Acta Linguistica Asiatica

Volume 1, Number 2, October 2011

ACTA LINGUISTICA ASIATICA
Volume 1, Number 2, October 2011

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

Issued by: Department of Asian and African Studies

For the publisher: Andrej Černe, the dean of the Faculty of Arts

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

ISSN: 2232-3317

Abstracting and Indexing Services:
COBISS, Directory of Open Access Journals, Open J-Gate and Google Scholar.

Publication is free of charge.

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

E-mail: mateja.petrovcic@ff.uni-lj.si

TABLE OF CONTENTS

FOREWORD

The present issue of ALA, the second in its new incarnation, brings two pieces of good news. The first is that it is now also included in the Directory of Open Access Journals (DOAJ), besides Open J-Gate and Google Scholar. The second is that in 2014, the Department of Asian and African Studies of the Faculty of Arts of the University of Ljubljana is going to host the 14th conference of the European Association of Japanese Studies.

The focus of this issue is on Chinese and Japanese. Papers devoted to Chinese are more theoretically oriented. In the first paper, David Ta-Chun SHEN argues about which theoretical devices are sufficient to explain the phenomenon in Mandarin where prepositions may or may not undergo the third tone sandhi. The second paper, by Jens KARLSSON, deals with adverbs *zai* and *you* in Modern Standard Chinese, showing the similarities of semantic content between the two adverbs and pointing out the main difference between them, i.e. the difference in viewpoint, and possible consequences of this fact.

On the other hand, papers devoted to Japanese look at various issues from an applied linguistic perspective. Nagisa MORITOKI, based on her experience of teaching Japanese in Slovenia, discusses the learner and the teacher's role and their possible strategies when dealing with the vast treasure-trove of information available on the Internet.

Next, Bor HODOŠČEK explores genre variation in the newest large-scale modern Japanese language corpus, the BCCWJ, and the usability of modifier-verb ratio as a genre classifier. In the last paper, Irena SRDANOVIĆ and co-authors discuss the issues involved in creating Japanese language word sketches, singling out in particular the lemmatizer, the tagger, the corpus and statistical methods used, and the sketch grammar that is specifically written for Japanese.

In this issue's Book Review, Mateja PETROVČIČ reviews the work by Xiaoqin SU on reflexivity in Chinese, *Reflexivität im Chinesischen: Eine IntegrativeAnalyse*.

Andrej Bekeš

RESEARCH ARTICLES

PF MERGER WOULD DO, TOO: A REPLY TO ZHANG (1997)*

David Ta-Chun SHEN
Department of English, National Taiwan Normal University
david.shen@std.ntnu.edu.tw

Abstract

The analysis for the phenomenon that prepositions may or may not undergo the third tone sandhi in Mandarin in Zhang (1997) is reviewed. She considers that this phenomenon is short of sound coverage and couches her analysis in the framework of Optimality Theory (OT). However, upon scrutiny, Zhang's analysis invites unnecessary questions. The postulation of two "constituent strength" constraints is with no foundation. It is difficult to grab the idea behind the constituent-strength concept even till now. Related to the concept, the non-specification of a prepositional phrase is not clear. Instead, the syntactic feature manifestation could mark a preposition's uniqueness. In addition, the misuse of the Generalized Alignment and stipulations toward the evaluations in OT are spotted, too. My synthetic approach, based on the extant and developing knowledge about constituency, PF merger, and Shih's (1997) foot formation, shows that for this phenomenon, no new device is needed.

Keywords

Mandarin, third tone sandhi, PF merger

Izvleček

Članek se osredotoča na analizo, ki jo predlaga Zhang (1997), in ponovno prouči trditev, da v primeru predlogov v sodobni kitajščini ne pride nujno do glasovne spremembe tretjega tona. Zhang (1997) namreč meni, da ta pojav ni zadostno podkrepljen z zvočnimi primeri in razvije svoj pristop v okviru optimalnostne teorije (OT). Kljub veliki natančnosti taka analiza sproža vsrto nepotrebnih vprašanj, kot je na primer smisel predpostavke o dveh omejitvah "moči konstituentov", predvsem pa je še danes težko razumeti ideje, ki stojijo za tem konceptom. V tem kontekstu je nezadovoljivo pojasnjena tudi tonska nedoločenost predložne zveze. Avtor v zameno ponudi pristop, ki temelji na manifestaciji sintaktičnih lastnosti in bi utegnil obvladati edinstvenost predložnih zvez. Poleg tega je bila ugotovljena napačna uporaba teorije splošnih formacij (GA) in pogojev za ocenjevanje v OT. S sintetičnim pristopom, ki temelji na razpoložljivem znanju o sintaktični strukturi, na strnitvi fonoloških struktur (PF zdruzitev) in na oblikovanju stopic po Shih (1997), avtor pokaže, da za razlago tega pojava ne potrebujemo novih sredstev.

Ključne besede

Mandarinščina, glasovna sprememba tretjega tona, PF združitev

* I would like to thank the two anonymous reviewers of *Acta Linguistica Asiatica*'s encouragement. Moreover, their comments help me fine-tune the article and its presentation and inspire me to think the argumentation more carefully. All remaining errors, however, are mine to blame.

Acta Linguistica Asiatica, 1(2), 2011.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.1.2.9-24

1. Introduction

The third tone sandhi (TS3) is the ever-lasting issue in the Mandarin phonology. The phenomenon *per se* is quite straightforward: the first third tone (T3) becomes a second tone (T2) when followed by another T3. However, the crux lies in when there are more than two consecutive T3s together. How does phonology interact with morphology, syntax, and semantics to derive the surface representations (SRs)? What roles do the above modules play? It is this interface characteristic that squeezes out abundance of the literature.

Under this context, Zhang (1997) has directed the focus to a sub-phenomenon of TS3 where she considered that all the previous literature has fallen short of a satisfactory analysis (Zhang, 1997, p. 297-304). This sub-phenomenon is termed the avoidance of TS3. And it is further divided into two types. In (1) below, the sentences all contain the structure [[$_{XP}$ σ^a σ^b] σ^c]. Only when σ^a is a preposition, could that syllable have its SR as T3 or (sandhied) T2. This type is called "'category dependency in avoidance of TS [tone sandhi]'" (Zhang, 1997, p. 295).

(1) Category dependency in avoidance of TS (Zhang, 1997, p. 293-295)

a. σ^a prepositional

α.	狗	比	馬	小。
	Gou	[[bi	ma]	xiao].
	dog	than	horse	small
UR	3	3	3	3[1]
SR	2	3	2	3
SR	3	2	2	3

"A dog is smaller than a horse."

β.	馬	往	北	走。
	Ma	[[wang	bei]	zou].
	horse	to	north	walk
UR	3	3	3	3
SR	2	3	2	3
SR	3	2	2	3

"The horse walked to the north."

[1] UR means underlying representation, i.e. underlying tone in the present study. The same applies to SR. And the numeral two and three stand for T2 and T3, respectively.

b. σ^a non-prepositional

α.	馬	很	少	吼。
	Ma	[[hen	shao]	hou].
	horse	very	seldom	roar
UR	3	3	3	3
SR	3	2	2	3

“Horses seldom roar.”

β.	有	兩	碗	米。
	You	[[liang	wan]	mi].
	have	two	bowl	rice
UR	3	3	3	3
SR	3	2	2	3

“There are two bowls of rice.”

The other type is called “structure dependency in avoidance of TS” (Zhang, 1997, p. 295). For this type, prepositions and other categories behave together. It is structure differences that make the avoidance of TS possible or not.

(2) Structure dependency in avoidance of TS (Zhang, 1997, p. 296)

a. $[[\sigma^c[_{XP}\ \sigma^a\ \sigma^b]]\sigma^d]$

α.	買	小	馬	好。
	[Mai	[xiao	ma]]	hao.
	buy	small	horse	good
UR	3	3	3	3
SR	3	2	2	3

“It is good to buy small horses.”

β.	比	小	狗	懶
	[bi	[xiao	gou]]	lan
	than	small	dog	lazy
UR	3	3	3	3
SR	3	2	2	3

“to be lazier than the small dog”

b. $[\sigma^e[[_{XP}\ \sigma^a\ \sigma^b][_{XP}\ \sigma^c\ \sigma^d]]]$

α.	鬼	打	傘	買	酒。
	Gui	[[da	san]	[mai	jiu]].
	ghost	take	umbrella	buy	wine
UR	3	3	3	3	3
SR	3	2	3	2	3

“The ghost bought wine with an umbrella.”

β.	馬	給	狗	洗	澡。
	Ma	[[gei	gou]	[xi	zao]].
	horse	for	dog	wash	bath
UR	3	3	3	3	3
SR	3	2	3	2	3

"The horse bathed the dog."

In (2a), σ^c preserves the T3 regardless of its part of speech. Reversely, in (2b), σ^a has to change to T2, but again, regardless of its part of speech.

Based on her analysis, Zhang (1997) concludes that the Optimality Theory (OT) is superior to a rule-based analysis. However, a closer look at the analysis would show that this conclusion has been formed on a shaky foundation. Moreover, with the progress of the Minimalist Program and the initiation of the Distributed Morphology, one is equipped with the post-syntactic movement/P(honological)F(orm) merger/morphological merger, among other things. This mechanism provides the way to explain how a syntactic non-constituent could form a TS domain, like the reading 2323 of (1a).

Therefore, the avoidance of TS need not be the result of constraint ranking. Actually, as will be demonstrated below, it is the result of syntax-all-the-way-down, PF merger, and Shih's (1997) prosodic formation algorithm.

The organization of this study is as follows. In section 2, two concerns toward Zhang's analysis will be raised. The avoidance fact will then be re-examined in section 3. Section 4 concludes this paper.

2. Two Concerns

In this section, two parts which motivate me to reanalyze the phenomenon are going to be discussed.

In order to decide which underlying T3 will surface up as T3, Cinque's (1993) null stress theory has been reinterpreted as the manifestation of "constituent strength" (Zhang, 1997, p. 304-305). Originally, Cinque's idea is to predict phrasal/sentential and compound main stress through the syntactic structure, which would then make language-particular stress assignment redundant. The gist of his theory is that the most embedded constituent will have the primary stress (Cinque, 1993, p. 245). Therefore, if a complement is present, it will receive the primary stress. If not, the head will receive the stress, instead. A specifier and/or pre-modifier will always be weak in stress. Borrowing this idea, Zhang (1997) has posited two relevant constraints as in (3).

(3) Constituent strength related constraints (Zhang, 1997, p. 306)

a. Parse Underlying Tone of an Absolutely Strong Node (PTAS)

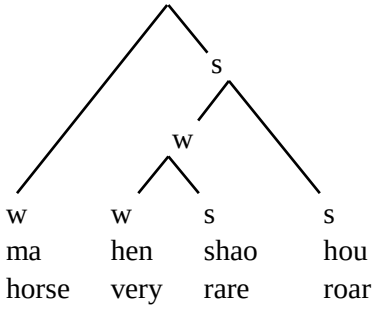
The underlying tone of a strong constituent which is not dominated by any w node must be parsed.

b. Parse Underlying Tone of a Relatively Strong Node (PTRS)

The underlying tone of a strong constituent which is dominated by at least one w node must be parsed.

In (3), *w* means that a constituent has weak strength. If a constituent has a strong strength, it bears a mark of *s*. And to be parsed means that a tone has to be unchanged. As an example, (4) would have the following constituent strength distribution.

(4) Constituent strength distribution of *Ma hen shao hou.* "Horses rarely roar." (Zhang, 1997, p. 305)



For *hen shao*, *shao* is strong because it is the head. *Hen* is therefore weak. Because *hen shao* and *ma* are pre-modifier and specifier respectively, they are both weak, while *hou* is the most embedded with two s's assigned. Because *hou* is not dominated by any w, *hou* will not violate PTAS only if it has T3 in the surface. Although PTAS is irrelevant to *shao*, PTRS is: *shao* is dominated by w once—PTRS will not be violated if it has T3 in the surface.

There is, however, a serious gap. What is the nature of the concept of constituent strength? The most probable possibility, stress, was denied as she had realized the complex interaction between stress and tone and tried to eschew this problem:

> …I will regard Cinque's contrast of s/w stress as contrast of s/w constituent strength rather than as a reflection of stress directly. (Zhang, 1997, p. 304)

The situation now goes back to the very beginning: what is the constituent strength? The innovative use of somebody else's idea or the creative use of one's very own idea is more than welcome for any scientific study, but this kind of use should not be taken for granted and should be reasoned. In the present case, a more elaborated

explanation is a must in order to make the use of constituent strength valid. Or, its use at that time was too novel to be defined, so it is the subsequent works (e.g. Wee, 2008) that are the places to look for. It seems not to be the case, either. This situation makes the status of constituent strength questionable. And one cannot help but think that the so-called constituent strength is merely a stipulation which tries to endorse itself through Cinque (1993) and at the same time hopes no controversy incurred through some vague rhetoric.

The other function of the constituent strength is to deal with the preposition-pertinent TS avoidance phenomenon. A prepositional phrase is not specified for s and w. Instead, a preposition and its complement are left blank. A prepositional phrase could be an sw or ws combination. It can be said that Zhang has utilized the non-specification for prepositional phrases. There are two questions about this device. Other than visualizing the special status of a preposition, what motives this move of non-specification? I consider that the special status of a preposition already can be manifested by the means of its feature make-up, i.e. [-N, -V]. The non-specification approach is nothing but another way to re-package the old information. The postulation of a new device should be motivated, otherwise it should be avoided. The non-specification is no superior to the feature manifestation, if not inferior. The other question is: why are there only two aforementioned combinations? What excludes the combinations of ss and ww when these two alternatives could also produce the desired results? This question has never been discussed in the text or in an endnote. The non-specification gives out the freedom and at the same time the unwanted logical possibilities. To resolve this with another constraint or qualification only makes things more complicated, compared with the commonly-assumed feature make-up practice. In sum, until Zhang is willing to provide more evidence on the constituent strength, I consider that it is not something to be recommended.

The second concern is her inconsistent use of the Generalized Alignment (GA). The given alignment constraint postulated by her is given as in (5).

(5) Disyllabic Constituent Alignment (Align-Di-L; Zhang, 1997, p. 308)

Align the left side of a TS domain with the left side of a disyllabic constituent when two or more TS domains occur.

In order to evaluate her use of GA, let's have the definition of it first.

(6) Definition of GA (McCarthy & Prince, 1993, p. 80)

Align (Cat1, Edge1, Cat2, Edge2) $=_{def}$
$\forall$ Cat1 $\exists$ Cat2 such that Edge1 of Cat1 and Edge2 of Cat2 coincide.

Where
Cat1, Cat2 $\in$ PCat $\cup$ GCat
Edge1, Edge2 $\in$ {Right, Left}

The most important information read from (6) is that the category 1 is universal and that the category 2 is existential. One immediate not-that-major problem which I have is why and/or how a TS domain and a disyllabic constituent qualify as a prosodic category or a grammatical category. They are not the typical members for each of the two categories, so their qualification should be argued. Instead, this has been just assumed without any comment. Next, the unconventional stipulation of "when two or more TS domains occur" in (5) is not founded. Within my limited knowledge toward the practice of GA, the only things needed are just what (6) depicts. From a hindsight perspective, the function of this stipulation merely tries to rescue the maximally changed 2*3 pattern from being non-harmonic.[2] This statement applies at least to (35), (39a-b), (45), (47), and (74) of Zhang (1997).

The more serious problem is when the universal/existential relationship is articulated more specifically. None of the only two possible alignment constraints for (5) which are permitted by (6), that is, Align (TS, L, disyllabic constituent, L) and Align (disyllabic constituent, L, TS, L), is consistently practiced.[3] Zhang's (1997, p. 328) (69) is going to be reproduced as a departure in order to see the chaotic use of GA. Even worse is, when the above two realizations are consistently applied, the wrong candidates would be chosen as the optimal ones, as shown in the undermentioned (8) and (9). In addition, in order for the optimal candidates to win out in Zhang's (69), she has stipulated that "it is worse to violate two equally ranked constraints than to multiply violate just one of those equally ranked constraints" (Zhang, 1997, p. 337, endnote 9). This tailor-for-particular-case is unsound and arbitrary. That the stipulation for her (69) makes no harm for her other tableaux is irrelevant and does not soften its arbitrariness in any extent.

(7) Reproduction of Zhang's (69)

ws w ws S	PTAS	*33[4]	PTRS	Align-Di-L	Max[5]
a. (23)(3)(223)		*!	*		*
b. (23)(2223)			*	*!	*
☞ c. (223)(223)			**		**
☞ d. (222223)			**		*
e. (23)(23)(23)			*	*!	

[2] An asterisk means that the 2s are more than one.

[3] Following Zhang's (1997) original alignment constraint formulation, I will not consider the R-to-R, L-to-R, or R-to-L edge pairings. As would be obvious from the following discussion, what truly matters are the universal-existential distinction and the constraint definition/evaluation stipulations. For more on the pairing of edge for GA, please refer to Kager (1999, p. 119).

[4] No sequential third tones (*33): No adjacent third tones are allowed. (Zhang, 1997, p. 307)

[5] Maximal Domain (Max): The maximal TS domain is two syllables in normal speaking rate but larger in more casual or faster style. (Zhang, 1997, p. 308)

(7) has been the tableau for the sentence *Lao gui xiang da san zou.* 'The old ghost wanted to walk with an umbrella.' (老鬼想打傘走。). The parentheses indicate TS domains and the underlines disyllabic constituents. Both (7c) and (7d) have been the optimal candidates because it has been assumed that when a suboptimal candidate is inferior to the optimal one only in the Max, the suboptimal is fine for the fast speech (Zhang, 1997, p. 308). The situation goes sour when the endnote 9 stipulation and the "when two or more TS domains occur" stipulation are abolished and either of which alignment constraints is carefully examined: the default (7c) is not directly available or even unavailable.

In (8), we first have Align (TS, L, disyllabic constituent, L) (abbreviated as Align-L (TS, disyllabic constituent)), which means that every TS's left edge coincides with some disyllabic constituent's left edge.

(8) (7) with Align-L (TS, disyllabic constituent)

ws w ws S	PTAS	*33	PTRS	Align-L (TS, disyllabic constituent)	Max
a. (23)(3)(223)		*!	*	*	*
☞ b. (23)(2223)			*	*	*
(☞) c. (223)(223)			**		**!
☞ d. (222223)			**		*
e. (23)(23)(23)			*	**!	

To help identify the suboptimal candidate, I have deliberately marked this kind of output by putting the pointing hand between the parentheses. (8c) is rescued through the high speech rate, which is contrary to Zhang's (1997, p. 327) own understanding that (8c) is the output for the moderate speed. (8b) is wrongly selected as being optimal with (8d) which should originally occur in the fast speech (Zhang, 1997, p. 327).[6]

In (9), we have Align (disyllabic constituent, L, TS, L) (abbreviated as Align-L (disyllabic constituent, TS)), which means that every disyllabic constituent's left edge coincides with some TS's left edge.

(9) (7) with Align-L (disyllabic constituent, TS)

ws w ws S	PTAS	*33	PTRS	Align-L (disyllabic constituent, TS)	Max
a. (23)(3)(223)		*!	*		*
(☞) b. (23)(2223)			*	*	*!
c. (223)(223)			**		*!*
d. (222223)			**	*!	*
☞ e. (23)(23)(23)			*	*	

[6] Precisely speaking, then, (7) is an unsolved problem for Zhang's (1997) analysis.

The optimal (9e) and the suboptimal (9b) are both not surface true. Interestingly, if the stipulations were revived, (9) would be the exact (7). Therefore, the superficial success of (7) is based on some unpersuasive manipulation of OT.[7]

Before I end this section, some thoughts about *33, Cl[8], and Max will be given in (10).

(10) Responses toward three more constraints in Zhang (1997)

a. For *33, the so-called property of "being violable" of this constraint is not due to the inherence from OT. It is due to foot boundary (Shih, 1997, p. 117).

b. As will be shown, I will view Cl as the outcome of PF merger.

c. For Max, the maximally changed realization is derived by seeing a given string as a single application domain for the TS3 rule (Shih, 1997, p. 86 and further comments therein).[9]

I hope that through this examination, I have demonstrated the problematic side of Zhang's constraint-based analysis. In the next section, I am going to show that a much simpler and null-invention analysis is feasible.

3. Resort to PF Merger

Importantly, in order to make everything else being equal, all Zhang's (1997) judgments will be assumed in this section, though mine differ from hers from time to time. Moreover, as mentioned in (10c), the maximally changed pattern will not be further discussed. In the present case, this kind of pattern is theoretically less interesting.

From (1), the focus would be the special status of a preposition. Presumably, this status can be attributed to its feature composition being [-N, -V]. I then propose that there is a PF merger in Mandarin as in (11).

(11) Optional preposition PF merger in Mandarin

In Mandarin, a [-N, -V] can PF merge with a preceding constituent and form a new constituent if PF adjacency is respected.[10]

---

[7] The closeness between (7) and (9) above may make one conclude that Align-L (disyllabic constituent, TS) is the one used in Zhang (1997), but this conclusion is false. One is more than welcome to verify this assertion of mine (for I had reached this conclusion once).

[8] Clitic Dependency (Cl): A clitic cannot be separated from the TS domain of the preceding verb or preposition head. (Zhang, 1997, p. 307)

[9] The constraint ranking of the six constraints is: PTAS, *33, Cl » PTRS, Align-Di-L » Max (Zhang, 1997, p. 312).

