *teishi*

Figure 1 shows that *teishi* is used more than three times as often as *sutoppu*, suggesting that Sino-Japanese is used with a much higher frequency.

As mentioned in the hypothesis, frequency comparisons between Sino-Japanese and western loanwords with similar meanings have been conducted in many studies, but the results have been disparate. For example, Miyata (2006) found that the frequency of the Sino-Japanese, 危険 *kiken,* was 30% lower than that of the western loanwords, リスク *risuku*. On the other hand, Chen (2014) showed that the frequency of the Sino-Japanese, 介護 *kaigo*, was 2.4 times higher than that of the western loanwords, ケア *kea*. In this study, *teishi* is 2.3 times more frequent than *sutoppu*, showing a similar result to the use of 介護 *kaigo* and ケア *kea* in terms of frequency relationship. Then, what causes the selection rate of Sino-Japanese for 危険 *kiken*/ リスク *risuku* to be low and that for *teishi*/ *sutoppu* high?

Firstly, as mentioned in the hypothesis, there is a difference in the contexts in which the words are used. In contexts closer to daily life, western loanwords are widely used to replace Sino-Japanese, while words that are frequently used in social and economic contexts, Sino-Japanese with high formality may be preferred.

Secondly, the difference may be caused by the part-of-speech nature of western loanwords. リスク *Risuku* is a pure noun that cannot be inflected, while *sutoppu* is a noun that can be inflected and used as a verb. In general, nouns refer to concrete objects, especially to specific, static objects with clear contours, so the one-to-one correspondence between an object and a word is strong and difficult to be replaced with other words. Verbs, on the other hand, are words that express the totality of changing motion, so their indicative content is usually broader than that of nouns. For example, if we compare the content implied by the noun 'oranges' with that implied by the verb 'eat', we find that the latter is much broader, more ambiguous, and less semantically specific. As a result, the strength of the one-to-one correspondence between verbs can be relatively lower, and as a result, a verb can be more easily substituted by another word. For this reason, the noun リスク *risuku* is rarely substituted

by 危険 *kiken*, while *sutoppu*, which can be used both as a verb and a noun, may frequently be substituted by *teishi*. This explanation also applies to the case of 介護 *kaigo* and ケア *kea*, which also have a high rate of Sino-Japanese selection.

Based on the above, the following two sections will look at conjugation types (parts of speech) (RQ2) and style (genre) (RQ3) separately.

4.2   The comparison of conjugation types

Next, we investigate the frequency of noun and verb uses of *sutoppu* and *teishi,* respectively. The results are shown as Figure 2 below.
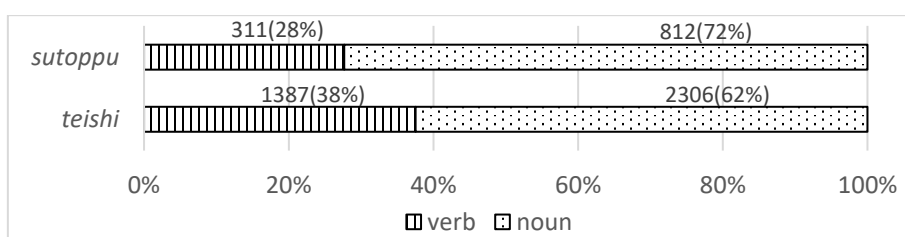
Figure 2: Frequency of the noun and verb uses of *sutoppu* and *teishi*

Figure 2 shows that both words are used more as nouns than verbs, however, the proportion of noun use of *sutoppu* was 72%, which is 10% higher than that of *teishi*.

Based on the BCCWJ's classification of verb conjugations, we can classify the verb uses of the two words as shown in Figures 3 and 4 below.
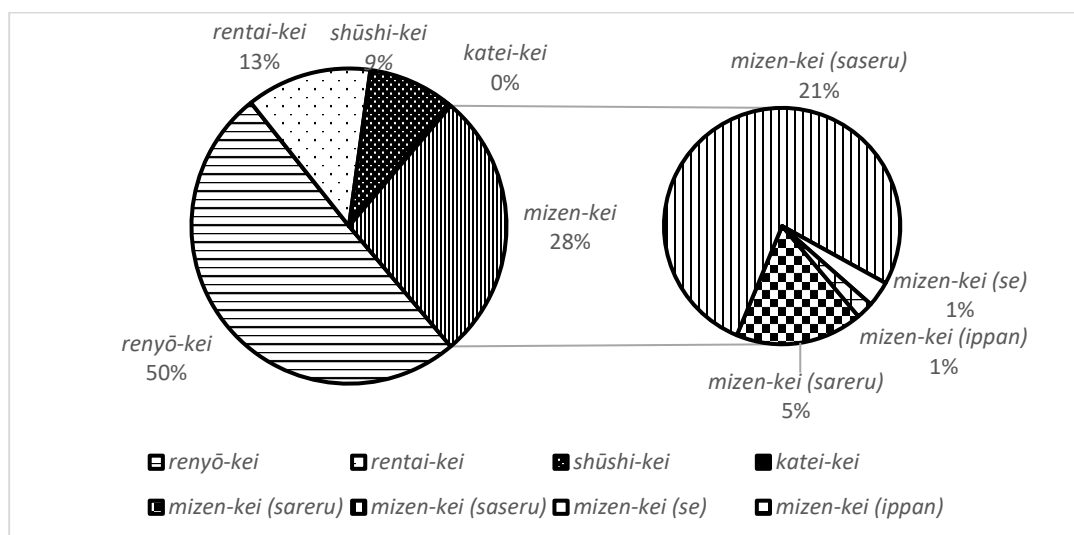
Figure 3: Conjugation types of *sutoppu* used as a verb
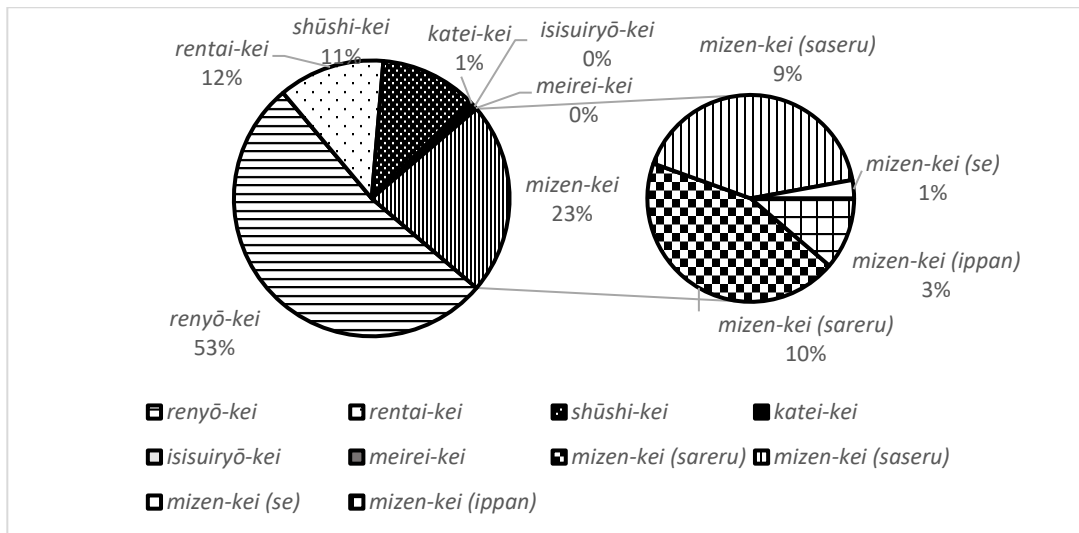
Figure 4: Conjugation types of *teishi* used as a verb

Figures 3 and 4 show that the causative-passive form of *sutoppu* is 21.22%, which is 126.47% more than *teishi*, and the passive form of *sutoppu* is 4.82%, which is 51.56% less than *teishi*. In other words, *sutoppu* is used mostly in the active form and implies the meaning of compulsory and intentional, while *teishi* is used mostly in the passive form and may include the meanings of being situational passive, or inevitable.

As stated in the hypothesis, the part-of-speech and conjugation types of the two words seem to be equivalent from the point of view of dictionaries, but the results suggest that, contrary to the hypothesis, *sutoppu* is used more often as a noun, and is more likely to include the meaning of compulsory. If this point could be added to the dictionary, it would make it easier to understand the usage of the two words.

4.3   The comparison of genres

Concerning RQ1 and RQ2 above, we investigated the differences between *sutoppu* and *teishi* from two perspectives, frequency, and conjugation types. The results show that (1) in terms of frequency, *teishi* is used more than twice as often as *sutoppu*, (2) in terms of the conjugation types, *sutoppu* has more noun usages than *teishi*, and *sutoppu* has been used in its causative form one time more than *teishi*, emphasizing its meaning of forceful. So what differences exist in the genre preferences of the two words? The following are the results of a survey of the frequency of the two words in each genre of the BCCWJ.

Table 3: Frequency and ratio of *sutoppu* and *teishi* used for each genre

Genres (Abbreviations in BCCWJ)		*sutoppu*		*teishi*	
		PMW	ratio (%)	PMW	ratio (%)
books	published book(PB)	7.64	6	35.93	5
	library book(LB)	7.37	6	24.46	4
	best-selling books (OB)	5.61	4	20.58	3
magazines (PM)		25.87	20	29.92	4
newspapers (PN)		15.33	12	66.41	10
blog (OY)		30.61	24	24.92	4
school textbooks (OT)		2.15	2	25.85	4
publicity newsletters of local governments (OP)		13.32	10	37.55	5
minutes of the national diet (OM)		9.21	7	42.72	6
bulletin-board (OC)		9.36	7	38.22	5
white papers (OW)		3.48	3	56.52	8
laws (OL)		0.00	0	293.75	42

Three inferences can be made from Table 3. To begin with, *sutoppu* is used much more in the 'blogs' and 'magazines' genres, and also seen in the 'newspapers' and ' publicity newsletters of local governments' genres, but not at all in the 'laws' genre. It seems that *sutoppu* was found to be favored in more casual contexts.

Furthermore, *teishi* is used more in the 'laws' genre and to some extent also in the 'newspapers' genre. In other words, the use of *teishi* is preferred in more official contexts. The hypothesis stated that *teishi* is more frequently used in more formal genres than *sutoppu* because western loanwords are more likely to be used in a more casual style and daily conversation, while Sino-Japanese are more likely to be used in a more stiff style and official contexts. On these two points, we can say that the hypothesis is supported.

Finally, As for bias toward specific genres, *teishi* is overwhelmingly used in the 'laws' genre, which is more biased than *sutoppu*. It has often been pointed out that western loanwords have the function of creating a sense of freshness and making positive use of stylistic differences, but the present results suggest that Sino-Japanese may also have the function of highlighting differences between highly formal styles and other genres.

4.4 The comparison of collocations

Last but not least, we will focus on the words that co-occur with the two words for comparison. We will first look at the case of noun usage.
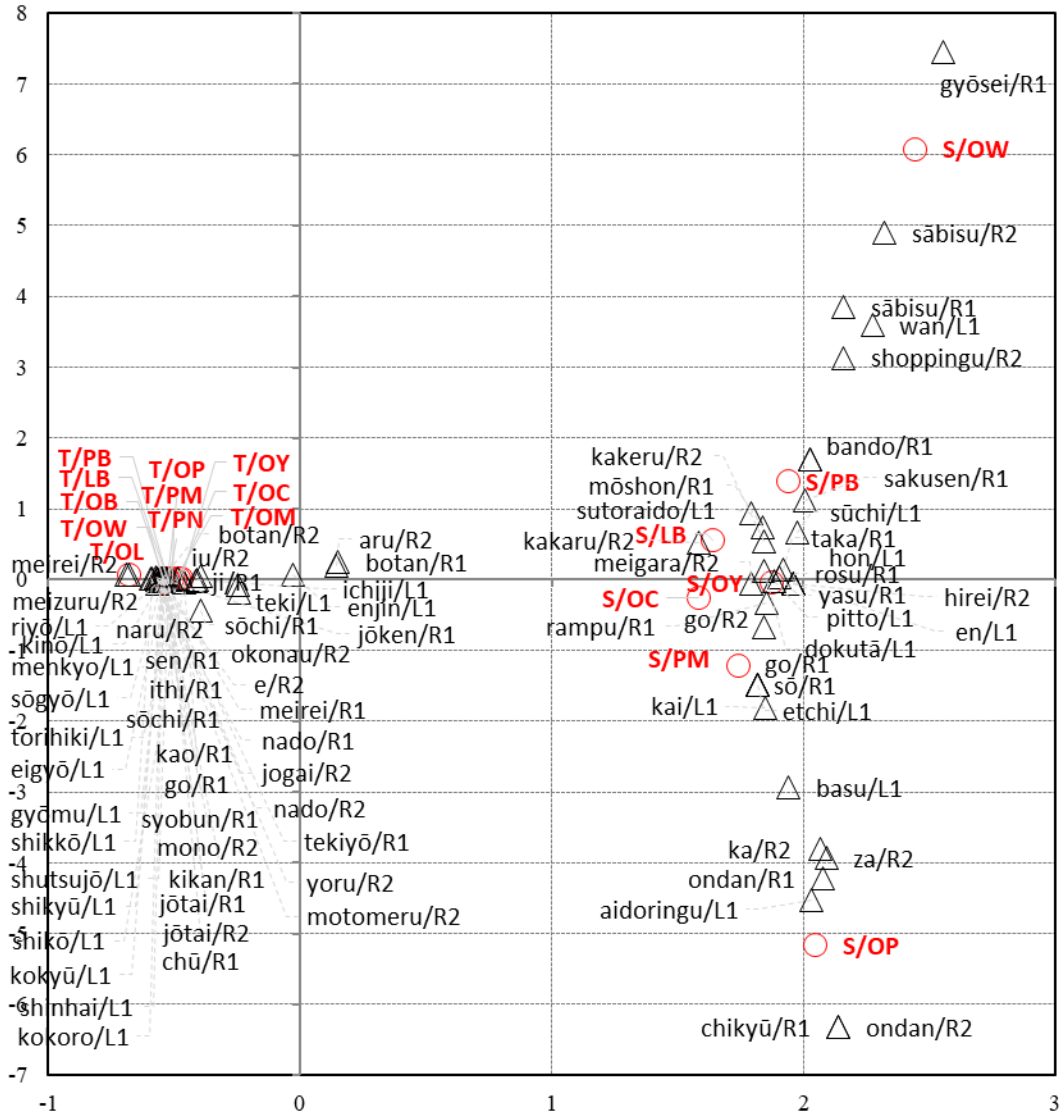
Figure 5: Scatter plots of feature words co-occurring with two words used as a noun

Figure 5 shows the results of the correspondence analysis which has the genres of *sutoppu* and *teishi* as first items and the 84 top co-occurring words as second items. The contribution of dimension 1 (vertical axis) and dimension 2 (horizontal axis) are 17.25% and 12.59%, respectively, which explains 29.83% of the total. As seen in Figure 5, *sutoppu* and *teishi* are separated on the left and right sides of the figure, and the words characteristic of the two words can be summarised in Table 4 below.

Table 4: Characteristic words for *sutoppu* and *teishi*

	sutoppu	*teishi*
left1	ワン *wan* (one), アイドリング *aidoringu* (idling), 数値 *sūchi* (numerical value), 円 *en* (yen), 本 *hon* (quantity unit), バス *basu* (bus), ビット *bitto* (bit), 回 *kai* (quantity unit), ドクター*dokutā* (doctor), エッチ *etchi* (etch), ストライド *sutoraido* (stride)	執行 *shikkō* (execution), 利用 *riyō* (use), 操業 *sōgyō* (operation), 呼吸 *kokyū* (breathing), 機能 *kinō* (function), 支給 *shikyū* (payment), 心 *kokoro* (heart), 心肺 *shinhai* (cardiopulmonary), 営業 *eigyō* (business), 業務 *gyōmu* (work), 免許 *menkyo* (license), 出場 *shutsujō* (one's turn to go on satge), 取引 *torihiki* (transactions), 思考 *shikō* (thinking),一時 *ichiji* (for a time), 的 *teki* (suffix), エンジン *enjin* (engine)
right1	行政 *gyōsei* (administration), サービス *sābisu* (service), 地球 *chikyū* (earth), 温暖 ondan (global warming), バンド *bando* (band), 作戦 *sakusen*(strategy), ロス *rosu* (loss), 高 *taka* (a limit high of a stock price), 安 *yasu* (a limit low of a stock price), ランプ *rampu* (lamp), モーション *mōshon* (freeze-fram), 層 *sō* (layer), 語 *go* (word), ボタン *botan* (button)	など *nado* (etc.), 条件 *jōken* (condition), 顔 *kao* (face), 提供 *tekiyō* (application), 処分 *shobun* (punishment), 措置 *sochi* (measure), 期間 *kikan* (period), 命令 *meirei* (command), 状態 *jōtai* (state), 位置 *ichi* (location), 中 *chū* (in), 線 *sen* (line), 後 *go* (after), 装置 *sōchi* (device), 時 *ji* (time)
right2	サービス *sābisu* (service), ショッピング *shoppingu* (shopping), 温暖 *ondan* (warm), ザ *za* (the), 化 *ka* (-ificaion), 比例 *hirei* (proportion), ゴール *gōru* (goal), かける *kakeru* (make), 銘柄 *meigara* (brand name), かかる *kakaru* (make), ある *aru* (be)	命令 *meirei* (command), 命ずる *meizuru* (command), など *nado* (etc.), もの *mono* (person), よる *yoru* (according to), 求める *motomeru* (demand), へ *e* (to), 除外 *jogai* (exclusion), 行う *okonau* (do), ボタン *botan* (button), 状態 *jōtai* (state), なる *naru* (become), 言う *iu* (say), する *suru* (do)

Regarding the word types of characteristic words, as hypothesized, *sutoppu* co-occurs with native Japanese (e.g. 高 *taka*, 安 *yasu*), Sino-Japanese (e.g. 数値 *sūchi*, 比例 *hirei*) and western loanwords (e.g.ドクター *dokutā*, モーション *mōshon*), but *teishi* mainly co-occurs with Sino-Japanese (e.g. 利用 *riyō,* 支給 *shikyū*).

Furthermore, in the words co-occurring with *sutoppu*, many words that express the characteristics of the point at which the *stop* occurs, such as 数値 *sūchi* (numerical value), 高 *taka* (maximum allowable single-day gain (stock exchange)), or 安 *yasu* (maximum allowable single-day loss (stock exchange)), whereas, in the words co-occurring with *teishi*, there are many words with continuity, such as 期間 *kikan* (period), 状態 *jōtai* (state), and 中 *chū* (middle). In other words, *sutoppu* often refers to a point of stopping and rarely includes the meaning of continuation of time, whereas *teishi* is allowed to mean the continuation of time.

Also, *sutoppu* is often used to emphasize the action of stopping something that is moving, or the act of trying to stop, such as ドクターストップ *dokutā sutoppu* (doctor stop: a doctor forbids a patient to do something they think will prevent a disease or a disability from worsening or leading to death) or ストップ地球温暖化 *sutoppu chikyū ondanka* (stop global warming) while *teishi* is often used to emphasize the state or the result of being stopped, such as 免許停止 *menkyo teishi* (license stop) or 営業停止 *eigyō teishi* (business stop), and to express the state of stopping bodily functions, such as 呼吸停止 *kokyū teishi* (stop breathing) or 心肺停止 *shinhai teishi* (cardiopulmonary stop).

The specific examples are given as follows.

(1) この問題にストップをかけるため，無駄なエネルギーを使用しない節約型の暮らしや，自然エネルギーへの転換などが求められています。(OP74_00002)

Kono mondai ni sutoppu wo kakeru tame, muda na enerugī wo shiyō shinai setsuyakugata no kurashi ya, shizen enerugī e no tenkan nado ga motomerareteimasu.

'In order to stop this problem, we need to save energy and switch to renewable energy sources.'

(2) 区は目標の達成に向け，照明設備や空調設備を高効率なものに更新し，装着可能なすべての庁有車にアイドリングストップ装置を装着など，率先して温室効果ガスの排出抑制にも取り組んでいます。(OP30_00002)

Ku wa mokuhyō no tassei ni muke, shōmeisetsubi ya kūchōsetsubi wo kōkōritsu na mono ni kōshinshi, sōchaku kanō na subete no chōyūsha ni aidoringu sutoppu sōchi wo sōchaku nado, sossenshite onshitsukōka gasu no haishutsu yokusei ni mo torikundeimasu.

'In order to achieve the target, the ward is taking the initiative to reduce greenhouse gas emissions by upgrading their lighting and air conditioning systems efficiency, and installing idling stop devices in all government-owned vehicles where they can be installed.'

(3) あと，不動産営業の労働実体は時間に直すと労働基準法以上に働いているケースが多く，そのまま手続きと会社が営業停止になってしまいます。(OC03_00938)

Ato, fudōsan eigyō no rōdō jittai wa jikan ni naosu to rōdō kijunhō ijō ni hataraiteiru kēsu ga ōku, sono mama tetsuzuki to kaisha ga eigō teishi ni natte shimaimasu.

'Also, in many cases, the actual labor of real estate salespeople is more than the labor standard law allows, if it is measured in hours. If the practise is not changed, the company will be suspended from business.'

(4) 特に，心肺停止状態の傷病者の救命率向上に資CPR（心肺蘇生法）の習得に主眼を置き，かつ訓練用人形等を用いた住民体験型の普及啓発活動の積極的な推進が求められている。(OW4X_00479)

Toku ni, shinhai teishi jōtai no shōbyōsha no kyūmeiritsu kōjō ni shiCPR (shinhai soseihō) no shūtoku ni shugan wo oki, katsu kunren-yō ningyō nado wo mochiita jūmin taikengata no fukyū keihatsu katsudō no sekkyokuteki na suishin ga motomerareteiru.

'In particular, there is a need to focus on the learning of CPR (cardiopulmonary resuscitation) to improve the lifesaving rate of injured people in cardiopulmonary arrest, and to actively promote hands-on educational activities using training dolls.'

The findings above can be summarised as in Figure 6.
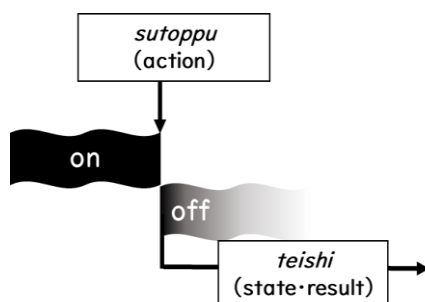
Figure 6: Illustration of the difference in noun usage between *sutoppu* and *teishi*

As shown in Figure 6, *sutoppu* emphasizes instantaneity and the action of stopping, whereas *teishi* allows for a certain amount of time and denotes the state of being stopped.

Next, we will look at cases where *sutoppu* and *teishi* are used as verbs. Figure 6 shows the results of the correspondence analysis with each genre of *sutoppu* and *teishi* as the first items and the 84 top co-occurring objects as the second items. As already mentioned, *sutoppu* is not used at all in the 'laws' genre, so we exclude the 'laws' genre

here. The contribution rates of dimensions 1 and 2 are 30.23% and 18.14%, which explain 50.24% of the total.

As can be seen from Figure 7, each genre of *sutoppu* and *teishi* was found to be divided on left and right. Summarizing their respective characteristic words , we obtain the following Table 5.

Figure 7: Scatter plots of feature words co-occurring with two words used as verbs

Table 5 shows that in the feature objects of *sutoppu,* there are many words that semantically denote 'something with high resistance or ongoing action', such as ゲーム *gēmu* (game), 物流 *butsuryū* (logistics), and 支援 *shien* (support). On the other hand, in the feature object of *teishi*, there are many nouns related to machines, such as 列車

ressha (train) and エンジン *enjin* (engine), and nouns related to bodily functions, such as 心臓 *shinzō* (heart) and 呼吸 *kokyū* (breathing), which express something with low resistance. Specific examples are given below.

Table 5: Characteristic words for *sutoppu* and *teishi*

Objects of *sutoppu*	Objects of *teishi*
心肺 *shiai* (match),	エレベーター *erebētā* (elevator),
ライン *rain* (line), 金 *kane* (money)	輸入 *yunyū* (import), 供給 *kyōkyū* (supply),
審議 *shingi* (deliberation),	思考 *shikō* (thought),
建設 *kensetsu* (construction),	回路 *kairo* (circuit (electric)),
物流 *butsuryū* (logistics),	処理 *shori* (processing),
反応 *hanno* (reaction)	呼吸 *kokyū* (breath), ポンプ *pompu* (pump),
工事 *kōji* (construction),	動き *ugoki* (movement), 列車 *ressha* (train),
研究費 *kenkyūhi* (research funds),	原子炉 *genshiro* (nuclear reactor),
それ *sore* (it), 支援 *shien* (support),	成長 *seichō* (growth),
融資 *yūshi* (loan),	車 *kuruma* (vehicle),
交通機関 *kōtsūkikan* (transportation),	輸出 *yushutsu* (export),
飛行機 *hikōki* (airplane),	機能 *kinō* (function),
行動 *kōdō* (behavior),	エンジン *enjin* (engine),
流れ *nagare* (flow), 句 *ku* (phrase),	活動 *katsudō* (activity),
連勝 *renshō* (consecutive victories),	自動車 *jidōsha* (automobile),
生産 *seisan* (production),	再生 *saisei* (regeneration),
支払い *shiharai* (payment),	運転 *unten* (operation),
時間 *jikan* (time),	作業 *sagyō* (operation),
発行 *hakkō* (publishing)	心臓 *shinzō* (heart), 支給 *shinkyū* (payment),
	サービス *sābisu* (services),
	事業 *jigyō* (business),
	使用 *shiyō* (use), 利用 *riyō* (use)

(5) 春先なのに数十年ぶりの大寒波がおとずれ，雪がふりはじめ，やがて猛吹雪にかわって，多くの交通機関がストップし，もちろん飛行機もとばず…(LBo2_00085)

   harusaki na noni sūjūnen buri no daikampa ga otozure, yuki ga furi hajime, yagate mōfubuki ni kawatte, ōku no kōtsū kikan ga <u>*sutoppu*</u> *shi, mochiron hikōki mo tobazu...*

   'Even though it was the beginning of spring, the first major cold wave in decades hit, and snow began to fall, which soon became a blizzard, <u>stopping</u> many transportation systems, and of course, flights.'