The product derived from the PF merger is seen as an immediate constituent for foot formation in Mandarin.[11] The algorithm for footing adopted is provided in (12).

(12) Mandarin foot formation (Shih, 1997, p. 98)

a. Join immediate constituents into disyllabic feet.[12]

b. Scanning from left to right, join monosyllabic constituents into disyllabic feet.

c. Join any remaining monosyllables to neighboring feet.

The optionality of (11) accounts for two realizations in (1a).[13] The non-specification of a prepositional phrase in regards of constituent strength is replaced by a preposition's applying the PF merger or not, which is like a (phonological) unary feature's presence or absence. A presupposition from this proposal is that the PF merger (11) has to take place before the foot formation (12). This is achieved by means of the extrinsic ordering or the organization of morphological structure earlier than phonological form (Halle & Marantz, 1993, p. 114). In sum, based on Zhang's (1997) data, the way to deal with the (avoidance of) TS can be visualized in (13).

(13) Flow chart of TS operation

narrow syntax constituency	→	(morphological structure) (prepositional PF merger)	→	phonological form foot formation

In (13), *morphological structure* and its association are put within the parentheses to indicate their optionality. That nothing new can be read from (13) is the beauty I am trying to argue for in this study. Thanks to the development of the Chomskian syntax

---

[10] The formal definition of PF merger in this paper is given in (i).

(i) PF merger (Halle & Marantz, 1993, p. 116)

> Merger…joins terminal nodes under a category node of a head…but maintains two independent terminal nodes under this category node.

Furthermore, the application of PF merger respects "PF adjacency" (Bošković, 2001, p. 84).

[11] A foot here is equal to what Chen (2000, p. 366 et seq.) terms a "minimal rhythmic unit".

[12] When one thinks more carefully on all TS3 data in the literature, immediate constituency cannot cover very well all of it. Because syntax conventionally only sees word and the bigger chunks, compounds, like *shuiguo* 'fruit' (水果; both syllables with T3), enter syntax without their sub-word structure being visible. Therefore, to say *shuiguo* can form a disyllabic immediate constituent is wrong—the term is a unit already. I believe that, due to this difficulty and other TS facts, Chen (2000, p. 380-386) argues that the TS3 rule in Mandarin is both lexical and post-lexical. And Chen's analysis suggests that Mandarin is a big support for the Lexical Phonology. However, against the backdrop of the interface inquiry, especially phase and multiple spell-out, Scheer (2008) argues that the Lexical Phonology has to go because it has some built-in redundancy. We are then left with a dilemma. As this question is beyond the scope of this paper, I will play vague and stick with Shih's (1997) algorithm for the time being and leave this issue for future pursue.

[13] I generalize the idea of optionality for phonology (Vaux, 2008, p. 41-44) to morphology.

and the Halle-Marantz morphology, we are more equipped than ever to cope with the interface issues. It is time for us to choose the appropriate ones among these weapons, not to create more. Below I am going to demonstrate with some instances from Zhang (1997) and see some refinement.

Let's first begin with the derivation of (1aα). If the PF merger does not occur, the derivation would be that in (14).

(14) Derivation of (1aα) without the PF merger

a. [gou3 [[bi3 ma3] xiao3]]
↓
b. (gou3((bi3ma3)xiao3)) or ((gou3(bi3ma3))xiao3)
↓
c. gou3 bi2 ma2 xiao3

In (14a), the constituency is given as the result of narrow syntax. In (14b), because *bi ma* is an immediate constituent, it forms a disyllabic foot. The remaining monosyllabic constituents are far apart, so (12b) is not applicable. (12c) will then include *gou* and *xiao* into the existent foot starting from either of them. The TS3 rule then applies cyclically outwards, which results in (14c).

If the PF merger occurs, the derivation would be that in (15).

(15) Derivation of (1aα) with the PF merger

a. [gou3 [[bi3 ma3] xiao3]]
↓
b. [[gou3 bi3] ma3 xiao3]
↓
c. (gou3 bi3)(ma3 xiao3)
↓
d. gou2 bi3 ma2 xiao3

(15a) is the same as (14a). (15b) is the application of PF merger, which forms *gou* and *bi* as post-syntactic constituent. This constituent will then be an immediate constituent and make a disyllabic foot. (12b) this time is applicable to the remaining string. After that, (12c) is of no use. The cyclic application of the rule produces the output (15d).

(14) and (15) basically complete my analysis. The gist of my analysis is plain: allow a preposition to have the freedom of PF merger in Mandarin, where PF merger is now a familiar and hot topic. Below the discussion continues with those (seemingly) problematic cases.

Under the present PF merger analysis, that the sentence *Gou bi wo xiao.* 'The dog is younger than me.' (狗比我小。; Zhang, 1997, p. 297) cannot have the SR 2323 is not understandable since *bi* could PF merge with *gou* and the two thus form an immediate constituent and then a disyllabic foot. To account for this, I revise (11) as (11') based on Zhang's (1997, p. 296-297) observation that there is a proform following *bi*.

(11') Optional preposition PF merger in Mandarin

> In Mandarin, a [-N, -V] can PF merge with a preceding constituent and form a new constituent if PF adjacency is respected, except when the given [-N, -V] is followed by an overt and adjacent D.

The revision is possible because under the D(eterminer)P(hrase) hypothesis (Abney, 1987; Baker & Hale, 1990) a proform is a D. The *ma* in (1aα) is also a DP under the DP hypothesis, but it lacks an overt and adjacent D for the preceding *bi*. So, there is no change for the derivation in (15) with (11'). As for the sentence *Gou bi wo xiao.*, the PF merger is not applicable (*wo* is an overt and adjacent D for the preceding *bi*), so only the SR 3223 is produced.

In Zhang's discussion of the constraint Cl, she has mentioned a verb-classifier-noun construction. To deal with the TS of this construction, I borrow the spirit of Xu (1999) and see the construction has another obligatory PF merger going on between a bare classifier and its preceding verb. Therefore, the verb phrase *xiang mai ba san* '(I) want to buy an umbrella' (想買把傘; Zhang, 1997, p. 307) have *mai* and *ba* as a post-syntactic constituent, which provides the reason why the SR 2323 is not derivable.[14]

For Zhang's (1997, p. 334) (80), Shih's (1997, p. 91-94) "initial cycle" is needed. Initial cyclicity makes flat all the more embedded foot structure below a chosen foot. Since the concern is not to derive all the possible TS outputs, in (16) below, only the most relevant footing will be given. The footing in (16) is the interaction of the application of PF merger and the "non-deterministic" (Shih, 1997, p. 98) nature of (12c). The non-determinacy gives an unfooted syllable to be incorporated with the foot in front of or in back of it. And in the present case, this unfooted syllable *ye* is incorporated into the foot following it.

[14] The PF merger analysis here is so similar to Shih's (1997, p. 110-112) clitic analysis. I, however, hesitate to make this connection. For one thing, cliticization seems to implicate that it must be applied, but from the above discussion, optionality has to be granted. (I think that to account for cross-Sinitic TS phenomena, this freedom should also be allowed.) For another thing, "[c]lictics can attach to material already containing clitics" (Zwicky & Pullum, 1983, p. 504), why shouldn't a proform cliticize to a host which has been cliticized by a preposition, if the two categories are both the target of cliticization? Whether this separation of PF merger on the one hand and cliticization on the other hand is valid or not, I leave it for future research.

(16) Footing of *Lao ma ye dei gei gou yao.*
"The old horse will also be bitten by a dog.' (老馬也得給狗咬。)
(lao ma)(ye (dei gei))(gou yao)

The usual application of the TS3 rule will give out the SR 2332323 or 2232323 (TS3 across foot boundary is optional (Shih, 1997, p. 85-86)), not Zhang's 2322323. However, the initial cycle allows us have two alternatives for the tri-syllabic foot's TS rule application: starting from the inner foot (without initial cyclicity) or the outer one (with initial cyclicity). The inner first gives the SR 2232323 (from 2332323); the outer first gives Zhang's SR 2322323. This strategy of Shih's also makes Zhang's (1997, p. 325, 332) (62) and (75) accountable.

I end this section with the string *bi gou xiao* 'smaller than dog' (比狗小; Zhang, 1997, p. 295). In Zhang's analysis, *bi* is a preposition and takes *gou* as its complement. Because there is nothing preceding *bi*, the PF merger does not function. The expected SR should only be 223. However, there is another realization provided by Zhang: 323. To derive this realization, I may borrow the idea from Xu (1992, p. 74-78): because *gou* is a noun (contra to a proform) in the middle of a prepositional phrase, the noun may initiate its sandhi domain, which provides the footing (bi (gou xiao)). From this footing, the realization 323 is then derivable. Or, if the SR 323 is realized as being contrastive, Shih's (1997, p. 112-116) "emphatic boundary" can be said to function. An emphatic boundary is a left foot parenthesis, that is (, established before an emphatic element. An emphatic left footing cannot be restructured by the footing algorithm, but it can undergo initial cyclicity. Once again, the probable use of Xu (1992) echoes with what has been mentioned earlier in the text—we have the tools already and it is time to make use of them.

4. Conclusion

Using Zhang (1997) as the starting point, I first argued that her analysis has been based on the questionable assumption of constituent strength and the misused GA, plus some stipulations about OT. Then, together with the narrow syntax, PF merger, and foot formation algorithm, I hope that I have displayed that what theoretical achievement we have now has prepared us more ready than ever to re-examine the previous literature and to re-shape the persistent questions into more simplex realizations. Along the way, a robust point has been re-ensured: in order to understand better the interface between phonology and syntax, the formal syntactic structure has to be more carefully looked after (cf. Lasnik & Lohndal, 2010, p. 48). Despite of the current interest in interface, to encapsulate all the joining forces into a parallel-evaluation model, like OT, cannot be right because "*what needs to be explained* becomes the *explanation*" (van der Hulst, 2004, p. 237, emphases in original). The interest should, instead, push us harder to scrutinize the interwoven forces and put each of them back to its place.

For my synthetic analysis, there are certainly things to work on as well. For example, can the two mergers for preposition and classifier be subsumed under more general principle(s)?[15] Another task can be: is the analysis presented here adaptable enough to operate on the other (Sinitic) TS systems? These and more are important issues ahead.

References

Abney, S. P. (1987). The English noun phrase in its sentential aspect. Doctoral dissertation, Massachusetts Institute of Technology.

Baker, M., & Hale, K. (1990). Relativized Minimality and pronoun incorporation. *Linguistic Inquiry, 21(2)*, 289-297.

Bošković, Ž. (2001). *On the nature of the syntax-phonology interface: Cliticization and related phenomena*. Amsterdam, the Netherlands: Elsevier.

Chen, M. Y. [陳淵泉] (2000). *Tone sandhi: Patterns across Chinese dialects*. Cambridge, England: Cambridge University Press.

Cinque, G. (1993). A null theory of phrase and compound stress. *Linguistic Inquiry, 24(2)*, 239-297.

Grimshaw, J. (2005). Extended Projection. In J. Grimshaw, *Words and structure* (pp. 1-73). Stanford, CA: CSLI Publications.

Halle, M., & Marantz, A. (1993). Distributed Morphology and the pieces of inflection. In K. Hale & S. J. Keyser (Eds.), *The view from building 20: Essays in linguistics in honor of Sylvain Bromberger* (pp. 111-176). Cambridge, MA: The MIT Press.

Kager, R. (1999). *Optimality Theory*. Cambridge, England: Cambridge University Press.

Lasnik, H., & Lohndal, T. (2010). Government-binding/principles and parameters theory. *Wiley Interdisciplinary Reviews: Cognitive Science, 1(1)*, 40-50.

McCarthy, J. J., & Prince, A. (1993). Generalized Alignment. In G. Gooij & J. van Marle (Eds.), *Yearbook of morphology 1993* (pp. 79-153). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Scheer, T. (2008). Spell out your sister!. In N. Abner & J. Bishop (Eds.), *Proceedings of the 27th west coast conference on formal linguistics* (pp. 379-387). Somerville, MA: Cascadilla Proceedings Project.

Shih, C. [石基琳] (1997). Mandarin third tone sandhi and prosodic structure. In J. Wang [王嘉齡] & N. Smith (Eds.), *Studies in Chinese phonology* (pp. 81-123). Berlin, Germany: Mouton de Gruyter.

Van der Hulst, H. (2004). Phonological dialectics: A short history of generative phonology. In P. van Sterkenburg (Ed.), *Linguistics today: Facing a greater challenge* (pp. 217-242). Amsterdam, the Netherlands: John Benjamins Publishing Company.

Vaux, B. (2008). Why the phonological component must be serial and rule-based. In B. Vaux and A. Nevins (Eds.), *Rules, constraints, and phonological phenomena* (pp. 20-60). Oxford, England: Oxford University Press.

[15] If classifiers could somehow be incorporated into Grimshaw's (2005, p. 4) Extended Projection for D, the two PF mergers might be collapsible: both of them are [-V, +N] with a preposition bearing a functional feature F2 and a D F1. And the functional feature might also hint at the obligatoriness of the two mergers. Whether this collapse is substantive or not, I leave it for future research.

Wee, L.-H. [黃良喜] (2008). Opacity from constituency. *Language and Linguistics, 9(1)*, 127-160.

Xu, D. [許德寶] (1992). Mandarin tone sandhi and the interface study between phonology and syntax. Doctoral dissertation, University of Illinois at Urbana-Champaign.

Xu, D. B. [許德寶] (1999). A syntactic account for the inseparability of the verb and the classifier in the structure V+C+N in tone sandhi. *Journal of the Chinese Language Teachers Association, 34(3)*, 77-90.

Zhang, N. [張寧] (1997). The avoidance of the third tone sandhi in Mandarin Chinese. *Journal of East Asian Linguistics, 6(4)*, 293-338.

Zwicky, A. M., & Pullum, G. K. (1983). Cliticization vs. inflection: English n't. *Language, 59(3)*, 502-513.

LINGUISTIC TEMPORALITY, LOGICAL MEANING AND NARRATIVE PERSPECTIVES: ADVERBS *ZAI* AND *YOU* IN MODERN STANDARD CHINESE

Jens KARLSSON
Lund University, Centre for Languages and Literature, Chinese Studies
Jens.Karlsson@ostas.lu.se

Abstract

In this paper is presented an inquiry into some aspects of the meaning and usage of two temporal adverbs *zai* (再) and *you* (又) in Modern Standard Chinese. A decompositional analysis of the semantic encoding of the adverbs is conducted, aiming to better explain their recorded differences in usage. First, a sketch of some of the fundamental features of linguistic temporality is provided in order to model the structure of temporal semantic information encoded in the adverbs. Non-temporal (logical) meaning such as assertion and inference is also shown to be an important aspect of the semantic content of the adverbs. Adverbs *zai* and *you* are shown to encode the same semantic content except for a difference in viewpoint; the first being prospective, the second retrospective. Concrete linguistic examples reflecting the intrinsic semantic encoding of the adverbs are raised and discussed. It is then argued that through combining the decompositional analysis with ideas concerning conceptual analogy, some issues raised by Lu and Ma (1999) regarding the usage of *zai* and *you* in past and future settings may be resolved.

Keywords

Chinese, temporal adverbs, decompositional analysis, linguistic temporality, conceptual analogy

Izvleček

Članek prouči nekatere pomene in načine uporabe časovnih prislovov *zai* (再) in *you* (又) v sodobni standardni kitajščini. Za boljše razumevanje razlik v uporabi služi dekompozicijska analiza semantičnih oznak prislovov. Avtor najprej na kratko predstavi osnovne značilnosti izražanja časa v jeziku, na podlagi česar izdela strukturo semantičnih informacij o času, ki jih nosijo prislovi. Tudi ne-časovni (logični) pomeni, kot sta na primer trditev (assertion) in sklepanje (inference), so se izkazali za pomemben del opisa semantike prislovov. Avtor pokaže, da prislova *zai* in *you* nosita iste semantične informacije in da je razlika med njima le v pogledu na situacijo – prvi je prospektiven in slednji retrospektiven. Članek v nadaljevanju izpostavi in prouči dejanske primere, ki odražajo notranje semantične zančilnosti prislovov. Nazadnje avtor pokaže, da je s kombinacijo dekompozicijske analize in idej o konceptualni analogiji mogoče razložiti nekatera vprašanja, ki sta jih v zvezi z uporabo prislovov *zai* in *you* v preteklih in prihodnjih situacijah že izpostavila Lu and Ma (1999).

Ključne besede

Kitajščina, časovni prislovi, dekompozicijska analiza, jezikovna časovnost, konceptualna analogija

Acta Linguistica Asiatica, 1(2), 2011.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.1.2.25-38

1. Introduction

There exists a pair of morphemes *zai* and *you* in Modern Standard Chinese (MSC) which are normally categorized as temporal adverbs.[1] Referring to a morpheme as "adverb" concerns its' grammatical function; referring to a morpheme as "temporal" concerns its' semantic information. The previous century saw much debate over the issue of word classes in Chinese, including some controversy as to whether the language can be said to have word classes at all. According to Lu (2003), this debate was especially vivacious during the 1930's, 50's and 80's. These discussions resulted more or less in a consensus among (Chinese) linguists, saying that words in Chinese are categorized into word classes according to (morpho)syntactic properties. The prevalent view on adverbs in this context is that they have the sole function of adverbial modifier. "Strict adverbs are words that conform with the two following criteria: (1) may modify verbs or adjectives; (2) may not modify nouns; may not act a subject, object or predicate." (Zhu, 1961, p. 70-71) This may be contrasted against adjectives for instance, which commonly assume the role of adverbial modifier in addition to several other grammatical roles, including both subject and predicate. Adhering to the view expressed by Zhu (1961), the morphemes *zai* and *you* may be considered prototypical adverbs, as their only grammatical function is acting as adverbial modifiers in a predicate clause (Karlsson, 2010).

The question of semantic information carried by temporal adverbs in MSC is not a matter of consensus in the same way as their syntactic function. In the following I present a model of the semantic core content encoded in *zai* and *you*. The two adverbs are shown to encode both temporal and non-temporal information. The structure of the temporal information is based on a sketch of the fundamental features of linguistic temporality, which is introduced in the first section. I identify the non-temporal information as logical meaning, the workings of which are introduced in the following section. I then discuss the meaning and function of the two adverbs, presenting empirical data to support the model of their semantic structure which is based on the sections on linguistic temporality and logical meaning. I then discuss some issues concerning the usage of *zai* and *you* in narrative contexts which intuitively seem to conflict with their normal usage. I argue for an explanation which considers the intrinsic semantic encoding of the adverbs and the application of associative thinking in the form of conceptual analogy.

[1] This paper is in principle an elaboration of ideas first developed in Karlsson (2010). It relies to considerable extent on Lu and Ma (1999) for empirical data, and Karlsson (2010) for theory and method.

2. Linguistic Temporality

According to Klein (1994), there is no real consensus concerning the nature of linguistic temporality, referring to "that concept of time which underlies the expression of temporal relations in natural languages" (p. 60).[2] A typical basic representation of linguistic temporality would probably be something akin to the visual conceptualization of time used in Comrie (1985) (Fig. 1).

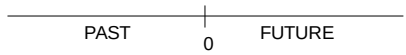


Fig. 1: Basic representation of time (Comrie, 1985, p. 2)

Comrie's representation includes a straight line where the past is located to the left and the future to the right of the present moment (0). It instantiates two of the fundamental features of linguistic temporality as identified by Klein (1994): 'origo' and 'linear order'. Some of the features identified by Klein (1994), like 'linear order' for instance, can be said to be directly derived with logical necessity from even more fundamental characteristics of temporality. Kant (1787) identifies in Critique of Pure Reason time as a "necessary representation, lying at the foundation of all our intuitions" (p. 39). The interesting consequence of this conclusion for the linguistic sciences lies in "the possibility of apodeictic principles of the relations of time, or axioms of time in general, such as: 'Time has only one dimension', 'Different times are not coexistent but successive' (...)" (p. 39). Unidimensionality and unidirectionality are according to this idea a priori given features of temporality, and it would seem that "the expression of temporal relations in natural languages" as referred to by Klein (1994, p. 60) well corresponds with this conception. Relating it to Comries (1985) representation in Fig. 1, it determines that the past, present and future must be arranged in sequential order, constituting different parts of the time axis.

The representation of time depicted in Fig. 1 is intended to show a basic deictic arrangement underlying all tense-systems found in natural languages. Claiming its' universal application corresponds well with the following assumption expressed by Smith (2005[3]): "The deictic pattern – in which Speech Time is central – is a linguistic universal, so far as we know" (p. 3). Reichenbach's (1947) system for a representation of the English tenses is built on the concept of temporal deixis, consisting of Speech

[2] Klein (1994) lists seven basic features of linguistic temporality which he believes to be "indispensable" (but not necessarily exhaustive). The features are 'segmentability','inclusion','linear order','proximity','lack of quality', 'duration', 'origo'. Some of these features are quite useful for the discussion in this paper.

[3] The year is given according to the printed publication as listed in the references; the page refers to the publication available online, which is also listed in the references.

Time (S), Event Time (E) and Reference Time (R). Although the system was devised with the English tenses in mind, it has proven to be useful for various accounts of temporal expressions cross-linguistically. Smith (2005) notes that "the notion of Reference Time is not dependent on tense, but is basic to temporal location in language. Indeed, it has explanatory value for Mandarin" (p. 10). Given the fact that the system is comprised of two basic features of linguistic temporality – deixis and sequence ('origo' and 'linear order' using Klein's (1994) terminologi) – it is not that surprising to find that it has application in the description of temporal relations other than tense-systems and in languages other than English. As will become evident, the semantic information of *zai* and *you* interacts intimately with the deictic temporal structure of sentences they appear in.

3. Logical and Pragmatic Meaning

In order to properly account for the meaning and function of many temporal adverbs, including *zai* and *you*, identifying the characteristics of their temporal semantic encoding is not enough. In addition to such information, they encode semantic content which is probably best described as logical meaning/information.

(1) 她 还 在 中国
ta hai zai Zhongguo
she still be (in) China
"She is still in China."

The information which is explicitly conveyed – asserted – by the temporal adverb *hai* is that "she" is *still* in China at the time when the sentence is uttered. But we may also deduce from the sentence that "she" has been in China for some time *prior* to when the utterance is made. This information is merely implicitly provided, or inferred. Inferred meaning is conveyed in different forms. Information inferred from the intrinsic meaning of words and propositions is usually labelled "entailment", and defined something like "information logically inferred from single propositions". Information pragmatically inferred from a certain context is usually labelled "implicature". Such information has to do with the extrinsic meaning of words and propositions (Korta & Perry, 2008; Peccei, 1999).

(2) This is an apple.[4]
Entailment: This is a fruit.

(3) -Are you coming to Agathon's this evening?[5]
-You know how I love listening to Socrates!
Implicature: Yes.

[4] Example taken from Karlsson (2010).

[5] Example taken from Karlsson (2010).

While the inferred information in (2) is logically entailed, the answer in (3) is taken to be affirmative despite the fact that there is nothing in the literal meaning of the utterance from which the listener can deduce an affirmative answer. Instead, the affirmative answer is deduced partly through the presupposition that Socrates will be present at the gathering at Agathon's. Presupposition can be seen as a third kind of inferred information belonging somewhere between entailment and implicature, as it may be divided into pragmatic and semantic presupposition. Inferred information such as the presupposition concerning Socrates in (3), on which the affirmative implicature relies, is derived completely from context and therefore a pragmatic presupposition. Semantic presuppositions provide inferred information directly from individual words and propositions stripped of further context. Semantic presuppositions can in turn be divided into existential and logical presuppositions (Simpson, 1993).

(4) Guan Yu doesn't serve Cao Cao anymore.
Existential presupposition: There exists (existed) someone by the name of Cao Cao.
Logical presupposition: Guan Yu used to serve Cao Cao.

Just as we may deduce as a logical presupposition in example (4) that Guan Yu used to serve Cao Cao, in example (1) we may deduce as a logical presupposition that "she" has been in China for some time already when the utterance is made, while the information that she is in China when the utterance is made is asserted. Without the presupposed information the utterance doesn't make sense; it is simply part of the intrinsic semantic encoding of the adverb. The kind of inferred information ascribed to the adverbs dealt with in this paper is all of the type logical presupposition.

4. Temporal Adverbs *zai* and *you*

The fact that the adverbs *zai* and *you* convey some sort of temporal notion is intuitively clear from looking at examples like the following.[67]

(5) 去 过 了 还 可以 再 去
qu guo le hai keyi zai qu
go GUO LE still may again go
"Having gone (there) before, you can still go a second time."

(6) 你 敢 再 赛 一 场 吗
ni gan zai sai yi chang ma
you dare again compete one CLF MA
"Do you dare compete one more time?"

[6] Examples 5-8 taken from *Xiandai Hanyu Babai Ci* (1999).

[7] Adverbs *zai* and *you* also express other temporal notions such as continuation, as well as some modal meanings. Due to limited space, these notions are not discussed in the present paper.

(7) 这 个 人 昨天 来 过 今天 又 来 了
zhe ge ren zuotian lai guo jintian you lai le
this CLF person yesterday come GUO today again come LE
"This person was here yesterday, and came again today."

(8) 你 又 生 我 的 气 了
ni you sheng wo de qi le
you again get/have I DE anger LE
"(Now) you became angry with me again."

It is clear from examples (5) through (8) that one salient feature of the semantic information carried by *zai* and *you* is the notion of repetition. In examples (5) and (7), the core predicates *qu* "go" and *lai* "come", modified by the adverbs *zai* and *you* respectively, are even explicitly provided in both clauses. In examples (6) and (8), the core predicates *sai* "compete" and *shengqi* "become angry" are only provided on one instance, but nevertheless the idea that the core predicate has been realized already (at least) once before is clearly conveyed. This indicates that the adverbs *zai* and *you* intrinsically encode the notion of repetition of the modified core predicate (as the core predicate is understood as being repeated despite only explicitly provided once). Repetition as a temporal phenomenon can be further analysed as the sequential arrangement of (at least) two separated points or stretches on the time axis. I will therefore argue that the adverbs *zai* and *you* intrinsically encode two separate reference times at which the core predicate modified by the adverbs occurs. I shall call these times E1 and E2.

While the instance of the core predicate directly modified by the adverb is explicitly asserted to be realized, the previous instance(s) of the core predicate is taken for granted to having been realized already before. Thus we see that the adverbs also encode non-temporal information as discussed in section 3. I argue that they encode an assertion that the core predicate modified by the adverb occurs at E2, and also encode a logical presupposition that the core predicate occurs (occurred) at E1.

We have seen so far that *zai* and *you* encode two identical sets of semantic notions: (1) A sequential arrangement of two separate times E1 and E2, at which the modified core predicate occurs. (2) Assertion directed at the realization of the core predicate at E2, and logical presupposition[8] directed at the realization of the core predicate at the E1. In Fig. 2, a schematic model of the shared semantic structure of *zai* and *you* is presented.

[8] Henceforth referred to only as presupposition for the sake of convenience.

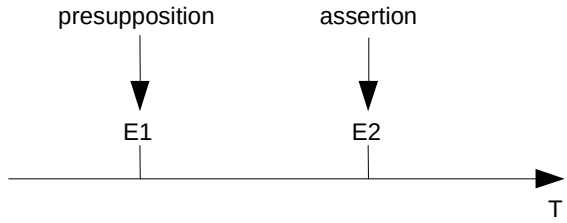


Fig. 2: Model of the shared semantic structure of *zai* and *you*

4.1 The Viewpoint Component

Despite the obvious similarities in the semantic encoding, it is well documented that *zai* and *you* display certain grammatical disparities. "The two can both be used for repeated acts. 'Zai' is used for acts which will be repeated, 'you' is used for acts that are already repeated." (*Xiandai Hanyu Xuci Cidian*, 1998, p. 719) "When expressing repetition or continuation of an action, 'zai' is used for unrealized ones [actions], 'you' is used for realized ones [actions]." (*Xiandai Hanyu Babai Ci*, 1999, p. 644) I argue that this difference must be attributed to an additional semantic component. The model in Fig. 2 cannot be complete, as it ascribes the exact same semantic structure to both adverbs, and can therefore not account for the systematic discrepancies in grammatical function noted in several works as cited above. Instead, I propose that these adverbs encode an additional semantic component in the form of a viewpoint, to which the semantic structure modelled in Fig. 2 is related positionally. A viewpoint can be understood simply as a deictic centre in the temporal structure; a "vantage point" on the time axis. In the case of *zai*, the viewpoint is located between E1 and E2. It is placed subsequent to E1 because the realization of the core predicate at that time is presupposed (and therefore naturally located prior to the viewpoint). The assertion encoded in the adverb is directed at E2, but the data suggest that the core predicate is typically understood as unrealized at that time. Therefore I argue that the viewpoint component in *zai* is located between E1 and E2. The presupposed information is taken for granted and need not any direct attention, so to speak. The asserted information however, is the focal point of the whole semantic structure, and therefore naturally the point towards which the viewpoint is aiming. Therefore the perspective of the viewpoint is prospective. Fig. 3 is a model of the semantic structure of *zai*.

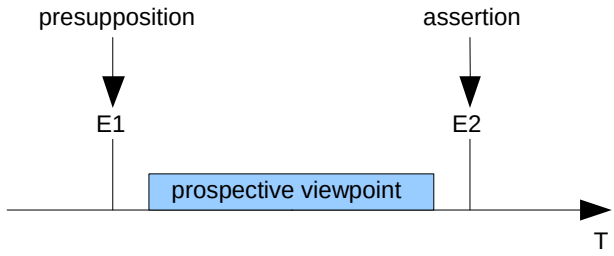


Fig. 3: Semantic structure of *zai*

With *you*, the situation is different. The data suggest that the core predicate is typically understood as realized at E2. Therefore I argue that the viewpoint component in *you* is placed subsequent to E2.[9] The asserted information is always the focal information of the semantic structure, and therefore the viewpoint is in the case of *you* retrospective. Fig. 4 is a model of the semantic structure of *you*.

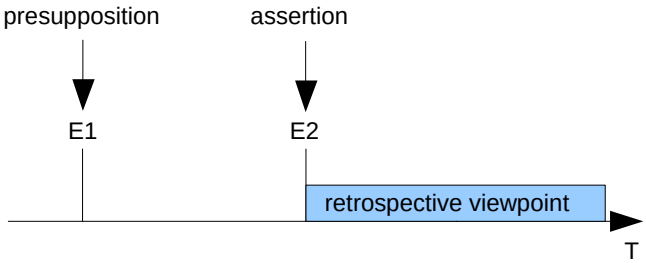


Fig. 4: Semantic structure of *you*

Due to the intrinsic arrangement of semantic components, with a prospective viewpoint located between E1, at which the core predicate is presupposed to be realized, and E2, at which the core predicate is asserted to be realized, the typical temporal structure of a basic declarative sentence with *zai* is one where E1 is located in the past, E2 located in the future, and the viewpoint coinciding with S. S being the default orientation point, the basic declarative sentence centres temporally around it, and the intrinsic semantic structure of *zai* is "distributed" in accordance, with the viewpoint coinciding with S, while E1 is located prior to S (in the past) and E2 is located subsequent to S (in the future). On the contrary, *you* is normally not compatible

[9] It seems that in examples similar to (8), the core predicate can be interpreted as being realized virtually at the time of the utterance. Therefore the viewpoint is really located subsequent to or no earlier than at E2, which is shown in Fig. 4.

with such sentences due to the fact that its viewpoint is retrospective and located after (or at) E2. Instead it is readily used in settings located wholly prior to S, i.e. in the past. These circumstances are exemplified in (9), where only *zai* and not *you* is grammatical, and (10), where only *you* and not *zai* is grammatical.[10]

(9) 明天 我 再（*又） 来 看 你
Mingtian wo zai (*you) lai kan ni
tomorrow I again come see you
"I'll come and see you again tomorrow."

(10) 妈 那 篇 课文 我 刚才 又（*再） 背 了 一 遍
Ma na pian kewen wo gangcai you (*zai) bei le yi bian
mum that CLF text I just now again learn LE one time
"Mum, I went through (in order to learn by heart) that text again just now."

4.2 Temporal Adverb *zai* in Past Settings

As noted by Lu and Ma (1999), *zai* may be used in a past setting, if the sentence depicts a hypothetical perspective.

(11) 昨天 如果 我 再（*又） 看 一 遍 就 记住 了
Zuotian ruguo wo zai (*you) kan yi bian jiu jizhu le[11]
yesterday if I again read one time JIU remember LE
"I would have remembered it had I only read it one more time yesterday."

Since S is the default orientation point of the sentence, the prospective viewpoint cannot normally be applied in a past setting; the prospective viewpoint is naturally directed towards a time subsequent to the default orientation point S. But the hypothetical perspective relativizes these circumstances, as E2 is never explicitly realized but merely hypothetically realized. Therefore *zai* can still be used to express a prospective viewpoint directed towards E2 even though E2 is located prior to S. The hypothetical perspective functions as a mitigating factor extenuating the inherent contradiction between the two concepts past narrative and prospective viewpoint. Fig. 5 shows a temporal interpretation of (11). The time of the utterance is S. The Reichenbachian reference time R is set by *zuotian* 'yesterday'. The temporal and logical structure intrinsically encoded in *zai* is distributed around R, and the prospective viewpoint (roughly) coincides with this time.[12]

[10] Examples (9) and (10) taken from Lu and Ma (1999).

[11] Example (11) taken from Lu and Ma (1999).

[12] The prospective viewpoint is located somewhere between E1 and E2, and E1 and E2 are located somewhere within the scope defined by R. The most important aspect of the figure is to show the relationship between the "outer" temporal reference structure (S and R) and the "inner" temporal and logical structure of *zai*. Since E2 is merely hypothetically realized, the semantic information of *zai* can be applied in a past setting.

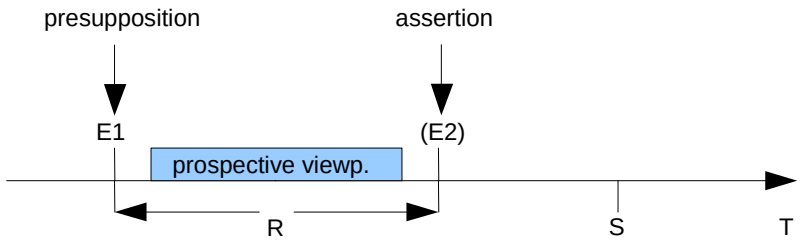


Fig. 5: Temporal interpretation of (11)

As a contrast, *you* is readily applied in a past setting, as the retrospective viewpoint then coincides with S, with E1 and E2 both located prior to S and understood as realized. The temporal interpretation of such an example (12) is given in Fig. 6.

(12) 我 昨天 又 看 了 一 遍
wo zuotian you kan le yi bian[13]
I yesterday again read LE one time
"I read it again one more time yesterday."[14]

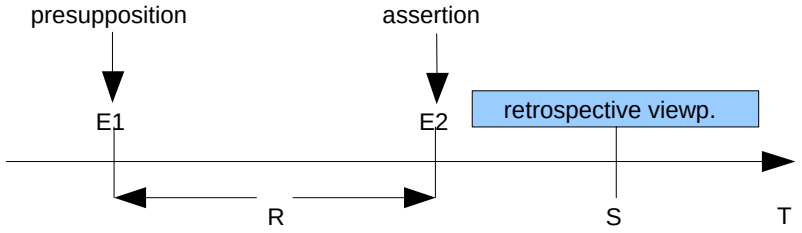


Fig. 6: Temporal interpretation of (12)

4.3 Temporal Adverb *you* in Future Settings

Also noted by Lu and Ma (1999), there are some circumstances under which *you* may be used in a future setting: (1) When the situation is understood as cyclic and recurrent. (2) When the situation depicts an undesirable scenario. Using *you* in a future

[13] Example from my own hand.

[14] Strictly speaking the more natural interpretation of example (12) is that E1 occurs prior to R, i.e. at a day earlier than *zuotian* 'yesterday', but most importantly the intrinsic semantic structure of *you* can readily be distributed around the "outer" temporal reference structure (S and R), as the retrospective viewpoint coincides with S, and E1 and E2 are both realized and located prior to S.

setting is in the normal case a violation of the intrinsic semantic encoding of the adverb, as the viewpoint is retrospective. It can thus not be properly applied when E2 is located subsequent to S. I shall argue that the model of the semantic encoding of *you*, paired with insights concerning conceptual analogy, help make significant progress towards explaining this problem.

4.3.1 Cyclic events

We have seen that since *you* encodes retrospective viewpoint, it is naturally compatible with situations to which such a viewpoint can be applied. The most obvious example is a situation located in the past. These circumstances can be somewhat relativized if the situation modified by *you* is located in the future, but still perceived as certain to occur in some sense. Cyclic events are perceived as certain to occur, even when located in the future. Example (13) shows that in sentences depicting cyclic events located in the future, only *you* is grammatical and not *zai*.

(13) 明天 又（*再） 是 星期天
mingtian you (*zai) shi xingqitian[15]
tomorrow again is sunday
"Tomorrow it's Sunday again."

Cyclic events occur again and again in accordance with a law of regularity and can therefore be anticipated with certainty even when the next occurrence is located in the future. I argue that the conceptual analogy between retrospectivity and certainty makes *you* compatible with situations depicting a future occurrence of a cyclic event. As shown in (13), *zai* is ungrammatical in such a context, indicating that the perception of recurrent regularity inhibits the application of a prospective viewpoint, since it requires a stronger notion of uncertainty. Similarly to the case of *zai* used in past settings, where the hypothetical perspective functions as a mitigating factor extenuating the inherent contradiction between the two concepts past narrative and prospective viewpoint, the certainty associated with cyclic events extenuates the inherent contradiction between the two concepts future narrative and retrospective viewpoint.

4.3.2 Undesirable scenarios

Lu and Ma (1999) present empirical data showing that *you* may be used to modify undesirable scenarios located in the future.[16] If the scenario is not undesirable, only *zai* and not *you* may be used. These circumstances are shown in (14) and (15).[17]

[15] Example from my own hand.

[16] There are no restrictions to the usage of *zai* in such contexts, as a future setting is its natural environment.

[17] Examples taken from Lu and Ma (1999); (14) slightly altered.

(14) 要是 明天 再（又） 吃 面条 我 就 吃倒 胃口 了
Yaoshi mingtian zai (you) chi miantiao wo jiu chidao weikou le
if tomorrow again eat noodles I JIU lose appetite LE
“If I have noodles again tomorrow I’m gonna lose appetite.”

(15) 如果 明天 再（*又） 吃 面条 就 好 了
Ruguo mingtian zai (*you) chi miantiao jiu hao le
if tomorrow again eat noodles JIU good LE
“It would be great if we are having noodles again tomorrow.”

I believe that the key to explaining why *you* may also modify situations depicting future undesirable scenarios lies in the realm of conceptual analogy, just as in the case with *you* modifying cyclic events located in the future. Retrospectivity and certainty are conceptually closely resemblant, as are certainty and unavoidability. If a situation is located in the future, and also certain to occur (like the next occurrence of a cyclic event), that means it is unavoidable.[18] Something which is perceived of as unavoidable is very likely to also be perceived as desirable to avoid. Such a situation is necessarily undesirable. This does not mean, of course, that the future occurrence of a cyclic event must be undesirable. The reason that *you* may modify such situations is simply due to the conceptual resemblance between retrospectivity and certainty, and there is no need to invoke additional analogous concepts in order to explain why *you* may do so. What is does mean, is that since there is a clear chain of conceptually interrelated notions linking the concepts of retrospectivity and undesirability to each other, a possibility is created for the viewpoint encoded in *you* to be transferred from the concept of retrospectivity onto the concept of undesirability through analogous association. This chain of interrelated concepts is visualized in Fig. 7.

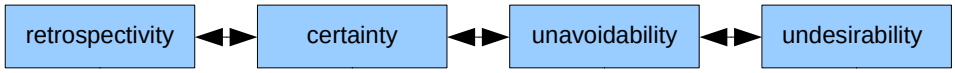


Fig. 7: Chain of conceptually interrelated concepts

The reason why *you* may be used in a future setting modifying undesirable scenarios is due to basically the same mechanism explaining why *you* may modify cyclic events located in the future. The later case is arguably somewhat more easily understood, as the conceptual analogy between retrospectivity and certainty seems rather direct. The concept of undesirability is probably conceptually less directly analogous to the concept of retrospectivity, and the connection is reached through additional intermediate concepts.

[18] Or in any case perceived of as unavoidable.

5. Summary

The inquiries in this paper showed how a decompositional analysis of the adverbs *zai* and *you* reveals that these morphemes encode both temporal and non-temporal information. The analysis also showed that the intrinsic semantic encodings of *zai* and *you* are virtually identical, as they encode the exact same semantic components. It was argued that the documented differences in grammatical use are due to a discrepancy in the arrangement of the semantic components; *zai* encoding a prospective viewpoint located between E1 and E2; *you* encoding a retrospective viewpoint located after or no earlier than at E2. Through combining the decompositional analysis of the adverbs with ideas concerning conceptual analogy, explanations to the usage of *zai* and *you* in unnatural contexts could be provided.

Abbreviations

CLF Classifier
DE Subordinator; nominalizer
GUO Experiential marker
JIU Connective
LE Verb/sentence-final particle
MA Interrogative particle

References

Comrie, B. (1985). *Tense*. Cambridge University Press.

Davis, W. (2010). Implicature. In *The Stanford Encyclopedia of Philosophy*, E Zalta, ed. (2010 edition) Retrieved from http://plato.stanford.edu/entries/implicature/.

Grice, P. (1989). Logic and Conversation. In *Studies of the Way of Words*. Harvard University Press.

Kant, I. (1787). *Critique of Pure Reason*, (revised edition). Retrieved from http://www.gutenberg.org/etext/4280.

Karlsson, J. (2010) *Temporal Adverbs in Modern Standard Chinese – A Decompositional Inquiry*. Lund University.

Klein, W. (1994). *Time in Language*. Routledge.

Korta, K. & Perry, J. (2008). Pragmatics. In *Stanford Encyclopedia of Philosophy*, E. Zalta (ed.), (fall 2008 edition). Retrieved from http://plato.stanford.edu/entries/pragmatics/.

Lu, J. & Ma, Z. (1999). (关于表重复的副词“又”“再”“还”) Guanyu biao chongfu de fuci “you” “zai” “hai” [Concerning the adverbs “you” “zai” “zai” expressing repetition]. In: (现代汉语虚词散论) *Xiandai Hanyu Xuci Sanlun* [*Various Treatments of Function Words in Modern Chinese*], (revised edition). Language and Literature Press.