(6) 現在，ヨーロッパの多くの国が，移民の流れをストップしようとしています。(PB35_00199)

   genzai, yōroppa no ōku no kuni ga, imin no nagare wo <u>*sutoppu*</u> *shiyō to shiteimasu.*

   'Currently, many European countries are trying to <u>stop</u> the flow of immigrants.'

(7) これを田辺海上保安部通信所が受信したが，感度が非常に微弱であり，さらに，浸水で発電機が停止しバッテリー使用による送信であったため約 3 分で通信が途絶えてしまったことなどから，L 号の遭難位置は野島埼の南方としか確認できなかった。(LBg5_00015)

kore wo tanabe kaijō hoambu tsūshin sho ga jushin shita ga, kando ga hijō ni bijaku de ari, sara ni, shinsui de hatsudenki ga teishi shi batterī shiyō ni yoru sōshin deatta tame yaku sampun de tsūshin ga todaete shimatta koto nado kara, L gō no sōnan ichi wa nojimazaki no nampō to shika kakunin deki nakatta.

'The *Tanabe* Coast Guard Station received the message, but the sensitivity was very weak, and the transmission was cut off after about three minutes because the generator was stopped by the flooding and a battery was being used for transmission.'

(8) けがや急病などで呼吸や心臓が停止してしまった場合，人工呼吸や心臓マッサージは，知識がないとできません。(OP15_00002)

kega ya kyūbyō nado de kokyū ya shinzō ga teishi shite shimatta baai, jinkō kokyū ya shinzō massāji wa, chishiki ga nai to dekimasen.

'If a person's breathing or heart stops due to injury or sudden illness, artificial respiration and cardiac massage cannot be performed without knowledge.'

The above examples show that *sutoppu* can mean to deliberately stop something that is of high resistance or a movement in progress, and similarly, in its noun usage, *sutoppu* emphasizes the action of stopping. In example (5), interruption of traffic should be avoided (which leads to psychological resistance to the interruption of traffic), but the snow unavoidably stopped the traffic, anyways. In example (6), it is considered that there is great resistance from the immigrants who try to come in with a strong will.

On the other hand, *teishi* often refers to stopping something with low resistance. In example (7), the subject of the action is a generator, and a generator does not have any intention of not wanting to stop, so *teishi* is chosen here. In example (8), the heart has already stopped spontaneously due to an injury or illness, and there is no particular resistance, so *teishi* has also been chosen here. Thus, *sutoppu* emphasizes sudden and forcible stopping which overcomes some kind of resistance, whereas *teishi* often describes things that have spontaneously stopped.

5　Conclusion

In summary, this paper has attempted to elucidate the use of the semantic functions of the western loanword *sutoppu* and its synonymous Sino-Japanese word *teishi* from four perspectives: (1) frequency, (2) conjugation types, (3) characteristically used genres, and (4) collocations. The findings of this paper can be summarized as follows.

Firstly, in terms of frequency, *teishi* is used more than twice as often as *sutoppu,* which may be caused by the contexts in which the words are likely to be used and the part-of-speech nature (see 4.1).

Secondly, in terms of conjugation types (parts of speech), both words are used as nouns more frequently than as verbs, and *sutoppu* is used as a noun 10% more than *teishi*. In addition, *sutoppu* is used mostly in the causative form, implying something compulsory or intentional. Whereas *teishi* is mostly used in its passive form to imply inevitability (see 4.2).

Thirdly, from the perspective of genres, *sutoppu* is less genre-biased and is preferred in more informal contexts. On the other hand, *teishi* has a greater genre bias and is preferred in more public contexts (see 4.3).

Finally, in the case of collocations, when the two words are used as nouns, *sutoppu* emphasizes instantaneity and describes the action of stopping, whereas *teishi* permits a certain duration of time and describes the state of being stopped. When the two words are used as verbs, the characteristic objects of *sutoppu* semantically express something with high resistance or ongoing action, while the characteristic objects of *teishi* have many words which express something with low resistance. In addition, *sutoppu* has the meaning of intentionally stopping something with great resistance or stopping an action in progress, emphasizing the action of stopping, whereas *teishi* often refers to the state of being stopped (see 4.4).

Since the first corpus-based dictionary *Collins Cobuild Dictionary of English*, compiled by John Sinclair and published in 1987, the way dictionaries were compiled which had relied heavily on the introspective judgment of native speakers had changed. At present, almost all English dictionaries for learners are corpus-based, but Japanese dictionaries for learners are few in number, and there are still no fully corpus-based dictionaries (Ishikawa, 2014). Many previous studies have pointed out that there is room for the contribution of corpora and corpus analysis methods in the development of Japanese dictionaries for learners (Sunakawa, 2011; Tanomura, 2010; Ishikawa, 2014). In this study, we expect to shed new light on the development of Japanese dictionaries for learners. Table 6 is an example of a dictionary description that utilizes the findings of the present study.

As shown in Table 6, the use of the two words can be made clearer by adding the information of genre and conjugation types, as well as specific usage preferences.

To conclude, this paper has elucidated the usage of *sutoppu* and *teishi* in terms of their semantic functions. However, there are some limitations. Firstly, only one pair of a western loanword and its synonymous Sino-Japanese word was observed. Secondly, the study does not incorporate native speaker reflections. We hope to address these points in our ongoing research.

Table 6: Proposal for a new dictionary description

sutoppu	● Characterisitic genres: ★ 'blogs', 'magazines' ☆'newspapers', 'white papers' × 'laws' ● When used as a noun: Meaning: Emphasizes an action which is <u>stopped instantaneously</u>. Examples of collocations: (1) ドクターストップ *dokutā sutoppu* (doctor stop) (2) アイドリングストップ *aidoringu sutoppu* (idling stop) (3) ワンストップ行政 *wansutoppu gyōsei* (one-stop administration) (4) ストップ高 *sutoppu taka* (a limit high of a stock price) ● When used as a verb: meaning: Emphasises <u>Intentionally or forcibly</u> stopping things or ongoing actions with <u>high resistance</u>. Examples of collocations: (1) 連敗をストップ *rempai wo sutoppu* (Stop the losing streak) (2) 供給をストップ *kyōkyū wo sutoppu* (stop the supply) (3) 支払いをストップ *shiharai wo sutoppu* (stop the payment)
teishi	● Characterisitic genres: ★'laws' ☆ 'newspapers' ● When used as a noun: Meaning: Emphasizes the the state of <u>being stopped</u>. Examples of collocations: (1) 利用停止 *riyō teishi* (utilization stop) (2) 出場停止 *shutsujō teishi* (exit stop) (3) 心肺停止 *shinhai teishi* (cardiopulmonary stop) (4) 停止状態 *teishi jōtai* (stop status) ● When used as a verb: meaning: Emphasizes stopping <u>naturally</u> with <u>low resistance</u>. Examples of collocations: (1) 活動を停止 *katsudō wo teishi* (stop activity) (2) 機能を停止 *kinō wo teishi* (stop function) (3) エンジンを停止 *enjin wo teishi* (stop engine) (4) 呼吸を停止 *kokyū wo teishi* (stop breathing)

References

Aroonmanakun, V. (2015). Quick or fast: A corpus based study of English synonyms. *LEARN Journal: Language Education and Acquisition Research Network*, *8*(1), 53-62.

Baba, T. (2018). The possibility of studies of stylistic features of words using "writing style annotation for the library subcorpus of the BCCWJ" (in Japanese). *Proceedings of Language Resources Workshop, 3*, 241-256.

Beh, E. J., & Lombardo, R. (2021). *An introduction to correspondence analysis*. John Wiley & Sons.

Benzécri, J. P. (1973), L'Analyse des Données, Vol. I, La Taxinomie; Vol. II, L'Analyse des Correspondances. Dunod.

Biber, D., Conrad, S., & Reppen R. (1996). Corpus-based investigation of language use. *Annual Review of Applied Linguistics, 16*, 115-136.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Chen, X. (2014). The Difference of the Meanings of Synonyms in Katakana and Kanji: The Semantic Difference of "Care" and "Kaigo" (in Japanese). *The Ritsumeikan literature review, 637,* 1438-1450.

Chung, S. F. (2011). A corpus-based analysis of "create" and "produce". *Chang Gung Journal of Humanities and Social Sciences, 4*(2), 399-425.

Evison, J. (2010). What are the basics of analysing a corpus. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics*. (pp. 122-135). Routledge.

Greenacre, M.J. (1984). *Theory and Application of Correspondence Analysis*. Academic Press.

Greenacre, M.J. (2017). *Correspondence Analysis in Practice* (3rd ed.). Chapman & Hall/CRC.

Gries, S. T., & Otani, N. (2010). Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME journal, 34*, 121-150.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.

Inoue, N., & Akano, I. (Eds.) (2013). *The Wisdom English-Japanese Dictionary* (3rd ed.). Sanseido.

Ishikawa, S. (2012). *A Basic Guide to Corpus Linguistics* (in Japanese). Kabushiki Kaisha Hitsuji Shobō.

Ishikawa, S. (2014). Cooccurrence network analysis for sophistication of dictionary definitions of Japanese synonymous words: possibility of corpus-based Japanese dictionaries. (in Japanese). *The Institute of Statistical Mathematics cooperative research report, 308*, 1-21.

Ishikawa, S. (2016). Japanese polite sentence-final markers: desu, desuyo, desune, and desuyone A corpus-based analysis with a focus on frequency, collocation, and functional grouping. In J. Szerszunowicz, B. Nowowiejski, P. Ishida & K. Yagi (Eds.), *Linguo-cultural research on phraseology, 3*, 537-554.

Jammassy, G. (2015). *A Handbook of Japanese Grammar Patterns for Teachers and Learners*. Kurosio Press.

Jinnouchi M. (2009). Nihongo gakusyūsya no katakana isiki to katakanago kyōiku (in Japanese). *Language and Culture, 11*, 47-60.

Kennedy, G. D. (1998). *An introduction to corpus linguistics*. Routledge.

Liu, D. (2010). Is it a chief, main, major, primary, or principal concern?: A corpus-based behavioral profile study of the near-synonyms. *International Journal of Corpus Linguistics, 15*(1), 56-87.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., & Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language resources and evaluation, 48*(2), 345-371.

Matumura, A., & Sanseidohenshūjo (Eds.) (2006). *Daijirinn* (3rd ed.). Sanseido.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction* (2nd ed.). Edinburgh University Press.

McEnery, T., & Hardie, A. (2014). *Gaisetsu Kōpasu Gengogaku: Shuhō, Riron, Jissen. [Corpus linguistics: Method, Theory, and Practice]*. (S. Ishikawa, Trans.). Hitsuji Shobō. (Original work published 2012).

Minamide, K., & Nakamura, M. (Eds.) (2011). *Genius Japanese-English Dictionary* (3rd ed.). Taishukan.

Miyata, K. (2007). Gairaigo "meritto" to sono ruigigo no imi hikaku shimbun wo shiryo to shite (in Japanese). *Kokyo Baitai no Gairaigo: Gairaigo Iikae Teian wo Aasaeru Chosa Kenkyu*. Kokuritsu Kokugo Kenkyusho.

Miyata, K., & Tanaka, B. (2006). Gairaigo risuku to sono ruigigo no imi hikaku: kison no ruigigo wo motsu gairaigo no sonzai riyu (in Japanese). *Proceedings of the Annual Meeting of the Association for Natural Language Processing, 12*, 600-603.

Mogi, T. (2015). A Corpus-based Study on Loanword Verbs in Japanese : A Case Study of maaku-suru (<mark) (in Japanese). *Kumamoto journal of culture and humanities, 106*, 83-95.

Moon, R. (2010). What can a corpus tell is about lexis?. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 345-358). Routledge.

Nihongogakkai. (Eds.) (2018). *The Encyclopedia of Japanese Linguistics* (1st ed.) (in Japanese). Tokyodo Shuppan.

Phoocharoensil, S. (2020). Collocational patterns of the near-synonyms error, fault, and mistake. *The International Journal of Communication and Linguistic Studies, 19*(1), 1-17.

Sato, T. (2013). Semantic analysis in research of loan words : A case study in synonymy of mudo (mood) and fun'iki (in Japanese). *Bulletin of Gakushuin Women's College, 15*, 45-46.

Shaw, E. M. (2011). *Teaching Vocabulary Through Data-driven Learning*. Brigham Young University.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Shinya, T. (2010). Syntactic analysis of synonyms, jokyo and jotai: a corpus-based quantitative comparison. *Mathematical Linguistics, 27*(5), 173-193.

Sugimoto, T. (2009). A corpus-based analysis of synonymous verbs: nejiru and hineru. *Studies in language and literature. Language, 55*, 109-122.

Sugimoto,T., & Iwabuchi, T. (Eds.) (1994) *New Japanese linguistic dictionary* (in Japanese). Oufu.

Sunakawa, Y. (2011). On the Use of Corpora in Teaching Japanese as a Foreign Language (in Japanese). *Journal of Japanese Language Teaching, 150*, 4-18.

Tanomura, T. (2010). Japanese corpus and collocation: possibility of application to dictionary description (in Japanese). *Gengo Kenkyu, 138*, 1-23.

Taylor, J. R. (2003). Near synonyms as co-extensive categories: 'high' and 'tall' revisited. *Language Sciences, 25*(3), 263-284.

Yamada, T., Shibata, T., Sakai, K., Kuramochi, Y., & Yamada, A. (Eds.) (2005). *Shimmeikai Kokugo Jiten* (6th ed.) Sanseido.

Yamashita, N., Hata, Y., & Todoroi, Y. (2018). Japanese Native Speakers' use of Katakana Words and Their Synonyms (in Japanese). *Memoirs of the Faculty of Education Kagawa University, Part I, 149*, 45-42.

Zhao, (2013). A corpus-based study of the synonymous adverbs yatto and yōyaku: their stylistic differences and their co-occurrence relations with predicates (in Japanese). *Nihongokenkyū, 33*, 15-30.

Zhou, Q.L. (2014). Analysis of characteristics of different etymological synonyms in Japanese : A Survey on College Students (in Japanese). *Bulletin of Shizuoka Sangyo University, 16*, 33-44.

The Roman Alphabet Within the Japanese Writing System: Patterns of Usages and Their Significance

Hironori NISHI

University of Memphis, United States
hnishi1@memphis.edu

Abstract

The present study explores the usages of the Roman alphabet within the writing system of Japanese. Japanese is typically said to have three types of characters in its writing system: *hiragana*, *katakana*, and *kanji*. However, the Roman alphabet is also commonly used in Japanese for various purposes along with other types of characters in Japanese. The present study argues that with the recent surge in electronic communication, the writing practice of Japanese is transitioning from vertical writing to horizontal writing, and this transition allows more foreign words and expressions written in the Roman alphabet to be used within Japanese without being converted into *katakana* loanwords. The present study also discusses the influence of the ever-increasing international interaction to the usages of the Roman alphabet within Japanese.

Keywords: Japanese, loanwords, Roman alphabet, *katakana*, writing system

Povzetek

Članek raziskuje uporabo latiničnega zapisa v pisnem sistemu japonščine. Za japonščino običajno pravimo, da ima v pisnem sistemu tri vrste znakov: hiragano, katakano in pismenke. Hkrati se v japonščini pogosto pojavlja tudi latinica, nameni za to pa so različni. V članku ugotavljamo, da nedavni porast elektronske komunikacije, s katero pisna praksa japonščine prehaja iz navpičnega v vodoravni zapis, omogoča, da se v japonščini uporablja več tujih besed in izrazov, napisanih v latinici, ne da bi jih pretvorili v izposojenke v katakani. Članek obenem obravnava še vpliv vedno večje mednarodne interakcije na uporabo latinice v japonščini.

Ključne besede: japonščina, prevzete besede, latinični zapis, *katakana*, pisni sistem

1   Introduction

The writing system of the modern Japanese language is typically said to consist of three major types of characters; the three types are *hiragana*, *katakana*, and *kanji*. However, one set of characters that is often overlooked in the writing system of Japanese is the Roman alphabet. In addition, with the recent surge of the use of the written language on the Internet, the Roman alphabet used within Japanese has also been increasing its importance. The present paper explores the usages of the Roman alphabet within the Japanese language, with a special focus on the linguistic, communicative, and social significance resulting from the usages of the Roman alphabet within the writing system of Japanese.

Before discussing the usages of the Roman alphabet within the writing system of Japanese, first, the present study provides a brief overview of the types of characters used in Japanese.

In the writing system of Japanese*, hiragana* characters are considered to be the most basic set of characters. After Chinese characters were introduced to Japan around the late fourth and early fifth centuries, *hiragana* was first created as a reading aid for Chinese characters around the ninth century (Hasegawa, 2015). *Hiragana* went through some modifications after its initial creation, and the modern set of *hiragana* used in present-day Japanese consists of 46 characters. *Hiragana* are phonograms, each of which represents a specific syllable, and are used for words in various lexical categories such as nouns, verbs, and adjectives. *Hiragana* is also used for particles and conjugational endings.

Another set of characters used in Japanese is *katakana*. *Katakana* characters were also developed from Chinese characters around the ninth century as a reading aid. In modern Japanese, *katakana* are mostly used for loanwords from non-Japanese languages and also for onomatopoeic expressions. Another major set of characters in Japanese is *kanji* characters. The literal meaning of *kanji* (漢字) is 'characters of the Han Dynasty of ancient China' (Hasegawa, 2015), and approximately 50,000 *kanji* characters exist in modern Japanese (Morton et al., 1992; Taylor & Taylor, 1995). However, most of those 50,000 characters are not in use in everyday life in Japan. The Agency of Cultural Affairs of Japan sets 2,136 characters as *joyo kanji*, which translates to English as 'commonly used kanji' (Agency of Cultural Affairs, 2010). The *Joyo kanji* list also serves as the guideline to select *kanji* characters that are to be included in school education as part of the government-set nationally standardized curriculum. *Kanji* characters are mostly used for words that were incorporated into Japanese from Chinese as well as Japanese nouns and the non-conjugational components of various lexical items such as verbs and adjectives.

The following sentences demonstrate the typical mixture of *hiragana*, *katakana*, and *kanji* in Japanese sentences.[1]

(1) 食堂でアイスクリームを買いました。
Shokudoo de aisukuriimu o kaimashita.
'I bought ice cream at the cafeteria.'

(2) 田中さんはロンドンにいます。
Tanaka san wa rondon ni imasu.
'Mr./Ms. Tanaka is in London.'

In (1), *shokudoo* (食堂) 'cafeteria' and the non-conjugational part of the verb *kaimashita* (買いました) 'bought' are written in *kanji*. Particles such as the location marking *de* (で), the object-marking *o* (を), and the conjugational component of the verb *kaimashita* are written in *hiragana*. *Katakana* is used for the loanword *aisukuriimu* (アイスクリーム) 'ice cream,' in accordance with the phonological transformation from the English *ice cream* to the Japanese *aisukuriimu*. As shown in (1), the three major types of characters are used within a continuous sentence in Japanese, and the choice of the type of characters is determined at the word level depending on the property of each word. Similarly, in (2), the personal proper noun *Tanaka* (田中) is written in *kanji*, the city name *rondon* (ロンドン) 'London' in *katakana*, and other components in *hiragana*.

2   The present study

The Roman alphabet is typically not included as a major character type when the writing system of Japanese is discussed. However, close observation of the written language in Japanese shows that the Roman alphabet is frequently used as part of the Japanese writing system, and the Roman alphabet can be considered to be one of the major character types used within the Japanese writing system. The present paper explores the usages of the Roman alphabet used within Japanese, and discusses the communicative and sociolinguistic significance surrounding the usages of the Roman alphabet in Japanese.

This paper's organization is as follows; the usages of the Roman alphabet are categorized into several categories, and example sentences are listed in each category

---

[1] The present study follows the convention of transliteration that is typically adopted in linguistic studies in North America for its Romanization of Japanese. In this convention, the long vowel *oo* as in *gakkoo* (がっこう) 'school' is Romanized as *oo*, not *ō* or *ou*. Similarly, *ee* as in *sensee* (せんせい) 'teacher' is Romanized as *ee*, not *ē* or *ei*.

to demonstrate the usages and illustrate the significance and implications of choosing to use the Roman alphabet instead of other types of characters in Japanese. The present study primarily focuses on qualitative analysis of examples, and example sentences are selected from various sources in Japanese from online and printed materials. Some selected examples from the *Balanced Corpus of Contemporary Written Japanese* (Maekawa, 2008) are also included in the present study.

As for the selection of data, the present paper mostly focuses on the usages of the Roman alphabet within Japanese sentences, not the Roman alphabet used at the word or the phrase level. Various studies report that English words are used with the original spelling in the Roman alphabet, especially in advertisements for commercial products and also in popular culture such as popular songs and titles of movies and TV shows (e.g., Honna, 1995; Kay, 1995; Kinjo, 1998; Kubota, 1998; Stanlaw, 2004; Daulton, 2008; Irwin, 2011; Sung & Mitsudo, 2016). However, when English words written in the Roman alphabet are used independently of other sentential components in Japanese, the examples may simply be considered as the usages of English, rather than English words in the Roman alphabet appearing within the Japanese language. For this reason, even though the increased use of English in Japan is an intriguing phenomenon and relevant to the present paper's topic, considering the abundance of existing literature on the use of English in Japan, this present paper attempts to contribute to the existing literature by focusing on the Roman alphabet used within Japanese along with other types of characters in Japanese.

3   The usage of the Roman alphabet in Japanese

Various types of usages of the Roman alphabet within the Japanese writing system are explored in this section.

3.1   Roman letters used for acronyms

One of the common types of usages of the Roman alphabet in Japanese is for abbreviated expressions originated in Western languages, most of which are from English. When loanwords from foreign languages are incorporated into the vocabulary pool of Japanese, they are conventionally written in *katakana*. However, when acronyms that are written in the Roman alphabet in foreign languages are used as loanwords in Japanese, the letter combinations in the original language are preserved when they are used in Japanese. Those acronyms in Roman letters are typically used as nouns in Japanese, both as common nouns and proper nouns. Tables 1 and 2 list some select examples of acronyms in the Roman alphabet that are used as common nouns and proper nouns in Japanese, respectively.

Table 1: Acronyms in Roman Letters used in Japanese [Common Nouns]

Acronyms	Original Forms
AED	automated external defibrillator
ATM	automated teller machine
CD	compact disc
DNA	deoxyribonucleic acid
DVD	digital video disc
LPG	liquefied petroleum gas
PR	public relations

Table 2: Acronyms in Roman Letters used in Japanese [Proper Nouns]

Acronyms	Original Forms
FBI	Federal Bureau of Investigation
IMF	International Monetary Fund
NHK	*Nippon Hoosoo Kyookai* 'Japan Broadcasting Corporation'
NTT	Nippon Telegraph and Telephone
OECD	Organisation for Economic Co-operation and Development
OPEC	Organization of the Petroleum Exporting Countries
USJ	Universal Studios Japan
WHO	World Health Organization

The following example demonstrates how acronyms in Roman letters are used mixed with other types of characters in Japanese.

(3) 現在では、オンライン・リアルタイム処理は鉄道や航空機の座席予約、銀行の ATM、工場での機械制御などで利用されている。

Genzai de wa, onrain-riarutaimu shori wa tetsudoo ya kookuuki no zaseki yoyaku, ginkoo no ATM, koojoo deno kikai seigyo nado de riyoo sarete iru.

'Currently, online real-time processing is used in reservation systems for trains and airplanes, bank ATMs, and machine control systems in factories.'

([BCCWJ: PB43_00081] *Gendai keeee to nettowaaku*, Kishikawa & Hoshino, 2004)

In (3), the acronym *ATM* 'ATM' is written in Roman letters and used with other components in characters that are unique to the Japanese language such as *hiragana*, *katakana*, and *kanji*. Also, since *ATM* is a commonly-known word among speakers of Japanese, no further explanations are provided in the text.

When acronyms in the Roman alphabet are used in Japanese sentences, the full Japanese translation of the form before abbreviation is often written in parenthesis

after the acronym. The following example demonstrates such listing of translation in Japanese. Translations for acronyms in Japanese tend to be provided when the acronym is a proper noun, or a less commonly known phrase.

(4) 当時の占領軍に該当する外圧といえば、いまなら IMF(国際通貨基金)である。

Tooji no senryoogun ni gaitoo suru gaiatsu to ie ba, ima nara IMF (kokusai tsuuka kikin) de aru.

'What is equivalent to the occupying force back then is the IMF (International Monetary Fund) in modern times.'

([BCCWJ: PB53_00657] *Shisan hookai*, Ota, 2005)

In (4), after *IMF* in Roman letters, *kokusai tsuuka kikin* (国際通貨基金) 'International Monetary Fund' is listed to provide further information about the acronym, and when the acronym reappears in later parts of the text, the full translation in parenthesis is not repeated.

Similarly, the translation in Japanese occasionally precedes the acronym and the acronym in parenthesis follows the Japanese translation.

(5) 健康とは「単に病気あるいは虚弱でないということではなく，身体的にも精神的にも社会的にも完全に良好な状態であること」と世界保健機関(WHO)は定義している。

Kenkoo to wa "tan ni byooki aruiwa kyojaku de nai to iu koto de wa naku, shintaiteki ni mo seeshinteki nimo shakaiteki ni mo kanzen ni ryookoo na jootai de aru koto" to sekai hoken kikan (WHO) wa teegi shite iru.

'The World Health Organization (WHO) defines health as "health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity."'

([BCCWJ: LBt4_00058] *Wakari yasui kaigo no tame no eeyoo to choori*, Yoshida & Sumii, 2005)

In (5), *WHO* in parenthesis follows its translation in Japanese *sekai hoken kikan* 'World Health Organization.' In later parts of the text, the acronym *WHO* is repeated in the main body of the text without being accompanied with the translation in Japanese. Examples (4) and (5) demonstrate typical ways of introducing acronyms in the Roman alphabet when they first appear in a written text in Japanese. When the acronym may not be familiar to the potential readers, the acronym is accompanied with its Japanese translation, and when it reappears in the text, it is used without being accompanied with the translation.

As for the way acronyms in the Roman alphabet are read with other components in the sentence in Japanese, the pronunciation is based on the way in which Roman letters are pronounced within the Japanese sound system, which can be written using *katakana*. For example, when the word *CD* 'compact disc' is pronounced in a Japanese sentence, it is pronounced as *shiidii*, not /siːdiː/ as in the original English pronunciation. The conversion from English pronunciation to the pronunciation that conforms to the Japanese sound system and can be accurately transcribed using *katakana* is the standard procedure for using acronyms in the Roman alphabet in Japanese, and even in the Japanese language textbooks for learners of Japanese, students are instructed to pronounce acronyms in *katakana* pronunciation. For example, in *Genki I* (Banno et al., 2011), which is an elementary-level Japanese language textbook, the word *CD* is introduced with *shiidii* (シーディー) in *katakana* as a pronunciation guide in the vocabulary list. It should also be noted that for acronyms that have unique reading patterns that are different from simply reading out the Roman letters, such as OPEC read as /oʊpek/ in English, the pronunciation conforms with the Japanese sound system and is enunciated as *opekku* (this could also be written as オペック) in Japanese.

3.2   Proper nouns written in the Roman alphabet in Japanese

In addition to proper nouns that are acronyms, certain types of proper nouns in non-Japanese languages are written in the Roman alphabet without being converted to *katakana*.

3.2.1   Roman alphabet used for reference purposes

The first types of examples discussed here are proper nouns such as personal names and place names originally written in the Roman alphabet. The following examples are from news articles that include proper nouns in foreign languages.

(6) フランス大統領府は 17 日、エマニュエル・マクロン(Emmanuel Macron)大統領が新型コロナウイルス検査で陽性と診断されたと発表した。今後 1 週間自主隔離を行うという。

Furansu daitooryoofu wa juunana nichi, emanyueru makuron (Emmanuel Macron) daitooryoo ga shingata korona uirusu kensa de yoosee to shindan sareta to happyoo shita. Kongo isshuukan jishu kakuri o okonau to iu.

'The Office of the President of France announced that President Emmanuel Macron has tested positive for the new-type coronavirus. He will self-quarantine for a week.'

(*Yahoo! News*, https://news.yahoo.co.jp/, accessed 12/17/2020)

(7) 気象当局によるとニューヨーク市の北 290 キロに位置するビンガムトン
(Binghamton)やペンシルベニア州リッチフィールド(Litchfield)で積雪が 1 メート
ルを超えた。

Kishoo tookyoku ni yoru to nyuuyooku shi no kita nihyaku kyuujukkiro ni ichi suru
bingamuton (Binghamton) ya penshirubania shuu ricchifiirudo (Litchfield) de sekisetsu
ga ichi meetoru o koeta.

'According to the Weather Bureau, snow accumulation exceeded one meter in
Binghamton, which is located 290 kilometers north of New York City, and also in
Litchfield, Pennsylvania.'

(*Yahoo! News*, https://news.yahoo.co.jp/, accessed 12/18/2020)

In (6), *Emmanuel Macron*, which is the proper noun for a French politician, is
written in *katakana* as in *emanyueru makuron* (エマニュエル・マクロン) and it is
followed by his name in the Roman alphabet in parentheses. Similarly, in (7),
Binghamton and *Litchfield*, which are proper nouns for city names in the United States,
are written in *katakana* as *bingamuton* (ビンガムトン) and *ricchifiirudo* (リッチフィ
ールド), followed by their original names in the Roman alphabet. It should be noted
that in (7), *nyuuyooku* (ニューヨーク) 'New York' and *penshirubenia* (ペンシルベニ
ア) 'Pennsylvania' are not followed by their Roman alphabet forms. The reason for this
inconsistency appears to be how well-known those proper nouns are among the
intended readers, who are speakers of Japanese that reside in Japan. Many foreign
place names are well-known in Japan, and major states in the United States such as
New York and Pennsylvania have been incorporated into the vocabulary pool of
Japanese as *katakana* loanwords, as demonstrated by the fact that *nyuuyooku* and
penshirubenia in *katakana* are both included in the *Daijisen Dictionary* (Matsumura,
2020), which is one of the leading Japanese monolingual dictionaries. However, less-
commonly-known city names such as *Binghamton* and *Litchfield* are unlikely to be
recognized by speakers of Japanese, and it appears that the forms in the original Roman
alphabet are provided in the parentheses as supplementary information that helps
readers make a connection between the *katakana* forms and the original forms.

In addition, personal proper nouns in languages that are written in the Roman
alphabet are also written in the original format in scholarly publications in Japanese.
The following excerpt includes such examples of foreign names written in the Roman
alphabet in a scholarly article.

(8) Gumperz (1982) は、異文化を持つ者同士の会話では、参加者が互いに異なる期待を持っているがために、誤解が生じるとし、Tannen (1993) は、文化により会話のスタイルが異なることを指摘している。

Gumperz (1982) wa, ibunka o motsu mono dooshi no kaiwa de wa, sankasha ga tagai ni kotonaru kitai o motte iru ga tame ni, gokai ga shoojiru to shi, Tannen (1993) wa, bunka ni yori kaiwa no sutairu ga kotonaru koto o shiteki shite iru.

'Gumperz (1982) claims that misunderstandings occur in conversations between participants with diverse cultural backgrounds because their expectations are different, and Tannen (1993) points out that conversation styles are different in different cultures.'

(*Nihonjin daigakusee no guruupu tooron ni okeru ketsuron seesee to shinkooyaku no yakuwari*, Otsuka, 2003)

(8) is from an article about linguistics, and the authors of the cited references are written in the Roman alphabet as in *Gumperz* and *Tannen*. This convention to use the Roman alphabet for references in scholarly publications seems to be developed from the necessity to maintain the original spelling information for the authors' names in the cited documents. If the authors of the cited documents were converted to *katakana*, it would be challenging for the readers to identify the cited documents that were written in non-Japanese languages. In addition, if each author attempts to convert non-Japanese names into *katakana*, the same author's name written in the Roman alphabet in the original language may be converted to *katakana* differently depending on the author's personal preferences, even though the conventions for converting English to *katakana* are mostly uniform among native speakers of Japanese. For those reasons, names in non-Japanese languages appear to be conventionalized to be written in their original form (the Roman or other Western alphabets) in scholarly publications in Japanese.[2]

Furthermore, it should also be noted that even when the author of the cited document has a name originally written in the Roman alphabet, when the cited document is published in Japanese with the author's name written in *katakana*, the author's name of the cited document is also written in *katakana*. Observe the following excerpt.

---

[2] Names in East Asian languages, especially Chinese names, are conventionalized to be written in *kanji* in scholarly documents in Japanese. However, names in Asian languages in the entertainment field tend to be written in *katakana*. For more details, see Kimura (2013), Mochizuki et al. (2014), etc.

(9) 林(2008)は、言語管理理論(Jernudd & Neustupný 1987, Neustupný 1985a, 1985b, 1994a，ネウストプニー1994b, 1995)に沿って、日本語非母語話者が何を留意したことで聞き返しが実施されたのかを明らかにしている。

Hayashi (2008) wa, gengo kanri riron (Jernudd & Neustupný 1987, Neustupný 1985a, 1985b, 1994a, neusutopunii 1994b, 1995) ni sotte, nihongo hi bigowasha ga nani o ryuui shita koto de kikikaeshi ga jisshi sareta no ka o akiraka ni shite iru.

'Hayashi (2008) follows the Language Management Theory (Jernudd & Neustupný 1987, Neustupný 1985a, 1985b, 1994a, *neusutopunii* 1994b, 1995), and reveals what triggered non-native speakers of Japanese to make reiteration requests.'

(*Kikikaeshi no sutorategii to mondai kaiketsu*, Hayashi, 2009)

In (9), *Neustupný* in the Roman alphabet, which refers to the Czech-Australian linguist J.V. Neustupný, is used for some of the citations, but *neusutopunii* (ネウストプニー) in *katakana* is also included in the citation list. *Neustupný* in the Roman alphabet and *neusutopunii* in *katakana* in (9) both refer to the same linguist, J.V. Neustupný. This inconsistency is resulting from the difference in the languages used in the cited documents. The cited documents with *Neustupný* in the Roman alphabet are written in English, and the author's name is also written in the Roman alphabet in those articles. On the other hand, the documents cited with *neusutopunii* in *katakana* are written in Japanese*,* and the author's name is also listed in *katakana* in the cited articles. Therefore, even when the author's name is written in the Roman alphabet in the original language if the same name is written in *katakana* in the cited document in Japanese, the name remains in *katakana* in academic citations.

3.2.2   Roman alphabet used for product names

Another type of usage of the Roman alphabet within Japanese is for specific proper nouns such as product and service names. Certain product names are marketed in the Roman alphabet without being converted into *katakana* for the Japanese market. In addition, those product names are also written in the Roman alphabet when they appear in Japanese sentences. The following excerpt is from a website about mobile phone services. The page includes information about the *iPhone*, which is also marketed as *iPhone* in the English-speaking market.

(10) iPhone のモデルを比較する: 容量やディスプレイ、サイズなどドコモから発
売中の iPhone のすべてのモデルの機能と仕様を比較できます。

iPhone no moderu o hikaku suru: Yooryoo ya disupurei, saizu nado dokomo kara hatsubaichuu no iPhone no subete no moderu no kinoo to shiyoo o hikaku dekimasu.

'Compare iPhone models: You can compare the storage space, display, and size of all iPhone models sold by Docomo.'

(https://www.nttdocomo.co.jp/iphone/, accessed 12/27/2020)

As shown in (10), the product name *iPhone* is written in the Roman Alphabet within a Japanese sentence without being converted into *aifon* (アイフォン) in *katakana*, and the readers are also expected to understand the product name without relying on *katakana*. However, it should also be noted that when *iPhone* in the Roman alphabet is read aloud in Japanese sentences, it is pronounced as *aifon*, conforming with the sound system of Japanese, as it would be transcribed in *katakana*.

Furthermore, the usage of the Roman alphabet for commercial products is not limited to products with foreign origins. The following excerpt is from a webpage about the connected service for Toyota vehicles.

(11) T-Connect のユーザー様専用のサービスサイトや設定・操作マニュアルなど
、お役に立つ情報をご紹介します。

T-Connect no yuuzaa sama senyoo no saabisu saito ya settee soosa manyuaru nado, oyaku ni tatsu joohoo o goshookai shimasu.

'[This page] introduces useful information for *T-Connect* users such as the service website, settings, and operation manuals.'

(https://toyota.jp/tconnectservice/user/?padid=ag341_from_owner_tconnect_user/, accessed 12/27/2020)

In (11), *T-Connect*, which is the proper noun for a service offered by Toyota, is written in the Roman alphabet in a Japanese sentence. This example shows that even though Toyota is a Japanese company and the service is also for the Japanese domestic market, the Roman alphabet is used for the service's proper noun and the assumption is that the usage of the Roman alphabet does not block communication for the Japanese-speaking audience. Also, the use of the Roman alphabet for *T-Connect* contrasts with the use of *katakana* for general loanwords from English such as *yuuzaa* (ユーザー) 'user' and *saabisu* (サービス) 'service' in the same sentence, which are considered to be already integrated into the vocabulary pool of Japanese as *katakana* loanwords.

3.2.3   Common nouns written in the Roman alphabet in Japanese: IT-related vocabulary

The usage of the Roman alphabet in Japanese may initially appear to be predominantly for acronyms and proper nouns; however, there are some somewhat conventionalized usages of the Roman alphabet for common nouns. This subsection explores those usages of the Roman alphabet for common nouns and expressions. (12) is an excerpt from the website about service schedules for *Tokyo Gas*, which is the provider of natural gas in the Tokyo metropolitan area.

(12) ガス作業希望時間が午前の場合: ガス作業希望日前日の 17:00 まで Web での
       変更が可能です。

   Gasu sagyoo kiboo jikan ga gozen no baai: Gasu sagyoo kiboobi zenjitsu no 17:00 made web deno henkoo ga kanoo desu.

   'If you prefer a.m. gas work: It is possible to change (the schedule) until 17:00 of the day before your preferred day.'

   (https://home.tokyo-gas.co.jp/procedure/moving/change.html, 2/22/2021)

In (12), *Web* 'web' is written in the Roman alphabet in a sentence in Japanese, even though its *katakana* counterpart *webu* (ウェブ) 'web' already exists as a *katakana* loanword in the vocabulary pool of Japanese. Also, *gasu* (ガス) 'gas,' which has been part of the Japanese vocabulary for a considerable length of time, is written in *katakana*.

The next excerpt also includes an IT-related term written in the Roman alphabet.

(13) 備北広域農業普及指導センターは、7 月 26 日に開催された新・農業人フェ
       ア（場所　東京ドームプリズムホール）に online で参加しました。

   Bihoku kooiki noogyoo shidoo sentaa wa, 7 gatsu 26 nichi ni kaisai sareta shin noogyoojin fea (basho tookyoo doomu purizumu hooru) ni online de sanka shimashita.

   'Larger Bihoku area Agriculture Advocacy and Training Center participated in the New Agriculture People Fair (location: Tokyo Dome Prizm Hall) online on July 26.'

   (https://www.pref.okayama.jp/site/587/674203.html, accessed 2/25/2021)

Example (13) was taken from the newsletter by the Okayama Prefecture Government. In it, *online* in the Roman alphabet is used instead of its *katakana* counterpart *onrain* (オンライン). This type of usage of the Roman alphabet instead of using loanwords in *katakana* is frequently observed for the vocabulary items related to the IT field, and other well-known IT-related terms such as *chat*, *spam*, and *bot* are also often written in the Roman alphabet in sentences written in Japanese. Those IT-related terms were not in common usage until recently, since only 25% of households in Japan

had personal computers at home in 1997, and even in 2005, the percentage was 65% (Okabe, 2009). It is speculated that specialized terminologies from foreign languages that were incorporated into Japanese recently may have a tendency to be written in the Roman alphabet without being converted to *katakana*, and the same tendency may also exist for non-specialized loanwords that were recently incorporated into Japanese.

3.2.4   Common nouns written in Roman alphabet in Japanese: general expressions

The use of the Roman alphabet for non-proper nouns is not limited to IT-related terminologies. Some commonly-recognized expressions in English are also written in the Roman alphabet in Japanese sentences without being converted into *katakana*. Observe the following two excerpts.

(14) 電気とガスをまとめておトクにしたいお客さま: 電気とガスをまとめると、ガスセット割がついて電気代が年間約 1,200 円 OFF。

Denki to gasu o matomete otoku ni shitai okyakusama: Denki to gasu o matomeru to, gasusetto wari ga tsuite denkidai ga nenkan yaku 1,200 en OFF.

'For customers who want to bundle electricity and gas and save money: If you bundle electricity and gas, it will be 1,200 yen off with the gas-set discount.'

(https://www.tepco.co.jp/ep/private/plan/index-j.html, accessed 2/21/2020)

(15) 経営学部の国際交流・交換留学ページを OPEN しました。これから経営学部の国際交流・交換留学制度については、このサイトを通じて最新の情報を発信していきます。

Keeee gakubu no kokusai kooryuu kookan ryuugaku peeji o OPEN shimashita. Korekara keeee gakubu no kokusai kookan ryuugaku seedo ni tsuite wa, kono saito o tsuujite saishin no joohoo o hasshin shite ikimasu.

'We have opened the webpage for international exchange and exchange programs for the Department of Management. The latest information about the international exchange and exchange study abroad programs will be uploaded on this website from now on.'

(https://www.meiji.ac.jp/keiei/exchange/topics/2010/6t5h7p0000001v4j.html, accessed 2/27/2020)

In (14), *OFF* in English is written in the Roman alphabet. In this particular example, even though the same information can be conveyed by using Japanese words such as *waribiki* 'discount' or the *katakana* loanword *ofu* (オフ) 'off,' *OFF* in the Roman alphabet is selected as the vocabulary item to convey the information to the reader. Similarly, in (15), which is from a university's website about its study abroad programs, *OPEN* in the Roman alphabet is used to form a past-tense verb *OPEN shita* 'opened,' even though *oopun* (オープン) 'open' in *katakana* is already a commonly-used

loanword in Japanese. Furthermore, the same information can also be conveyed with Japanese vocabulary such as *dekita* 'to be made, to be completed' in (15).

The usages of the Roman alphabet in (14) and (15) seem to contrast with the cases we observed in (12) and (13). The IT-related terms included in (12) and (13) were loanwords that were newly introduced to the vocabulary pool of Japanese. However, general expressions in English such as *off* and *open* have been part of the Japanese vocabulary pool for a long time, and as mentioned earlier, *katakana* are typically used when words from non-Japanese languages are used in Japanese in conformity with the phonological system in Japanese. However, the role of *katakana* loanwords is not simply limited to enabling the usage of foreign words within Japanese. *Katakana* loanwords are often chosen even when Japanese equivalents already exist in the vocabulary pool of Japanese, and they have various sociolinguistic functions that are unique to loanwords, especially when the loanwords are from European languages including English. For example, Kay (1995) argues that "[l]oanwords are often associated with a sophisticated, Western lifestyle, and may be used in place of Japanese words of equivalent meaning because of their foreign appeal" (p. 74). As Kay argues, loanwords are often used to portray certain sophisticated and advanced images, even though the same information can be expressed with non-loanword Japanese vocabulary. Those loanwords used for the purpose of displaying sophisticated images are written in *katakana* in most cases; however, by using the Roman alphabet instead of *katakana*, the intended sophisticated image seems to be even more amplified, since the Roman alphabet carries more sophisticated appeal compared to *katakana,* which are considered to be fully part of the Japanese writing system.

In addition, even without an intention to portray a sophisticated image, using the Roman alphabet can be simply eye-catching in the text where the rest is predominantly written in Japanese characters. Examples of the usage of the Roman alphabet such as what we observed in (14) and (15) could also be from the writers' desire to make the phrases more eye-catching for the audience, and since almost all Japanese readers would know basic English expressions such as *off* and *open*, the understandability of the sentences will not be sacrificed by using the Roman alphabet as in (14) and (15).

3.2.5   Proper nouns in the Roman alphabet with preserved English structures

Some English expressions written in the Roman alphabet contain expressions that are not typically considered to be included in the vocabulary pool of Japanese as *katakana* loanwords. Observe the following excerpt.

(16) iPhone 12 Pro を購入: Apple Trade In を利用すると、最大 40,000 円割引に
なります。

iPhone 12 Pro o koonyuu: Apple Trade In o riyoo suru to, saidai 40,000 en waribiki ni narimasu.

'Purchasing iPhone 12 Pro: If you make use of *Apple Trade In*, you can save up to 40,000 yen.'

(https://www.apple.com/jp/shop/buy-iphone/iphone-12-pro, accessed 12/28/2020)

In (16), *Apple Trade In* is written in the Roman alphabet. *Apple Trade In* is a discount service for purchasing new phones with which buyers can use the value of their old phones towards the payment for the new phone, and the same service is also marketed as *Apple Trade In* in the English-speaking market. What is noticeable here is that the English expression *trade-in* is not typically considered to be a loanword used in Japanese. For example, *trade-in* in English would be *toreedoin* (トレードイン) in *katakana*, but it is not included in the *Daijisen* Japanese monolingual dictionary (Matsumura, 2020). *Apple Trade In* in (16) is an example that requires readers to understand an English phrase that is not incorporated into the vocabulary pool of Japanese. Needless to say, *trade-in* is a basic phrase for speakers of English; however, even though *trade-in* in (16) is used as part of a proper noun, (16) can be interpreted as an example in which a neologism from English written in the Roman alphabet is incorporated into a Japanese sentence mostly written in Japanese characters.

The following excerpt also includes an example of the Roman alphabet used as part of a proper noun. (17) is from an announcement about a music and dramatic reading event featuring the *Shingeki no Kyojin* series, a well-known manga and anime series in Japan. *Shingeki no Kyojin* is known as *Attack on Titan* in the English-speaking market.

(17) 2017 年 10 月 29 日に東京国際フォーラムホールで開催される「進撃の巨人」Reading & Live Event Orchestra 「Attack 音体感２」の各公演のチケット一般販売が決定いたしました!

2017 nen 10 gatsu 29 nichi ni tookyoo kokusai fooramu hooru de kaisai sareru "Shingeki no Kyojin" Reading & Live Event Orchestra "Attack on taikan 2" no kaku kooen no chiketto ippan hanbai ga kettee itashimashita!

'We announce that the tickets for each performance of "Attack on Titan" Reading & Live Orchestra "Attack Sound Experience 2," which will be held at Tokyo International Forum Hall on October 29, 2017, will be on sale to the general public!'

(https://shingeki.tv/news/archives/3381/, accessed 1/15/2021)

In (17), "*'Shingeki no Kyojin' Reading & Live Event Orchestra 'Attack on taikan 2'*" is the proper noun for the event, and the whole name is repeatedly used on other pages of the website and also in other advertisement materials. As shown in (17), *Reading & Live Event Orchestra* is written in the Roman alphabet, and the readers are expected to read the English component in the Japanese sentence. In addition, *Attack* is also in the Roman alphabet, which appears to be taken from *Shingeki no Kyojin*'s English title *Attack on Titan.* Unlike the example of the use of the Roman alphabet in (16), *katakana* versions of the English words included in (17), which are *reading*, *live*, *event*, *orchestra*, and *attack* are all included in the *Daijisen* Japanese dictionary and considered to be part of the modern Japanese vocabulary. However, the usage of English in the Roman alphabet in (17) demonstrates that English is presented as is, and the event is marketed with the assumption that the target audience would understand English in the Roman alphabet without relying on *katakana*.

3.2.6   Other miscellaneous usages of the Roman alphabet in Japanese

It should be noted that there are some other miscellaneous usages of the Roman alphabet in texts that are predominantly written in Japanese. The present paper does not discuss those miscellaneous usages in detail, but those usages are also considered to be fully integrated into the writing system of the Japanese language.

One type of usage of the Roman alphabet that was not discussed in earlier parts of the present paper is the Roman alphabet simply used as symbols, especially in listing and bullet points. Various types of characters are used as the bullet points for listing in Japanese. The characters commonly used for listing in Japanese are numbers in Arabic numerals and *kanji* (*1, 2, 3,…* or 一, 二, 三,...), *katakana* characters (ア, イ, ウ,… or イ, ロ, ハ,...[3]), and the Roman alphabet in alphabetical order (*a, b, c,…*). In addition, Roman numerals in the Roman alphabet (*I, II, III,…*) are also used for listing in Japanese. Regarding the usage of the Roman alphabet as symbols, login IDs and passwords on Japanese websites are typically in the Roman alphabet and/or Arabic numerals, even when the entire webpage is predominantly in Japanese. In addition, URLs and email addresses used in Japan are predominantly in the Roman alphabet and Arabic numerals.

Finally, one type of usage of the Roman alphabet that is prevalent in Japan but often overlooked is the usage of Romanized Japanese for typing Japanese on electronic devices. There are several different methods to type Japanese on a computer such as the Romanization input method and the *kana* input method when the QWERTY keyboard is used. When the Romanization input method is selected, the intended text in Japanese is typed in the Romanized format on the QWERTY keyboard. On the other hand, with the *kana* input method, each *hiragana* or *katakana* character is assigned to

---

[3] The usage *of katakana* for listing in the ア, イ, ウ,… order is based on the *hiragana*/*katakana* ordering system in modern Japanese. The イ, ロ, ハ,… order is based on an ordering system used in classical Japanese, which was dominant prior to the implementation of the modern order.

a key on the QWERTY keyboard, and those characters are input directly without being Romanized. However, even though the two input methods, the Romanization method and the *kana* method, are both available on computers sold in Japan, a survey by Endo (2015) shows that 93.1% of the respondents answered that they use the Romanization method when using computers with the QWERTY keyboard. Therefore, even though Japanese texts typed on computers are mostly in Japanese characters such as *hiragana*, *katakana*, and *kanji*, the majority of them have gone through Romanization during the input process. The usage of the Roman alphabet for typing is typically not discussed in relation to the writing system of Japanese, but it has become an essential component in the writing practice in Japan and has become the primary means of writing in modern day Japan. As for the input methods for Japanese on electronic devices, it should be noted that the input methods used on devices without physical QWERTY keyboards do not involve the Romanization process. For example, a survey by Nagasawa (2017) shows that when Japanese college students type Japanese on smartphones, more than 90% of them use the *flick* method or the *toggle* method, both of which do not require the Romanization process for typing Japanese characters.[4] In addition, a study by Noborimoto et al. (2021) indicates that high school students in Japan can type Japanese faster on smartphones compared to typing Japanese on a computer using the QWERTY keyboard. However, what Nagasawa and Noborimoto et al. have found does not indicate that Japanese people stopped using the Roman alphabet for writing Japanese electronically since the usage of computers with QWERTY keyboards are still very common in Japan, but we should also be aware that Japanese texts typed on smartphones are typically input without going through the Romanization process.

4   Discussion

The present study has explored the usages of the Roman alphabet within the writing system of modern Japanese. As the excerpts explored in this paper demonstrated, even though the Roman alphabet is not typically listed as a character type in Japanese, it has already been an essential component in the writing practice in Japan.

It seems that there are several major factors that are relevant to the usage of the Roman alphabet in modern Japanese, and those factors have also been contributing to the increased usage of the Roman alphabet within Japanese. The factor that is discussed first is the shift from vertical writing to horizontal writing in modern Japanese. Yanaike (2003a, 2003b) reports that horizontal writing originally did not exist in Japanese, but it emerged due to the interaction with the West during the late Edo period and the early Meiji period. Despite the emergence of horizontal writing in the late 19th century, the writing practice in Japan remained mostly vertical, especially in

---

[4] For the details about the *flick* and the *toggle* input methods, see Nagasawa (2017).

printed media such as books, magazines, and newspapers. As of the early 2020s, the majority of printed media in Japan remain in the vertical writing style.