Lu, J. (2003). (现代汉语语法研究教程) *Xiandai Hanyu Yufa Yanjiu Jiaocheng* [*A Course in Modern Chinese Grammar Research*]. Peking University Press.

Lü, S., ed. (1999). (现代汉语八百词) *Xiandai Hanyu Babai Ci* [Eight Hundred Words in Modern Chinese]., (revised and enlarged edition). Commercial Press. Original edition 1980.

Peccei, J. S. (1999). *Pragmatics*. Routledge.

Reichenbach, H. (1947). *Elements of Symbolic Logic*. Macmillan.

Simpson, P. (1993). *Language, Ideology and Point of View*. Routledge.

Smith, C. S. and Erbaugh, M. S. (2005). Temporal Interpretation in Mandarin Chinese. In *Linguistics: an interdisciplinary journal of the language sciences*, 43 , no. 4. Retrieved from http://uts.cc.utexas.edu/~carlota/papers/S%26E%202005.pdf.

Xiandai Hanyu Xuci Cidian (现代汉语虚词词典) [*A Dictionary of Function Words in Modern Chinese*] (2004). Peking University Press. Hou, X., ed. Original edition 1998.

Zhu, D. (1961). (现代汉语语法研究) *Xiandai Hanyu Yufa Yanjiu* [*Research on Modern Chinese Grammar*]. Commercial Press.

THE LANGUAGE TEACHER'S ROLE IN THE AGE OF THE INTERNET*

Nagisa MORITOKI
University of Ljubljana, Faculty of Arts
nagisa.moritoki@guest.arnes.si

Abstract

The Internet can have a strong influence on students learning the Japanese language in Slovenia, as well as in other parts of Europe. Almost all freshmen have come into contact with Japanese pop culture via the Internet. The aim of this paper is to discuss the teacher's role in overcoming certain problems associated with learning the Japanese language in the age of the Internet. First, looking at a general survey of the current situation surrounding teaching Japanese language in Slovenia, we identify the advantages and disadvantages of using the Internet when learning the language. However, the disadvantages of the Internet that lead to learner problems are, in fact, the problems that we also face in daily communication. So, as a teacher, I propose following three strategies to lead the learner: first, let the learner's interests stimulate him to explore a wider and deeper world; second, lead the learner to reconstruct his world; and third, lead the learner to self expression so that he can be understood by the listener and improve his communication skills. Such are teacher's strategies for interactive communication based on individual standpoint versus a world view, which has emerged in teaching Japanese language when the learner seeks language skills not solely for practical purposes as in Slovenia. Considering this, I additionally propose for Common European Framework of Reference (CEFR) ideology that those strategies aim to achieve "an expertise of the relationship with the Other" (Zarate, Gohard-Radenkovic, Lussier, & Penz, 2004, p. 11).

Keywords

The Internet, language teaching, teacher's role, socio-cultural proficiency, CEFR

Izvleček

Medmrežje ima lahko močan vpliv na študente, ki se učijo japonskega jezika v Sloveniji kot tudi v drugih evropskih državah. Zadnje čase se vpišejo skoraj vsi novi študentje na univerzo, že oboroženi z znanjem japonske pop kulture, pridobljenjim skozi medmrežja. Cilj pričujočega prispevka je, da proučimo učiteljevo vlogo pri premagovanju težave, ki jih povzroča obdobje medmrežja študentom japonskega jezika. Najprej si ogledamo rezultate splošnega vprašalnika o sedanjem položaju učenja japonskega jezika v Sloveniji in identificiramo prednosti in slabosti uporabe medmrežja pri učenju jezika. Slabosti medmrežja, ki povzročajo težave učencem, so pravzaprav težave, s katerimi se soočimo tudi mi v vsakodnevni komunikaciji. Zato kot učitelj

---

* This paper is based on my presentation "'Krikku wo koeta ibunka rikai (〈クリック〉を超えた異文化理解, Cross cultural understanding over the last click)" on the symposium "Cross cultural understanding" at Gunma University, Japan, on February 8th, 2011. I thank all the participants who shared this topic to discuss at the symposium.

ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.1.2.39-52

predlagam strategijo naslednjih treh postopkov, s katerimi laže in bolj spretno vodimo učence: prvič, pustimo prostor za učenčevo zanimanje, ki ga lahko spodbuja k raziskovanju širšega in globljega sveta okoli sebe; drugič, naj učenec skuša rekonstruirati svoj svet; in tretjič, pustimo učenca k samoizražanju in pri tem mu pomagajmo, da pravilno razvije svoje komunikacijske spretnosti, da se ga razume. Takšna je učiteljeva vloga za medsebojno sporazumevanje na osnovi posameznega stališča in pogleda na svet. Takšno razmišljanje je aktualno tudi v pedagogiki japonskega jezika, ko učenec išče jezikovne spretnosti ne le za praktične namene. Dodatno predlagam, da bodimo pozorni na ideologijo Skupnega evropskega okvirja referenc (CEFR), ki z omenjeno strategijo skuša doseči 'ekspertizo o odnosih z Drugim' (Zarate et al., 2004, str. 11).

Ključne besede

Internet, poučevanje jezika, vloga učitelja, družbeno-kulturna usposobljenost, CEFR

1. Introduction

In Slovenia, just as in the other parts of the world, many languages are spoken in daily life and taught at school. Looking at the current condition of teaching Japanese as a foreign language in Slovenia, the course of Japanese studies – Japanese language is taught as the main subject – is one of the most popular courses at the University of Ljubljana. Since 1995, when the course was founded, more than 200 students have graduated and, at present, approximately 200 students are enrolled in the course. Who might have enough interest in Japanese culture and language to enter the course? In the last several years almost all such freshmen already have some familiarity with, and enjoy, Japanese pop culture, such as anime (cartoon films), manga, and J-pop songs, all of which are available via the Internet. On the other hand, experiences of the students enrolled in the course of Japanese studies are mostly made in a virtual world, for example, artificial situations for exercises in the classroom, homepages and blogs on the Internet, as Japanese language is rarely spoken in Slovene daily life and only a small number of graduates get jobs related to Japan. The question arises, therefore, what the present aim of Japanese language teaching in higher educational institutions in Slovenia is, considering that CEFR (2001) stresses the importance of communicative proficiency.

In this paper I focus on the problems of teaching Japanese in Slovenia and on possible solutions, which could be shared with other educational institutions in other countries where Japanese language is not typically spoken as a foreign language. First, in chapter 2, I present a general survey of the Slovene situation, focusing mostly on the course of Japanese studies at the University of Ljubljana. After that, I address advantages and disadvantages that we face and clarify the problems in chapter 3. Then, in chapter 4, I propose three strategies for the teacher to solve these problems. It is not to let language learners pacify themselves with the situation of "virtual" life, but to lead them to a wider, deeper, and more substantial world where they can think for themselves and interact with each other. Finally, in chapter 5, I conclude with the proposition of redefining the role of language learning in Europe. How can learners

benefit from Japanese instruction even if they do not use the language later in their daily life? This is my proposal for CEFR, from the classroom of a minor and therefore "powerless" foreign language in Europe.

2. The Slovene environment surrounding Japanese language teaching

Slovenia is a small country located on "the sunny side of the Alps" in central Europe with a population of approximately two million people, 96% of them Slovene (Eurostat, European Commission, 2010). The official language is Slovene. English, German and other languages are studied as foreign languages in elementary and secondary school. A tourist can communicate in those languages in Ljubljana, the capital city, or several big cities in Slovenia without any problem. Moreover, because Slovenia is surrounded by Italy, Austria, Hungary and Croatia, those languages – Italian, German, Hungarian and Croatian – are familiar to residents near the border, and words and expressions are loaned from these languages in daily conversation. People over 40 years old learned the Serbo-Croatian language during compulsory education. We can say that the Slovene people are efficiently accustomed to using foreign languages.

Due to the geographic distance between Slovenia and Japan, Japanese is not a commonly spoken foreign language for Slovene people. They do not generally hear Japanese spoken on the street, nor hear it on the television. There are now more than 100 Japanese residents in Slovenia, with numbers rising after the Japanese embassy was established in 2006. However, ordinary Slovene people do not have much contact with Japanese people. While two Japanese restaurants are thriving in the fashion of worldwide healthy foods, only a few companies have business trade with Japan. We can say that Slovenia has neither a strong relationship with Japan nor a vivid image of it. I firmly believe that the environment of Japanese language teaching in Slovenia has not adopted the idea that language is a "communicative event" (for example, de Beaugrande, 1996, Hopper, 1998, and so forth), which is the current linguistic main stream in the age of post-structuralism after the Swiss linguist Ferdinand de Saussure.

3. The learner's access to Japan and related problems

The course of Japanese studies started in 1995 at the University of Ljubljana; its predecessor of more than 10 years was the biannual intensive course of Japanese language organized by the Slovenian Oriental Society ("Slovensko orientalistično društvo") (Shigemori Bučar & Bekeš, 2005). By 2011 more than 200 students have graduated, and are either in the graduate course of Japanese studies, or working in Slovenia or Japan, translating Japanese literature and so on. Some of them successfully got jobs teaching the Japanese language, in newly established courses in private

language schools. Let us now briefly examine the environment surrounding Japanese learners in Slovenia in order to point out features and problems.

3.1 Growing individual access to Japan by Internet

Europeans' interests in Japan and Japanese culture have changed over time, and so has the paradigm of Japanese language learning (Sasaki 2010). In the 1980s, Japanese culture attracted only a small fraction of the population. If they talked about Japan, it might have been about sports such as jūdō and karate, orientalism such as ikebana and bushidō, or geisha, which most people viewed "from the opposite bank" but with no real personal experience. It was little more than "something different."

After that, in the 1990s, with Japan's economic growth, Japanese pop culture started generating more interest worldwide. Animated films, movies by the filmmaker Kitano Takeshi and the literature of Murakami Haruki were launched on the international stage and became popular among ordinary people.[1] Successful Japanese companies established branches in chief cities in Europe and throughout the world and Japanese tourists thronged abroad. This is a drastic change in support of the point that Japanese culture has become more familiar to the general public of Slovenia. Japan is no longer interesting only for a small group of people, but for us "on the same bank." In this decade Japan has become widely known in the world and the number of Japanese language learners has risen rapidly. The establishment of the course of Japanese studies at the University of Ljubljana was also a part of this age.

One more dramatic change for the reception of Japanese culture abroad happened due to the rapid spread of the Internet into standard homes (Statistical office of the Republic of Slovenia, 2010). The Internet delivers various fields of information: in addition to Japanese traditional culture, sports, popular literature and films, there are the Japanese subcultures of J-pop songs, anime and manga, which are now available through mass production and accessible on the Internet. The questionnaire that I carried out on the freshmen in the autumn of 2005 and 2010 showed that almost all of the respondents had experienced Japanese subculture via the Internet before entering the university.

[1] For instance, modern literature works by Murakami Haruki began to be translated into European languages in 1990s. These are the first translations published in each of the following languages: English in 1985 (Pinball, 1973, Kōdansha), French in 1990 (La Course au Mouton Sauvage, Seuil), Italian in 1992 (Sotto il segno della pecora, La Gaja scienza), Spanish in 1992 (La caza del carnero salvaje, Anagrama) and German in 1997 (Wilde Schafsjagd, Suhrkamp). The first Slovene translations were published in 2004: "Divja jaga za ovco" (Založba Blodnjak) and "Ljubi moj sputnik" (Založba Mladinska knjiga), both translated from the English version. The first direct translation from Japanese was first published in 2005: "Norveški gozd" (Založba Sanje), translated by Nika Cejan, a graduate of the course of Japanese studies at the University of Ljubljana. In the field of modern films, Kitano Takeshi received "the Golden Lion" at the Venice Film Festival for "Hana-bi" in 1997.

3.2 Advantages of the Internet for Japanese learners

For Japanese language learners, Internet exposure brings with it its pros and cons. In this section we will look at two advantages the Internet brings to Japanese learners in Slovenia.

First I must point out that the Internet has removed barriers that previously blocked a person's access to information. Thanks to the Internet, information is now equally available to anyone with Internet access. Thus, the traditionally spatial, temporal and financial barriers have been removed.

There are a few libraries in Europe that are proud of the quality of their book collections on Japanese studies. They are based on well-organised library plans for synchronic and diachronic views and, of course, they depend on enough finance to fulfill the plans. However, just because there are good libraries, this does not mean that one can freely get one's target document. Even if a person knows that the document is in the library, several problems still lie in front of him: geographic distance between him and the library, and time and money required to reach there. For instance, not every European scholar who is researching the history of Japanese teaching in 18th century Russia can go to St. Petersburg to search through the rich abundance of documents. Only a few lucky scholars can go to St. Petersburg and devote themselves to research at the library with such a rich collection. However, the availability of Internet access has brought St. Petersburg closer to the researcher. Now, one can freely access the desired document as long as the document is digitalized and made available over the Internet and the researcher knows how to use the Internet or at least find the information, where the document can be found. The time and money required to get from Slovenia to St. Petersburg no longer poses an obstacle. In the world of Internet, there are significantly less conventional limitations for the person who is eager to obtain information.[2]

The second advantage of the Internet is that we can access to a great amount of information in the digital world, from government papers, to private blogs or twitter, to other media such as texts or movies. One can download a digital book of Japanese classical literature at home. An enormous amount of information is ready to be accessed on the Internet.

Considering the advantages of the Internet, we can fairly say that knowledge is equally available to everyone who wishes to know. Japanese language learners are no exception. On this point, we are one step closer to opinions such as of Mey (1993/1996)[3] that "education is for the rich", using Brecht's words.

---

[2] The barrier free information that I mention above is for the person who actively pursues this knowledge. The untouchable enclosure on the Internet, of course, exists, as people have information that they do not want to disclose.

[3] The citation is from the Japanese edition (Mey, 1996, p. 313).

3.3 Disadvantages of the Internet for Japanese language learning

The advantages of the Internet, as discussed above, could, however, simultaneously lead to disadvantages for the Japanese learner.

First, the Internet allows every person to access to a great amount of information online, but it also poses the hazard that a learner might be drowned in the vast ocean of information. The flow of the information via the Internet often causes a person to get lost in front of the screen or be at a loss when deciding how reliable different documents are. On the other hand, library users can feel confident that the research material has been manually selected and is worth consulting. There are no such selection processes, nor quality control censors on the Internet. As a result, it is often difficult to distinguish reliable sources from unreliable ones, and the qualified one from the unqualified one. It is difficult to judge what is suitable, especially in a foreign language, where one is not a proficient speaker. Why? Because it is difficult for a language learner to understand written texts, especially when he does not have enough social and cultural background in the target language to support his understanding of the content. For example, a learner can understand the sentence, "It was pointed out to me that I had slippers for the toilet and I was laughed at by my friends," but cannot understand why the person was laughed at if he does not know Japanese people have a custom of changing their slippers when they go to the toilet at home. This understanding is something more than linguistic comprehension. Haruhara mentions that language proficiency is embedded organically in socio-cultural proficiency (Haruhara, 2009, p. 17)[4]. His mention can be acknowledged when considering the relationship between language and socio-culture. So, it is necessary for a learner to cultivate socio-cultural proficiency, as well as language proficiency. Here we can address the language teacher's first role: to cultivate both linguistic and socio-cultural proficiency, which helps the learner to correctly understand the speaker's intention of the sentences and also evaluate the appropriateness of information when selecting from an ocean of information.

The second disadvantage of the Internet is that everything is done by the user's click. In chapter 3.2 we considered the advantage of the Internet as a means of access to the information on demand, however, it is to the user's disadvantage that he may never come across information that is of no interest to him. He can see what he wants; but he does not see what he does not want. And he might be completely satisfied with what he finds. The matter is different in the library. A library user is very likely to pick up a book that he has not looked for, and so the book exists substantially even if he is not consciously aware of it. Thus, a role of the teacher is to inform the learner of the

[4] I will discuss in another paper about the differences between 'language proficiency and socio-cultural proficiency' (Haruhara, 2009) and 'plurilingualism as a competence and plurilingualism as a value' (Beacco & Byram, 2003), considering the arguments of 'Nihon jijō' and 'cultural literacies' in the Japanese language learning.

amount of information behind him and, therefore, stimulate his curiosity in the right direction.

The third disadvantage of the Internet is that the amount of extra-linguistic information needed to help communication is extremely limited. One advantage of the Internet is that everyone can retrieve information anytime and from anywhere, but it also means that the receiver of the information does not necessarily share the same time and place as the sender of the information. In everyday communication, we usually share the same location and time zone, even in the case of talking over the telephone we share the time. To understand the speaker, we rely heavily on that what is not easily written down: linguistic intonation, pause, speed or visual information, such as the speaker's [5] gestures, expressions on his face, and eye movements. After considering the sum of these linguistic and extra-linguistic cues, we understand the speaker's intention and judge his veracity and what he feels. On the other hand, those elements that are difficult to write down are left out on the Internet, so one has to infer a writer's intention only from what is written, sometimes with the help of emoticons and the layout of the page. So there is a higher risk of misunderstanding. The writer takes a gamble, as well. He does not know whether the reader understands his intention of the sentence correctly or not, because of the dyschronism of the Internet. Because of this same reason, the writer might forget whom he is sending the message to when he writes, failing to effectively communicate through one-way utterances. As we have seen above, communication on the Internet is not two-way interaction as in "usual" conversation.

In summary, the three disadvantages of the Internet are: too much information; targeted information retrieval without serendipitous encounters; and the dyschronism of the writer and the reader. We find these problems with the Internet, yet we find that they are problems that are not entirely specific to the Internet. They also appear in a slightly different form in daily conversation when we get lost in an ocean of too much information and fail with one-way expression, which can easily lead to misunderstanding. We might say that the above disadvantages of the Internet are reflections of the problems we have in daily conversation.

4. Three strategies to lead a learner towards successful communication

Although it is true that the Internet which most of the university studenta are using everyday is not the best tool for teaching Japanese language, it is worth discussing the disadvantages of the Internet, including the problems of communication that I pointed out above. It is because usual communication shares these problems and teaching the Japanese language must provide for not only linguistic proficiency but also socio-

[5] A writer or speaker is the one who sends information, while a reader or hearer receives that information. In the main text, I distinguish between those words depending on how the communication is carried out – if it is written or spoken.

cultural proficiency for interactive communication. In this section I propose three strategies with which a teacher should lead learners, so that the learners will be able to acquire both proficiencies to successfully communicate in daily life.

4.1 Widening and deepening the learners' background knowledge

As already discussed, a learner's active actions – clicks on the Internet – are promoted by his interests. These interests must hold his attention, if he is to be stimulated, and most importantly, in order to achieve a high level of linguistic competence to communicate effectively in Japanese, the learner needs a well-balanced and sufficiently comprehensive amount of knowledge about the society and culture in which the language is used. I consider here two paths through which his interests should be led in order to effectively construct such a world.

Examining these two directions of interest, let us review one example from the class of Modern Japanese Culture in the academic year 2010/11 at the University of Ljubljana. The participants, mostly university students in their third year in the course of Japanese studies, made a Web journal for the students in their second year[6], so that the second-year students could review what they learned so far and could learn about a part of Japanese life that is not mentioned in their school textbooks. One of the participants, who had stayed in Japan for four weeks on a short visit program, reported in her article that the Japanese put their slippers on when they enter a house. This is true, but is this anything more than the information that can be found in any travel guide?

What can a teacher do for the students in the classroom? What we did together was to try and recall any related information, and to consider, what other information could help us to a better understanding, and think about how and where such information could be found. The first step was to visually interpret the theme by observation, for example, where in their house the Japanese take their shoes off; where they put them back on; where the border is between taking slippers off and putting shoes on; how the Japanese use their slippers in some other places and so on. The next step was to consider causal relationships within this topic by consideration, for example, why they take their shoes off at home; what they would do and think if someone did not take off their shoes upon entering the house; what the historical background is; whether there are other influences from Europe and so on. Here, I call the first method of observation the "horizontal" direction and the second one of consideration the "vertical" direction. Both directions of thinking are needed, because without them one would have "the frog in the well" scenario where the learner would be stimulated only by his own interests. Furthermore, both directions are important: observation without consideration does not lead the learner to understand. Without consideration of causal relations, the learner cannot understand and appreciate the significance of similar cases. Consideration

[6] The journals are on the Web site: http://www.flickr.com/photos/kulturologija2010/sets/.

without observation leads to impractical arguments. Though the example of the slippers was a tiny and simple one, I emphasize that the teacher can lead the learner to acquire wider and deeper extent of knowledge even with such a simple beginning.

4.2 Giving learners cues to reconstruct their world

Here is another example of a certain Japanese student, who had been to Slovenia several times. She once told the class about her disappointing experience at a Japanese restaurant in the center of Ljubljana because it was too expensive. She continued to complain that it was a tourist trap and she would never go there again. It is true that the restaurant is expensive. However, it is no excuse to stop thinking. Her unpleasant experience gave us an opportunity to examine the matter and its background once again. A complaint based on a person's conflict or misunderstanding often gives us to understanding the matter from a new standpoint if the misunderstanding is cleared.

In the case of the Japanese restaurant, the student seemed not to have enough information to judge why the price was so expensive. So our class first examined the price of Japanese dishes in various Japanese restaurants in Europe. Then we discussed economy and management; how strong the economic relationship is between Slovenia and Japan; importing the materials needed to create Japanese cuisine in Slovenia where there is no direct flight from Japan; and the location of the restaurant. We talked about the image of Japan in Slovenia too. They are all horizontal observations and vertical considerations of the world based on the price of Japanese food. After the discussion, we reached the conclusion that the Japanese restaurant in Ljubljana is really expensive, but it should be reasonable, considering the current situation surrounding to restaurant.