However, due to the recent development of IT technology, the writing practice has largely shifted towards horizontal writing, since IT-based content such as websites, emails, text messaging, and social media are predominantly based on horizontal writing. Due to this shift from vertical writing to horizontal writing, it has become significantly easier to integrate words written in the Roman alphabet into Japanese compared to doing so in vertical writing. For example, in order to integrate the English word *open* in a Japanese sentence written horizontally, it can be simply written as *open* horizontally as shown in (a) in figure 1. On the other hand, when the sentence is written vertically, the whole alphabet sequence for *open* must be turned by 90 degrees as in (b), or each letter in the Roman alphabet must be written vertically, following the pattern of other Japanese characters as in (c). The sentences in (a), (b), and (c) all read *mise ga open shita* 'the store opened.'

(a) 店が open した。

(b) 店が open した。

(c) 店が open した。
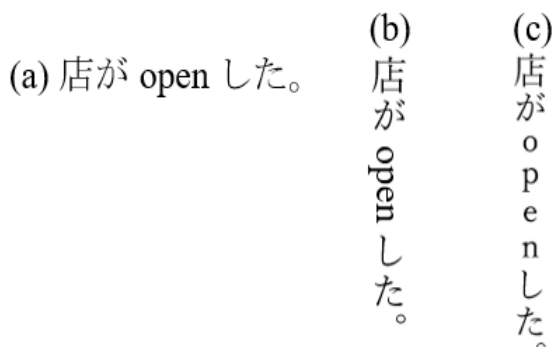
Figure 1: Horizontal and Vertical writing styles in Japanese
(Mise ga *open* shita. 'The store opened.')

As demonstrated in Figure 1, using the English word *open* in the Roman alphabet may not impose a significant problem in horizontal writing, but when the text is written vertically, it is difficult to use *open* in the Roman alphabet without modifying how the word is written in English sentences. As mentioned earlier, due to the recent development of IT technology, horizontal writing is becoming more common, and the opportunities to use the Roman alphabet within Japanese are more widely available because of the shift to horizontal writing.

Another relevant factor is the burden of typing *katakana* versions of English words compared to directly typing English words in the Roman alphabet. For writing English words within Japanese sentences, depending on the complexity of *katakana* combinations and the original spelling in English, the amount of work required for typing the word may be lower when the word is written in the Roman alphabet. For instance, in order to type in the word *website*, when *website* is written in the Roman

alphabet, the word can simply be input as *website*. However, if *website* is converted into *katakana*, the writer must input *webusaito* in Romanized Japanese and then convert it to *katakana*. Similarly, in order to type in the word *open*, it can be simply input as *open* in the Roman alphabet, but to type in *oopun* (オープン) in *katakana*, the typing process involves the long vowel symbol (ー) and the whole word must be input as *o-pun,* the hyphen corresponding to the long vowel symbol.

In addition, the use of the Roman alphabet in Japanese clearly relates to the recent globalization and the increased exposure to written information from outside Japan due to the development of IT technology. According to a survey by *W3Techs* (2021), 61.7% of the web content in the top 10 million websites was in English as of July 2021, while the content in Japanese was 1.9% in those top 10 million websites. As this discrepancy between English and Japanese web content shows, users of the Internet can access a significantly larger amount of information through English, and when the words or expression used in English is included in sentences written in Japanese, writing those in the Roman alphabet may be the most convenient option if they are not fully conventionalized as *katakana* loanwords in Japanese. In addition, with the recent surge of the popularity of social media such as *Facebook*, *Twitter, Instagram*, and *TikTok*, Internet memes tend to be shared globally in a timely manner, and users with various language backgrounds also comment on those shared contents. Christiansen (2016) reports that even novice-level speakers of English tend to communicate in English in online global communities, and such experiences by Japanese speakers may increase the use of the Roman alphabet when they write in Japanese.

Finally, it must be noted that the pronunciation of words written in the Roman alphabet within Japanese is still subject to an adaptation to the Japanese phonological system, and it is likely that such *katakana*-like pronunciation will continue to be prevalent for words written in the Roman alphabet in the foreseeable future. As Martin (2004) points out, *katakana* characters are used as pronunciation guides in English education in Japan. In addition, the conventions for converting English words into their *katakana* counterparts are highly standardized across speakers of Japanese (Quackenbush, 1977; Kay, 1995; Irwin, 2011; Nishi & Xu, 2013; Tsujimura, 2013; Hasegawa, 2015, etc.). For those reasons, speakers of Japanese who received school education in Japan are mostly capable of converting English words into *katakana* in a uniform manner, and they are also capable of applying the conventions to English words they have never seen. Therefore, even with the increased usage of the Roman alphabet in written Japanese, it is expected that the phonological realization of words written in the Roman alphabet will follow the *katakana* conventions when they are read aloud within Japanese.

5   Conclusion

The present study has explored the usages of the Roman alphabet within the writing system of Japanese. As demonstrated by the examples in this paper, the Roman alphabet has been integrated into the modern writing system of Japanese, and the use of the Roman alphabet is expected to increase in the future, especially with the recent development of IT technology and also due to the ever-increasing interaction with languages and cultures outside of Japan. As Yanaike (2003a, 2003b) reports, horizontal writing emerged in Japanese in the 19th century because of the interaction with the West, but the writing practice mostly remained vertical for paper-based printed materials. However, with the recent surge of electronic communication, the writing practice in Japanese is currently going through the transition from vertical writing to horizontal writing, and the transition also allows writers to use foreign words originally in the Roman alphabet without converting them into *katakana* loanwords. In addition, because of the Japanese government's recent policy change, English has been included as an official subject in elementary schools throughout Japan from the 2020–2021 academic year, which used to be from the first year of middle school (Ministry of Education, Culture, Sports, Science and Technology, 2016). This indicates that speakers of Japanese are exposed to English, which uses the Roman alphabet, from younger ages compared to the past. Considering the shift from vertical writing to horizontal writing, the influence of foreign languages written in the Roman alphabet, and the rapidly increasing global communication, it is highly likely that we will see more words written using the Roman alphabet within the writing system of Japanese in the future.

References

Agency for Cultural Affairs 文化庁 (2010). *Jooyoo kanji hyoo* 常用漢字表 *[Joyo kanji list].* Retrieved November 20, 2020, from https://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/kijun/naikaku/pdf/joyokanjihyo_20101130.pdf

Banno, E., Ikeda, Y., Ohno, Y., Shinagawa, C., & Tokashiki, K. (2011). Genki: *An integrated course in elementary Japanese I*. Tokyo, Japan: Japan Times.

Christiansen, T. (2016). The Internet as a global speech community: Towards Plurilingualisms and English Lingua Franca. *Lingue e Linguaggi*, *19*, 77-96.

Daulton, F. E. (2008). *Japan's built-in lexicon of English-based loanwords*. Clevedon/Buffalo/Toronto: Multilingual Matters.

Endo, S. 遠藤諭 (2015). *Nihonjin wa otona ni naru to roomaji nyuuryoku ni naru rashii* 日本人は"大人"になるとローマ字入力になるらしい *[Japanese people use Romanized input when they become adults]*. Retrieved December 2020, from *Weekly Ascii*. https://weekly.ascii.jp/elem/000/002/631/2631699/.

Hasegawa, Y. (2015). *Japanese: A linguistic introduction*. Cambridge: Cambridge University Press.

Hayashi, R. 林里香 (2009). Kikikaeshi no sutoratejii to mondai kaiketsu: Nihongo hibogowasha ni yoru choosee keekaku dankai no kinoo to hyoogen keeshiki no sentaku 聞き返しのストラテジーと問題解決 - 日本語非母語話者による調整計画段階の機能と表現形式の選択 [Requests for clarification and problem solving : Functions and forms chosen by non-native speakers in the stage of adjustment plan]. *Chiba Daigaku Daigakuin Jinbun Shakaikagaku Kenkyuuka Kenkyuu Purojekuto Hookokusho* 千葉大学人文社会科学研究科研究プロジェクト成果報告書, *218*, 1-17.

Honna, N. (1995). English in Japanese society: Language within language. *Journal of Multilingual & Multicultural Development*, *16*(1-2), 45-62.

Irwin, M. (2011). *Loanwords in Japanese*. Philadelphia: John Benjamins Publishing Company.

Kay, G. (1995). English loanwords in Japanese. *World Englishes*, *14*(1), 67-76.

Kimura, M. 木村麻衣子 (2013). Chuugokujin dantai chosha mee tenkyo deeta no hyooki no sooi: Chuugoku, nihon, kankoku o chuushin ni 中国人・団体著者名典拠データの表記の相違：中国, 日本, 韓国を中心に [Differences in descriptions of Chinese personal and corporate name authority data: A Comparison between China, Japan and South Korea]. *Library and Information Science, 69*, 19-46.

Kinjo, F. 金城ふみ子 (1998). Daigaku kookoku ni okeru katakana hyookigo oyobi arufabetto hyookigo no shiyoo jookyoo: Choosa hookoku 大学広告におけるカタカナ表記語及びアルファベット表記語の使用状況 - 調査報告 [The use of katakana and the alphabet in ads for colleges]. *Bulletin of Center for Japanese Language, Waseda University* 早稲田大学日本語研究教育センター紀要, *10*, 97-118.

Kubota, R. (1998). Ideologies of English in Japan. *World Englishes*, *17*(3), 295-306.

Maekawa, K. (2008). Balanced Corpus of Contemporary Written Japanese. *IJCNLP 2008*, 101.

Martin, A. (2004). The 'katakana effect' and teaching English in Japan. *English Today, 20*(1), 50-55.

Matsumura, A. 松村明 (2020). *Daijisen* 大辞泉. Tokyo: Shogakukan 小学館. Retrieved December 2020, from https://dictionary.goo.ne.jp/jn/.

Ministry of Education, Culture, Sports, Science and Technology 文部科学省 (2016). *Yoochien, Shoogakkoo, Chuugakkoo, Kootoogakkoo oyobi Tokubetsu Shien Gakkoo no Kaizen oyobi Hitsuyoona Hoosaku tou ni Tsuite* 幼稚園、小学校、中学校、高等学校及び特別支援学校の学習指導要領等の改善及び必要な方策等について [On the improvement of the curricula for kindergartens, elementary schools, middle schools, high schools, and special education schools]." Retrieved December 2020, from http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/1380731.htm.

Mochizuki, K., Murao, S., Katayama, S., Fukuda, S., & Fujii, Y. 望月圭子・村尾誠一・片山晴一・福田翔・藤井嘉章 (2014). *Daigakusee no tame no akademikku raitingu* 大学生のためのアカデミックライティング [Academic writing for college students]. Tokyo: Tokyo University of Foreign Studies 東京外国語大学.

Morton, J., Sasanuma, S., Patterson, K., & Sakuma, N. (1992). The organization of the lexicon in Japanese: Single and compound kanji. *British Journal of Psychology*, *83*(4), 517-531.

Nagasawa, N. 長澤直子 (2017). Daigakusee no sumaatofon to PC de no moji nyuuryoku hoohoo: Wakamono ga PC yori mo sumaatofon o kononde shiyoo suru riyuu no ichikoosatsu 大学生のスマートフォンと PC での文字入力方法 - 若者が PC よりもスマートフォンを好んで使用する理由の一考察 [How Japanese college students type on

smartphones and computers: Why they prefer to use smartphones over computers]. *Computer & Education* コンピュータ＆エデュケーション, *43*, 67-72.

Nishi, H. and Xu, J. (2013). Teaching katakana loanwords to learners of Japanese: Current issues and pedagogical suggestions. *2013 CAJLE Conference Proceeding*s, 182-189.

Noborimoto, Y., Takahashi, J., & Horita, R. 登本洋子・高橋 純・堀田龍也 (2021). Kookoosee no PC, sumaatofon no moji nyuuryoku no hayasa ni kansuru choosa 高校生の PC・スマートフォンの文字入力の速さに関する調査 [Survey on character input speed on personal computers and smartphones by high school students]. *Japan Journal of Educational Technology* 日本教育工学会論文誌, *44*, 29-32.

Okabe, Y. 岡部勇太 (2009). Nihon shakai ni okeru seekatsusha no joohoo media to shite no pasokon riyoo no rekishiteki suii: Shijoo shugi shakai ni okeru seekatsusha niizu no han'ee o shiten ni shite 日本社会における生活者の情報メディアとしてのパソコン利用の歴史的推移 - 市場主義社会における生活者ニーズの反映を視点にして [Historical changes of the usage of personal computers as information media in Japanese society: From the perspective of consumer demands in capitalist society]. *Hakusan Sociological Review* 白山社会学研究, *16*, 86-103.

Otsuka, A. 大塚淳子 (2003). Nihonjin daigakusee no guruupu tooron ni okeru ketsuron seesee to shinkooyaku no yakuwari 日本人大学生のグループ討論における結論生成と進行役の役割 [The role of the leader in group discussion in the case of Japanese university students]. *Nihongo Nihon Bunka* 日本語・日本文化, *29,* 147-159.

Quackenbush, H. C. (1977). English Loanwords in Japanese: Why are they difficult for English-speaking students?. *The Journal of the Association of Teachers of Japanese*, *12*(2/3), 149-173.

Stanlaw, J. (2004). *Japanese English: Language and Culture Contact*. Hong Kong: Hong Kong University Press.

Sung, Y. & Mitsudo, H. 成儒彬・光藤宏行 (2016). Eego hyooki rogo taipu wa nihonsan inryoo shoohin no inshoo o koojoo saseru 英語表記ロゴタイプは日本産飲料商品の印象を向上させる [English logotype improves the impression of Japanese beverages]. *Proceedings of the Japanese Society for Cognitive Psychology* 日本認知心理学会発表論文集, 48.

Taylor, I., & Taylor, M. (1995). *Writing and literacy in Chinese, Japanese and Korean*. Amsterdam: John Benjamins.

Tsujimura, N. (2013). An Introduction to Japanese Linguistics: Third Edition. Hoboken: Wiley-Blackwell.

W3Techs. (2021). *Usage Statistics of Content Languages for Websites*. Retrieved July 2021, from https://w3techs.com/technologies/overview/content_language

Yanaike, M. 屋名池誠 (2003a). *Yokogaki toojoo: Nihongo hyooki no kindai* 横書き登場 - 日本語表記の近代 *[Development of horizontal writing: Modern Japanese writing system]*. Tokyo: Iwanami Shoten 岩波書店.

Yanaike, M. 屋名池誠 (2003b). Yokogaki no seeritsu: Nihongo hyooki no epokku 横書きの成立 - 日本語表記のエポック [Introduction of horizontal writing: An epoch in Japanese writing system]. *Annals of the Institute for Comparative Studies of Culture, Tokyo Woman's Christian University* 東京女子大学比較文化研究所紀要, *64,* 23-40.

Liushu-based Instruction and Its Effects on the Motivation and Intended Learning Efforts: The Case of Laos Learners of Standard Chinese

GUO Qingli
University of Malaya, Malaysia
17221874@siswa.um.edu.my

CHEW Fong Peng
University of Malaya, Malaysia
fpchew@um.edu.my

Abstract

This study examines the effects of *Liushu*-based instruction on the motivation and intended learning efforts. Beginners of Standard Chinese from Laos were assigned to the experiment group and the control group. The *Liushu*-based instruction in the experiment group was carried on for ten weeks. It was found that *Liushu*-based instruction has a positive effect on learners' motivation, especially for Standard Chinese learners' Ideal L2 Self and L2 learning experience. In addition, Ideal L2 Self showed a mediation effect between *Liushu*-based instruction and intended learning efforts. The article also provides several suggestions for the use of *Liushu* in Chinese character teaching.

Keywords: *Liushu*-based Instruction, Standard Chinese learners, motivation, intended learning efforts

Povzetek

Članek preučuje učinke poučevanja po metodi Liushu na motivacijo in načrtovana učna prizadevanja. Učenci začetniki standardne kitajščine iz Laosa so bili uvrščeni v poskusno in kontrolno skupino. Poskusna skupina je navodila po metodi Liushu izvajala deset tednov. Ugotovljeno je bilo, da ima poučevanje po metodi Liushu pozitiven učinek na motivacijo učencev, zlasti za učencem standardne kitajščine Idealnega L2 sebe ter za L2 učne izkušnje. Poleg tega je Idealni L2 jaz izkazal posredniški učinek med navodili, ki temeljijo na metodi Liushu, in načrtovanimi učnimi napori. V članku je na koncu podanih tudi nekaj predlogov za uporabo metode Liushu pri poučevanju kitajskih pismenk.

Ključne besede: poučevanje po metodi *Liushu*, učenci standardne kitajščine, motivacija, predvidena prizadevanja za učenje

1   Introduction

Recently, more and more people in Laos choose to learn languages for a variety of purposes and needs. As a result of the Belt and Road initiative, China and Laos have developed greater economic cooperation, which has stimulated the Laos labor market. Accordingly, Standard Chinese as a second language is becoming increasingly popular in Laos due to the demand for Standard Chinese professionals (Tao, Lin & Zhang, 2020).

For many learners, Chinese character learning is a key factor affecting their progress, efficiency, and level. The success or failure of Standard Chinese learning largely depends on the success or failure of Chinese character learning (Li, 2009). However, Ye (2013) stated that the influence of the mother tongue makes it difficult for Standard Chinese learners to learn Chinese characters. A high number of homophones confuse Standard Chinese learners due to the complicated relationship between sound, form, and meaning of the characters. Chen (2020) also indicated that Standard Chinese learners from Laos had difficulty writing Chinese characters since the Chinese language, a logographic writing system, is different from Lao script, which belongs to the alphabetic writing system.

In the Han dynasty of Ancient China, scholars summarized six types of Chinese characters as *Liushu*. The work *Shuo Wen Jie Zi Xu* (说文解字·叙) describes the six types of Chinese characters, which are pictographs (*xiangxingzi* 象形字), ideographs (*zhishizi* 指事字), ideological compounds (*huiyizi* 会意字), semantic-phonetic compounds (*xingshengzi* 形声字), mutually explaining characters (*zhuanzhuzi* 转注字), and phonetic loan characters (*jiajiezi* 假借字). *Liushu* illustrates the association between Chinese characters' glyphs and their meaning (Chen & Fu, 2014). Therefore, *Liushu*-based instruction is recognized as an effective way of Chinese character learning, especially, for the recognition of the sound and meaning of Chinese characters. For example, students can identify the meaning and sound of semantic-phonetic compounds by the meaning element and the sound element. A detailed example is stated as follows: the meaning of *hú* 湖 'lake' is related to its meaning element *sān diǎn shuǐ* 氵 'water-related', which means water, and the pronunciation of *hú* 湖 'lake' is similar to its sound element *hú* 胡 '(surname) Hu'.

Studies have shown that *Liushu*-based instruction can improve Chinese character achievement (e.g. Qi, 2017, Qiao, 2011, Su & Li, 2019). Liu (2011) also proposed that *Liushu*-based instruction can strengthen Standard Chinese learners' learning motivation. Motivation, which is the effort learners put into the learning of a second language (L2) due to the need or desire to do so (Ellis, 1994, p. 509), plays an influential role in second language acquisition (SLA) (Dörnyei, 2005). However, few studies examine the impact of *Liushu*-based instruction on learners' motivation.

The importance of motivation prompted the development of various models, constructs, and systems. Recently, Dörnyei's (2009) work on the L2 Motivational Self

System (L2MSS) attracted the attention of researchers, in which learners' motivation can be further divided into their Ideal L2 Self (IL2S), Ought-to L2 Self (OL2S), and L2 Learning Experience (L2LE). Based on this system, researchers have investigated the relationship between the three L2MSS components and learner's Intended Learning Efforts (ILEs). In English, learning positive associations between L2MSS components and ILEs have been reported (i.e., in Alshahrani, 2016; Ryan, 2009). Therefore, in this study, the researcher seeks to bridge the gap with an analysis of the relationships between *Liushu*-based instruction, L2MSS components, and ILEs among Standard Chinese learners.

2   Literature review

2.1   *Liushu*-based instruction

Liushu contains six rules summed up by the ancients analyzing the method of making Chinese characters (Chen, 1982). According to *Liushu*, Chinese characters were divided into pictographs, ideographs, ideological compounds, semantic-phonetic compounds, mutually explaining characters, and phonetic loan characters. 1) Pictographs depict specific objects like pictures, such as *rì* 日 'sun' ; 2) Ideographs represent abstract ideas. For example, *shàng* 上 'up' and *xià* 下 'down' are marked with symbols above or below the main line to indicate up and down, respectively; 3) ideological compounds combine the meanings represented by the components to create a new meaning. For example, three *rén* 人 'person' make up the *zhòng* 众 'people'; 4) Semantic-phonetic compounds are created by combining a semantic part related to the character's meaning and a phonetic part related to the character's pronunciation. For example, *sān diǎn shuǐ* 氵 'water-related' means water which is the semantic part of *hú* 湖 'lake', and the pronunciation of the phonetic part *hú* 胡 '(a surname) Hu' is the same as the pronunciation of *hú* 湖 'lake'; 5) Mutually explanatory characters refer to the mutual conversion between synonyms with the same radical. For example, both *kǎo* 考 '(the original meaning) elder' and *lǎo* 老 'elder' mean elder in ancient Chinese. However, the meaning and usage of *kǎo* 考 have all been transferred to *lǎo* 老 in modern Chinese, while the modern Chinese character *kǎo* 考 has lost its original meaning; 6) Phonetic loan characters mean that there is no such Chinese character, but for expression, the existing Chinese character is borrowed to serve as the new meaning without creating a new Chinese character. For example, *lìng* 令 'official' of *xiànlìng* 县令 'magistrate' is used to represent the meaning of *lìng* 令 'order' as in *mìnglìng* 命令 'order'.

Xu (2009) proposed that applying *Liushu* to Chinese character teaching can form three specific teaching methods: image display, character configuration analysis, and systematic induction. a. The method of image display is primarily employed for pictographs. By communicating the relationship between the image and the text, the vivid image can help students understand Chinese characters' meaning in connection

with real-life; b. Character configuration analysis is mainly used for ideographs and ideological compounds. Like storytelling, this method combines vivid life scenes with Chinese characters; c. A method of system induction is commonly used for semantic-phonetic compounds. For example, teachers can introduce characters *zhāng* 彰 'evident' (evident), *zhāng* 漳 '(proper name) Zhang', *zhāng* 樟 'camphor', *zhāng* 蟑 '(part of the word) cockroach', *zhāng* 璋 'ancient stone ornament', *zhàng* 嶂 'cliff', *zhàng* 障 'to block', or *zhàng* 瘴 'malaria' through *zhāng* 章 'chapter'. This method can help learners better master the sounds of Chinese characters and their meanings systematically.

However, there have been a lot of changes in the graphical forms of Chinese characters (including the simplification of characters), resulting in a different perception of Chinese characters based on *Liushu*. Li (2012) proposed that the main value of *Liushu* does not lie in theory but in the application. In other words, *Liushu* should serve teaching rather than bringing it a burden. For example, in the teaching of the Chinese character *yào* 药 (藥) 'medicine', it is important to let students know what the components of *yào* 药 are (i.e., *cǎo* ⁺⁺ 'grass' and *yuē* 约 'to make an appointment' and the functions that these components have (i.e., *cǎo* ⁺⁺ 'grass' is the semantic part, whereas *yuē* 约 is the phonetic part, instead of making students distinguish *yuē* 约 'to make an appointment' and *yuè/lè* 乐 'music/happy'. Another example is the semantic-phonetic compound *tīng* 听(聽) 'to hear'. According to the interpretation of *Handian* 漢典, the traditional character *tīng* 聽 is made of the semantic part *ěr* 耳 'ear' and *dé* 德 'virtue' and the phonetic part *tǐng* 壬, while the simplified character *tīng* 听 'to hear' is composed of the semantic part *kǒu* 口 'mouth' and the phonetic part *jīn* 斤 'catty'. It would be more feasible to use simplified characters to explain the composition of Chinese characters for Standard Chinese learners since simplified characters are the main focus of their learning.

As Qiao (2011) stated, analyzing all Chinese characters in their traditional form would undoubtedly make the situation more complex. However, some simplified characters lose their ideographic function. For example, we can identify the part inside of *guó* 国(國) 'country' as *huò* 或 '(originally meaning) country' only through traditional character. Therefore, in the teaching process, teachers can use traditional characters as an auxiliary means to teach Chinese characters to make it easier for students to understand the meaning and structure of simplified characters. That is, using the close relationship between simplified characters and traditional characters to reasonably explain simplified characters. In addition, some Chinese characters have added or subtracted meanings to suit the use of the language, leading to mutually explanatory characters and phonetic loan characters as stated above. It is worth noting that mutually explanatory characters and phonetic loans are commonly found in ancient texts (e.g. *kǎo* 考 means 'older' in ancient texts), and they are not recommended for assisting Chinese character teaching (Chen & Fu, 2014). Therefore, this study does not involve mutually explanatory characters and phonetic loan characters.

2.2 L2 Motivational Self System (L2MSS)

L2 Motivational Self System (L2MSS) was developed by Dörnye (2005, 2009), and mainly consists of the following three components: IL2S, OL2S, and L2LE. IL2S, which is correlated with integrativeness, represents what students wish to be. OL2S indicates what others expect the student to achieve, which means that it reflects the expectations of others, including their parents and peers. It represents an external motivation, which is related to instrumental motivation. Besides, L2LE is one's experience in a learning environment that is primarily influenced by teachers, courses, and classmates. It is mainly reflected in learners' evaluation of textbooks, teachers, peers, and class. It is worth noting that in Dörnyei's model, it is most common to discuss the relationship between the L2MSS components and ILEs. ILEs refer to the effort that learners make toward a goal, as stated by Moskovsky et al. (2016). For example, learners show a willingness to spend more time and energy on a certain course.

Although many studies have been conducted to examine the relationship between Standard Chinese learners' motivation and ILEs, the results are miscellaneous. Dörnyei & Chan (2013) explored the relationship between self-guides (i.e., IL2S and OL2S) and ILEs in Standard Chinese learning among Standard Chinese learners from Hong Kong. The correlation coefficient showed that the IL2S (in English and Standard Chinese learning) had a positive effect on ILEs ($p<0.001$). Also based on Dörnye's L2MSS, the structural model results revealed that both IL2S and OL2S can better predict learners' motivational behavior and the two self-guides together explain 62% of the significant difference in motivational behavior in Wong (2018). Li and Zhang (2021) used multiple regression analysis to examine the predictive ability of IL2S, OL2S, and L2LE on ILEs. The results show that the main components (IL2S, OL2S, and L2LE) of L2MSS can directly or indirectly predict ILEs (R^2=0.55). Among them, IL2S, directly or indirectly, is the strongest predictor (β=0.44) with L2LE as a mediator. In general, the three components of L2MSS have a strong predictive effect on ILEs.

Standard Chinese learning is recently gaining popularity in the countries of Southeast Asia. In Laos, for example, since the early 21st century, Standard Chinese learning has gradually emerged as an important part of higher education with the support of the Confucius Institute Headquarters (Zhang, Lu, & Zhejing, 2021). However, little research was conducted on the subject of Standard Chinese teaching and learning in Laos (Tao, Lin, & Zhang, 2020). Therefore, to fill the gap, this current study intends to bridge the gap with an experiment on *Liushu*-based instruction in Chinese character teaching and provide a more comprehensive picture of *Liushu*-based instruction, three components of L2MSS, and ILEs. Accordingly, this study aims to find answers to the following questions:

1. Is there any difference in the motivation between the control group and the experimental group in the pre-test?
2. Is there any difference in ILEs between the control group and the experimental group in the pre-test?
3. Is there any difference in the motivation of the control group between the pre-test and post-test?
4. Is there any difference in ILEs of the control group between the pre-test and post-test?
5. Is there any difference in the motivation of the experimental group between the pre-test and post-test?
6. Is there any difference in ILEs of the experimental group between the pre-test and post-test?
7. Is there any difference in motivation between the control group and the experimental group in the post-test?
8. Is there any difference in ILEs between the control group and the experimental group in the post-test?
9. Do L2MSS components act as mediators in the relationship between *Liushu*-based instruction and ILEs?

Then, nine null hypotheses were developed based on the research questions.

1. H_{01}: There is no significant difference in the motivation between the control group and the experimental group in the pre-test.
2. H_{02}: There is no significant difference in ILEs between the control group and the experimental group in the pre-test.
3. H_{03}: There is no significant difference in the motivation of the control group between the pre-test and post-test.
4. H_{04}: There is no significant difference in ILEs of the control group between the pre-test and post-test.
5. H_{05}: There is no significant difference in the motivation of the experimental group between the pre-test and post-test.
6. H_{06}: There is no significant difference in ILEs of the experimental group between the pre-test and post-test.
7. H_{07}: There is no significant difference in the motivation between the control group and the experimental group in the post-test.
8. H_{08}: There is no significant difference in ILEs between the control group and the experimental group in the post-test.
9. H_{09}: L2MSS components do not act as mediators in the relationship between *Liushu*-based instruction and ILEs.

3   Methodology

3.1   Research design

The study was a quasi-experimental study using the pre-test and post-test of control and experiment groups. The quasi-experimental design refers to a planned study, including a series of intentional changes to process elements and observations of the effects (Chua, 2016). The experiment involved six classes (two Accounting classes and four Information Technology classes). Human intervention was conducted to ensure homogeneity between both groups. As a result, each group was composed of Accounting students and Information Technology (IT) students.

3.2   Participants

There are a total of 217 Standard Chinese learners majoring in Accounting and IT at H College in Laos. These 217 students are assigned to nine classes with 10-30 students in each class. 47 students (18 males, 29 females) majoring in Accounting were assigned to three classes, and 170 students (105 males, 65 females) majoring in IT were grouped into six classes. Even though Standard Chinese was a mandatory course for them, most Standard Chinese students are beginners.

Six classes with two Accounting classes and four IT classes were selected for this experiment. The final sample size was 133 students, with 68 (37 males, 31 females) in this experiment and 65 (42 males, 23 females) in the control group. 31 of the participants were majoring in Accounting, and 102 were majoring in IT.

3.3   Instruments of study

L2 Motivational Self System Questionnaire (L2MSSQ) was utilized during the data collection process. The L2MSSQ used in this study is a five-point Likert scale adapted from Moskovsky et al. (2016) and Taguchi et al. (2009) to measure the motivation for SLA. The questionnaire also contains two parts. Part one consists of demography and part two consists of 41 items designed to measure learners' IL2S, OL2S, L2LE, and ILEs.

3.4   Validity and reliability of the instrument

In order to check the validity related to the face and the content of the instrument in this study, the questionnaire was handed over to three experts for review. Some modifications were made to adapt to the Standard Chinese learners in this study. For example, according to the experts' suggestions, the items with similar meanings in the questionnaire were also deleted.

To check the reliability of the instrument, the questionnaire was distributed to 39 respondents not participating in the experiment. Results showed the overall Cronbach coefficient for the L2MSSQ was 0.895. Each of its three components, IL2S, OL2S, and L2LE, had Cronbach coefficients of 0.833, 0.817, and 0.750, respectively. Besides, ILEs had a Cronbach coefficient of 0.772. It can be seen Cronbach's alpha values for all subscales are higher than the lowest acceptable value of 0.60 (Pallant, 2010), indicating that the questionnaire has achieved internal consistency.

Finally, in the questionnaire, a demographic survey and a five-point Likert scale are provided. The items in a demographic survey are matric number, gender, age, major, class, length of learning Standard Chinese, and level of Standard Chinese. The items in a five-point Likert scale are IL2S (9 items), OL2S (9 items), L2LE (15 items), and ILEs (8 items).

3.5   Intervention procedure

The experiment was carried out for 10 weeks. The experiment group underwent 10 sessions held once a week. Each session lasted 90 minutes and a total of 20-30 minutes per session was allocated to teaching Chinese characters. Throughout the experiment process, 10 lessons from the course book were taught to both groups. Each of these lessons contained 2-4 Chinese characters for learning. The main purpose was to help learners learn Chinese characters by using the *Liushu*-based instruction. Therefore, the criteria for selecting Chinese characters were 1) Chinese characters planned for the next ten weeks according to the syllabus; 2) Chinese characters that can be used as content words since it is more difficult to explain the Chinese characters' meaning using *Liushu* if it is a function word. For example, *māo* 猫 'cat' was chosen instead of the *le* 了 'a modal particle'. As a result, the topics and Chinese characters of each lesson were set as follows:

1. 九月去北京旅游最好
   Jiǔyuè qù Běijīng lǚyóu zuì hǎo
   'September is the best time to visit Beijing'
   characters: *māo* 猫, *tā* 它, *yào* 要, *zuì* 最

2. 我每天六点起床
   Wǒ měitiān liù diǎn qǐchuáng
   'I get up at six every day'
   characters: *gāo* 高, *máng* 忙, *yào* 药, *shàng* 上

3. 左边那个红色的是我的
   Zuǒbiān nàgè hóngsè de shì wǒde
   'The red one on the left is mine'
   characters: *xià* 下, *hóng* 红, *sòng* 送

4. 这个工作是他帮我介绍的
   Zhège gōngzuò shì tā bāng wǒ jièshào de
   'He recommended me for this job'
   characters: *gěi* 给, *wèn* 问, *cháng* 长, *liǎng* 两

5. 就买这件吧
   Jiù mǎi zhèjiàn ba
   'Take this one'
   characters: *yú* 鱼, *yī* 衣, *mǎi* 买, *mài* 卖

6. 你怎么不吃了
   Nǐ zěnme bù chī le
   'Why don't you eat more'
   characters: *mén* 门, *wài* 外, *yáng* 羊

7. 让我想想再告诉你
   Ràng wǒ xiǎng xiǎng zài gàosù nǐ
   'Let me think about it and I'll tell you later'
   characters: *děng* 等, *bái* 白, *hēi* 黑, *guì* 贵

8. 题太多，我没做完
   Tí tài duō, wǒ méi zuòwán
   'There are too many questions, I did not finish all of them'
   characters: *cuò* 错, *cóng* 从, *dǒng* 懂, *wán* 完

9. 你穿得太少了
   Nǐ chuān de tài shǎo le
   'You wear too little'
   characters: *xuě* 雪, *jìn* 近, *jìn* 进, *chuān* 穿

10. 你看过那个电影吗
    Nǐ kànguò nàgè diànyǐng ma
    'Have you seen that movie'
    characters: *wán* 玩, *qíng* 晴

The researcher drafted Chinese character teaching plans to assist the instructor in teaching Chinese characters. The procedures of the lesson were simple routine. The participants in the control group were not given any form of instruction. However, in the experiment group, the teaching steps of each Chinese character included an introduction, reading, writing, character shape evolution or Chinese character structure analysis, and sentence making.

During the introduction phase, students were asked to think of Chinese characters related to the picture (for example, a picture of a fish). With pictographs and ideographs, the teacher encouraged students to discover the similarities between Chinese characters and things.

The instructor would then lead the students to read Chinese characters by practicing the pronunciation of initials and finals. In the case of semantic-phonetic compounds, the teacher instructed students on how to pronounce the sound elements.

Following that, students were given an animated presentation of the writing sequence and then practiced writing Chinese characters, an important step also for the control group.

The next step involved pictographs and ideographs. Students learned the development and evolution of Chinese characters by looking at the changes in the form of Chinese characters from ancient times to the present. They learned the construction principles of ideological compounds and semantic-phonetic compounds by splitting the components (meaning elements and sound elements) of Chinese characters.

The final step was also a routine step in the control group. In it, students used the Chinese characters that they have learned to make sentences based on pictures and other prompts.

3.6   Data analysis

The data were processed and analyzed with SPSS. To answer the questions posed by this study, independent sample t-tests were utilized, along with paired sample t-tests, to examine any differences in learners' L2MSS and ILEs before and after the experiment. Regression analyses were conducted to examine the mediating effect of L2MSS Components between *Liushu*-based instruction and ILEs.

4   Findings

An independent sample t-test was conducted to examine the existence of a statistically significant difference in motivation and ILEs between the control group and the experiment group in the pre-test. The t-test results are displayed in Table 1 and Table 2.

Table 1: The control and experimental group's motivation in the pre-test

	Control		Experiment		t	p
	M	*SD*	*M*	*SD*		
IL2S	30.95	5.95	29.32	5.30	1.671	0.097
OL2S	30.32	6.03	31.03	6.16	-0.668	0.505
L2LE	46.35	5.20	46.72	6.71	-0.351	0.726
Overall	107.63	13.52	107.07	15.87	0.218	0.828

IL2S: Ideal L2 Self; OL2S: Ought-to L2 Self; L2LE: L2 Learning Experience

Table 2: The control and experimental group's ILEs in the pre-test

	Control		Experiment		t	p
	M	SD	M	SD		
ILEs	28.20	4.43	27.76	5.43	0.505	0.614

ILEs: Intended Learning Efforts

Table 1 indicates that there is no significant difference in motivation (t=0.218; p=0.828) between the control group (*M*=107.63, *SD*=13.52) and experiment group (*M*=107.07, *SD*=15.87). Therefore, it was determined that the two groups were identical and H_{01} was accepted.

In terms of IL2S, the mean score of the control group (*M*=30.95, *SD*=5.95) is slightly higher than that of the experimental group (*M*=29.32, *SD*=5.30). This indicates that the control group has a stronger Ideal L2 Self. Becoming an ideal speaker of Standard Chinese is the main source of motivation for the control group.

For OL2S and L2LE, the mean scores (*M*=30.32, *SD*=6.03) and (*M*=46.35, *SD*=5.20) of the control group are slightly lower than those of the experimental group with (*M*=31.03, *SD*=6.16) and (*M*=46.72, *SD*=6.71). In other words, expectations from parents and peers have become a strong motivation for the experimental group to learn Standard Chinese. At the same time, the progress of the experimental group with the current learning experience is more satisfying.

Table 2 shows that there is no significant difference in ILEs (t= 0.505; p=0.614) between the control group (*M*=28.20, *SD*=4.43) and experiment group (*M*=27.76, *SD*=5.43). Therefore, H_{02} was accepted.

After a 10-week experiment, a paired *t*-test was conducted in the control group and the experiment group to examine the changes in motivation and ILEs.

Table 3: Motivation of the control group in the pre-test and post-test

	Pre-test		Post-test		t	p
	M	SD	M	SD		
IL2S	30.95	5.95	30.65	5.35	0.469	0.640
OL2S	30.32	6.03	29.54	5.87	1.082	0.283
L2LE	46.35	5.20	47.62	7.74	-1.217	0.228
Motivation	107.63	13.52	107.80	17.01	-0.086	0.932

Table 4: ILEs of the control group in the pre-test and the post-test

	Pre-test		Post-test		t	p
	M	SD	M	SD		
ILEs	28.20	4.43	27.98	5.09	0.22	0.379

As shown in Table 3 and Table 4, no significant differences occurred in motivation (*t*=-0.086, *p*=0.932) and ILEs(*t*=0.22, *p*=0.379) in the control group, which indicates that the conventional teaching method did not produce significant positive or negative changes in learners' motivation and ILEs. Thus, H_{02} and H_{04} were accepted.

Table 5: Motivation of the experiment group in the pre-test and post-test

	Pre-test		Post test		t	p	Cohen's d
	M	SD	M	SD			
IL2S	29.32	5.30	31.12	6.73	-2.502	0.015*	0.303
OL2S	31.03	6.16	30.41	5.71	0.775	0.441	
L2LE	46.72	6.71	49.06	8.22	-2.233	0.029*	0.271
Motivation	107.07	15.87	110.59	17.94	-1.593	0.116	

*. Significant at level 0.05

Table 6: ILEs of the experiment group in the pre-test and post-test

	Pre-test		Post test		t	p
	M	SD	M	SD		
ILEs	27.76	5.43	28.03	5.24	-0.373	0.710

According to Table 5, there is a statistically significant difference between the pre-test (*M*=29.32, *SD*=5.30) and the post-test (*M*=31.12, *SD*=6.73) of experiment group students in IL2S at *t*= -2.502, *p*=0.015, Cohen's *d*= 0.303, and between pre-test (*M*=46.72, *SD*=6.71) and post-test (*M*=49.06, *SD*=8.22) in L2LE at *t*= -2.233, *p*=0.029, Cohen's *d*= 0.271. This result shows that there are positive changes in learners' IL2S and L2LE in the experiment group who learn Chinese characters by using *Liushu*-based instruction. However, there is no significant difference between the pre-test (*M*=27.76, *SD*=5.43) and the post-test (*M*=28.03, *SD*=5.24) of experiment group learners in ILEs. Therefore, H_{05} was rejected and H_{06} was accepted.

Table 7: Motivation of the control and the experimental group in the post-test

	Control		Experiment		*t*	*p*
	M	*SD*	*M*	*SD*		
IL2S	30.65	5.35	31.12	6.73	-0.446	0.656
OL2S	29.54	5.87	30.41	5.71	-0.870	0.386
L2LE	47.62	7.74	49.06	8.22	-1.042	0.299
Motivation	107.80	17.01	110.59	17.94	-0.919	0.360

Table 8: ILEs of the control and the experimental group in the post-test

	Control		Experiment		*t*	*p*
	M	*SD*	*M*	*SD*		
ILEs	27.98	5.09	28.03	5.24	-0.050	0.960

As shown in Table 7, there is no significant difference in motivation after the experiment between the control group (*M*=107.80, *SD*=17.01) and the experiment group (*M*=110.59, *SD*=17.94) at *t*=-0.919, *p*=0.360. As far as the three components of motivation are concerned, the scores of the experimental group are generally higher than those of the control group, but there is no significant statistical difference. Also, Table 8 shows no significant difference in ILEs between the control group (*M*=27.98, *SD*=5.09) and the experiment group (*M*=28.03, *SD*=5.24) at *t*=-0.050, *p*=0.960 after the experiment. Therefore, H_{07} and H_{08} were accepted.

Next, regression analysis was conducted to examine the mediating effect of the L2MSS components between *Liushu*-based instruction and ILEs.

Table 9: The result of mediating effect of L2MSS components
between Liushu-based instruction and ILEs

	c X to Y β	a X to M β	b M to Y β	a*b M β	c' X to M to Y β	a*b 95%BootCI	a*b／c % of Total effect
IL2S	0.480	2.102*	0.175*	0.367	-0.112	-0.001~0.088	76.46%
OL2S	0.480	0.167	0.340**	0.057	-0.112	-0.066~0.081	
L2LE	0.480	1.077	0.156**	0.168	-0.112	-0.030~0.070	

*. Significant at level 0.05;

**. Significant at level 0.01;

X: Independent variable (*Liushu*);

Y: Dependent variable (ILEs);

M: Mediator (IL2S, OL2S, L2LE);

95%BootCI: 95% confidence interval

As seen in Table 9, the total effect of *Liushu*-based instruction on ILEs is 0.480, which is not significant. The influence of *Liushu*-based instruction on IL2S is significant with β=2.102, p<0.05, as well as it is significant the impact of IL2S on ILEs with β=0.175, p<0.05. Both a and b are significant, which indicates that IL2S has a significant mediating effect at β=0.367. However, the direct effect of *Liushu*-based instruction on ILEs is not significant (β=-0.112, p>0.05). The results show that *Liushu*-based instruction indirectly predicts the level of ILEs through the IL2S of L2MSS components. In other words, IL2S mediated 76.46% of the total effect between *Liushu* and ILEs.

OL2S has the strongest impact on ILEs at β=0.340, p<0.01, followed by L2LE (β=0.156, p<0.01). However, the result did not show a significant effect of *Liushu*-based instruction on OL2S and L2LE at β=0.167, p>0.05 and β=1.077, p>0.05, respectively. That is, at least one of a and b is insignificant. Thus, the researcher used bootstrap to test the significance of a*b. 95% BootCI represents the 95% confidence interval calculated by Bootstrap sampling. The 95% confidence interval containing zero, indicates that the mediation effect is not significant. As seen in Table 9, neither IL2S nor L2LE shows any significance. It can be concluded that IL2S and L2LE have no significant mediating effect between *Liushu*-based instruction and ILEs.

In short, although *Liushu*-based instruction has no direct influence on ILEs, it can indirectly influence ILEs through IL2S, which is the mediator. Thus, H_{09} was rejected.

5   Discussion and conclusion

This study investigated the impact of a 10-week *Liushu*-based instruction on Standard Chinese learners' motivation and ILEs in Laos. Both the control group and the experimental group have almost identical motivation and ILEs in the independent sample *t*-test results of the pre-tests. The post-test results, however, show a slight change in the average scores in the motivation of the control group, while it increases from 107.80 to 110.59 in the experimental group. Furthermore, it is more obvious that the paired-sample t-tests showed that the experimental group had significant differences in IL2S and L2LE between the pre-test and the post-test. IL2S, meaning something that somebody wants to become, represents a person's desire and as such is a powerful motivation for a student to learn a second language. In this process, *Liushu*-based instruction integrates the cultures to enhance learners' interest in learning Chinese characters (Li, 2018). This may encourage learners to have a stronger desire to become the so-called ideal learners of Standard Chinese. L2LE is one's experience in the learning environment influenced by teachers, courses, and teaching strategies. This result is in line with the view of Chen and Fu (2014) who believe that the application of *Liushu* can deepen learners' understanding and memory of Chinese characters, increase their interest in Chinese character teaching, and strengthen its effect.

The mediation test showed that *Liushu*-based instruction is likely to be a significant predictor of IL2S of L2MSS components. Although the paired *t*-test showed the L2LE changed significantly and positively in the experiment group, *Liushu*-based instruction was not a major predictor of L2LE in the regression analysis. This is an important finding because it emphasizes the need to deepen our understanding of *Liushu*-based instruction in motivational processes. A second key finding was the predictive influence of IL2S, OL2S, and L2LE on ILEs. OL2S is a stronger predictor than IL2S and L2LE. In other words, OL2S, which indicated the motivation from others' expectations, played a larger effect on ILEs among learners. This finding is different from the results of Li and Zhang (2020) and Wong (2018) who found that IL2S exerted more influence on ILEs than OL2S among Tibetan learners and learners of Standard Chinese in Hong Kong, respectively. As Kormos & Kiddle (2011) stated, OL2S may have more relevance in the context where language education places great pressure on learners' performance. This means that in Laos, the pressure exerted by others (e.g., parents, peers) motivates learners to put more effort into learning Standard Chinese. Unfortunately, the *Liushu*-based method did not have a positive impact on learners' OL2S. To rephrase this result, motivation derived from external pressures or other people's expectations could not change easily with changes in teaching methods. As a result, *Liushu*-based instruction can indirectly influence ILEs only through IL2S, which is motivated by the learner's own desire for a second language. This study enhances understanding of the link between the instruction with *Liushu*-based instruction, L2MSS components, and ILEs among Standard Chinese learners in Laos.

This study demonstrates that *Liushu*-based instruction is considerably beneficial for creating positive effects on learners' IL2S and L2LE, which are the two components of motivation. It furthermore shows that *Liushu*-based instruction is good support in teaching Chinese characters. Teachers may use the source and structure of the characters involved in *Liushu* to make their teaching style more interesting, which is also the reason for the improvement of students'IL2S and L2LE. In addition, IL2S was found to mediate 76.46 % of the total effect of the relationship between *Liushu*-based instruction and ILEs, which means that teachers should pay more attention to the cultural functions carried by *Liushu* to improve students' IL2S and motivate them to spend more efforts on learning Standard Chinese.

However, it is important to note that applying *Liushu*-based instruction may be challenging for teachers of Standard Chinese. Although Chinese characters' composition has a strong basis, modern Chinese characters have undergone great changes in form, which may cause learners to get bored. This requires Standard Chinese teachers to have a deep knowledge of Chinese characters to identify the relationship between their form and meaning. According to Gao (2018), only Standard Chinese teachers with sufficient knowledge can ensure good teaching results. It is important for teachers to understand the nature and characteristics of Chinese characters, as well as

the evolution of Chinese characters, and teachers of Standard Chinese should possess enough professional knowledge to explain these phenomena.

As a second recommendation, teachers should adjust *Liushu*-based instruction to the characteristics of learners. As stated by Gao (2018), learners of Standard Chinese that have relatively weak receptivity could easily get exhausted by even some simple Chinese learning. Adding the knowledge of *Liushu* (e.g., the evolution of characters) would likely increase their learning burden to the limits. Furthermore, in the current study, the learning time of Chinese characters was much longer in the experimental group than in the control group. Learning Chinese characters is the most time-consuming part of learning Standard Chinese for beginners. It is possible, however, that the use of *Liushu* for Chinese character learning is limited for intermediate and advanced learners since vocabulary and grammar learning will take more time and effort with the continuous increase of learning content. The learners' Standard Chinese levels should therefore be considered when assigning Liushu to learners.

Finally, teachers should limit the use of *Liushu* by considering the characteristics of Chinese characters. Chinese characters that are very different in shape or meaning from their ancient Chinese equivalents should no longer be explained by *Liushu*. As Fei (1998) stated, modern Chinese characters have undergone significant changes in form due to the emergence of simplified Chinese characters. It is true that *Liushu* can help students understand cultural meanings hidden in the original forms, however, it does little to help them memorize the characters in their simplified form if their present meaning had already lost the connection with the original meaning. Therefore introducing *Liushu* might even increase their workload. Last but not least, for learning pictographs, ideographs, and ideological compounds, it is best to focus on teaching simple components to avoid blindly tracing the source and causing more harm than good to the learners.

Acknowledgments

References

Alshahrani, A. A. S. (2016). L2 motivational self-system among Arab EFL learners: Saudi Perspective. *International Journal of Applied Linguistics and English Literature*, *5*(5), 145–152. https://doi.org/10.7575/aiac.ijalel.v.5n.5p.145

Chen, J. 陈俊权. (2020). Xiaoxue yuwen "Liushu" yujing xia de "Siti" shizi jiaoxue yu shuxie yanjiu 小学语文"六书"语境下的"四体"识字教学与书写研究. *Zhongguo Jiaoyu Xuekan 中国教育学刊, S1*,41–43.

Chen, Y. 陈亚川.(1982). Liushushuo, jiantizi yu hanzi jiaoxue 六书说、简体字与汉字教学. *Yuyan Jiaoxue Yu Yanjiu 语言教学与研究, 1*, 85–104+160.

Chen, Y. 陈羿竹, & Fu, Y. 傅亚庶. (2014).Guanyu Liushu lilun yingyong zai duiwai hanzi jiaoxue zhong de yanjiu 关于六书理论应用在对外汉字教学中的研究. *Hubei Minzu Xueyuan Xuebao (Zhexue Shehui Kexue Ban)湖北民族学院学报(哲学社会科学版)*,*2*,113-117.

Chua, Y. P. (2016). *Mastering research methods* (2nd ed.). Mcgraw-Hill Education.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. London, UK: Lawrence Erlbaum.

Dörnyei, Z. (2009). The L2 motivational self system. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 9-42). Bristol, UK: Multilingual Matters.

Dörnyei, Z., & Chan, L. (2013). Motivation and vision: An analysis of future L2 self images, sensory styles, and imagery capacity across two target languages. *Language Learning*, *63*(3), 437–462. https://doi.org/10.1111/lang.12005

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Fei, J. 费锦昌.(1998). Duiwai hanzi jiaoxue de tedian, nandian jiqi duice 对外汉字教学的特点、难点及其对策.*北京大学学报（哲学社会科学版） Journal of Peking University(Philosophy and Social Sciences),3*,3-5. https://doi:CNKI:SUN:BDZK.0.1998-03-017.

Gao, T. 高婷婷. (2018).*Liushu zai duiwai hanzi jiaoxue zhong de yingyong yanjiu 六书在对外汉字教学中的应用研究*. (Shuoshi Xuewei Lunwen 硕士学位论文,Zhejiang Keji Xueyuan 浙江科技学院).

Kormos, J., & Kiddle, T. (2011). Systems of goals, attitudes, and self-related beliefs in second-language-learning motivation. *Applied linguistics*, *32*(5), 495-516.

Li, J. 李静. (2018). Cong *Liushu* lilun kan duiwai hanzi jiaoxue 从六书理论看对外汉字教学. *Wenxue Jiaoyu (Shang)文学教育（上）*,*5*,146-147.

Li, M., & Zhang, L. (2021). Tibetan CSL learners' L2 Motivational Self System and L2 achievement. *System, 97,* 102436.

Li, Q. 李泉.(2009). Guanyu jianli guoji hanyu jiaoyu xueke de gouxiang 关于建立国际汉语教育学科的构想. *Shijie Hanyu Jiaoxue 世界汉语教学,3,* 399-413.