The discussion was not intended to force the Japanese student to consent to the idea that the restaurant was not expensive, but instead to give her and other students another perspective of the world. At first, her experience was only that she ate Japanese food in the restaurant and unexpectedly paid a lot. Through the classroom discussion, she and her classmates could effectively realize the circumstances surrounding the restaurant. What must be pointed out is that the participants in the class could relate to each other the circumstances they knew and successfully create another picture of the Japanese restaurant.

According to Sunakawa (2007), language learning has the following four aspects: 1. elevation of language ability; 2. expansion of the world inhabited; 3. ego formation as a subject; and, 4. reinforcement and elevation of literacy as practical reception ability (p. 150). I support his idea in principle, but what is more important, as I found in the case of the Japanese restaurant, is not only the expansion of the world inhabited, but also the "reconstruction" of it. The students already knew that there is no direct flight from Japan to Slovenia or were aware that the owner wants to attract a certain type of clientele to the restaurant, but until the discussion they had not considered those matters in relation to the price of the food. Only after they had found new correlations among those matters were their thoughts were perceived and new phase of their

conscious world emerged in front of them. A teacher never needs to show his new view, but needs instead to help a learner find a standpoint In which the learner can have his own new view of the world.

4.3 Making a substantial relationship with the world

We have hitherto considered two strategies in section 4.1. (the horizontal and vertical paths of observation and consideration) and 4.2. (reconstruction of the learner's conscious world). When a learner gets a newly reconstructed world that is based on a wider and deeper understanding, the third strategy of the teacher is to lead the learner to express his ideas and thoughts out loud. It can be said that a small but substantial contribution of the learner's perspective adds to this world too. The learner's idea does not exist for other people until it is expressed, just as the information on the Internet, which does not exist until he clicks on the desired links.

The learner has made a relationship with the world because the first two strategies were concluded only within the learner's mind. But it is not enough to make a relationship with the conscious world if the relationship is only one of two ways. There really is no relationship if it is only one-way. The ideas and thoughts of the speaker need to be understood when they are expressed, or at least a speaker has to try to be understood, in order to make a relationship with others. Talking to oneself is not communication, since communication is an interactive event. As Mey (2001) explained, "the reader is party to the textual discourse as much as is the author" (p. 793), therefore, there is no communication without a reader or listener who understands.

Here again is another example from the class of Modern Japanese Culture when the participating students made Web journals for second-year students and during this process learned to understand the role of the reader. Third-year students selected topics for the second-year students and then individually wrote articles in Japanese at home. However, it often went wrong. The most frequent failure of the writer, even Japanese students who participated in the project, was to write long lists resembling a catalogue on some topic. It must have been hard work for them to make such a long list in a foreign language, and this should be applauded. However, they needed to write in an article format for second-year students. Who can read such long lists?

What a second-year university student might be interested in is not a long list of information but what a friend sitting next to us does and thinks, and "a big frame that helps us interpret the purpose of other's acts or communication" (FitzGerald, 2002/2010, p. 26). As for slippers in the Japanese house, an interesting article is not an encyclopedic description (FitzGerald, 2002/2010, p. 26) that the Japanese take their shoes off and put their slippers on when they enter the house, but it is, for example, the individual episode of a student who forgot to change her slippers when she came out from the toilet. The first mistake a learner is inclined to make is to forget the listener

and express inorganic contents. A learner, as a writer or a speaker, should always be aware of the existence of the reader or listener.

The learner then has to be aware of his own role and how he might better understand the speaker. In daily conversation, where the speaker transfers ideas or thoughts to the listener, sharing the same place and the same time, we usually choose words, phrases and strategies, even contents while paying attention to the listener and his understanding. If the speaker assumes that the listener does not understand what he has said, the speaker would explain in a different way to help him understand. This reminds me again of Mey's statement, "the reader is party to the textual discourse as much as is the author" (Mey, 2001, p. 793). However, it is difficult for a speaker to be aware of the listener when they do not share the same place, and time. The language teacher needs to teach a learner to be conscious of the listener.

In the class of Modern Japanese Culture a few years ago, one project required students to prepare a presentation about Japan for a Slovene elementary school. I expected university students to find different ways of expression for a different audience. During the preparation phase, topics that might interest the pupils were decided on by common consent. However, some students could not prepare suitable expressions that could be understood by elementary school children who knew Japan only through anime cartoons. I am convinced that their failure did not stem from their lack of Japanese language skills, as they also prepared presentations in Slovene language, but whether they could identify and successfully communicate with their target audience. For example, they introduced the following geographic features of Japan: "the land of Japan is approximately 37,800 square kilometers and its population is 130 million." The pupils had no concept of such big numbers. After the main discussion, extra explanations were added, comparing Slovenia's size to that of Shikoku island. The university students showed pupils maps of both Slovenia and Shikoku island, and they reported that the population of Japan is 60 times larger than that of Slovenia, so that their young subjects could make direct correlations. This is one example where the third-year students tried to make themselves understood, and succeeded.

Learners sometimes forget who they are talking to, particularly when they have to express themselves in a foreign language, because they are preoccupied in trying to remember words and phrases and they forget how to express themselves effectively. It is not a rhetorical effort, but a strategy to make oneself correctly understood by the listener, as intended. Furthermore, the tendency for the speaker to forget the listener would be high when the listener does not sit in front of the speaker, such as in report writing or on the Internet. Where there is no listener, there is no interaction. Even when they do not share time and space, the listener exists. With that said, expression must be delivered so that the listener understands the speaker correctly and the speaker knows how to reach his audience.

5. Conclusion – A contribution to language learning

In this paper, we discussed what a teacher can do for a learner in the age when more than 95% of freshmen in Slovenia have Internet access. In chapter 3, we saw advantages and disadvantages of the Internet and I pointed out that they are not only features seen with Internet usage but also problems found in every day communication. I proposed in chapter 4, a language teacher's three strategies: 1. leading the learner's interests to a wider and deeper conscious world; 2. leading the learner to reconstruct his world; and 3. leading the learner to make substantial relationships with the world – for better interaction. Why should a language teacher follow these strategies? Because learning a foreign language gives us an opportunity to think about communication not only in the target language but also in any language universally.

One might think that teaching Japanese language in Europe, especially where the relationship to Japan is as remote as in Slovenia, could lead to an impractical theory. Furthermore, it could be problematic in an age when CEFR seems to set practical communicative proficiency as the main learning goal. This learning goal is sometimes mistakenly presented as the only learning goal of language education in any institution, including university courses of Japanese studies. However, I would like to disagree. Regardless of whether or not a learner will use Japanese after his graduation, the language learner should enlighten himself with regard to many important questions and their possible solutions. Especially when the language, like Japanese, is rarely spoken in the country. Perhaps this gives the learner less practical language proficiency, but a deeper understanding of the significance of communication and of the ways of communication which are for "an expertise of the relationship with the Other" (Zarate, Gohard-Radenkovic, Lussier, & Penz, 2004, p. 11) is achieved. Such an understanding helps the learner live substantially in our modern world.

References

Beacco, J. C. & Byram, M. (2003). *Guide for the development of language education policies in Europe – From linguistics diversity to plurilingual education: Main version* (Revised edition). Strasburg Cedex, France: Council of Europe Publishing.

Beaugrande, R. de. (1996). *New foundation for a science of text and discourse: Cognition, communication, and the freedom of access to knowledge and society. The series advances in discourse processes, Volume LXI*. Norwood, New Jersey, the United States of America: Ablex Publishing Corporation.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment (CEFR)*. Cambridge, the United Kingdom: Cambridge. Retrieved from http://www.coe.int/t/dg4/linguistic/cadre_en.asp. Framework_EN.pdf

Eurostat, European Commission. (2010). *Europe in figures - Eurostat yearbook 2010: Population.* Retrieved from http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/publication?p_product_code=KS-CD-10-220.

FitzGerald, H. (2010). *How different are we? – Spoken discourse in intercultural communication.* (Y. Murata, Y. Shigemitsu, M. Ōtani & Y. Ōtsuka, Trans.). Tokyo, Japan: Hituzi shobō. (Original work published 2002).

Haruhara, K. (2009). Shakai bunkateki purofishienshī towa nanika: shakaiteki kōshō wo kanō ni suru kōkyōteki purofishienshī shiron [What is socio-cultural proficiency?: An essay on common proficiency which makes social negotiation possible]. In O. Kamata, H. Yamauchi & R. Tsutsumi (Eds.), *Purofishienshī to nihongo kyōiku [Proficiency in teaching Japanese as a second Language]* (pp. 69-97). Tokyo, Japan: Hituzi shobō.

Hopper, P. (1998). Emergent grammar. In M. Tomasello (Ed.), *The new psychology of language* (pp. 155-176). New Jersey, United States: Lawrence Erlbaum Associates.

Mey, J. L. (1996). *Pragmatics: An introduction.* (H. Sawada & M. Takaji, Trans.). Tokyo, Japan: Hituzi shobō. (Original work published 1993).

Mey, J. L. (2001). Literary pragmatics. In D. Schiffrin, D. Tannen & H. E. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 787-798). Oxford, the United Kingdom: Blackwell Publishing.

Sasaki, M. (2010). Nihongo kyōshi yōsei kōza no jūjitsu wo motomete – Minkan kyōshi yōsei kōzayō no jiko hyōkahyō shian. (Enriching Japanese language teacher training programs – Proposal of self-assessment tools for private institutions). *The Journal of J. F. Oberlin University. Studies in Language and Culture 1*, 89-106.

Shigemori Bučar, C., & Bekeš, A. (2005). Tečaj japonskega jezika pri Slovenskem orientalističnem društvu: pouk japonskega jezika v Ljubljani do ustanovitve Oddelka na FF. *Azijske in afriške študije IX, 1*, 50-55.

Statistical office of the Republic of Slovenia. (2010). *Usage of information and communication technologies in households and by individuals, Slovenia, 2010 - final data.* Retrieved from http://www.stat.si/eng/novica_prikazi.aspx?id=3462.

Sunakawa, Y. (2007). "Gengo no kakutoku, shūtoku" to "nichijōseikatsu sekai no kakutoku kakujū" no ittaisei ni tsuite [On the unity of "Language acquisition and learning" and "Acquisition and expansion of the world of daily life"]. In M. Sasaki, H. Hosokawa, Y. Sunakawa, I. Kawakami, M. Kadokura, & H. Segawa (Eds.), *Henbō suru gengo kyōiku – tagengo, tabunka shakai no riterashīzu towa nanika [Changing language education – What are 'literacies' in the plurilingual and pluricultural society?]* (pp. 141-164). Tokyo, Japan: Kurosio.

Zarate, G., Gohard-Radenkovic, A., Lussier, D. & Penz, H. (2004). *Cultural mediation in language learning and teaching.* Council of Europe. Strasburg Cedex, France: Council of Europe Publishing.

WORD CLASS RATIOS AND GENRES IN WRITTEN JAPANESE: REVISITING THE MODIFIER VERB RATIO

Bor HODOŠČEK
Tokyo Institute of Technology,
Graduate School of Decision Science and Technology,
Department of Human System Science
hodoscek.b.aa@m.titech.ac.jp

Abstract

This paper explores the variability of genres in the Balanced Corpus of Contemporary Written Japanese using the modifier-verb ratio proposed by Kabashima and Jukaku (1965). Using bagplots to quantify the relation between noun and modifier-verb ratios for each genre, an attempt is made to classify genres on the scale of descriptive to summative according to Kabashima and Jugaku (1965). Our initial analysis confirms previous research results, while at the same time uncovering some contradictions in the ratios of the genre of magazines.

Keywords

BCCWJ, MVR, bagplot, genre, Long Unit Words

Izvleček

V članku bom raziskoval variabilnost žanrov v referenčnemu besedilnemu korpusu Balanced Corpus of Contemporary Written Japanese preko razmerja med dvema besednima vrstama – modifikatorjem in glagolom – prvič predstavljenega v delu Kabashime in Jugaku (1965) pod kratico MVR. Z uporabo statistične grafične metode bagplota raziščem relacijo med samostalniki in razmerjem med modifikatorji in glagoli, ter na podlagi te relacije klasificiram besedila v različnih žanrih po lestvici od opisnostni do povzetnostni. Analiza potrjuje večino prejšnih rezultatov, kot hkrati odkriva nekatera protislovja v relaciji med besednimi vrstami v žanru revij.

Ključne besede

BCCWJ, MVR, bagplot, žanri, dolge besedne enote

Acta Linguistica Asiatica, 1(2), 2011.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.1.2.53-62

1. Introduction

The modifier-verb ratio, commonly abbreviated as MVR, was proposed in 1965 by Kabashima and Jugaku as part of a study on Japanese stylistics. More recently, as more sophisticated language processing tools and larger, more varied, corpora have become available, word class ratios have begun to be reexamined by studies like Fujiike, Konishi, Ogura, Ogiso, & Koiso (2011). In this paper, we will attempt to explore the variability between genres in the Balanced Corpus of Contemporary Written Japanese, as well as outline methodological issues in word class frequency extraction.

2. Materials and Methodology

This section introduces the corpus used for the proceeding analysis, including issues with sampling, balance and representativeness. Furthermore, the status of word units in a language without clear word boundaries and issues pertaining to word class classification and aggregate word class ratios like the modifier-verb ratio are explained.

2.1 The Balanced Corpus of Contemporary Written Japanese

Developed as part of the National Institute for Japanese Language (NINJAL) priority area research project "Japanese Corpus" [1], the "Balanced Corpus of Contemporary Written Japanese" (BCCWJ)[2] is a large scale 100 million word corpus constructed during the five year period 2006-2011. Aiming to represent contemporary standard written Japanese, it consists of written texts published in the 30 year interval from 1976 to 2005, except for several special-purpose corpora, which were by necessity collected only after the project had started. One aspect contributing to the balance of the corpus is the use of several different sampling methods that aim to faithfully represent the large body of contemporary written material in Japanese, as well as support different research goals and practical applications (Maekawa, 2007). Furthermore, the corpus consists of three sub-corpora, and the three sub-corpora consist of one or more genres (Table 1). One reason for including various specific purpose sub-corpora is because they consist of written material not in ordinary circulation, but without which the corpus could not be considered balanced. For example, with the advent of the Internet, a significant amount of language exchange happens under the radar, so to speak, of traditional media.

[1] Project homepage accessible at: http://www.ninjal.ac.jp/english/products/bccwj/.

[2] The 2009 project-internal (ryōikinai) edition is used here.

Table 1: BCCWJ sub-copora, genres, sampling period and LUW token counts used in this study

Sub-corpus	Genre	Sampling period	LUW tokens
Publication	Books	2001-2005	30,090,054
	Newspapers		1,018,274
	Magazines		2,363,513
Library	Books	1986-2005	25,509,426
Specific purpose	White papers	1976-2005	3,961,882
	Bestseller books		3,896,429
	Minutes of the Diet		4,659,941
	Textbooks	2005-2007	1,175,215
	Yahoo! Chiebukuro	2005	5,294,245
	Yahoo! Burogu	2008	2,723,005
			80,691,984

The Publication sub-corpus consists of randomly extracted samples from the population of all books, magazines, and newspapers in Japan. The Library sub-corpus is randomly sampled from the population of all books cataloged at more than 13 metropolitan libraries in Tokyo. The Specific purpose sub-corpus contains an unrelated set of corpora, ranging from government white papers, government-approved standard primary and secondary education textbooks, transcribed dialog from Japan's National Diet, and bestselling books, to blog and Q&A-style message board posts from Yahoo! Burogu and Yahoo! Chiebukuro, respectively. All three sub-corpora contain the genre of books, and as we are more concerned with characterizing inter-genre differences than intra-genre ones, we treat all book corpora as one corpus, thus lowering the total genre count to 8. In order to extract word class information from the corpora we introduce the concept of word units, natural language processing tools, and associated dictionary developed for the BCCWJ project.

2.2 Word units and UniDic

The concept of word units in Japanese is not as straightforward as in many Western languages, in large part because the Japanese writing system does not employ spaces to delimit word boundaries. One of the goals of the BCCWJ project was to provide a standardized unit of language that could accommodate diverse research goals and applications (Maekawa, 2007). The basic word unit taken up by the project was the Short Unit Word (SUW), which represents a relatively short unit corresponding to one morpheme. The second unit, called the Long Unit Word (LUW), is defined both in terms of SUW's and the phrasal bunsetsu unit in Japanese (Figure 1).

Bunsetsu	今回、		この	ホテルを		使って		大型夜景鑑賞イベントを					企画した。			
LUW	今回	、	この	ホテル	を	使っ	て	大型夜景鑑賞イベント				を	企画し		た	。
SUW	今回	、	この	ホテル	を	使っ	て	大型	夜景	鑑賞	イベント	を	企画	し	た	。
Reading	Konkai	,	kono	hoteru	o	tuka	tte	oogata	yakei	kansyo	ibento	o	kikaku	si	ta	.

Figure 1: Relationship between SUW's, LUW's and bunsetsu
(Source: Yomiuri shimbun (evening edition), 2004/4/28; BCCWJ sample ID: PN4c_00026)

SUW's are most simply defined as the dictionary entries contained in the morphological parser dictionary UniDic, which is a hierarchical dictionary specifically constructed for morphological parsing of the BCCWJ corpus (UniDic, 2010; NINJAL, 2011). The top-level word classes of the word class hierarchy contained in UniDic are nouns (名詞 /meisi/), pronouns (代名詞 /daimeisi/), verbs (動詞 /dousi/), i-adjectives (形容詞 /keiyōsi/), na-adjectives (形状詞 /keizyōsi/), adverbs (副詞 /hukusi/), prefixes (接頭辞 /settōzi/), suffixes (接尾辞 /setubizi/), interjections (感動詞 /kandōsi/), particles (助詞 /josi/), auxiliary verbs (助動詞 /zyodōsi/), pre-nominals (連体詞 /rentaisi/), conjunctions (接続詞 /setuzokusi/), symbols (記号 /kigō/), punctuation (補助記号 /hozyokigō/), and space characters (空白 /kūhaku/) (UniDic, 2010). As LUW's are constructed from SUW's, the word class hierarchy is identical for the most part. One important difference, especially for this study, is that ambiguous SUW's that, depending on context, can be either a noun or adverb, or a noun or verb, are disambiguated in the process of becoming LUW's. This enables the computation of more accurate modifier-verb ratios that are also more comparable to the ones calculated by Kabashima and Jugaku (1965).

2.3 Modifier-verb ratio

According to Kabashima and Jugaku, texts can be classified on a scale ranging from summative (要約的 /yōyakuteki/) to descriptive (描写的 /byōsyateki/) (1965). Summative texts convey only the bare minimum – the skeleton or framework of what they are describing. In contrast, descriptive texts specify in detail what they are describing, making the reader feel as if he is part of the situation described. For descriptive texts, Kabashima and Jugaku (1965) further characterize them as active (動き描写的 /ugokibyōsyateki/) or static (ありさま描写的 /arisamabyōsyateki/).

The modifier-verb ratio was first introduced in Kabashima and Jugaku (1965) as a quantitative method of classifying texts into these categories using only the ratio of modifier to verb counts. Kabashima and Jugaku define modifiers as consisting of adjectives, adverbs and pre-nominals (1965, p. 122). Having classified words into word

classes and calculated ratios based on each word class, one can then calculate the modifier-verb ratio by dividing the ratio of modifiers with the ratio of verbs in a text:

$$MVR = 100 \times \frac{modifiers}{verbs}.$$

In sum, texts with a high noun ratio and low modifier-verb ratio tend to be summative, while those with low noun ratios tend to be descriptive (see Figure 1). Furthermore, descriptive texts with low modifier-verb ratios tend to be active, while those with high modifier-verb ratios tend to be static.

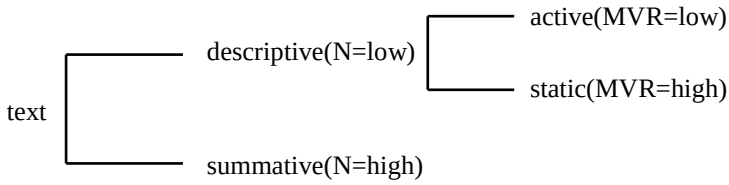


Figure 2: Categorization of texts using noun and modifier-verb ratios (adapted from Kabashima p. 25; N and MVR information added by author)

2.4 Extraction of modifier-verb ratios

Modifier-verb ratios for all samples in the BCCWJ are computed as follows. First, using plain text samples from the BCCWJ, we split sentences based on a set of common sentence delimiters (from the set of half- and full-width characters ". !?. 。 ！？"), except for when such a delimiter is encountered inside a quotation. We then morphologically analyze the sentences into SUW's using the morphological parser MeCab version 0.98 and UniDic version 2.1.0 (Kudo, 2010; UniDic, 2010; NINJAL, 2011). In the next step, the SUW data is fed into the LUW analyzer Comainu, version 0.53a, which produces morphologically parsed Japanese text with LUW's as the base unit. Finally, following Kabashima and Jugaku (1965), all word classes are coded and counted into the five classes of modifiers (M), nouns (N), verbs (V), interjections (I), and other (O). Using these frequency counts, it is straightforward to calculate the modifier-verb ratio as MVR = 100 x M/V. Special care has to be taken for samples that contain no verbs, and for these we do not calculate the modifier-verb ratio, but treat them as outliers (572 and 492 samples from Yahoo! Chiebukuro and Yahoo! Burogu, respectively).