Li, Y. 李运富. (2012). "Liushu" xingzhi ji jiazhi de chongxin renshi"六书"性质及价值的重新认识. *Shijie HanyuJiaoxue 世界汉语教学, 1*,94-105.

Liu, Y. 刘雅芬.(2011). Sanweiyiti—cong renzhi jiaodu lun hanzi xingshengzi de jiaoxue 三位一体——从认知角度论汉字形声字的教学.*Haiwai Huawen Jiaoyu 海外华文教育,1*, 58-77.

Moskovsky, C., Assulaimani, T., Racheva, S., & Harkins, J. (2016). The L2 Motivational Self System and L2 Achievement: A Study of Saudi EFL Learners. *Modern Language Journal*, *100*(3), 641–654. https://doi.org/10.1111/modl.12340

Pallant, J. (2010). *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using SPSS*. Milton Keynes: Open University Press.

Qi, W .齐威.(2017). Zao zi fa zai duiwai hanzi jiaoxue zhong de yingyong 造字法在对外汉字教学中的应用. *Yishu Keji 艺术科技, 7*, 358+379.

Qiao, C. 乔春梅.(2011). *Zaozifa zai duiyue chuji hanzi jiaoxue zhong de yunyong 造字法在对越汉字教学中的运用*. (Shuoshi Xuewei Lunwen 硕士学位论文,Huazhong Shifan Daxue 华中师范大学). https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD2012&filename=1012261280.nh

Ryan, S. (2009). Self and Identity in L2 Motivation in Japan: The Ideal L2 Self and Japanese Learners of English. In *Motivation, Language Identity and the L2 Self* (pp. 120–143). https://doi.org/10.21832/9781847691293-007

Su, J.苏锦锦,& Li, W.李炜.(2019). Naotu zhuji tuidong xia "zaozifa"lilun zai duiwai hanzi jiaoxue zhong de yingyong 脑图助记推动下"造字法"理论在对外汉字教学中的应用 *Jiaoyu Guancha 教育观察, 35*, 26-28+35.

Taguchi, T., Magid, M., & Papi, M. (2009). The L2 Motivational Self System among Japanese, Chinese and Iranian Learners of English: A Comparative Study In Z. Dörnyei & E. Ushioda (Eds.). In *Motivation, language identity and the L2 self* (Issue May 2014). Bristol: Multilingual Matters. Retrieved from https://doi.org/10.21832/9781847691293-005

Tao, J. 陶家骏, Lin, Q. 林齐倩, & Zhang, T. 张唐凌.(2020). "Yidai yilu" changyi xia de Laowo "hanyu re" yanjiu "一带一路"倡议下的老挝"汉语热"研究. *Guoji Hanyu Jiaoyu(Zhongying Wen)国际汉语教育（中英文）, 2*, 91-99+90.

Wong, Y. K. (2018). Structural relationships between second-language future self-image and the reading achievement of young Mandarin learners in Hong Kong. *System*, *72*, 201–214. https://doi.org/10.1016/j.system.2017.12.003

Xu, L. 徐丽亚.(2009). Liushu lilun yu duiwai hanzi jiaoxue 六书理论与对外汉字教学. *Wenxue Jiaoyu (Xia)文学教育（下）(08)*,146-147.

Ye, J.叶剑锋. (2013). Laowo xuesheng xuexi hanzi de nandian ji duice tanxi 老挝学生学习汉字的难点及对策探析. *Xin Xibu(Lilun Ban)新西部（理论版）, 7*, 88+90.

Zhang, C.张成霞, Lu, L.卢露萍, & Zhejing, B. 哲京·巴侯涛. (2021). Laowo hanyu jiaoxue xianzhaung ji fazhan yanjiu 老挝汉语教学现状及发展研究. *Shijie Jiaoyu Xinxi 世界教育信息, 2*, 63-69.

Exceptions vs. Non-exceptions in Sound Changes: Morphological Condition and Frequency

LIU Sha
Fukuoka University, Japan
liusha@fukuoka-u.ac.jp

Abstract

This paper takes an approach different from most previous studies by firstly comparing exceptions with non-exceptions in the diphthongization of [i] to [ei] in Mandarin (AD 1324–Present) to locate factors that explain exceptions. Then it focuses solely on non-exceptional morphemes in this process by comparing morphemes at the forefront of this process with those undergoing it later to examine factors to explain morphemes leading this sound change. Statistical analysis shows that morphemes with the highest frequency among all related morphemes tend to be exceptions to diphthongization of [i] to [ei], and morphemes with high frequency among those non-exceptional morphemes tend to undergo this process earlier. In addition, the factor of frequency change, a proposal of this paper, is statistically significantly correlated with morphemes that lead diphthongization. The morphological condition has been rejected as statistically significant both for explaining exceptions to sound change and leading morphemes in sound change.

Keywords: diphthongization, leading morphemes, frequency, frequency change, exceptions

Povzetek

Članek uporablja nekoliko drugačen pristop v primerjavi z večino prejšnjih študij; najprej primerja izjeme z neizjemami pri diftongizaciji [i] v [ei] v mandarinščini (1324 AD–danes) z namenom poiskati dejavnike za razlago izjem. Nato se osredotoči izključno na neobičajne morfeme v tem procesu s primerjavo morfemov, ki so vodilni v spremembi, in morfemov, ki so spremembam podvrženi relativno pozno. Na ta način preučimo dejavnike in okolja, ki spodbujajo diftongizacijo [i] v [ei] . Rezultati statističnih analiz kažejo, da so morfemi z najvišjo pojavnostjo med vsemi sorodnimi morfemi ponavadi izjeme pri diftongizaciji [i] v [ei], morfemi z visoko pojavnostjo med temi neobičajnimi morfemi pa so ponavadi podvrženi temu procesu prej. Poleg tega je dejavnik sprememb pojavnosti, ki je predlog tega prispevka, statistično značilno povezan z morfemi, ki so vodilni pri diftongizaciji. Morfološki pogoj je bil zavrnjen kot statistično pomemben tako za razlago izjem pri glasovni spremembi kot tudi za vodilne morfeme v obravnavani glasovni spremembi.

Ključne besede: diftongizacija, vodilni morfemi, pojavnost, sprememba pojavnosti, izjeme

1   Introduction

Previous studies have proposed morphological conditions and frequency to account for exceptions to sound change, and morphemes that are at the forefront of sound change (see e.g., Grimm, 1822; Postal, 1968; King, 1969; Antilla, 1972; Melchert, 1975; Vincent, 1978; Guy, 1991a; Pintzuk, 1991; Santorini, 1992, 1993; Kroch, 1994; Donohue, 2005; Bermúdez-Otero, 2007). However, disagreements concerning the role of the two mentioned factors in sound change are not rare. Although Postal (1968) and King (1969), among others, claim that the morphological condition has a role in sound change, Renwick et al. (2014) state that their results show little support for such a claim. Even scholars who claim that the morphological condition has a role in sound change cannot agree on its exact role: some claim that bound morphemes lead sound change, while others claim that it is free morphemes that lead it. Arguments concerning the frequency factor and sound change are more complex. A long-standing debate exists between authors claiming that the frequency factor takes a part in sound change and others claiming that sound change is independent of any frequency effects. The debate is further complicated by the fact that even authors who claim that frequency has a role in sound change cannot agree with each other on its exact role: some suggest that high-frequency morphemes lead sound change, while others argue that low-frequency morphemes are at the forefront of sound change, and yet there are those who claim that either high-frequency morphemes or low-frequency morphemes are the ones that lead sound change.

In addition to the debate presented above, the present study notices another problem: scholars use different types of data to probe into the role of morphological condition and frequency in sound change. Some scholars compare exceptions and non-exceptions in a sound change process, while others compare morphemes that undergo a sound change process earlier with morphemes that undergo the same sound change later. This may be the reason that different scholars arrive at different conclusions: they had focused on different phases of sound change. This prompted the present study to inspect a sound change process from two aspects: (1) exceptions and non-exceptions in this sound change, and (2) morphemes that lead this sound change and morphemes that undergo it later. The paper firstly compares exceptions with non-exceptions in this sound change process to locate factors to explain exceptions in sound change. Then the paper compares morphemes that were at the forefront of this sound change with morphemes that underwent it later to locate factors to account for morphemes leading this sound change.

This paper takes the diphthongization of [i] to [ei] in Mandarin (AD 1324–Present) as its data source. What makes data from Chinese interesting is that Chinese is a language with a long history and thus can provide various kinds of data for sound change discussion. In addition, not much attention has been paid to Chinese in sound change study (see e.g., Wang, 1969; Chen and Wang, 1975). Previous studies mainly

take languages in the Indo-European language family as their focus. Since the sound change mechanism is supposed to be universal, it appears reasonable and necessary to examine data from languages belonging to various language families. In this respect, Mandarin is a good candidate since it belongs to the Sino-Tibetan language family and has not been widely discussed. Diphthongization of [i] to [ei] is a sound change process with exceptions that can provide data serving the proposal of the present study. Further, diphthongization of long high vowels is common in the world's languages and thus possible to be compared with parallel processes in other languages.

This paper is structured as follows. Section 2 discusses factors in the sound change in previous studies and presents a new proposal. Section 3 briefly sketches diphthongization of [i] to [ei] in Mandarin (AD 1324–Present). Section 4 carries out statistical analysis to locate factors to explain exceptions to the diphthongization of [i] to [ei] in Mandarin. Section 5 carries out another statistical analysis to examine factors to explain the morphemes leading this process. Section 6 puts forward some conclusions and issues relating to future research.

2   Previous studies

Previous studies explain exceptions to sound change in terms of various factors. Among them, morphological condition and frequency are the two most debated (see e.g., Chafe, 1968; King, 1969; Postal, 1968; Antilla, 1972; Campbell, 1974). In this section, the paper firstly reviews the two factors and disagreements concerning them, and then gives its own proposal.

2.1   Morphological condition

Postal (1968) and King (1969) can take the credit for noticing the possible role of morphological conditions in sound change although their claims have been questioned from many perspectives without resolution (Melchert, 1975). Following them, Zwicky (1970) discusses the auxiliary reduction in English. He (1970) notices that the rule which renders [i:y e:y u:w] to [i e u] and [a:y a:w] to [ay aw] applies only to pronominal forms ending in vowels (e.g., he, me, who) and followed by a contracted auxiliary other than [z]. A few years later, Rochet (1974) considers the change eN > aN in Old French and claims that this process was initiated as a morphological conditioned change. Malkiel (1976) discusses the diphthongizations in late Old Spanish, *ié > i* and *ué > e*, and argues that a set of morphological conditions can explain this phenomenon better than phonological conditions (see also Cerrón-Palomino, 1974; Johnson, 1982).

2.1.1   Exceptions vs. non-exceptions

Postal (1968, p. 247) focuses on Mohawk, the language spoken by Mohawk people, and notices that [e] is regularly inserted into [kw] sequences except when the [k] is "the first person morpheme and the [w] the first element of the plural morpheme." In his view (1968, p. 240), this language example shows that "nonphonetic morphophonemic and/or superficial grammatical structure" could also condition sound change (reviewed in Fudge, 1972; see also King, 1969). King (1969) is concerned with the final schwa deletion in Yiddish. According to him (1969), this rule does not apply when the final schwa is in an adjective inflectional ending. King (1969) concludes that this is evidence of morphologically conditioned sound change (reviewed in Robinson and van Coetsem, 1973; see also Antilla, 1972). Vincent (1978, p. 420) refers to the word-final schwa deletion in Spanish, and points out that the word-final schwa following a VC sequence is dropped, except in "the first person singular preterite of a number of irregular verbs …, the third person singular present indicative of all second and third conjugation verbs …, and the first and third singular present subjunctive of all first conjugation verbs." Morphological conditions in King (1969) and Vincent (1978) are concerned with word class in a rough term: King (1969) is concerned with adjectives and Vincent (1978) is concerned with verbs. More recently, Crowley (1997, p. 243) focuses on Southern Paamese and Northern Paamese, two languages of Central Vanuatu, and reports "a correspondence of Southern Paamese /l/ to Northern Paamese /i/, /l/, or zero" in all word classes except verbs. According to Crowley (1997, p. 244), this is "a clear example of a sound change that does not involve purely phonological conditioning factors, but also … grammatical conditioning." More specifically, this example shows that at least some sound changes apply only in some word classes (Crowley, 1997).

What complicates the picture is that some scholars claim that it is morphosyntactic structure, instead of word class, that can explain exceptions to sound change. Donohue (2005) uses the voicing of voiceless stops in Palu'e to show that morphological conditions can explain exceptions to this sound change: bound grammatical morphemes seem to have fewer exceptions than free lexemes. Donohue (2005, p. 441) goes on with sound changes in Modern Indonesian and Bali-Vitu (Austronesian, Oceanic) to further support his claim, and concludes that sound change "depends as much on morphosyntactic information as it does on … phonotactic constraints, (phonological) conditioning environments, or changes in related sounds …." Bybee (2002) focuses on word-final /t, d/ deletion in American English and concludes that bound morphemes can affect the deletion process. Guy (1991b, p. 2) also focuses on word-final /t, d/ deletion in American English and gives a more detailed conclusion: "underived or monomorphemic words such as *mist*, *pact*, undergo deletion at a higher rate than inflected forms such as past tense verbs like *missed*, *packed*" (see also Labov et al., 1968; Fasold, 1972; Guy, 1991a). Baranowski and Turton (2020) report a result similar to Guy (1991b) for word-final /t, d/ deletion in British English.

It is at this point important to point out that although all scholars referred to in this subsection claim that morphological condition has a role in explaining exceptions to sound change, they hold different views concerning its exact effect. To exemplify, although Crowley (1997) claims that word class can explain exceptions to sound change, Donohue (2005) suggests that morphosyntactic structure, that is the distinction between free morpheme and bound morpheme, is more correlated with exceptions to sound change.

2.1.2 Early application vs. late application

This subsection gives its attention to morphological conditions and non-exceptions in sound change. *Early application* and *late application* mean that morphemes do not undergo a sound change simultaneously: some morphemes undergo it earlier and some morphemes undergo it later. Phillips (1983, 2001, 2006) accounts for this from the perspective of morphological conditions and classifies words into two categories, function words and content words. Function words refer to a wide range of words that normally receive low sentence stress, such as adverbial conjunctions, auxiliary verbs, determiners, prepositions, quantifiers, and so on. Content words mainly include adjectives, adverbs, nouns, verbs, etc. Phillips (1983, 2006) gives an example of strengthening sound change, the change from /d/ to /t/ in Old High German Isidor, and points out that it affected function words last. He (1983, 2006) thus concludes that content words tend to be affected by strengthening sound changes first and function words by weakening sound changes first. Donohue (2005) also tries to explain morphemes leading sound change from the perspective of a morphological condition, but his approach is completely different from that of Phillips (1983, 2001, 2006): Donohue (2005) examines the voicing of voiceless stops in Palu'e and shows that free morphemes tend to lead sound change and bound morphemes follow. Similar to the previous section, disputes are over whether word class or morphosyntactic structure can explain morphemes leading sound change.

2.1.3 Disputes over morphological conditions

Though intriguing, the role of morphological conditions in sound change has been challenged by scholars like Jasanoff (1971), Blevins and Lynch (2009), Brown (2013), and so on. Jasanoff (1971) states that what appears to be morphologically conditioned is in fact regular sound change partially obscured by analogy. Blevins and Lynch (2009, p. 111) claim that the sound change discussed in Crowley (1997) applies to all word classes including verbs, but "phonological and morphological aspects of verbal inflectional paradigms" restore the change in verbs later and give rise to "the apparent exceptionality." Renwick et al. (2014) also focus on word-final /t, d/ deletion in British English as Guy (1991b) and Baranowski and Turton (2020), but they (2014) claim that their results show little support for the role of any specific morphological condition.

What is most surprising is that completely opposite conclusions concerning the role of morphological conditions have been drawn from the same phenomenon, word-final /t, d/ deletion in British English.

At the same time, some scholars adopt the middle way by claiming that no conclusion can be drawn yet and further investigation is necessary (Sihler, 2000; Campbell, 2013; Manker, 2015). For example, although Manker (2015, p. 287) states that many examples in Phillips (2006) are "actually … influenced by the 'most common phonetic environment' where certain word classes happen to be used in the favorable phonetic environment for the change more often," Manker (2015) does not give any clear cut answers to this issue. Instead, he (2015) suggests that the possibility of sound changes influenced by morphological factors cannot be absolutely ruled out and needs further extensive investigation, a view in line with Sihler (2000), Campbell (2013), etc.

To sum up, controversies concerning the morphological condition are twofold. The first controversy is whether it has a role in explaining exceptions to sound change and leading morphemes in sound change. The second one is whether word class or morphosyntactic structure can account for exceptions and leading morphemes.

2.2   Frequency

Morphological conditions are not the only factor put forward in previous studies. The frequency factor is perhaps the most widely argued: a long-standing debate exists between scholars claiming that frequency is relevant and others claiming that it is not. What is interesting is that more than a century has passed, and yet there has not been consensus on the frequency effect, and so it is still not well understood. In what follows, this subsection first reviews the frequency effect and exceptions to sound change in previous studies, then reviews the frequency effect and leading morphemes in sound change, and finally reviews disputes over the frequency effect.

2.2.1   Exceptions vs. non-exceptions

The frequency effect has been brought to view since the 19th century. Grimm (1822) discusses the relationship between high-frequency auxiliary verbs and their irregularity. Thomsen (1879) gives a few frequent Romanic verbs and the fact that they are exceptions to normal phonetic development. Jespersen ([1922] 2007, p. 267) more plainly expresses Thomsen's (1879) ideas in English as "words which from their frequent employment are exposed to far more violent changes than other words, and therefore to some extent follow paths of their own." Vilhelm Thomsen himself gives a similar explanation in his work (Thomsen, 1920). More recently, Labov (1989, p. 44) focuses on Philadelphia a-tensing and reports that "the most common words … show the least tendency to shift to the tense class." Bermúdez-Otero (2007, p. 512) also states that "the words with the very highest token frequency may exceptionally

withstand the change." Another related research includes Van Bergem (1995). He (1995) discovers that frequency influences the reduction of a pre-stressed vowel in Dutch: the high-frequency words, such as *minuut* ('minute'), *vakantie* ('holiday'), and *patat* ('chips'), are more likely to have a schwa in the first syllable than the phonetically identical low-frequency terms, e.g., *miniem* ('marginal'), *patent* ('patent'), *vakante* ('vacant'). Fidelholtz (1975) reports a similar tendency for the reduction of pre-stressed vowels in English words.

In contrast, other studies suggest that low-frequency words tend to be exceptions to sound change. To exemplify, Bybee (2002) studies the deletion of word-final /t/ and /d/ in American English and finds that the deletion rates in low-frequency words are statistically lower than in high-frequency words. Coetzee and Kawahara (2013, p. 62) observe two language phenomena, English *t/d*-deletion and geminate devoicing in Japanese loanwords, and argue that "*t/d*-deletion usually applies at higher rates to words of higher frequency," and frequency and rate of Japanese geminate devoicing are positively correlated.

2.2.2   Early application vs. late application

Concerning the frequency effect and leading words in sound change, Hooper (1976) focuses on schwa deletion in English and concludes that frequent words tend to lead this change. Furthermore, Hay and Foulkes (2016) focus on the ongoing change in the pronunciation of word-medial intervocalic /t/ in New Zealand English and report that frequent words lead this change (see also Pierrehumbert, 2001; Duncan, 2011).

Not all scholars hold a similar view. To name a few, Hay et al. (2015, p. 83) conclude the study of regular pronunciation changes in New Zealand English over a 130-year period with the expression that "low-frequency words were at the forefront of … changes and higher frequency words lagged behind."

A further dimension of debate is that some scholars claim that different sound changes are led by words of different frequencies. Phillips (1984, 2001, 2006) states that the most frequent words lead sound changes motivated by physiological factors, such as vowel reduction, deletion, assimilation, etc., while the least frequent words lead sound changes that arise from phonological segmental and sequential constraints of the language, such as unrounding in Middle English, diatone formation in Modern English, and others (see also Ogura, 2012).

2.2.3   Disputes over the frequency factor

Fruehwald et al. (2013) express their doubt concerning the role of frequency in sound change. They (2013, p. 219) focus on the Middle High German final stop fortition and claim that this change progresses "in frequency in every context at the same rate over time," the so-called *constant rate effect* (see also Kroch, 1989, 1994; Pintzuk, 1991;

Santorini, 1992, 1993; Dinkin, 2008). Other scholars more plainly claim that sound change is independent of frequency effects. For example, Zellou and Tamminga (2014, p. 18) study the co-articulatory vowel nasality in Philadelphia English and conclude that "the changes in nasality are independent of an observed frequency effect." Similarly, Labov (2010) examines the role of frequency in several different phonetically gradual changes and gives the conclusion that the role of frequency is minimal, even if not zero, a view agreed on by Kiparsky (2014). Attention has also been paid to language processes which have been used to support the role of frequency in sound change. To exemplify, word-final /t, d/ deletion in American English is used by Bybee (2002) to argue for the role of frequency in sound change as noted in Section 2.2.1. Walker (2012) focuses on word-final /t, d/ deletion in Canadian English, and reports that his initial results show the correlation between frequency and deletion. However, he (2012) further states that only phonological and morphological factor groups have emerged as statistically significant after he considered more factor groups. Abramowicz (2007) suggests that since scholars like Bybee (2002) and Phillips (1983, 2006) have used the variation word-final /t, d/ deletion in English to argue for the role of frequency, it is reasonable to expect the variable *ing*, that is *g*-dropping as in *walkin'* or *livin'*, to show similar effects in terms of frequency. However, Abramowicz (2007) concludes that his study does not show much frequency effect. Tamminga (2014, p. 457) argues against the frequency effect from another perspective: he questions the legitimacy of using word-final /t, d/ deletion in English to discuss the role of frequency in sound change since "[t]here has never been any evidence … that coronal stop deletion is a change in progress in any North American dialect." In other words, Tamminga (2014, p. 457) claims that the data of word-final /t, d/ deletion in English are "stable variation," but they have been used as evidence for "change in progress." Tamminga (2014, p. 458) further explores the adjective, conjunction, discourse marker, and preposition forms of *like* and claims that "frequency effects fail to arise."

2.3   Present study proposal

In sum, disputes in previous studies involve the following two questions. (1) Can morphological conditions and frequency account for exceptions to sound change? (2) Can morphological conditions and frequency account for words leading sound change? If answers to the above two questions are *yes*, then the following two questions should also be brought forward. (1) In terms of morphological condition, does word class or morphosyntactic structure have a role in sound change? (2) In terms of frequency, do high-frequency words or low-frequency words carry the tendency to be exceptions to sound change, and also, do they tend to lead sound change? To answer all these questions, the present study proceeds step by step; it first compares exceptions with non-exceptions in the diphthongization of [i] to [ei] in Mandarin and then compares words that have led this process with words that have undergone it

later. A comparison between exceptions and non-exceptions by the use of statistical analysis may present hints concerning factors to explain exceptions to sound change. A statistical analysis of words leading sound change with those undergoing it later may present factors correlated with words at the forefront of sound change.

In addition, the present paper notices a factor that awaits exploration about its role in sound change: previous studies have made little reference to the frequency change factor. The frequency change is calculated by the subtraction of the frequency of a word in the time concerned with the frequency of the same word in an earlier period. If the frequency factor is correlated with sound change, the frequency change factor may also be associated with sound change. To exemplify, Pierrehumbert (2001), Duncan (2011), among others claim that sound change usually affects the most frequent lexical items first. Following their logic, lexical items with increased frequency seem more likely to lead a certain sound change than lexical items with decreased frequency, since lexical items with increased frequency are more active and more accessible to a related sound change than lexical items with decreased frequency. In sum, this paper supposes that the frequency change factor also constitutes a desideratum for research.

3   Diphthongization of [i] to [ei] in Mandarin

This paper takes the diphthongization of [i] to [ei] in historical Chinese as its language sample. The diphthongization is a part of the Middle Chinese Great Vowel Shift (Chen 1976, Li 1999), which began no later than the end of the 16th century and finished no later than the beginning of the 19th century (Trigault, [1626] 1957; Edkins, 1857; Luo, 2008). Chart (1) graphically presents the Middle Chinese Great Vowel Shift (Chen, 1976; Li, 1999)

(1) The Middle Chinese Great Vowel Shift (Chen, 1976, p. 195)

$$ei, \ddot{\imath} \leftarrow i \qquad\qquad u \rightarrow ou$$
$$e \qquad\qquad o$$
$$a = \alpha$$
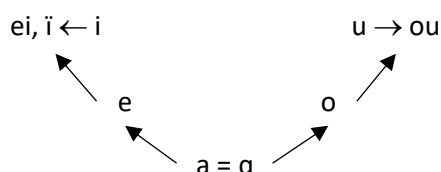
According to Chart (1), the general upward movement pushed the original high vowels *i and *u to undergo diphthongization and "became /əi/ and /əu/ respectively…. Eventually, /əi/ and /əu/ emerged as [ei] and [ou]" (Chen, 1976, p. 194). Due to the limited space, the present paper exclusively focuses on *i. The diphthongization of [i]

to [ei] applied with certain conditions. However, diphthongization was "exceptionless with regard to" [i] in the syllables [-ui], [vi], and [fi] in Middle Chinese (AD 601–AD 1336) according to Chen (1976, p. 200). The consonant [v] gradually turned to [u] in the Ming dynasty (1368–1644), almost simultaneously with diphthongization of [i] to [ei] (Luo, 2008). As a result, the three syllables [-ui], [vi], and [fi] in Middle Chinese were "obligatorily realized as [-uei] and [fei] respectively" in Mandarin (AD 1324–Present) (Chen, 1976, p. 200). In other words, the syllables [-ui], [vi], and [fi] are not in the Mandarin syllable inventory. The condition was more complex with [i] in the syllables [bi], [mi], [pi], and [pʰi] in Middle Chinese (Chen, 1976). Here are some related examples.

Table 1: Diphthongization of [i] to [ei] in Mandarin (Baxter & Sagart, 2014)

Middle Chinese (AD 601–AD 1336)	Mandarin (AD 1324–Present)	Gloss
[fi]	[fei]	非 'not'
[gwijH]	[ku̯eɪ]	柜 'box'
[pjie]	[pei]	卑 'low, humble'
[pjij]	[pi]	庇 'cover, protect'
[mijH]	[mei]	媚 'love, flatter'
[mij]	[mi]	麋 'a kind of deer'

Historical Chinese is usually divided into the following three phases: Old Chinese (1250 BC–AD 600), Middle Chinese (AD 601–AD 1336), and Mandarin (AD 1324–Present) (Wang, 1957, 1985; Pulleyblank, 1984, 1991; Shi, 2002; Handel, 2015; Pan and Zhang, 2015; Peyraube, 2020; Shen, 2020). All transcriptions of Middle Chinese and Mandarin in this paper are given according to Baxter and Sagart (2014), with reference to *Zhongyuan Yinyun* (Rhymes of the Central Plain; Zhou, [1324] 1996) and reconstructions of *Zhongyuan Yinyun* in Pulleyblank (1984, 1991) and Chou (1993). Tone marks are omitted since they are irrelevant to the present study.

As shown in Table 1, [i] in the syllables [fi] and [gwijH] was diphthongized to [ei] in Mandarin. However, [i] in the syllables [bi], [mi], [pi], and [pʰi] has a more complicated pattern. To exemplify, the two morphemes 'low, humble' (卑 [pjie]) and 'cover, protect' (庇 [pjij]) have similar pronunciations in Middle Chinese, but they have different pronunciations in Mandarin: the morpheme 'low, humble' has been diphthongized into [pei], while the morpheme 'cover, protect' remains [pi]. Similarly, [i] in the morpheme 'love, flatter' (媚 [mijH]) in Middle Chinese is realized as [ei] in Mandarin, while [i] in the morpheme 'a kind of deer' (麋 [mij]) remains [i] in Mandarin. Chen (1976) suggests that diphthongization of [i] in the syllables [bi], [mi], [pi], and [pʰi] is a highly irregular process in the sense that neither manner of articulation of the bilabial initials, prosodic

features, nor fine distinctions among these syllables in Middle Chinese could explain why [i] has been diphthongized into [ei] in some syllables, while has remained [i] in other syllables. In the next section, statistical analysis is used to locate factors to account for these exceptions to diphthongization of [i] to [ei], i.e., those morphemes that remain [i] in Mandarin.

4   Exceptions vs. non-exceptions

The paper firstly relies on Baxter and Sagart (2014), with reference to *Zhongyuan Yinyun* (Rhymes of the Central Plain; Zhou, [1324] 1996) and reconstructions of *Zhongyuan Yinyun* in Pulleyblank (1984, 1991) and Chou (1993), to locate morphemes of [fi], [-ui], [vi], [bi], [mi], [pi], and [p^{h}i] in Middle Chinese. Then the paper uses the CCL Corpus (Center for Chinese Linguistics PKU) to look for related information concerning both morphemes that have been diphthongized into [-ei] and those that remain [-i] in Mandarin. The CCL Corpus is composed of two databases: an Old Chinese and Middle Chinese database and a Modern Mandarin database. It also permits searching for data according to Chinese dynasties. Another benefit of the corpus is its capacity, over 470 million Chinese characters from a wide range of sources.

Diphthongization of [i] to [ei] began no later than the end of the 16th century and finished no later than the beginning of the 19th century (Chen, 1976; Luo, 2008; Shen, 2020). As the time phase of diphthongization was mainly within the Qing dynasty (1644–1912), data from the Qing dynasty were extracted. In addition, data from the Ming dynasty (1368–1644) were also extracted to calculate frequency change from the Ming dynasty to the Qing dynasty.

Altogether this paper locates 201 related morphemes after the exclusion of obsolete morphemes. Among them, 109 morphemes have been diphthongized to [ei] in Mandarin, and 92 morphemes remain [i] in Mandarin. The factors for statistical analysis, their respective factor levels, and statistical analysis results are reported below in Table 2.

Table 2: Data and results for the binary logistic regression model
(exceptions vs. non-exceptions)

Factor	Factor level	Estimate	Std. Error	$|z|$	p	
Intercept		2.46	0.34	7.17	<0.00	*
Morphosyntactic structure (1644–1912)	Free Bound	0.27	0.43	1.21	0.63	
Word class (1644–1912)	Function Content	-1.37	0.91	1.50	0.79	
Normalized frequency (1644–1912)	Numerical	-0.00	0.00	0.13	0.90	
Frequency dummy (1644–1912)	Low Medium High	-0.08	0.04	1.96	0.04	*
Frequency change from between 1368 and 1644 to between 1644 and 1912	Numerical	-0.14	0.46	0.73	0.12	
Frequency change from between 1368 and 1644 to between 1644 and 1912 dummy	Decrease Increase	-0.17	0.38	0.46	0.65	

Notes: * = *p* < 0.05.
|*z*| stands for the absolute value of *z* as given in the GraphPad Prism version 8.0.0 for Windows.

4.1 Factors and factor levels in Table 2

As stated in Section 2.1.1, Donohue (2005) claims sound change advances in bound grammatical morphemes more completely than in free lexemes. Bybee (2002), Guy (1991b), Baranowski and Turton (2020), among others, claim that bound morphemes can affect /t, d/ deletion in English. Thus, the factor of *morphosyntactic structure (1644–1912)* with two levels *free* and *bound* was constructed to test whether the morphosyntactic structure has a role in the sound change. The time period *1644–1912* means that the data were extracted from the language dating to the Qing dynasty (1644–1912). *Free* and *bound* respectively mean that a related morpheme is mainly used either as a free or as a bound morpheme. "If a morpheme can stand alone in an utterance to represent a … part of speech …, it is free. If it must be augmented with additional language material …, it is bound" (Packard, 2015, p. 264; see also Chao, 1968; Hsieh, 2016). To exemplify, the morpheme 'love, flatter' (媚) in Table 1 is a free morpheme because it can represent an adjective, while the morpheme 'box' (柜) is a bound morpheme since it must be used with another morpheme in a word.

The factor of *word class (1644–1912)* with two levels, namely the *content* and the *function* was configured to examine the contradictory claims concerning the role of

word classes in the sound change discussed in Crowley (1997), Blevins and Lynch (2009), and others in Section 2.1.3. Adjectives, nouns, verbs, etc., are classified as content words; adverbial conjunctions, auxiliary verbs, determiners, prepositions, quantifiers, etc are classified as function words, which is in line with the dichotomy of words in Phillips (1983, 2001, 2006). It may appear ideal to classify words into adjectives, adverbial conjunctions, determiners, nouns, prepositions, verbs, etc, however, for statistical analysis, we shall avoid too many factor levels.

As noted in Section 2.2.1, the frequency factor is claimed to be associated with exceptions to sound change by some scholars, although disagreement exists concerning whether high-frequency words or low-frequency words lead sound changes (Grimm, 1822; Bybee, 1985, 2000, 2002; Pierrehumbert, 2001; Bermúdez-Otero, 2007; Smith, 2012; Hay and Foulkes, 2016). The raw numerical data of *frequency between 1644 and 1912* were examined in the first place. The cross-tabulation analysis carried out on the GraphPad Prism version 8.0.0 for Windows (hereafter the GraphPad software) revealed that it was rejected as a statistically significant factor for diphthongization (p = 0.63). Thus the raw data of frequency between 1644 and 1912 were normalized on the GraphPad software and reported as *normalized frequency (1644–1912)* in Table 2.

The factor of *frequency dummy (1644–1912)* was configured following the debate in Section 2.2 concerning whether low-frequency words or high-frequency words lead sound changes. It was also partly configured in line with claims in Wedel et al. (2013) and Liu (n.d.). Wedel et al. (2013) claim that the relative frequency of minimal pair members, instead of the absolute frequencies, is a significant predictor of phoneme merger. Liu (n.d.) compares the relative frequencies of all morphemes involved in palatalization in Mandarin and finds that relative frequency is statistically significantly correlated with it. The present study does not refer to pair members and thus does not refer to the relative frequency of pair members. Instead, the present study refers to all morphemes involved in diphthongization. Therefore, it took into account the relative frequencies of all related morphemes following Liu (n.d.). The factor *frequency dummy (1644–1912)* has three factor levels: *low*, *medium*, and *high*. Each level takes one-third of the data: one-third of the data with the lowest frequencies in this column is marked as *low*; another one-third with the highest frequencies is marked as *high*; the remaining one-third between *low* and *high* is *medium*. As a result, *low* here does not refer to a frequency lower than a specific count. Instead, it means that the frequency of a certain morpheme is among the lowest frequencies of all morphemes involved in the diphthongization of [i] to [ei].

The next factor, namely the *frequency change from between 1368 and 1644 to between 1644 and 1912* (henceforth *frequency change*) was introduced due to the possibility that morphemes with either increased or decreased frequencies may have undergone diphthongization at different rates. For example, the frequency of the

morpheme 'not' (非 [fei]) in Table 1 is 5633 in the period from 1368 to 1644 and 14457 in the period from 1644 to 1912. Thus the frequency change for the morpheme 'not' (非 [fei]) is 8824, where the positive number means that comparing the first and the second period, its frequency has increased. It is possible to normalize the raw data of frequency change by adding all numbers with the absolute of the most negative. In this way, the most negative number will become zero, and all the other numbers become positive. However, the focus of the frequency change factor is partly on whether related morphemes have increased or decreased frequency. Therefore, the present study will use raw data instead of normalized data.

The factor of *frequency change from between 1368 and 1644 to between 1644 and 1912 dummy* (hereafter *frequency change dummy*) was introduced because it seems that whether frequency has decreased or increased is also a factor in sound change. The raw numerical data for the above factor *frequency change* were converted to categorical data with two levels: *decrease* and *increase*, with *decrease* as the reference level. The morpheme 'not' (非 [fei]) in Table 1 is exemplified again: its frequency increased 8824 times from between 1368 and 1644 to between 1644 and 1912, so it is marked as an *increase* for the factor of *frequency change dummy*.

4.2 The binary logistic regression results

The statistical relationship between the six factors in Table 2 and the dependent variable diphthongization of [i] to [ei] was assessed using multiple logistic regression in the GraphPad software. Model selection was guided by AIC (Akaike Information Criterion; Akaike, 1974; Burnham and Anderson, 2004), calculated probability (*p*-value), and VIF (Variance Inflation Factor; Rawlings et al., 1998; James et al., 2017). The dependent variable has two categories: *undiphthongized* and *diphthongized*, with *undiphthongized* as the reference level. *Undiphthongized* means that a related morpheme remains [i] in Mandarin, while *diphthongized* means that a relevant morpheme has been diphthongized to [ei]. A *p*-value smaller than 0.05 was considered statistically significant.

As shown in Table 2, the *frequency dummy (1644–1912)* is the only factor that has emerged as statistically significant (*p* = 0.04). Its negative coefficient indicates that the possibility of undergoing diphthongization decreases from the low level, medium level, to the high level (Estimate = -0.08). In other words, the higher the frequency of a morpheme is, the less likely for it to undergo diphthongization. Morphemes with the highest frequencies tend to be exceptions to diphthongization. Other factors like *morphosyntactic structure*, *word class*, and *frequency change* have been rejected as statistically significant.

5   Early application vs. late application

This section considers the morphemes that have been diphthongized to [ei] in Mandarin to explore factors that account for morphemes leading this sound change process.

5.1   Data source

This paper relies on works compiled at the beginning of diphthongization to look for morphemes that were at the forefront of this process and then carries out statistical analysis. It may seem that works compiled by Chinese scholars in the 17th century are the best choice since diphthongization began no later than the end of the 16th century (Chen, 1976; Luo, 2008; Shen, 2020). However, the choice is not as straightforward as it appears to be: dictionaries and books compiled by Chinese scholars before the 20th century use *fanqie* (反切), a traditional method of indicating the pronunciation of a Chinese character by using two other Chinese characters. For example, the pronunciation of the character 唐 might be represented as the following: 徒郎. It roughly means that the initial of 唐 is the same as that of 徒, and the final of 唐 is the same as that of 郎. This representation makes it circular and thus difficult to understand the pronunciations of characters and morphemes they represent since the Chinese writing system is a representative logographic system, not a phonographic system like English. In contrast, dictionaries compiled by missionaries to China use a Romanization system and can provide a relatively clear picture of the pronunciation of Mandarin during the time concerned.

Aid to the Eyes and Ears of Western Literati (*Xiruermuzi*; Trigault, [1626] 1957) is a dictionary that contains the first Romanization system of the Chinese written language, and is an essential guide to the pronunciation of Chinese characters (Wang, 2016; Li, 2020).

As noted in Section 4, this paper locates 201 morphemes related to the diphthongization of [i] to [ei]. In this section, we focus solely on the 109 morphemes that have been diphthongized to [ei] in Mandarin. The paper uses *Aid to the Eyes and Ears of Western Literati* (Trigault, [1626]1957) to locate morphemes that were at the forefront of diphthongization of [i] to [ei]. Then the paper uses the CCL corpus to look for related information concerning all the 109 morphemes and locate factors in morphemes leading the diphthongization.

5.2   Statistical analysis results: early application vs. late application

The dictionary *Aid to the Eyes and Ears of Western Literati* (Trigault, [1626] 1957) was compiled during the time of the Ming dynasty (1368–1644). Accordingly, the focus of

this research is on the data from the Ming dynasty, and frequency change is assumed to be a factor in sound change. To be able to calculate the frequency change, the data related to frequency that came before the Ming dynasty is needed. The Ming dynasty was preceded by the Yuan dynasty (1271–1368). However, the Yuan dynasty lasted for less than one hundred years and there is little data to work on. As a result, this study extracted data from both the Yuan dynasty and the Southern Song dynasty (1127–1279). To simplify it for readers who are unfamiliar with Chinese history, the paper henceforth refers to *frequency change from the Southern Song and Yuan dynasties to the Ming dynasty* as *frequency change from between 1127 and 1368 to between 1368 and 1644*. The data and statistical analysis results are shown in Table 3.

Table 3: Data and results for the binary logistic regression model
(early application vs. late application)

Factor	Factor level	Estimate	Std. Error	$|z|$	p	
Intercept		-0.69	0.215	4.10	0.02	*
Normalized frequency (1368–1644)	Numerical	0.01	0.01	1.37	0.21	
Frequency dummy (1368–1644)	Low Medium High	0.09	0.02	1.98	0.03	*
Frequency change from between 1127 and 1368 to between 1368 and 1644	Numerical	0.61	0.27	1.34	0.15	
Frequency change from between 1127 and 1368 to between 1368 and 1644 dummy	Decrease Increase	0.52	0.28	1.89	0.04	*

Notes: * = $p < 0.05$.
$|z|$ stands for the absolute value of z as given in the GraphPad software.

The raw data of frequency (1368–1644) does not show the normal distribution and were thus normalized on the GraphPad software, and reported as *normalized frequency (1368–1644)*.

The binary logistic regression analysis carried out on the GraphPad software revealed that the factors of *frequency dummy (1368–1644)* and *frequency change from between 1127 and 1368 to between 1368 and 1644 dummy* have statistically significant correlation with morphemes leading diphthongization of [i] to [ei] (p = 0.03, 0.04). The positive value of the coefficient of *frequency dummy (1368–1644)* suggests that this factor has an additive effect on diphthongization (Estimate = 0.09): the possibility of leading diphthongization increases from the low level, medium level, to the high level of *frequency dummy (1368–1644)*. In a similar vein, the positive value of the coefficient of *frequency change from between 1127 and 1368 to between 1368 and 1644 dummy*

indicates that a morpheme that has an increased frequency tends to undergo diphthongization first (Estimate = 0.52). To sum up, the two factors to account for morphemes leading diphthongization of [i] to [ei] are *frequency* and *frequency change*.

5.3   Frequency: exceptions vs. early application

Another statistical analysis was carried out to examine differences between the frequencies of the exceptional morphemes to diphthongization of [i] to [ei] in the time frame of 1644 to 1912 and the frequencies of the leading morphemes in the time frame of 1368 to 1644. The two sets of data were firstly normalized on the GraphPad software. A Mann-Whitney test carried out on the same software shows that statistically significant differences exist between the two sets of data (p = 0.04). Descriptive statistical analysis of the raw data was also carried out on the GraphPad software. The mean frequency of the exceptional morphemes is 926.8, while the mean frequency of the leading morphemes is 658.3. The highest frequency of the exceptional morphemes is 8229, while the highest frequency of the leading morphemes is 5662. Both the mean frequency and the highest frequency of the exceptional morphemes are about 1.4 times higher than those of the leading morphemes. The present study cannot draw any conclusions on whether the multiple of 1.4 may be universal or whether it changes from one sound change process to another. To do so, more sound change processes within one language as well as sound processes across languages need to undergo the analyses. The topic proves to be a boon for later research.

6   Conclusion

Word class and morphosyntactic structure, or morphological condition in a broader term, have been rejected as statistically significant both in accounting for exceptions to diphthongization and leading morphemes in diphthongization. This suggests that morphological condition does not have a role in diphthongization.

Previous studies mainly focus on whether high-frequency morphemes or low-frequency morphemes lead sound change. The present study reveals that the role of the frequency factor in sound change is more complex than the debate in previous studies. The frequency factor has been proven as statistically significant both for exceptions to diphthongization of [i] to [ei] and leading morphemes in it. What proved interesting is the following two points: (1) high-frequency morphemes tend to be exceptions to diphthongization; (2) among morphemes that underwent diphthongization, high-frequency morphemes tend to lead diphthongization. More plainly, morphemes with the highest frequency tend to be exceptions to diphthongization. High-frequency morphemes among non-exceptional morphemes

tend to lead diphthongization, although the frequencies of these high-frequency morphemes tend to be less than the frequencies of exceptional morphemes.

Another statistically significant factor to explain leading morphemes in diphthongization is frequency change: a morpheme that has an increased frequency tends to undergo diphthongization earlier. This factor has not emerged as statistically significant to account for exceptions to diphthongization. Put differently, frequency change is not correlated with exceptions to diphthongization but correlated with early application of diphthongization. This factor has generally been overlooked in previous studies.

It can be debated that the conclusions in this paper are based on one language sample and cannot be applied to all languages. To answer this question, research into a parallel process in another language is clearly called for. Such an analysis constitutes an exciting area for future research.

Acknowledgments

References

Abramowicz, Ł. (2007). Sociolinguistics meets Exemplar Theory: Frequency and recency effects in (ing). *U. Penn Working Papers in Linguistics, 13*(2), 15–16.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Anttila, R. (1972). *An introduction to historical and comparative linguistics*. New York: Macmillan.

Baranowski, M., & Turton, D. (2020). TD-deletion in British English: New evidence for the long-lost morphological effect. *Language Variation and Change, 32*(1), 1–23.

Baxter, W. H., & Sagart, L. (2014). *Old Chinese: A new reconstruction*. New York: Oxford University Press. https://ocbaxtersagart.lsait.lsa.umich.edu

Bermúdez-Otero, R. (2007). Diachronic phonology. In P. de Lacy (Ed.)*, The Cambridge handbook of phonology* (pp. 497–517). Cambridge: Cambridge University Press.

Blevins, J., & Lynch, J. (2009). Morphological conditions on regular sound change?: A reanalysis of *l-loss in Paamese and Southeast Ambrym. *Oceanic Linguistics, 48*(1), 111–129.

Brown, E. L. (2013). Word classes in phonological variation: Conditioning factors or epiphenomena? In C. Howe, S. Blackwell & M. Lubbers Quesada (Eds.)*, Selected proceedings of the 15th Hispanic linguistics symposium* (pp. 179–186). Somerville, MA: Cascadia Proceedings Project.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research,* 33, 261–304.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form.* Amsterdam: John Benjamins.

Bybee, J. L. (2000). The phonology of the lexicon: Evidence from lexical diffusion. In M. Barlow & S. Kemmer (Eds.)*, Usage-based models of language* (pp. 65–85). Stanford: CSLI.

Bybee, J. L. (2002). Lexical diffusion in regular sound change. In D. Restle & D. Zaefferer (Eds.)*, Sounds and systems: Studies in structure and change* (pp. 59–74). Berlin: Mouton de Gruyter.

Campbell, L. (1974). On conditions on sound change. In J. Mathieson Anderson & C. Jones (Eds.)*, Historical linguistics: Proceedings of the First International Conference on Historical Linguistics*, Vol. 2, Theory and description in phonology (pp. 88–96). Amsterdam: North Holland.

Campbell, L. (2013). *Historical linguistics* (3rd edn). Cambridge, MA: MIT Press.

Cerrón-Palomino, R. (1974). Morphologically conditioned changes in Wanka-Quechua. *Studies in the Linguistic Sciences, 4*(2), 40–75.

Chafe, W. L. (1968). The ordering of phonological rules. *International Journal of American Linguistics,* 34, 115–136.

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

Chen, M. Y. (1976). From Middle Chinese to modern Peking. *Journal of Chinese linguistics, 4*(2/3), 113–277.

Chen, M. Y., & Wang, W. S.-Y. (1975). Sound change: Actuation and implementation. *Language, 51*(2), 255–281.

Chou, F.-K. (1993). *A pronouncing dictionary of Chinese characters in Archaic & Ancient Chinese, Mandarin & Cantonese*. Beijing: Zhonghua Book.

Coetzee, A., & Kawahara, S. (2013). Frequency biases in phonological variation. *Natural Language and Linguistic Theory, 31*(1), 47–89.

Crowley, T. (1997). *An introduction to historical linguistics* (3rd edn). New York: Oxford University Press.

Dinkin, A. (2008). The real effect of word frequency on phonological variation. *University of Pennsylvania Working Papers in Linguistics,* 14, 97–106. http://itre.cis.upenn.edu/myl/papers/Dinkin2007.pdf (accessed December 14, 2020).

Donohue, M. (2005). Syntactic and lexical factors conditioning the diffusion of sound change. *Oceanic Linguistics, 44*(2), 427–442.

Duncan, L. C. (2011). Variation in Finnish loan words: Evidence from Google. In Ain Haas & Peter B. Brown (Eds.), *Proceedings of the XIVth, XVth, and XVIth Conferences of the Finno-Ugric Studies Association of Canada: The Uralic World and Eurasia* (pp.107–126). Providence: Rhode Island College.

Edkins, J. (1857). *A grammar of the Chinese colloquial language, commonly called the Mandarin dialect*. Shanghai: London Mission.

Fasold, R. W. (1972). *Tense marking in Black English: A linguistic and social analysis*. Arlington, VA: Center for Applied Linguistics.

Fidelholtz, J. (1975). Word frequency and vowel reduction in English. *Chicago Linguistic Society,* 11, 200–213.

Fruehwald, J., Gress-Wright, J., & Wallenberg, J. C. (2013). Phonological rule change: The constant rate effect. In S. Kan, C. Moore-Cantwell & R. Staubs (Eds.)*, NELS 40: Proceedings of the 40th annual meeting of the north east linguistic society*, Vol. 1 (pp. 219–230). California: CreateSpace Independent Publishing Platform. https://www.pure.ed.ac.uk/ws/portalfiles/portal/14416788/Fruewald_Gress_Wright_W allenberg_Phonological_Rule_Change.pdf (accessed December 12, 2020).

Fudge, E. C. (1972). Review of Postal 1968. *Journal of Linguistics,* 8, 136–156.

GraphPad Prism. Version 8.0.0 for Windows. San Diego, California: GraphPad Software, 2019. Computer software.

Grimm, J. (1822). *Deutsche Grammatik*, Vol. 1. Göttingen: Dietrich.

Guy, G. R. (1991a). Contextual conditioning in variable lexical phonology. *Language Variation and Change, 3*(2), 223–239.

Guy, G. R. (1991b). Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change, 3*(1), 1–22.

Handel, Z. (2015). Old Chinese phonology. In W. S-Y. Wang & C. Sun (Eds.)*, The Oxford handbook of Chinese linguistics* (pp. 68–79). New York: Oxford University Press.

Hay, J. B., Pierrehumbert, J. B., Walker, A. J., & LaShell, P. (2015). Tracking word frequency effects through 130 years of sound change. *Cognition,* 139, 83–91.

Hay, J. B., & Foulkes, P. (2016). The evolution of medial /t/ over real and remembered time. *Language, 92*(2), 298–330.

Hooper, J. B. (1976). Word frequency in lexical diffusion and the source of morphophonological change. In W. M. Christie (Ed.)*, Current progress in historical linguistics* (pp. 95–105). Amsterdam: North-Holland Publishing Company.

Hsieh, S.-K. (2016). Chinese linguistics: Semantics. In S.-W. Chan (Ed.)*, The Routledge encyclopedia of the Chinese language* (pp. 203–214). New York: Routledge.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to statistical learning* (8th edn). New York: Springer.

Jasanoff, J. (1971). A generative approach to historical linguistics. *Romance Philology,* 25, 74–85.

Jespersen, O. ([1922] 2007). *Language: Its nature and development*. London: Routledge.

Johnson, S. (1982). Morphological influences on sound change. In A. Ahlqvist (Ed.)*, Papers from the 5th international conference on historical linguistics* (pp. 171–175). Amsterdam: John Benjamins.

King, R. D. (1969). *Historical linguistics and generative grammar*. Englewood Cliffs, NJ: Prentice-Hall.

Kiparsky, P. (2014). New perspectives in historical linguistics. In C. Bowern & B. Evans (Eds.)*, The Routledge handbook of historical linguistics* (pp. 64–102). New York: Routledge.

Kroch, A. S. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change,* (1), 199–244.

Kroch, A. S. (1994). Morphosyntactic variation. In K. Beals (Ed.)*, Proceedings of the thirtieth annual meeting of the Chicago Linguistics Society*, Vol. 2 (pp. 180–201).

Labov, W. (1989). Exact description of the speech community: Short a in Philadelphia. In R. W. Fasold & D. Schiffrin (Eds.)*, Language change and variation* (pp. 1–57). Amsterdam: John Benjamins.

Labov, W. (2010). *Principles of linguistic change,* Vol. 3: Cognitive and cultural factors. Oxford: Blackwell.