Kabashima and Jugaku (1965)'s methodology differs slightly from the one used here, for they took random sentences as the sampling method from each book, giving them the average word class ratio of each book. In contrast, for at least the library and

publication sub-corpora in the BCCWJ, the sampling was done on a whole body of material, with samples taken in fixed- and variable-length chunks from random books and other media, and thus while possibly misrepresenting the sampled work, the samples, when taken together, offer a representative sample of the body sampled (Maruyama et al., 2010).

3. Results

Having defined modifier-verb ratios and explained their extraction procedure in the previous section, this section plots the relationship between noun and modifier-verb ratios using bagplots. In addition to visual information, several summary statistics based on bagplots are also provided and used to quantify genre based on their distributional properties.

3.1 Bagplots of noun and modifier-verb ratios

Compared to Fujiike et al. (2011), whose study was based on smaller sample sizes and used a scatterplot to visualize noun and modifier-verb ratios, we plot the relationship between noun and modifier-verb ratios for each genre using a bivariate visualization method called a bagplot. The bagplot, first proposed by Rousseeuw et al. (1999), is a 2D generalization of a boxplot used to analyze the relationship between two variables. According to Rousseeuw et al. (1999), "the bagplot visualizes the location, spread, correlation, skewness, and tails of the data." More specifically, it consists of a bag containing 50% of the data points, a fence (computed by magnifying the bag by a factor of 3) separating inliers from outliers, and a loop containing points outside the bag but inside the fence. In addition, the central point for each bagplot is defined as the point with the highest halfspace depth and is denoted by a star-shaped point, while any point outside the fence is considered an outlier.

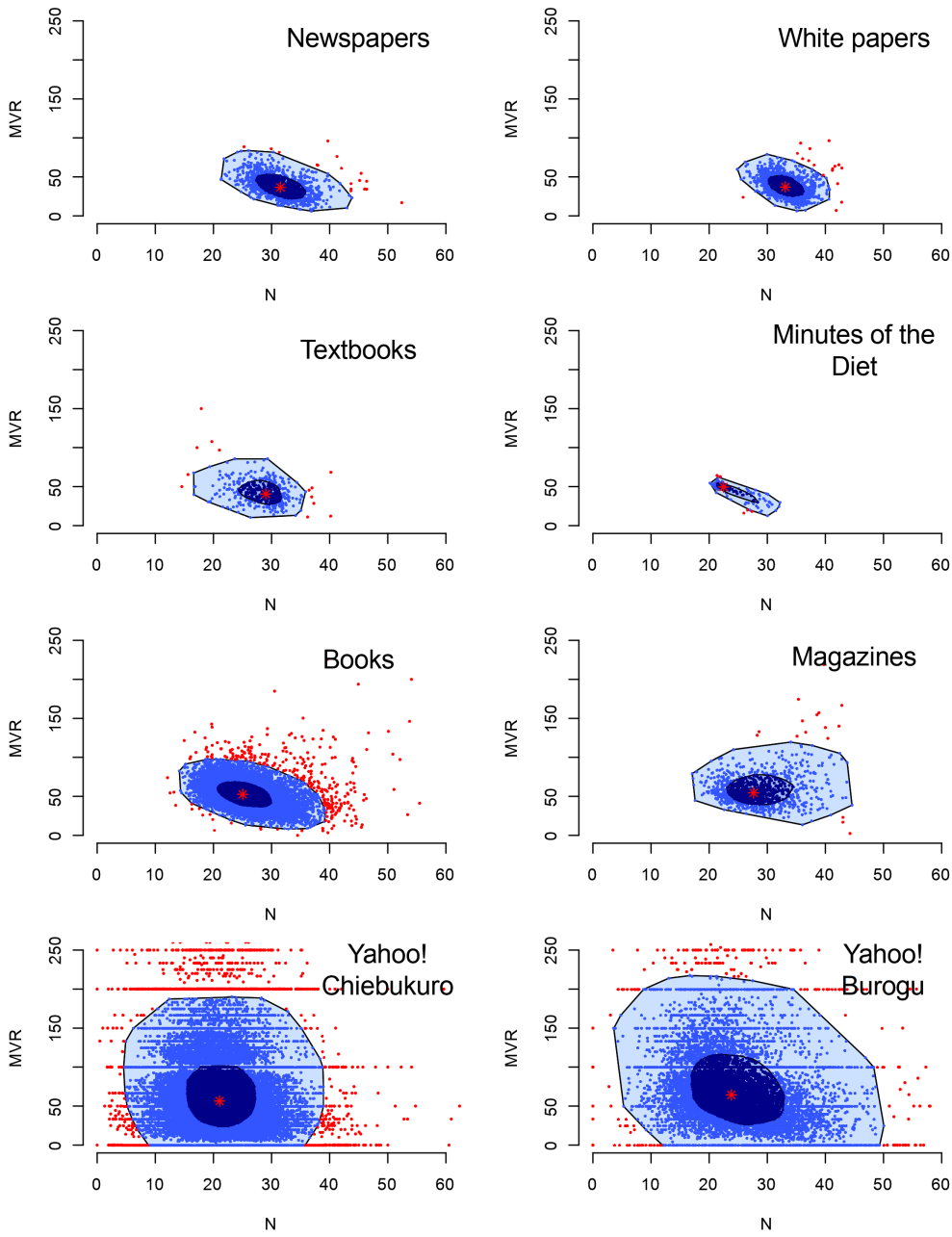


Figure 3: Bagplots of noun ratio (N) to modifier-verb ratio (MVR) for each genre.

Intuitively, the location of each genre is the halfspace median, here denoted by the central star-shaped point; the spread is represented by the size of the bag; the correlation between N and MVR by the orientation of the bag; skewness by the shape of the bag and loop; tail by points near the fence and beyond.

This study uses the ratios computed with the method outlined in subsection 2.4 together with the aplpack package for the statistical programming environment R to plot the bagplots (Wolf and Bielefeld, 2010; R Development Core Team, 2010). Figure 3 shows bagplots plotting the relationship between noun and modifier-verb ratios for each genre, here limited to noun ratios of 0-60% and modifier-verb ratios of 0-200 for clarity.

Table 2: Summary of bagplot statistics for each genre, sorted by MVR median

Genre	Bag (%)	Fence (%)	Outliers (%)	N median	MVR median
Newspapers	51,17	47,45	1,37	31,50	36,35
White papers	49,47	49,40	1,13	33,07	36,93
Textbooks	49,69	47,62	2,69	29,00	40,82
Minutes of the Diet	49,06	48,43	2,52	22,46	49,86
Books	52,36	46,22	1,41	25,00	52,86
Magazines	50,33	48,14	1,53	27,63	54,96
Yahoo! Chiebukuro	53,91	41,35	3,66	21,01	56,82
Yahoo! Burogu	48,51	43,28	3,71	23,76	64,77

Furthermore, in order to enable easier comparison and quantitative assessment of each genre, we also tabulate the halfspace median of N and MVR, the percent of samples inside the bag, the percent of samples outside the bag but inside the fence, as well as the percent of outliers for each genre (Table 2).

4. Discussion

In general, the negative correlation of noun ratios with modifier-verb ratios observed by Kabashima and Jugaku (1965) was re-confirmed for the BCCWJ in general, as can be seen from the general bottom-down facing orientation of the bags. The genre of Magazines was the only exception to this tendency, and merits further investigation. Interestingly, Magazines have both a relatively high noun ratio as well as a high modifier-verb ratio, a combination not adequately treated in Kabashima and Jugaku (1965).

Although both Internet corpora, Yahoo! Burogu and Yahoo! Chiebukuro, showed the biggest bag and fold areas, as well as the highest outlier percentages, Yahoo! Chiebukuro, in particular, has the highest concentration of samples in the bag and least between the fence and bag, suggesting a different distribution of word classes than other genres.

Not surprisingly, White papers, Newspapers, and Textbooks have the highest noun ratios, as well as the lowest modifier-verb ratios, classifying them as summative texts, while Books has an average noun ratio, but higher than average modifier-verb ratio.

5. Conclusion

This paper has hoped to shed some light on the ways in which the modifier-verb ratio can be applied to the study of genres. Issues in the extraction of word class ratios pertaining to word units and the morphological parser dictionary UniDic were touched on. Finally, using the bagplot to visualize the distributions of word classes inside genres and between genres was attempted, but revealed that further study was necessary to uncover the causes of variation and deviations from previous research in connection with the positive correlation observed for Magazines.

6. Future Work

The advent of the BCCWJ, its inclusion of various sampling targets and genres, in particular, lowers the barriers for the comparative study of genres in Japanese, as well as decreases the likelihood of overextending generalizations from one narrow genre, such as newspaper Japanese, onto the whole of Japanese. Modifier-verb ratios are a relatively simple index of variety observable in and between genres, and more needs to be done to extend Kabashima and Jugaku's analysis to new written genres. Additionally, comparisons with other measures like lexical density, as for example in Halliday (2009, p. 75-77), or the multidimensional feature approach outlined in Biber and Conrad (2009) should be attempted.

Another issue not adequately addressed here is the role of topic in genre studies. As a word class based measure, the modifier-verb ratio should be relatively robust to topic changes in an otherwise situationaly homogeneous genre. However, this should be qualified by, for example, using the Nippon Decimal Classification (NDC) library classification system supplied for the book corpora, as well as the topical information available for the Internet corpora to further understanding of intra-genre variation phenomena.

Finally, though the BCCWJ contains material sampled from a relatively short timespan of up to 30 years, we cannot be sure that any inference we make from the data is not attributable to diachronic differences.

References

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge Textbooks in Linguistics.

Fujiike, Y., Konishi, H., Ogura, H., Ogiso, T., & Koiso, H. (2011, Mar.). Tyōtan'i ni motozuku 'gendai nihongo kakikotoba kinkō kōpasu' no hinsihiritu ni kansuru bunseki. In *Proceedings of the 17th Annual Meeting of The Association for Natural Language Processing* (Vol. 17, pp. 663-666). Toyohashi, Japan.

Halliday, M. (2009). Methods - techniques - problems. In M. Halliday & J. J. Webster (Eds.), *Continuum Companion to Systemic Functional Linguistics* (pp. 59-86). Continuum International Publishing Group.

Kabashima, T., & Jugaku, A. (1965). *Buntai no kagaku* [Stylistics]. Sogeisha.

Kudo, T. (2011, July 20). MeCab: yet another part-of-speech and morphological analyzer. Retrieved from http://mecab.sourceforge.net/

Maekawa, K. (2007). Kotonoha and BCCWJ: development of a balanced corpus of contemporary written Japanese. In *Proceedings of the First International Conference on Korean Language, Literature, and Culture* (Vol. 2, pp. 158 −177). Corpora and Language Research. Seoul.

Maruyama, T., Yamazaki, M., Kashino, W., Sano, M., Akimoto, M., Inamasu, S., & Oyauchi, Y. (2010). Outline of sampling method in the balanced corpus of contemporary written japanese (4): Corpus design and the result of sampling. In *Tokutei ryōiki kenkyū 'nihongo kōpasu' heisei 21 nendo kōkai waakusyoppu (kenkyuseikahōkokukai) yokōsyū* (pp. 37-46).

NINJAL [National Institute for Japanese Language and Literature]. (2011). Tokuteiryōiki kenkyū nihongo kōpasu kenkyū seika hōkoku [Priority-Area Research "Japanese Corpus": Research Report] [DVD media containing UniDic and Comainu]. Tokyo: General Headquarters, Priority-Area Research "Japanese Corpus".

R Development Core Team. (2011, July 20). *R: a language and environment for statistical computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from R Foundation for Statistical Computing: http://www.R-project.org

Rousseeuw, P. J., Ruts, I., & Tukey, J. W. (1999). The bagplot: a bivariate boxplot. *The American Statistician*, 53(4), 382-387. doi:10.2307/2686061

UniDic. (2011, July 20). Keitaiso kaiseki zisyo UniDic. Retrieved from http://www.tokuteicorpus.jp/dist/

Wolf, P., & Bielefeld, U. (2011, July 20). *Aplpack: another plot package: stem.leaf, bagplot, faces, spin3r, and some slider functions*. R package version 1.2.3. Retrieved from http://CRAN.R-project.org/package=aplpack

JAPANESE WORD SKETCHES: ADVANTAGES AND PROBLEMS

Irena SRDANOVIĆ
University of Ljubljana, SI
irena.srdanovic@gmail.com

Naomi IDA
Meiji University, JP
idanaomi2002@yahoo.co.jp

Chikako SHIGEMORI BUČAR
University of Ljubljana, SI
chikako.bucar@guest.arnes.si

Adam KILGARRIFF
Lexical Computing Ltd., UK
adam@lexmasterclass.com

Vojtěch KOVÁŘ
Masaryk University, CZ
xkovar3@fi.muni.cz

Abstract

In this paper, we present results of an evaluation of Japanese word sketches and address in detail issues that were observed by the evaluators. A word sketch presents a list of salient collocates of a word, organized by the grammatical relations holding between the word and its collocate. The word sketch functionality is incorporated into the Sketch Engine corpus query system and has been created for more than twenty languages so far, including Japanese. The issues that have been discovered in the evaluation of word sketches in Japanese are to be addressed for further enhancement of the word sketch functionality. Other tools and resources which are combined for use and influence the performance of the word sketches should also be looked over. We divide the issues into the following: 1) the lemmatizer and tagger in use, 2) the sketch grammar that is specifically written for Japanese, and 3) the corpus and statistical methods.

Keywords

word sketches, Japanese collocations, evaluation, corpus, language technologies

Izvleček

V prispevku predstavljamo rezultate ocenjevanja japonskih besednih skic in podrobno prikazujemo probleme in težave, ki smo jih opazili ocenjevalci. Besedna skica je seznam izstopajočih kolokacij neke besede, ki ga organizirajo slovnične relacije med besedo in drugimi besedami, ki skupaj sestavaljajo kolokacije. Funkcije besedne skice so vgrajene v korpusno orodje Sketch Engine in na voljo trenutno že v več kot dvajsetih jezikih, med njimi tudi v japonščini. Problemi in težave, ki smo jih odkrili med ocenjevanjem besednih skic v japonščini, moramo dalje proučiti za okrepitev funkcij besednih skic. Problemi in težave so pri naslednjih: 1) pri sistemu ugotavljanja osnovne oblike besede in označevalcu besednih vrst v rabi; 2) v slovnici za skice, ki je napisana posebej za japonščino; 3) pri korpusu in statističnih metodah.

Ključne besede

besedne skice, kolokacije v japonščini, evalvacija/ocenjevanje, korpus, jezikovne tehnologije

Acta Linguistica Asiatica, 1(2), 2011.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.1.2.63-82

1. Introduction

The word sketches automatically summarize a list of salient collocates of a word, organised by the grammatical relations holding between the word and its collocate. By *collocate*, we refer to the word that joins with the headword to form a collocation. For any headword, a list of its collocates is a list of the words that it combines with to give its collocations (Kilgarriff et al 2010). The word sketches were first prepared for the English language and used for the compilation of the Macmillan English Dictionary for Advanced Learners (Rundell 2002). Later on they were integrated into the Sketch Engine corpus query tool (Kilgarriff et al 2004), created for numerous languages, and used on a large scale for lexicography by a number of publishers. The word sketches for Japanese were first prepared in 2008, employing 400 million-word web corpus that is tokenised and POS-tagged using the ChaSen toolset[1] and English translation of POS tags. The word skech grammar written for Japanese covers more then 50 collocational and grammatical relations for the Japanese nouns, verbs, adjectives and adverbs (Srdanović et al 2008a).

The first formal quantitative evaluation of word sketches is performed for four languages, Dutch, English, Japanese and Slovene, as part of the Sketch-Eval mini-project. The evaluation is undertaken from a user perspective, with the main question being "is the collocation suitable for inclusion in a published collocation dictionary". The background of the evaluation method and the results for all four languages is described in Kilgarriff et al (2010). In this paper, we concentrate on results of the Japanese word sketches evaluation and discuss the issues discovered by the evaluators.

2. Japanese word sketches

The creation of Japanese word sketches required preparation of the following components:[2]

- A corpus.
  At the time of creation of the Japanese word sketches, there was no publicly available corpus that could be used inside the SkE tool. Therefore, a large-scale Japanese language web corpus, named JpWaC, was created. Its size is 7.3GB or 400 million words.
- Language processing tools used for processing the corpus data: tokeniser, lemmatiser and part-of-speech (POS) tagger.
  The morphological analyzer and part-of-speed tagger ChaSen was used for processing the JpWac data. The ChaSen tagset is quite detailed, with 88 tags,

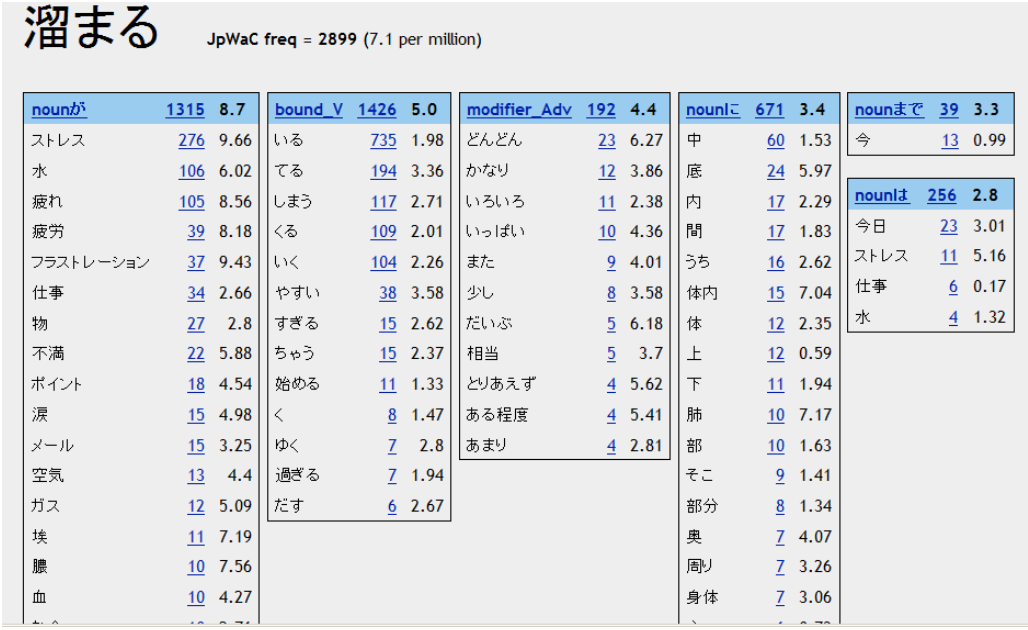
[1] http://chasen.naist.jp

[2] For details on the creation of the Japanese web corpus JpWaC and the Japanese word sketches, refer to Srdanović et al. (2008a).

and uses a fairly “narrow”, or fine-grained, tokenization: it splits inflectional morphemes from their stems. It uses the IPADIC dictionary.

- A sketch grammar[3] with a specified POS tagset for the language.
  The Japanese sketch grammar uses English translations of the ChaSen POS tagset. The Japanese sketch grammar defines 22 grammatical patterns covering more than 50 collocational relations for nouns, adjectives, verbs and adverbs.

The corpus was prepared in Sketch Engine format and installed in the system. The sketch grammar is also loaded into the system. Using the components described above, the system automatically selects frequent and statistically salient collocations. The statistics are based on the Dice coefficient.[4]



溜まる JpWaC freq = 2899 (7.1 per million)

nounが	1315	8.7
ストレス	276	9.66
水	106	6.02
疲れ	105	8.56
疲労	39	8.18
フラストレーション	37	9.43
仕事	34	2.66
物	27	2.8
不満	22	5.88
ポイント	18	4.54
涙	15	4.98
メール	15	3.25
空気	13	4.4
ガス	12	5.09
埃	11	7.19
膿	10	7.56
血	10	4.27

bound_V	1426	5.0
いる	735	1.98
てる	194	3.36
しまう	117	2.71
くる	109	2.01
いく	104	2.26
やすい	38	3.58
すぎる	15	2.62
ちゃう	15	2.37
始める	11	1.33
く	8	1.47
ゆく	7	2.8
過ぎる	7	1.94
だす	6	2.67

modifier_Adv	192	4.4
どんどん	23	6.27
かなり	12	3.86
いろいろ	11	2.38
いっぱい	10	4.36
また	9	4.01
少し	8	3.58
だいぶ	5	6.18
相当	5	3.7
とりあえず	4	5.62
ある程度	4	5.41
あまり	4	2.81

nounに	671	3.4
中	60	1.53
底	24	5.97
内	17	2.29
間	17	1.83
うち	16	2.62
体内	15	7.04
体	12	2.35
上	12	0.59
下	11	1.94
肺	10	7.17
部	10	1.63
そこ	9	1.41
部分	8	1.34
奥	7	4.07
周り	7	3.26
身体	7	3.06

nounまで	39	3.3
今	13	0.99

nounは	256	2.8
今日	23	3.01
ストレス	11	5.16
仕事	6	0.17
水	4	1.32

Figure 1 Japanese word sketch example for the verb tamaru (溜まる “to accumulate [intr.]”), partial results

[3] The sketch grammar is a mini-grammar of syntactic patterns. It is based on regular expressions over POS tags and enables the system to automatically identify possible collocations. See *Corpus Querying and Grammar Writing* on the Sketch Engine website, http://www.sketchengine.co.uk

[4] Refer to *Statistics used in the Sketch Engine* on the Sketch Engine website, http://www.sketchengine.co.uk.