Labov, W., Cohen, P., Robins, C., & Lewis, J. (1968). *A study of the nonstandard English of Black and Puerto Rican speakers in New York City (Cooperative Research Report No. 3288).* Washington, DC: U.S. Office of Education.

Li, W.-C. (1999). *A diachronically-motivated segmental phonology of Mandarin Chinese*. New York: Peter Lang.

Li, Y. (2020). *The Chinese writing system in Asia: An interdisciplinary perspective*. New York: Routledge.

Liu, S. (n.d.) *Factors in sound change: A quantitative analysis of palatalization in Mandarin*.

Luo, C. 罗常培 (2008). Hanyu yinyunxuede wailai yingxiang 汉语音韵学的外来影响 (Foreign Influences on Chinese phonological study). In *Luochangpei wenji 罗常培文集 (The collected linguistic works of Luo Changpei)*, Vol. 8. Jinan: Shandong Education.

Malkiel, Y. (1976). Multi-Conditioned sound change and the impact of morphology on phonology. *Language, 52*(4), 757–778.

Manker, J. (2015). Phonetic sources of morphological patterns in sound change: Fricative voicing in Athabascan. In *UC Berkeley Phonology Lab Annual Report 2015* (pp. 243–294).

Melchert, H. C. (1975). 'Exceptions' to exceptionless sound laws. *Lingua, 35*(2), 135–153.

Ogura, M. (2012). The timing of language change. In J. M. Hernández-Campoy & J. C. Conde-Silvestre (Eds.)*, The handbook of historical sociolinguistics* (pp. 427–450). West Sussex: Blackwell.

Packard, J. L. (2015). Morphology: Morphemes in Chinese. In W. S-Y. Wang & C. Sun (Eds.)*, The Oxford handbook of Chinese linguistics* (pp. 263–305). New York: Oxford University Press.

Pan, W., & Zhang, H. (2015). Middle Chinese phonology and Qieyun. In W. S-Y. Wang & C. Sun (Eds.)*, The Oxford handbook of Chinese linguistics* (pp. 80–90). New York: Oxford University Press.

Peyraube, A. (2020). Ancient Chinese. In S.-W. Chan (Ed.)*, The Routledge encyclopedia of the Chinese language* (pp. 1–17). New York: Routledge.

Phillips, B. S. (1983). Lexical diffusion and function words. *Linguistics,* 21, 487–499.

Phillips, B. S. (1984). Word frequency and the actuation of sound change. *Language,* 60, 320–342.

Phillips, B. S. (2001). Lexical diffusion, lexical frequency, and lexical analysis. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 123–136). Amsterdam: John Benjamins.

Phillips, B. S. (2006). *Word frequency and lexical diffusion*. New York: Palgrave Macmillan.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 137–157). Amsterdam: John Benjamins.

Pintzuk, S. (1991). *Phrase structures in competition: Variation and change in Old English word order* (Doctoral dissertation). University of Pennsylvania, Philadelphia, Pennsylvania.

Postal, P. (1968). *Aspects of phonological theory*. New York: Harper and Row.

Pulleyblank, E. G. (1984). *Middle Chinese: A study in historical phonology*. Vancouver: University of British Columbia Press.

Pulleyblank, E. G. (1991). *Lexicon of reconstructed pronunciation in Early Middle Chinese, Late Middle Chinese, and Early Mandarin*. Vancouver: University of British Columbia Press.

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (1998). *Applied regression analysis: A research tool* (2nd edn). New York: Springer.

Renwick, M., Baghai-Ravary, L., Temple, R., & Coleman, J. (2014). Deletions in big data? The phonetics of word-final (t,d) in the Audio BNC. Oral presentation at BAAP 2014 (British Association of Academic Phoneticians).

Robinson, O. W. III, & Coetsem, F. van. (1973). Review article of King 1969. *Lingua,* 31, 331–369.

Rochet, B. (1974). A morphologically-determined sound change in Old French. *Linguistics: An International Review,* 135, 43–56.

Santorini, B. (1992). Variation and change in Yiddish subordinate clause word order. *Natural Language & Linguistic Theory,* 10, 595–640.

Santorini, B. (1993). The rate of phrase structure change in the history of Yiddish. *Language Variation and Change,* 5, 257–283.

Sihler, A. (2000). *Language history: An introduction*. Amsterdam: Benjamins.

Shen, Z. (2020). *A phonological history of Chinese*. New York: Cambridge University Press.

Shi, Y. (2002). *The establishment of Modern Chinese grammar: The formation of the resultative construction and its effects*. Amsterdam: John Benjamins.

Smith, K. A. (2012). Frequency and language change. In A. Bergs and L. J. Brinton (Eds.)*, English historical linguistics: An international handbook*, Vol. 2 (pp. 1531–1546). Berlin: Mouton de Gruyter.

Tamminga, M. (2014). Sound change without frequency effects: Ramifications for phonological theory. In R. E. Santana-LaBarge (Ed.)*, Proceedings of the 31st West Coast Conference on Formal Linguistics* (pp. 457–465). Somerville, MA.

Thomsen, V. (1879). Andare-andar-anar-aller: En kritisk-etymologisk undersøgelse. *Det philolhist*, 197–214. Kjøbenhavn: Samfunds mindeskrift.

Thomsen, V. (1920). Den gotiske sprogklasses inflydelse på den finske. En sproghistorisk undersøgelse. In *Samlede afhandlinger*, Vol. 2 (pp. 49–264). Copenhagen: Kristiana.

Trigault, N. ([1626] 1957). *Xiruermuzi (Aid to the eyes and ears of Western literati)*. Beijing: Wenzigaigechubanshe.

Van Bergem, D. (1995). Acoustic and lexical vowel reduction. Amsterdam: IFOTT.

Vincent, N. (1978). Is sound change teleological? In J. Fisiak (Ed.)*, Recent developments in historical phonology* (pp. 409–430). New York: De Gruyter Mouton.

Walker, J. A. (2012). Form, function, and frequency in phonological variation. *Language Variation and Change, 24*(3), 397–415.

Wang, L. 王力 (1957). *Hanyu shigao 汉语史稿 (A sketch of the history of Chinese)*. Beijing: Science.

Wang, W. S.-Y. (1969). Competing changes as a cause of residue. *Language,* 45, 9–25.

Wang, W. S.-Y. (2016). Chinese Linguistics. In S.-W. Chan (Ed.)*, The Routledge encyclopedia of the Chinese language* (pp. 152–183). New York: Routledge.

Wedel, A., Jackson, S., & Abby, K. (2013). Functional load the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and Speech, 56*(3), 395–417.

Zellou, G., & Tamminga, M. (2014). Nasal coarticulation changes over time in Philadelphia English. *Phonetics*, 47, 18–35.

Zhou, D.Q. 周德清 ([1324] 1996). Zhongyuan Yinyun 中原音韵. In *Chuanshi cangshu jingku 传世藏书经库, Vol. 2 Language and word*. Hainan: Hainan International.

Zwicky, A. M. (1970). Auxiliary reduction in English. *Linguistic Inquiry,* 1, 323–336.

Word-Prosodic Typology: The Traps of Seemingly Similar Japanese and Slovene

Nina GOLOB
University of Ljubljana, Slovenia
nina.golob@ff.uni-lj.si

Abstract

The article briefly describes the historical development of language prosodic typology, introduces the two word-prosodic prototypes proposed by Hyman, and explains the positioning of pitch-accent languages on the lexical level. It points out the false similarity between Japanese and Slovene that was created with the introduction of the feature [±culminative] and proposes to expand it with the feature [±eliminative], which phonetically justifies the difference between pitch-accent systems and the stress-accent prototype.

Keywords: prosodic typology, features, pitch-accent languages, Japanese, Slovene

Povzetek

Članek na kratko opisuje zgodovinski razvoj prozodične tipologije jezikov, predstavi dva prozodična prototipa na besednem nivoju, ki ju je predlagal Hyman, in pojasnjuje položaj tonemskih jezikov na leksikalnem nivoju. Opozarja na lažno podobnost japonščine in slovenščine, ki je nastala z uvedbo značilnosti [±kulminativen], in predlaga njeno razširitev s funkcijo [±eliminativ], ki fonetično utemeljuje razliko med tonemskimi sistemi in jakostno-naglasnim prototipom.

Ključne besede: prozodična tipologija, razločevalne lastnosti, tonemski jeziki, japonščina, slovenščina

1   Introduction

Prosodic typology classified world languages by setting two opposite prototypes, tone languages such as Cantonese, Yoruba, etc. with the feature [+tonal] and stress languages such as English, Turkish, etc. with the feature [–tonal]. In the history of prosodic research, research on tone languages and their prototype progressed quickly and successfully implemented the binary tonal distinction *high* (H) and *low* (L) on each segment, leaving out the so-called pitch-accent languages. Stress on the other hand was phonetically elusive and was considered a mental construct. The already marginal phonological status of stress was weakened even further when the binary tone system proved to be applicable for intonation studies. This approach blurred pitch-accent languages such as Japanese, Swedish, etc. with stress languages because they share a common property; that is the feature [+culminative] also called *accent*.

Intonation phonology became the means of comparison among languages. The ToBI models, the transcription and annotation tools of prosodic events which would include both intonation and voice flow segmentation in units of study, define whether languages differ in the types of tones or/and tonal inventories they have, and consequently devide languages to tone languages, accent languages, and languages with no lexical specification of prosody.

Though ToBI models are indispensable in computer technology, which requires automated analysis of large speech corpora annotated with standardized annotation strings, Jun (2005, p. 437) points out that comparisons of prosodic systems based on phonetic descriptions show certain limitations. One very important limitation is that the similarities shown in the surface realization do not guarantee the same underlying distinctive prosodic features or structures and may be entirely accidental (also Gussenhoven, 2007; Ladd, 2008 [1996]; Hyman, 2011). The types of tones cannot distinguish stress-accent languages from lexical pitch-accent languages because the autosegmental-metrical model (AM model) does not specify whether pitch accent is a lexical property or a postlexical property.

As an example of such coincidence, Gussenhoven (2007, p. 256) points at the surface similarity between English and Tokyo Japanese H*L to write that 'while phonologically comparable, the pitch accents of Japanese and English have very different morphological statuses'. In Japanese, they form part of the underlying phonological specification of morphemes, along with the vowels and consonants. In English, on the other hand, pitch accents are intonational and therefore morphemically independent of the words they come with, and are chiefly used to express the information status of the expression. Closely related to this is also the false similarity of surface representations of different accent patterns in declarative intonation presented for Japanese and Slovene (Golob, 2011).

Therefore, this research will return to the so-called broad-stroke typology, where phonological systems are treated level-ordered, cf. the prosodic property of an utterance is a combination of prosody at the lexical level and prosody at the post-lexical level, with the former constraining the latter and the latter including the prosodic features of the former.

The structural approach, where there is a clear distinction between word-level tones and stress at the lexical level, is indispensable for practically any interdisciplinary research involving accounting for the structural properties of phonological systems (and their interface with morphology and syntax), predicting the effects that stress (but not tone) can have on segments, tracing linguistic change, conducting fieldwork on understudied and endangered languages, and last but not least, explaining foreign accents in second language acquisition.

2   Structural approach to prosody and difficulties in L2 acquistion

Foreign accents in second language production are caused by interference from the phonological system and phonetic realization of the speaker's first language.

Within the area of prosody, several studies have reported that lexically linked prosodic features in L1 are more likely to be transferred to L2 prosody and are more difficult to suppress than the post-lexical ones (Jun & Oh, 2000; Ueyama, 2000; Mennen, 2007; Golob, 2021).

Furthermore, sudden changes at the paralinguistic level of L2 speech, such as the inclusion of prosodic focus or the use of emotional speech is reported to destroy the already correctly adopted lexical or intonational prosody (Golob, 2008; van Maastricht et al. 2016; Kim, 2018).

The above findings show that processes that contribute to the foreign-accentedness in second language production can best be explained level-ordered or in other words, through the recognition of the properties involved at different prosodic levels and their mutual interactions.

3   The aim of this study

From teaching experience to Slovene students of Japanese as well as based on the findings about foreign accents in second language production, this study will introduce and evaluate the present word-prosodic typology proposed by Hyman (2006, 2009) through the results of a large acoustic survey that was recently conducted by Golob (2021).

4   Word-prosodic typology

Prosodic properties conveyed in an utterance are a combination of prosodic features at the word level and those at the phrase level and that postlexical prosody is constrained by the lexical prosody, and postlexical prosodic information contains information about the lexical prosody.

Prosodic typology revisited its foundations (word-prosodic typology) to redefine stress-accent prototype by the properties that would describe both the underlying distinctive prosodic features and their organization. Hyman (2006, p. 231) proposed an additional [+obligatory] saying that stress-accent languages meet the following two central criteria:

11. obligatoriness: every lexical word has at least one syllable marked for the highest degree of metrical prominence (primary stress);
12. culminativity: every lexical word has at most one syllable marked for the highest degree of metrical prominence.

By setting the stress language prototype more clearly, classification of languages according to the properties of their subsystems became more straightforward. Pitch-accent systems convey the features [−obligatory] and [+culminative], and based on numerous researches Hyman reckons them as 'mixed, ambiguous, and sometimes analytically indeterminate systems' that do not constitute a coherent prosodic type but instead 'freely pick-and-choose properties from the tone and stress prototypes' (Hyman, 2009, p. 213).
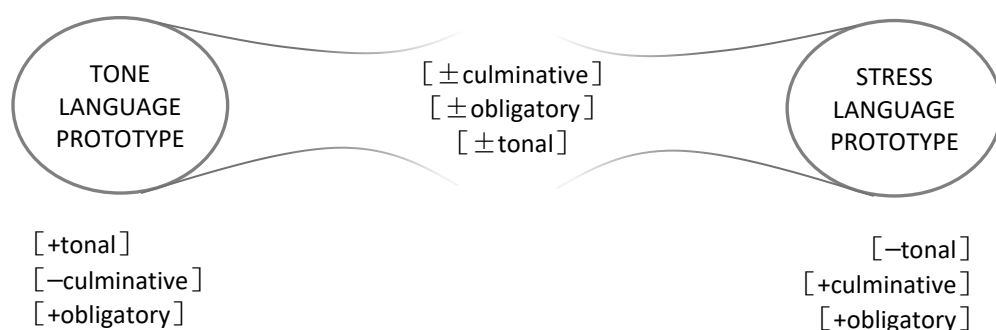

Figure 1: Word-prosodic typology according to Hyman (taken from Golob, 2021, p. 20)

The basic distinction [±obligatory] satisfies the difference between Tokyo Japanese as a pitch-accent language and a variation of Standard Slovene as a stress-accent language. Both languages convey the features [+culminative].

Tokyo Japanese is known as a typical pitch-accent or non-stress language in the literature as opposed to a stress-accent language like English (McCawley, 1978;

Beckman, 1986). It is also classified as a word-pitch language as opposed to a 'tone language' like Mandarin Chinese or an intonation language like English (Pike, 1948). It carries a distinctive lexical pitch accent, which is marked phonetically by the tonal change from H to L (Pierrehumbert & Beckman, 1988; Kubozono, 2008). Tokyo Japanese accent/tone is culminative, a property it shares with stress-accent systems. However, the lexicon is divided into tonic words (accented type) with H-L tonal change and atonic words (unaccented type) that convey no such H-L change. In other words, any given word in theory gets *n+1* accent possibilities, *n* being equal to the number of full moras, tone-bearing units in Tokyo Japanese (Labrune, 2012).

(1)		'pillow'+nom.	'heart'+nom.	'mirror'	'fish'
	a. accentual	maˀkura ga	kokoˀro ga	kagamiˀ ga	sakana ga
	b. tonal	MAkura ga	koKOro ga	kagaMI ga	sakana ga
		H L	H L	H L	

Some major Japanese dialects are reported to deviate from the standard pitch accent, mainly differing in the number of tonal patterns involved, and few of them are accentless (Uwano, 1999; Kubozono, 2012)

Just as Tokyo Japanese was the base for the so-called standard language, Standard Slovene was also constructed upon dialects. There are two prosodically distinct dialect types, the tonal or pitch-accent Slovene and the non-tonal or stress-accent Slovene (Toporišič, 2004 [1976]; Šuštaršič & Tivadar, 2001). The pitch-accent Slovene has distinctive tones, namely the acute (a long rising tone) and the circumflex (a long falling tone) that appear on long stressed vowels. In the absence of a long vowel stress falls on the final syllable still carrying tones.

Fixed stress is the norm in Slovene. It is obligatory on every lexical word. In stress-accent Slovene stressed syllable is prominent in the sense that it is longer and conveys higher tone and greater dynamics compared to unstressed syllables (Lehiste, 1970; Bhaskararao & Golob, 2006).

(2)		'saussage'	'cupboard'
	a. accentual	klo*bása	o*mâra
	b. tonal	klobasa	omara
		L H	H L

In overall, the three standard languages, the Tokyo pitch-accent Japanese, the pitch-accent Slovene, and the stress-accent Slovene are described with the following prosodic features according to Hyman's word-prosodic typology.
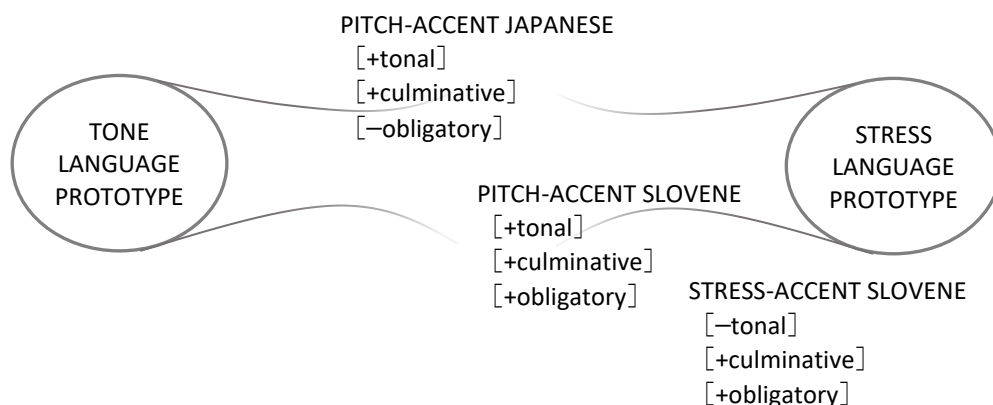
Figure 2: Japanese and Slovene according to Hyman's word-prosodic typology

According to Figure 2, by introducing the feature [±obligatory] the three language systems became prosodically distinct, which, from the point of view of phonology, could be completely satisfactory. However, the feature [±obligatory] alone does not make it possible to understand the prosodic typological differences between the two languages. It seems that it is not directly applicable neither to prosodic function nor to the nature of the stress language prototype, and would only partially or not at all help explain processes that appear during seond language acquisition.

In the following section, we will therefore introduce a bidirectional Japanese – Slovene L1 and L2 study that was conducted by Golob (2021), and of which acoustic measurements indicate a prosodic property of a stress-accent language, which is very obvious and is a current topic in phonetic research, but has been overlooked in discussions on prosodic typology features.

5   Stress language prototype revisited: the feature [±eliminative]

Golob (2021) conducted an acoustic experiment on Japanese and Slovene as native languages (L1) to show that, although the [+culminative] feature is common to both languages, there is a difference in the parameters responding to it as well as the way they respond. Furthermore, based on the "Integrated Contrastive Model" (Rasier & Hiligsmann, 2007) she observes how acoustic parameters respond to the feature [+culminative] in Japanese and Slovene as second languages (L2) to show that the prosodic mechanism at the word level is the most uncompromising in a language that establishes the overall prosodic circumstance.

Measured acoustic parameters, namely vowel formants, duration, fundamental frequency, and intensity match the four prosodies reported by Pfitzinger (2006),

thought to be essential for the linguistic aspect of prosody (vs. para-linguistic, extra-linguistic).

In general, the results for L1 Japanese and L1 Slovene show clear trends and support previous results. They serve as the benchmark for the L2 Japanese and L2 Slovene results and point out some new and interesting trends.

In L1 Japanese, pitch is the only prosodic feature that shows a systemic and uniform response, namely that accented vowels have a statistically higher pitch than the following vowels. In L1 Slovene, on the other hand, the pitch showed violent reactions but due to unclear tendencies, we consider it to be strongly structured. In other words, we assume that factors at higher metrical levels influence acoustic pitch values. The other three parameters in L1 Slovene show uniform responses; accented vowels are statistically longer than the following vowels, they show no apparent vowel reduction compared to the unaccented vowels, and they are statistically pronounced with higher intensity than the following vowels. The intensity response was rated as less reliable, with data showing statistical significance in three out of five informants.

Results for the second languages provide further important insights. L2 Japanese shows no correspondence to the [+culminative] feature, the deviation in the acoustic data is negligible for all speakers. On the other hand, L2 Slovene shows much more prosodic activity. The pitch showed violent responses as in L1 Slovene but the trend is unclear and requires further investigation. On the other hand, vowel formants are the only parameter that does not respond to the [+culminative] feature, and no vowel reduction is observed. In this context, the L2 Slovene manifestation of the duration response deserves further attention. Four out of five informants showed statistically greater duration on accented vowels and at the same time no vowel reduction, suggesting that Japanese speakers of Slovene used the segmental long-short distinction found in their native language to respond to the [+culminative] feature.

The above results suggest that the interpretation of word-level syntagmatic prominence in the case of stress language prototype needs to be reconsidered, and as suggested, should be defined bidirectionally. To rephrase, a part of a phonological word is prominent, either because the parameters of the outstanding part are in some way superiorized compared to those of the rest of the word (maximizing the paradigmatic opposition), or because the parameters of the rest of the word are in some way inferiorized (minimizing the paradigmatic opposition), or both.

The [+culminative] feature represents the former process, namely the superiorization of one part of a phonological word. As for the minimalization process to fulfill the insufficiency with the conventional typological features, Golob (2021) proposed a new prosodic typological feature called [±eliminative], the actual prosodic role of which should yet be investigated.

References

Beckman, M. E. & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook 3*, pp. 255-309.

Bhaskararao, P. & Golob, N. (2006). What matters in Slovene accent? An acoustic comparison of stress and pitch accents. Paper presented at the *Slovene International Phonetic Conference (SloFon 1)*, Ljubljana.

Golob, N. (2008). Speaking emotions in Japanese. *Asian and African Studies, 12*(3), pp. 57-70.

Golob, N. (2011). Acoustic prosodic parameters in Japanese and Slovene: Accent and intonation. *Acta Linguistica Asiatica, 1*(3), pp. 25-44.

Golob, N. (2021). *Phonetic evidence for an internal structure of the prosodic module: Japanese and Slovene based on the Integrated contrastive model* (in Japanese). PhD. Tokyo University of Foreign Studies.

Gussenhoven, C. (2007). Intonation. In P. de Lacy [ed.] *The Cambridge Handbook of Phonology*, pp. 253-280. Cambridge: Cambridge University Press.

Hyman, L. M. (2006). Word prosodic typology. *Phonology, 23*(2), pp. 225-257.

Hyman, L. M. (2009). How (not) to do phonological typology: The case of pitch-accent. *Language Sciences, 31*, pp. 213-238.

Hyman, L. M. (2011). Tone: Is it Different? ***The Handbook of Phonological Theory***, pp. 197-239. Oxford: Blackwell.

Hyman, L. M. (2018). What is phonological typology? In L. M. Hyman & F. Plank [Eds.] *Phonological Typology*. Berlin: Mouton De Gruyter.

Jun, S. A. (2005). *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Jun, S. A., & Oh, M. (2000). Acquisition of second language intonation. *The Journal of the Acoustical Society of America, 107*(5), 2802-2803.

Kim, J. (2018). Heritage speakers' use of prosodic strategies in focus marking in Spanish. *International Journal of Bilingualism*, 1-19.

Kubozono, H. (2008). Japanese accent. In Sh. Miyagawa & M. Saito [eds.] *Handbook of Japanese Linguistics*. Oxford: Oxford University Press.

Kubozono, H. (2012). Varieties of pitch accent systems in Japanese. *Lingua 122(13)*, pp. 1395-1414.

Labrune, L. (2012). *The phonology of Japanese*. Oxford: Oxford University Press.

Ladd, D. R. (2008[1996]). *Intonational phonology*. Cambridge: Cambridge University Press.

McCawley, J. (1978). What is a tone language? In V. A. Fromkin [ed.] *Tone: A linguistic Survey*, pp. 113-131. New York: Academic Press.

Mennen, I. (2007). Phonetic and phonological influences in non-native intonation: An overview for language teachers. *QMUC Speech Science Research Centre Working Paper*s WP-9, 1-17.

Pfitzinger, H. R. (2006). Five dimensions of prosody: Intensity, intonation, timing, voice quality, and degree of reduction. *Proceedings from the conference Speech prosody (SP 2006)*, pp. 105-108.

Pierrehumbert, J. & Beckman, M. E. (1988). *Japanese tone structure*. Cambridge: The MIT Press.

Pike, K. L. (1948). Tone languages: A technique for determining the number and type of pitch contrasts in a language, with studies in tonemic substitution and fusion. Ann Arbor: University of Michigan Press.

Rasier, L. & Hiligsmann, P. (2007). Prosodic transfer from L1 to L2. Theoretical and methodological issues. Paper presented at the *Symposium on Discourse Prosody Interfaces*, Geneva.

Tivadar, H. & Šuštaršič, R. (2001). Otvorena pitanja standardnoga slovenskog izgovora. *Govor, 18*(2), 113-122.

Toporišič, J. (2004 [1976]). *Slovenska slovnica*. Maribor: Obzorja.

Ueyama, M. (2000). *Prosodic Transfer: An Acoustic Study of L2 Japanese & L2 English*. PhD. UCLA.

Uwano, Z. (1999). Classification of Japanese accent systems. In S. Kaji [Ed.] *Proceedings of the Symposium 'Cross-Linguistic Studies on Tonal Phenomena, Tonogenesis, Typology, and Related Topics'*. ILCAA, Tokyo, pp. 151-186.

van Maastricht, Krahmer, E. & Swertz, M. (2016). Prominence Patterns in a Second Language: Intonational Transfer From Dutch to Spanish and Vice Versa. *Language Learning, 66*(1), 124-158.