Table 1 Types of collocational relations for verbs in the Japanese word sketches (14 different types of relations)

POS	Grammar sketch pattern	Type of relation	Example	Example transcription
Verb (14)	modifier_Adv	Adv modifying V	にこにこ笑う	*nikoniko warau*
	noun は	noun_*wa*＋V	彼は笑う	*kare wa warau*
	noun が	noun_*ga*＋V	鬼が笑う	*oni ga warau*
	bound_V	bound verbs connecting to free verbs	わらっちゃう	*warattyau*
	V_bound	free verbs connected to bound verbs	連れて行く	*turete iku*
	noun で	noun_*de*＋V	鼻で笑う	*hana de warau*
	noun に	noun_*ni*＋V	最後に笑う	*saigo ni warau*
	noun から	noun_*kara*＋V	（心の）底から笑う	*(kokoro no) soko kara warau*
	noun まで	noun_*made*＋V	最後まで笑う	*saigo made warau*
	noun を	noun_*wo*＋V	（人の）失敗を笑う	*(hito no) sippai wo warau*
	noun へ	noun_*he*＋V	公園へ行く	*kooen e iku*
	Coord	coordinate relation	笑う・泣く	*warau - naku*
	Suffix	V+*suffix*	笑いっぱなし	*waraippanasi*
	Prefix	*prefix*＋V	超笑う	*tyoo warau*

Figure 1 gives a word sketch for the Japanese verb *tamaru* (溜まる“to accumulate [intr.]”). Different grammatical relations, such as noun-particle-verb collocates (for example, noun が for noun+*ga*+verb collocates), bound verbs that appear with the verb, adverbs modifying the verb etc., are displayed in order of their significance, revealing the most frequent and most salient sets of collocations (see the first and second columns with numbers of frequency and saliency respectively).

Table 1 shows what types of collocational relations are covered in the Japanese word sketches for words classified as verbs (all together fourteen). An example for each of the relations is given.

3. Evaluation: method and results

Since the creation of the Japanese word sketches we have undertaken various kinds of evaluation. Srdanović et al (2008a) describes a comparison of a newspaper corpus and the JpWac corpus, showing that newspaper data are more specific both in

terms of form (being written mainly in the past tense and not using the formal predicate form *masu/desu*) as well as content, with a high proportion of news-specific and politically-oriented nouns. On the other hand, JpWaC contains more informal and interactional material, and more diverse content. Later on, the study by Srdanović et al (2008b) explored the appearance of adverb-final modality forms in various corpora, confirming that the newspaper corpus, as well as some other corpora, is more specific in nature than the web corpus. This study shows that the web corpus is the most similar to the Balanced Corpus of Contemporary Written Japanese (Maekawa et al 2010). Srdanović & Nishina (2008) evaluate the functionality by comparing its results to the first and only collocation dictionary for Japanese language students (Himeno 2004). The comparison of randomly selected items in the dictionary with the word sketch for the same word clearly shows the much wider spectrum of collocational and grammatical relations as well as a richer variety of collocations in the sketches. We see great potential for using word sketches for future dictionary compilations

In this section we present the evaluation methods and results for Japanese word sketches in the Sketch-Eval project (Kilgarriff et al. 2010).

3.1 Type of evaluation

Sketch-Eval is a quantitative type of evaluation, undertaken from a user perspective. It measures precision, which is the percentage of the answers given that are correct.[5] It is measured by examining the word sketch responses with the critical question being "is the collocation suitable for inclusion in a published collocation dictionary". Here, the Oxford Collocations Dictionary (OCD 2009) is proposed as a model and a reference point for what we wish to produce automatically. A number of human experts evaluated a sample of dictionary entries and a set of their collocates in the word sketch. For each language, forty-two headwords were sampled and twenty most salient collocates[6] for each of the headwords were inspected. Four languages were included in the Skech-Eval, among which was also Japanese.

[5] The information sciences distinguish between evaluating precision and recall. Kilgarriff et al (2010) describe it as follows: "Precision is the percentage of the answers given that are correct. Recall is the percentage of all correct answers that are found. If my word sketch for *flour* contains only *sift* and *sieve*, it has 100% precision, since all the given collocates are correct, but low recall, since there are many other collocates it does not give. As a response gets bigger, precision usually falls off (since some incorrect answers creep in) but recall improves (as more of the correct answers are included). Changing the size of the answer is a matter of adjusting the 'precision/recall tradeoff."

[6] This was subject to the constraint that no more than two thirds relate to any single grammatical relation, to provide variety of collocational relations. Later on it was realized that it might have been better to have lower number of collocates for medium and low-frequency words. (Kilgarriff et al. 2010)

Table 2 Randomly extracted sample list and reserve list of words for evaluation of the Japanese word sketches

	Sample list			Reserve list		
	Nouns	Verbs	Adjectives	Nouns	Verbs	Adjectives
Common (top 2999)	急 研究 完成 男性 緑 評価	生まれる 扱う 支払う 忘れる	よろしい っぽい 素晴らしい 大きい	心配 箱 積極 プロセス 地区 建設	ちゃう 知れる 語る 受け入れる	重い 長い 忙しい 無い
Mid (3000-9999)	欠席 蓄積 マスター 俳句 情勢 有力	まつ づく 資する 溜まる	黒い おとなしい こい 親しい	フォント 刑事 澄 蝶 包装 メス	溢れる 拒む 隠る しむ	柔らかい むずかしい きつい とんでもない
Low (10,000-30,000)	クレイ 方角 近鉄 走り 苑 人妻	駆け込む やせる 書き留める 滅ぶ	むつかしい 仲良い くすい 腹立たしい	水槽 グローバリゼーション 青島 懇話 射精 モグラ	対する 振舞う 吸い上げる くぼむ	若々しい 面倒くさい 耐え難い 香ばしい

We took a sample from the 30,000 commonest nouns, verbs and adjectives in the corpus, in a ration of roughly 2:1:1, and with the sample structured as in Table 2. Within these constraints, the sampling was random. Table 2 shows our automatically extracted random sample (and reserve) list of Japanese words for the evaluation of the Japanese word sketches. The reserve list is also available to provide a replacement for a misanalyzed word in the sample list. In case of Japanese, the reserve list is also used instead of narrowly analyzed morphemes, such as adjectival suffix っぽい, *-ppoi*, "-ish, like" or for words that may be typically written in another orthography, such as まつ, *matu* (typically written as 待つ "to wait").

For the Japanese SkE evaluation, it was specific that noun-verb-adjective proportion in the selected word list was quite different from other languages. This is because the Japanese tagset ChaSen includes under the noun tag a) nouns being part of

so called "*suru* verbs" (*suru*-V), formed as noun + verb "*suru*",[7] and b) adjectives ending in *–na*, derived from nouns.[8]

3.2 Evaluation categories and evaluators

A customised version of the Sketch Engine was prepared in which word sketches contained only the twenty highest-scoring collocates for each word, and in which each collocate was associated with a menu with the following items:

- Good
- Good but wrong grammatical relation
- Maybe (e.g. not striking collocate)
- Maybe (specialised vocabulary)
- Bad

Three evaluators performed the Japanese SkE evaluation: two of them being native speakers of Japanese, language teachers and linguists, and the third one being a non-native speaker, language teacher, linguist and lexicographer.

A screenshot of the evaluators' word sketch interface is shown in Figure 2. Evaluators selected the relevant item from the menu and choices were stored in a database.

In order to rule out 'unclear' data, we distinguished those instances where all evaluators agree from those where they disagree, and based our results only on the agreement cases. We noted that agreement on the boolean decision, "good or not good" was substantially higher than agreement on finer-grained categories, so we merged "Good" and "Good-but" as "good" and all other categories as "bad".

[7] *Suru* is a verb which means "to do" in general and is highly productive to derive verbs of foreign origin, e.g. *kekkon* "marriage (noun)" vs. *kekkon suru"*to get married (verb)", *kopii* "copy (noun)" vs. *kopii suru* "to copy (verb)".

[8] There are two types of Japanese adjectives: adjectives that end in *-i*, e.g. *ookii* "big", and adjectives that end in *–na*, e.g. *genki-na* "healthy, vigor, good", derived from the noun *genki* "health".

Rubric: **G** = Good **Gb** = Good but wrong grammatical relation **M** = Maybe (not striking collocate)
Ms = Maybe (specialized vocab) **B** = Bad

Gramrel	Collocation	Rating					Freq
		G	Gb	M	Ms	B	
modifier_Ai	高い	◉	○	○	○	○	3388
modifier_Ai	正しい	◉	○	○	○	○	208
modifier_Ana	多元的	◉	○	○	○	○	12
modifier_Ana	定性的	◉	○	○	○	○	12
modifier_Ana	正当	◉	○	○	○	○	107
modifier_Ana	適正	◉	○	○	○	○	68
modifier_Ana	厳正	◉	○	○	○	○	10
particle	に当たって	◉	○	○	○	○	73
prefix	再	◉	○	○	○	○	938
pronomの	読者	◉	○	○	○	○	253
pronomの	一定	◉	○	○	○	○	178
suffix	損	○	○	○	◉	○	134
suffix	額	◉	○	○	○	○	591
suffix	益	○	○	○	◉	○	62

Figure 2 An example of evaluators' word sketch interface
for the word *hyooka* (評価) "evaluation, assessment"

3.3 Results

Table 3 shows the total number of collocations evaluated by all assessors, the number for which all evaluators agreed, and for these, the number that were good and the number that were bad (where "bad" includes "maybe"). The total number of collocations evaluated by all assessors is slightly less than the maximum possible of 20 collocates for each of 42 headwords, owing to a different evaluator's choice on replacement of a sample word with a word from the reserve list, and from a range of minor omissions. For Japanese there was three-way agreement among lexicographers for less than half of the data, so we also give figures for two-out-of-three agreement, which provides more reliable evaluation results. Two-out-of-three agreement refers to the number of cases when any of two out of three evaluators agreed.

Table 3 Japanese word sketch evaluation results

Agreement	Total colls assessed	Evaluators all agreed on	Good	Bad	% good
Three-way agreement	747	294	278	16	94.56%
Two-out-of-three agreement		690	600	90	86.95%

In total 747 collocations were judged by all evaluators. All three evaluators agreed on 294 instances, and any of two out of three evaluators agreed on 690 instances. Of those, all three evaluators agreed that 278 collocates were "good" and 16 "bad", which results in 94.56% of collocates being good. Calculating the results that any of two out of three evaluators agreed, shows that 600 collocates were evaluated as "good" and 90 collocates as "bad", which gives result of 86.95% of collocates being good.

On examining possible reasons for the high proportion of disagreement between evaluators, we discovered that one out of three evaluators had a noticeably different approach towards collocations that are basically good, but not complete as a semantic or syntactic unit. If an evaluator uses concordance examples to check instances of a collocation candidate in the corpus, they will see the full collocation, which appeared as incomplete in the word sketch, and from that point it can be regarded as good and useful for a dictionary editor. In the evaluation process, for this type of collocations, a new selecting choice such as "Good but not complete" would be preferred. On the other hand, "Good but wrong grammatical relation" could be excluded as a selection choice for Japanese word sketches. Other possible reasons for disagreements in evaluation are related to treatment of collocations with technical terms in specialized fields, polysemic keywords, orthography issues etc. Some of the issues with examples are presented and discussed in detail in the following section.

Comparing Japanese word sketch evaluation results to other three languages, Slovene (71.1% good), English (70.7% good), Dutch (66.3% good) (Kilgarriff et al 2010), we can say that Japanese word sketches showed the best performance. For the other three languages too, two thirds or more of the collocations on which the assessors agreed were of publishable quality, which confims the word sketches being a valuable resource for lexicographic work in any of four languages.

4. Discussion

The quality of the word sketches functionality depends on the quality of its components: the corpus, POS-tagger, sketch grammar etc, which are described in section 2. Although the evaluation results for Japanese word sketches show a high precentage of good collocations, there are some issues that were discovered during the evaluation. Japanese word sketches have problems especially with some POS-tagger

and morphological analyzer issues, some overlooked grammar rules/relations, corpus junk and problems on the level of orthography. This section addresses these issues in detail.

4.1 Issues with the tagger and morphological analysis

ChaSen's analysis and tagset are very fine-grained which causes the following common problems: the excessive analysis of *suru* verbs (*suru*-V) and adjective on *–na* (*na*-Adj) (4.1.1), excessive analysis of derived and compound nouns (4.1.2), and problems related to the Japanese writing system (4.1.3).

4.1.1 Excessive analysis of *suru* verbs (*suru*-V) and adjective on *–na* (*na*-Adj)

The Japanese tagset ChaSen analyses *suru*-V into its most basic form that is tagged as noun, N.Vs [名詞−サ変接続] and the verb form *suru*, which is tagged as V.free [動詞−自] and further on analyzes into various segments depending on its inflectional forms. Similarly, the tagset analyses *na*-Adj into the basic noun form, N.Ana [名詞−形容動詞語幹], and the derivational suffix *-na*. The tagset does not distinguish if the tagged noun appears only as a noun or it is a part of its derived form of *suru*-V or *na*-Adj. This issue is already obvious from the automatically extracted sample list of nouns, verbs and adjectives (see Table 2), where adjectives ending in *-na* and *suru*-V were included under the noun POS category. This kind of analysis brings some consequences in the word sketch search and results. For example, it is not possible to search for word sketch results of *suru*-V or *na*-Adj independently from their basic noun category, which hides the real frequency of each of these two types of words, the noun and its *suru*-V pair, or the noun and its *na*-Adj pair. Also, the word sketch results provide lists of collocations that might be real collocates of only one of the word types. However, the issue with the real collocates is related to the current version of the sketch grammar for Japanese and could be overcome to some extent with corrections in the grammar (see also 4.2.3)

The consequences become more noticeable especially in the case of *suru*-V, when the meaning of the noun is quite different to that of the *suru*-V that is derived from the noun (1), (2).

(1)	N	マスター	*masutaa*	"teacher, leader"
	suru-V	マスターする	*masutaa-suru*	"to master"
(2)	N	左右	*sayuu*	"left and right"
	suru-V	左右する	*sayuu-suru*	"to control, affect, influence"

In the examples above, the meanings of the noun and that of the verb are very different but the word sketch search is possible only for the form マスター, *masutaa*, or 左右, *sayuu*, whether or not it is only a noun, or a derived verb form. Besides some corrections in the sketch grammar file, the issue is possible to resolve with an

additional retagging, as shown in Table 4. The table suggests that *suru*-V and *na*-Adj are tagged with one separate tag, and thus differentiated from the nouns that never appear with *suru* or *-na*.

Table 4 Current and desired tagging of *suru*-V and *na*-Adj

POS (example)	Current tagging	Desired tagging
suru-V (*eigo wo masutaa suru* "to master English")	*masutaa*[N.Vs] *suru*[V.*suru*]	*masutaa_suru*[*suru*-V]
noun, without possible *suru* form (*masutaa ga kita* "the teacher has come")	*masutaa*[N.Vs]	*masutaa*[N.Vs]
na-Adj (*genkina ko* "healthy child")	*genki*[N.Adj]+*na*	*genkina*[*na*-Adj]
noun, without possible *-na* form (*genki ga nai* "not to be well")	*genki*[N.Adj]	*genki*[N.Adj]

4.1.2 Excessive analysis of derived and compound nouns

ChaSen divides derived and compound nouns into morphemes and treat them as different segments, separating their non-standalone prefixes and suffixes too. This kind of narrow analysis causes incomplete and misleading collocational results, as shown in the following examples.

(3) 優秀な 研究 | 者[9]
yuusyuu-na *kenkyuu* | *sya*
excellent research | -er
"excellent researcher"

The word sketch shows the noun *kenkyuu* "research" collocating with the noun *yuusyuu* "excellent"[10]. This is incomplete collocational information: it is the noun *kenkyuu-sya* "researcher", derived from the noun *kenkyuu* "research", that collocates with *yuusyuu-na.*

(4) 財界 の 有力 | 者
zaikai *no* *yuuryoku* | *sya*
financial circles possessive particle influential | person
"an influential person in financial circles"

[9] Inappropriate separation of derived or compound noun is denoted by vertical bar (|).

[10] The *na*-Adj *yuusyuu-na* is tagged as N, in the form of *yuusyuu*, as we mentioned above.

Similarly, the word sketch shows the word *yuuryoku* “influential”[11] collocating with the noun *zaikai* “financial circles”. In reality, it is *yuuryoku-sya* “influential person” that collocates with *zaikai*.

(5) スポーツ | 用品　を　扱う
supootu | *yoohin*　*wo*　*atukau*
sport | goods　object particle　deal with
“to deal with sport goods”

The POS tagger divides compound nouns into segments, as shown in the example (5), which can be regarded as semantically incomplete from the point of view of collocational relations. In word sketches, only parts of compound nouns are shown as collocates. For example, the word sketch for *atukau* “to deal with” shows that the verb collocates with the noun *yoohin* “goods”, which can be regarded as incomplete. The complete collocation is *supootu yoohin* “sport goods”.

Similarly, there are examples of suffixes that are tagged as nouns (歳 *sai*, 代 *dai* etc.) and appear as separate words in the word sketches:

(6) 50 | 歳　の　男性
50 | *sai*　*no*　*dansei*
fifty | age　possessive particle　man
“a fifty-year-old man”

(7) 50 | 代　の　男性
50 | *dai*　*no*　*dansei*
fifty | generation　possessive particle　man
“a man in his fifties”

(8) 使用 | 料　を　支払う
siyoo | *ryoo*　*wo*　*siharau*
use | fee　object particle　pay
“to pay rental fee”

-sai in (6), *-dai* in (7) and *-ryoo* in (8) are the nouns with suffix role and inevitably require another noun, respectively 50 and *siyoo* “use, rental” to be complete in their usage.

Although the narrow analysis causes some obvious problem in finding appropriate collocations, it also offers the possibility of exploring behavior of suffixes and prefixes in detail. This type of information has also been a subject of interest for dictionary makers, language learners and language specialists, for example see Vance (1991).

[11] The *na*-Adj *yuuryoku-na* is tagged as N, in the form of *yuuryoku*.

What is encouraging from the lexicographers' point of view is that the complete and correct collocational relations for the majority of the above examples can be easily found from the combination of word sketch and corpus examples, searchable from the word sketch interface. However, when someone looks only at the list of collocates, the results are misleading, which gives rise to instability in the evaluators' judgments. This matter should be tested in detail among different POS taggers and dictionaries for further enhancement of the functionality.

4.1.3 Orthography issues: words that can be written in *hiragana*, *katakana* or *kanji*

This issue is peculiar to the Japanese language. The Japanese writing system uses a combination of three sets of letters: Chinese characters (*kanji*) and Japanese syllabic alphabets (*kana*: *hiragana* and *katakana*). There is a fair amount of fluctuation and overlap in the use of characters and *kana* (Seeley 1991). This problem is not yet very well addressed in the current natural processing tools, including ChaSen, and this is reflected in the word sketches. On the one hand, since there are two or three different orthographies for the same word, such as *subarasii* (9), the information about collocates is dispersed among different orthography variations, and therefore some collocates are missed. On the other hand, when there are two or more different words that are identical when written in *kana*, such as *matu* (10), results for the three words are mixed together.

(9) *i*-Adj: *subarasii* "wonderful, splended" 素晴らしい, すばらしい, スバラシイ

(10) V or N: *matu* まつ "to wait/pine/end", 待つ "to wait", 松 "pine", 末 "end"

Since the creation of the Japanese word sketches, some research progress has been made with the development of UniDic, the new dictionary for morphological analysis. The dictionary can be used with ChaSen or MeCab. The research on its accuracy reveals slightly better results in its combination with MeCab than with ChaSen.[12] The dictionary also provides some improvements in dealing with the Japanese orthography and pronunciation issues by providing canonical forms, word forms, writing variants, speech variants and accent.[13]

We plan to use the new set of tools for the new version of Japanese word sketches, which is expected to improve the performance of the functionality. This will include also the usage of canonical orthographic forms for each Japanese lemma, which means that each lemma will have only one, its typical, orthographic form.

[12] http://www.tokuteicorpus.jp/dist/

[13] The UniDic contained 150,000 words (canonical forms) (July 2009). (http://www.tokuteicorpus.jp/dist/) .

4.1.4 Other tagger issues

There are some other tagger issues that were revealed during the evaluation, such as errors in morphological analysis or a different inflectional form of an actual collocate than of its lemma form displayed in the word sketch

For example, the tagger analyses the proper noun 京急蒲田 *Keikyuu kamata*, a train station name, into three different elements, and the word sketch wrongly presents 急 *kyuu* and 蒲田 *kamata* as collocations

Other examples of wrong morphological analysis are in case of sayings, idioms, such as 急がば回れ *isogaba maware* “slow and steady wins the race”. The morpheme 急 *kyuu* “sudden, quick” is again regarded as a separate lemma and as a collocate of the potential form 回れる *mawareru* “to be able to go around”

As for different inflectional forms, we find cases such as (11), where the lemma of the collocate is a dictionary form, while the actual collocate is only in negation. Another example (12) is a variant of the adjective *yorosii*, an honorific variant of the word *yoi/ii* “good, nice”, which is present in a highly frequent set phrase.

(11)	忘れる	いける	→	忘れ（ては）いけない，（て）いけない
	wasureru	*ikeru*	→	*wasure(tewa) ikenai, (te) ikenai*
	Forget	able to go	→	“You should not forget”
(12)	よろしい	お願い	→	よろしくお願いします
	Yorosii	*onegai*	→	*yorosiku onegai simasu*
	Good	request	→	“I would appreciate your favor”

4.2 Issues with the word sketch grammar

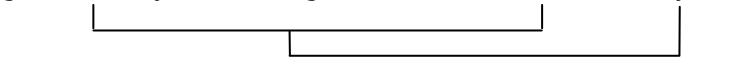
The evaluation of word sketch results revealed some issues in the word sketch grammar that are general for sketch grammar syntax and in use for all the languages (4.2.1), and some issues that could be improved by additions or changes in the grammar rules for Japanese (4.2.2, 4.2.3 and 4.2.4).

4.2.1 The reach of collocation: more than two collocates

In this section we list examples of collocates for which an additional collocational element in the phrase is lacking. This issue is related to the general limitation of sketch grammar syntax for identifying all parts of collocations of three or more words.

(13)	グローバリゼーション	が	もたらす
	guroobarizeesyon	*ga*	*motarasu*
	globalization	subject particle	bring
	“the globalization brings”		

(13') グローバリゼーション が もたらす 影響
guroobarizeesyon *ga* *motarasu* *eikyoo*



Globalization subject particle bring influence
"the influence that the globalization brings"

The word sketch shows the noun *guroobarizeesyon* "globalization" collocating with the verb *motarasu* "to bring" (13). However, *guroobarizeesyon ga motarasu* "globalization brings", without an object, is not a complete phrase. The noun *eikyoo* is the head of this noun-modifying clause and is related to two words, the noun *guroobarizeesyon* and the verb *motarasu*. In the web corpus, we find many occurrences of the example *guroobarizeesyon ga motarasu eikyoo* "the influence that the globalization brings", which suggests that the collocational relation is fully established only when all of three elements are present

Another similar example is a collocational relation between nouns without the head noun.

(14) グローバリゼーション の 負 （の 側面）
guroobarizeesyon *no* *hu* *(no* *sokumen)*
globalization poss. particle negative (poss. particle side)
"the negative side of globalization"

According to word sketch results, *guroobarizeesyon* and *hu* are collocates (14). However, this can be regarded as incomplete and with no semantic and syntactic relation established, without the pivot noun *sokumen*.[14]

These kinds of examples cannot be easily judged as correct or bad collocates. Depending on an evaluator's judgment, the semantic or syntactic relation of such collocates can be brought to question. However the word sketch results give a good hint to lexicographers to further check the collocates in corpus examples and search for the full collocational relation. Also, a method for identifying collocations longer than two words is currently being developed, for use in future evaluations.

[14] Similar example in "A Quantitative Evaluation of Word Sketches" (Kilgarriff et al. 2010):

"if the system lists a word which only collocates with the headword within a three-or- more-word unit, as *put* is a collocate for *cat* only in the context of *out* ("put the cat out"), is the collocate good or bad? Our decision was to treat it as good, as it is enough to signal to say to a lexicographer that there is a collocation to be included in a collocation dictionary, even if the system has not found all of it. But it was not a decision that human evaluators were comfortable with."

"Multi-word items were a recurring concern, as it did not seem natural to the evaluators to mark *cat* as good at *put* when the word sketch gave no indication that *out* was also needed: this was the evaluators' most often –voiced concern."

The example above is a problem of the grammatical level, while the problem of (15) in this paper is a problem of the lexico-syntactic level.

4.2.2 Distance of collocates

(15) 日本総領事館 に 亡命 を 求めて 駆け込んだ
nihon-sooryoozikan ni boomei wo motomete kakekonda
the Japanese Consulate General to exile obj.part. want rushed
"He rushed to the Japanese Consulate General for exile"

(16) 最寄り の 交番 に 息 を 切らして 駆け込む
moyori no kooban ni iki wo kirasite kakekomu
nearest of police box to breath object particle grasp to rush
"to rush into the nearest police box grasping for breath"

The word sketch shows the verb *kakekomu* "to rush into [the past tense form *kakekonda*]" collocating with *boomei* "exile" in the example (15), *iki* "breath" in the example (16). However, they are not direct collocates since they are syntactically positioned at a distance, in distinct clauses (here an inserted clause functioning as an adverbial clause). In both cases above, another verb with a predicate function exist as a direct collocate of the nouns in question, that is *motomeru* in *boomei wo motomete* "to seek for an exile" and *kirasu* in *iki wo kirasite* "to grasp for a breath". The issue is closely related to the problem of too wide grammatical relations (4.2.4) and the sketch grammar should be corrected for the grammatical relation in question to include only the first predicate verb as a collocate and exclude elements from another clause and thus limit the syntactic range (reach) where collocation extends to. However, the possibility to reach so distant collocates with the sketch grammar can be a valuable source for lexical information. This kind of distant co-occurrence of words could be added to the sketch grammar as a different type of collocational relation, called, for example, "distant collocates".

4.2.3 Missing collocates / sketch grammar relations

As mentioned above, the current sketch grammar for Japanese has a set of 22 collocational patterns, which covers various collocational relations for nouns, verbs, adjectives and adverbs: 16 different types of collocational relations for nouns, 14 for verbs, 7 for adjectives ending in *–i*, 11 for adjectives ending in *-na*, and 1 for adverbs. However, Sketch-Eval and other word sketch evaluations have descovered that a few sketch grammar relations are missing in the current sketch grammar.

The most important one from the point of view of overall lexical coverage is the inclusion of *suru*-V in various collocational relations for verbs. As mentioned in section 4.1, this issue is related to the tagging of a part of *suru*-V as nouns, which is currently overlooked in some sketch grammar relations. The correction of the grammar would resolve some of the issues.

For example, if we search for collocations of the noun *masutaa* "teacher, leader", we get collocational relations for nouns only, where the noun *masutaa* is the keyword; we do not get collocational relations for verbs, where the noun part of the verb *masutaa suru* "to master" is the keyword. Thus, the word sketch provides collocational relations

such as *masutaa wo yobu* “to call a teacher”, but does not provide collocations such as *tsukaikata wo masutaa suru* “to master the usage”, The same is true, if we exchange the keywords, for example, if someone searches for collocates of *tsukaikata* “usage”, *suru*-V, such as *masutaa suru* will not be in the collocate list of verbs. This deficiency can be overcome with corrections in the sketch grammar, but with the current tagset the collocational results for both. The noun and the *suru*-V would be still present without distinction in the same word sketch page.

Another missing point of the sketch grammar is for the collocational type noun_noun, or noun_noun_noun, which would cover compound noun collocations such as *Nihongo kyooiku* (日本語教育) “Japanese language teaching” or *Nihongo kyooiku gakkai* (日本語教育学会) “The society for Japanese language teaching”.

These types of deficiencies were not exposed in the Sketch-Eval type of evaluation since they relate to recall rather than precision. We plan to overcome the issues mentioned above in the new vesion of the sketch grammar for Japanese.

4.2.4 Too wide sketch grammar rules

The word sketch evaluators noticed that some collocations are bad or good but not so striking since sketch grammar rules are too wide for a few collocational relations. This is especially in the case of collocational relations for bound nouns and coordinating relations.

For example, one case of coordinating relations that is evaluated as poor is *subarasii* “great, superb” and *kazuooi* “many, numerous”, which actually is not a type of coordinate relation, as can be seen in the corpus example (17).

(17)	素晴らしい	映画	を	数多く	作り だす
	subarasii	*eiga*	*wo*	*kazuooku*	*tukuridasu*
	great	movie	object particle	many	create

“to create many great movies”

The sketch grammar rule for this kind of relations needs more contstraints for better results.

Too wide sketch grammar rules are related to the already described issue of collocates in different clauses, in other words, distant collocates (4.2.2).

4.3 Issues with the corpus and statistical methods

There are two main issues that appear in relation to the corpus and the statistical measurement:

- Page duplicates: when the same pages (or their copies) appear a number of times in the corpus. The result of this is that the same information appears a

number of times in the corpus, and the word sketch results wrongly present that some collocations are frequent, which they are not.

- Salience related problems happen when some collocates appear very frequently but only from one source, which is from one web page in the case of web corpora. Therefore, the results would be more accurate if the current statistical measurement took into account that collocates that appeared multiple times from one source were less salient.

Examples of collocation candidates that result from page duplicates or only one source and that are marked as bad are: *kesseki ga rongai* (欠席が論外) “absense is out of the question”, *wakawakasii midori* (若々しい緑) “youthful green”, *Terayama no haiku* (寺山の俳句) “Terayama's haiku”, *tikuseki to seitoo* (蓄積と正統) “accumulation and legitimacy”.

In addition, evaluators indicated that they would welcome a genre classification of the corpus, which would make the word sketches more usable in the field of Japanese language education. This kind of classification could be well applied in lexicography and other fields too.

The Japanese web corpus that is currently in use was created four to five year ago. Therefore, a new up-to-date web corpus with more thorough check on page duplicates or on other possible junk information would be welcome.

5. Conclusion and further work

The evaluation of Japanese word sketches inside a mini-project Sketch-Eval, though with a small number of evaluators, proved to be helpful both for system developers and system users, and especially promising for further research and activities on both sides and in collaboration. The evaluation results show a high percentage of good collocations, and we can conclude that Japanese word sketches could be a very useful resource for creation of collocational dictionaries. However, there are some issues that were discovered during the evaluation and which call for further enhancement of the functionality and for improvement of various components used by the tool (morphological analyzer/POS tagger, corpus, sketch grammar).

One of the basic questions to confront was the range of “collocation” and what set of words can be regarded as a collocate. Since word sketches were able to reach quite far in the search for collocates, sometimes neglecting our sense of syntax and semantics in a very general sense, we were newly confronted with sets of words which surprised us but opened our eyes. The word sampling and word sketches processed for the test evaluation suggested that linguists and language teachers should not limit themselves to their intuition and limited way of logic, but should be constantly ready for unusual (and actually at times usual for non-native learners') points of view. Of course, we are also aware that there are some language-specific categories and parts of

speech for which it is necessary to develop an umbrella category with subcategories in order to expect more effective results with word sketches.

A similar evaluation project with corrected POS tags, morphological analysis and evaluation categories, as well as increased number and variety of judges will probably show better results and offer yet new ideas for dictionaries, textbooks and teaching methods. Prior to this evaluation, a new version of Japanese word sketches will be created with a novel set of components, which would include a new and more up-to-date web corpus for Japanese, another morphological analyzer and an improved sketch grammar.

The issues related to POS taxonomy and Japanese orthography are actually problems which are very well present in the contemporary Japanese language grammar and orthography system. Word sketches reflect these problems "faithfully".

Aknowledgement

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536, in the National Research Programme II project 2C06009 and by the Czech Science Foundation under the project P401/10/0792.

References

Himeno, M. (2004). *Nihongo hyoogen katuyoo ziten*. Kenkyusha

Kilgarriff, A., Rychly, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Proceedings of EURALEX*. France: Université de Bretagne. 105-116. (available at http://promethee.philo.ulg.ac.be/engdep1/download/bacIII/sketch-engine.pdf)

Kilgarriff, A., Kovář, V., Krek, S., Srdanović, I., Tiberius, C. (2010). A Quantitative Evaluation of Word Sketches. *Proceedings of the XIV Euralex International Congress.* Leeuwarden : Fryske Academy. 7pp. (available at http://nlp.fi.muni.cz/publications/kilgarriff_xkovar3_etal/kilgarriff_xkovar3_etal.pdf)

Maekawa, K., Yamazaki, M., Maruyama, T., Yamaguchi, M., Ogura, H., Kashino, W., Ogiso, T., Koiso, H., Den, Y. (2010). Design, Compilation, and Preliminary Analyses of Balanced Corpus of Contemporary Written Japanese. *Proceedings of LREC 2010*, Malta. 1483-1486.

Oxford Collocations Dictionary for Students of English (OCD). (2009). Oxford University Press

Rundell, M, ed. (2002). *Macmillan English Dictionary for Advanced Learners*. London: Macmillan.

Seeley, C. (1991). *A History of Writing in Japan.* University of Hawai'i Press, Honolulu. 243pp.

Srdanović, E. I., Erjavec T. & Kilgarriff, A. (2008a). A web corpus and word-sketches for Japanese. *Sizen gengo syori (Journal of Natural Language Processing*) 15/2. 137-159. (also available at http://www.jstage.jst.go.jp/article/imt/3/3/3_529/_article)

Srdanović, I, Bekeš, A., Nishina, K. (2008b). Distant collocations of adverbs and modality forms observed in various Japanese language corpora. *Tokutei ryooiki kenkyuu 'Nihongo*

koopasu', Tokyo: Monbukagakusyoo kagakukenkyuuhi tokuteiryooiki kenkyuu 'Nihongo koopasu' Sookatu ban (*Workshop of the Priority Area Research "Japanese corpus"),* Tokio. 223-230.

Srdanović, E.I., Nishina, K. (2008). Koopasu kensaku tuuru Sketch Engine no nihongoban to sono riyoo hoohoo (The Sketch Engine corpus query tool for Japanese and its possible applications), *Nihongo kagaku* (*Japanese Linguistics*) 23. 59-80.

Vance, T. J. (1991). *Instant vocabulary through prefixes and suffixes.* Power Japanese series. Kodansha International. 128pp.

BOOK REVIEW

SU, X. (2011). *REFLEXIVITÄT IM CHINESISCHEN: EINE INTEGRATIVE ANALYSE: MIT ZWEI ANHÄNGEN VON HANS-HEINRICH LIEB.* (XIV + 293 PP.). FRANKFURT AM MAIN: PETER LANG. PAPERBACK.

Mateja PETROVČIČ

Introduction

This book was published in the Linguistics Series of European University Studies, and is written in German. As the book's title suggests, this monograph is primarily a comprehensive analysis of reflexivity in spoken Standard Chinese in the framework of Integrational Linguistics.[1] The author demonstrates that Chinese marks reflexivity only phonologically, with the use of reflexive pronoun(s), and argues that *ziji* (自己) is the only reflexive pronoun in Standard Chinese.

Different languages distinguish between referential and non-referential reflexive pronouns, and the author briefly demonstrates this with German *sich*. Referential uses denote semantic reflexivity (*inhaltliche Reflexivität*), whereas non-referential uses represent formal reflexivity (*formale Reflexivität*). Su asserts that the Chinese reflexive pronoun *ziji* is always referential and that there is no formal reflexivity in Chinese.

Since the research mainly focuses on the word *ziji*, not only in its reflexive usage but also in relation to intensifying and contrastive meanings and effects, this monograph could also be considered as a comprehensive research on *ziji* in Standard Chinese.

Summary

The volume is organized into five sections (A-E) and numbered chapters (0-16), followed by two appendixes written by Hans-Heinrich Lieb, the founder of the Integrational Linguistics.

The book opens with an introductory chapter, outlining the subject and scope of research, explaining the reason for the author's specific selection of examples, and sketching the structure of the book.

[1] The term "Integrational Linguistics" refers to the theory developed by Hans-Heinrich Lieb and should not be confused with the approach proposed by Roy Harris. (Sackmann, 2000, p. 472)

Acta Linguistica Asiatica, 1(2), 2011.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.1.2.85-88

Chapter 1 presents the issue of reflexivity in Chinese, from the distributional properties of *ziji* to its intensifying and contrastive effect. The author also raises the question of how many reflexive pronouns there are in Chinese. Two concepts relevant to further discussion are introduced at this stage, i.e. long-distance binding and blocking effect. This is linked to the problem of identifying the antecedent of the reflexive pronoun *ziji*, especially in situations when more possible antecedents appear in a sentence.

Chapter 2 gives a brief overview of the relevant research history; Chapter 3 focuses on the antecedent identification in the scope of generative linguistics; Chapter 4 finally closes the Section A with the findings in more semantic, pragmatic or functional oriented theoretical frameworks.

Sections B to E focus on Integrational Linguistics. Chapters 5 to 7 first introduce the reader to this theoretical framework in general and point out some properties of Chinese idiolect systems, which the author considers relevant for the following discussion.

Section C then starts with the actual investigation of reflexivity in Chinese. Chapter 8 deals with the concepts of reflexivity and intensification related to *ziji*. Author argues that reflexivity in Chinese should not be understood on the word-level of the reflexive pronoun *ziji*, but on the sentence-level. Su also stresses that the reflexive pronoun *ziji* should be clearly distinguished from its homophone, the intensifier *ziji* with completely different properties.

Because sentence-level is relevant for the antecedent selection, Chapter 9 first provides some assumptions and definitions of predicate, subject and object in Chinese. The author further stresses the difference between ellipsis (*Ellipse*), empty complement (*leeres Komplement*) and complement suppression (*Komplementunterdrückung*). Related to the above questions are also copula verbs, relational verbs and pivotal constructions in Chinese, which are of considerable importance for further discussion and described in the last parts of Chapter 9.

Chapter 10 briefly analyses sentences with the particles *ba* (把), *bei* (被) and the three *de*'s (的, 得 and 地). A more detailed investigation of Chinese *ba/bei* sentences in the scope of integrational linguistics is carried out in the appendix B at the end of the book.

Section D investigates highly controversial problem in Chinese, i.e. how an antecedent of the reflexive pronoun *ziji* is selected when several potential antecedents appear in the sentence. This is a difficult task in a language like Chinese with no pronoun-antecedent agreement. The author limits her discussion to the carefully selected examples with just one reflexive *ziji* and maximal three potential antecedents. Chapter 11 focuses on antecedent identification for subjects, Chapter 12 links to antecedent identification for objects, while Chapter 13 takes under consideration some restrictions for antecedent selection. Chapter 14 further discusses sentences with blocking effect mentioned already in Chapter 1.

Section E (Chapters 15 and 16) is the concluding part of Su's research. Chapter 15 summarizes all significant findings and conclusions, whereas Chapter 16 presents her hypotheses about reflexivity in every idiolect system.

The book is written for general linguists, preferably interested in integrational linguistics, experts in Standard Chinese, researchers of other languages who are working in the area of reflexivity, and is also a valuable piece of work for contrastive studies comparing closely related features such as reflexivity, intensification and contrast in a language.

Evaluation

This monograph represents an exceedingly detailed description of reflexivity in Chinese and deserves a readership far beyond the German academic community. The author very systematically introduces the reader to the theoretical framework and some basics of Chinese language, so readers with not much knowledge of Standard Chinese can also understand and benefit from this book.

Reflexivity in language is a complex and challenging issue, which is difficult to present in a simple linear way. However, the author skillfully interweaves various theoretical concepts and actual usages of the reflexive pronoun *ziji*, starting with very simple sentences and gradually proceeding towards constructions that are more complex and even ambiguous.

To discuss a selected topic in a scope of a chosen theoretical framework is already an immense project. Su, however, was faced with the additional task of providing explanation for some other phenomena in Chinese, which are closely related to the issue of reflexivity, but have not been discussed in Integrational Linguistics. It might be observed from the References in this monograph and the homepage of Integrational Linguistics that only a few studies about Chinese language have been done so far. In order to provide a comprehensive explanation of *ziji* on the sentence-level, the author had to first define several crucial concepts such as predicate, subject, direct and indirect object (or in Su's terms "first" and "second object") in Chinese. Since there are already several definitions and interpretations of the concerned concepts in the existing literature, the presented explanation might receive critical feedback.

In the subchapter of personal and reflexive pronouns, the author provides a persuasive argumentation, why *benren* (本人), *benshen* (本身) and *zishen* (自身) should not be treated as reflexive pronouns, as was sometimes claimed in the previous literature. However, speaking of personal pronouns, the author does not even mention *zanmen* (咱们), although it could be easily defined in the same way as the other personal pronouns. It is unclear whether the author considers *zanmen* as an expression

which does not belong to Standard Chinese, or is the omittance of this personal pronoun just a slip.

Throughout the entire book, the author provides numerous examples in Chinese. Moreover, from the Section C onward, most examples are also equipped with very illustrative structural diagrams. Their designing and formatting must have been a time-consuming job.

A potential limitation of the volume is that almost all examples are intentionally very similar and rely on the author's language sense as a native speaker of Chinese. When striving for a detailed theoretical explanation, the author seems to overlook ambiguous meanings of some examples or applies slightly unconvincing explanations (e.g. 3.6/12.5). Although the development of author's hypotheses and related argumentations seem to be very plausible, it is desired that the findings are confirmed in more diverse contextual environments, for not only reflexive usage of *ziji*, but also for intensifying meaning and its contrastive effects.

On the whole, the monograph offers numerous contributions concerning the issue of reflexivity in Standard Chinese. From the perspective of Integrative Linguistics, the author develops a detailed new analysis of the Chinese *ziji*, that will undoubtedly afford linguists in this area insightful knowledge evoking future research.

References

Sackman, R. (Last modified: December 11, 2004). *The Homepage of Integrational Linguistics*. Retrieved August 10, 2011, from http://userpage.fu-berlin.de/sackmann/+en/main-en.html.

Sackmann, R. (2000). Numeratives in Mandarin Chinese. In P. M. Vogel & B. Comrie (Eds.), *Approaches to the typology of word classes* (pp. 421-478). Berlin: Walter de Gruyter.