University of Ljubljana
FACULTY OF ARTS

Acta Linguistica Asiatica

Volume 10, Issue 1, 2020

ACTA LINGUISTICA ASIATICA
Volume 10, Issue 1, 2020

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

Issued by: Department of Asian Studies

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

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

ISSN: 2232-3317

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

Publication is free of charge.

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

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

TABLE OF CONTENTS

RESEARCH ARTICLES

FOREWORD

Any scientific discipline undoubtedly encounters different challenges in their development over time. However, with the rise of modern technologies, such challenges expanded to new dimensions.

In linguistics, corpus studies have already proven their advantages, and many researchers and other users enjoy the richness of different corpora, spreading enthusiasm and courage to treat linguistics interdisciplinarily. At the same time, minority languages and poorly studied languages are also gaining researchers' attention. Modern technologies further motivated different translation tools, which globalize the world with an enormous speed and seems to dicrease the relevance of language study and foreign language learning. However, the needs for interest in languages remain high. Though perhaps in a different form.

This issue comprises the above ideas in five articles.

Mayuri L. DILIP and **Rayesh KUMAR** coauthored the article “Clitic or Agreement Restriction in Santali: A Typological Analysis”, in which they investigated the syntactic configuration of pronominal number marking in Santali, the third most-spoken Austroasiatic language, from syntactic, morphological and prosodic perspective.

The following article “Contextual factors and language: an analysis of order placements” was written by **Andrej BEKEŠ**. It tackles Japanese in a specific social context, namely written ordering requirements on a crowdsourcing website, and reports negative correlation between the level of *added value* of such requirements and the quality of linguistic expression, comparing it to the Grice’s maxims of conversation.

ITO Hideaki's article “Orthoepic Competence Descriptors in Japanese Language Education: CEFR Levels B1 to C2” overviews the Common European Framework of Reference for Languages (CEFR) and the JF Standard for Japanese-Language Education to discuss the benefits of their merging. In particular, the author re-examines CEFR descriptors for levels B1 to C2 in a non-alphabetically transcribed Japanese, which have not yet been attempted, and combine them with the results for levels A1 and A2 to present descriptors for levels A1 to C2 in overall.

KIM Yu Young in his article “Development and Operation Results of Japanese Accent Perception Test Based On e-learning System” introduced the [AJ-phonetic Test] system, an online tool for practicing perception of Japanese word accent, presented its benefits through a longitudinal study on Korean learners of Japanese.

Finally, **Miha PAVLOVIČ** wrote an article entitled "Grammar Errors by Slovenian Learners of Japanese: Corpus Analysis of Writings on Beginner and Intermediate Levels". In it he first presents how to construct such a corpus, and then presents his analysis of grammar errors that a collection of 182 written texts written by Japanese learners contained.

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

ALA editorial board

RESEARCH ARTICLES

CLITIC OR AGREEMENT RESTRICTION IN SANTALI: A TYPOLOGICAL ANALYSIS

Mayuri J. DILIP
Indian Institute of Technology Madras, India
mayuri.dilip@gmail.com

Rajesh KUMAR
Indian Institute of Technology Madras, India
thisisrajkumar@gmail.com

Abstract

This paper investigates the syntactic configuration of pronominal number marking in Santali. Syntactic, morphological and prosodic restrictions show that pronominal number markers have properties of an affix as well as a clitic. A marker is an affix due to the fact that it cannot participate in a binding relation with other arguments. A pronominal number marker also functions as a clitic since it is attached to prosodically the most prominent constituent. The arguments that trigger object agreement do not manifest one particular case, but instead they indicate a dissociation between case and object agreement. On the other hand, the argument with subject agreement manifests nominative case only, indicating an association between nominative case and subject agreement. Both subject and object agreement are sensitive to case that indicates a property of an affix. Keeping in view the distribution of the pronominal number markers, we analyze feature checking of the two parameters, namely agreement and case in Santali.

Keywords: clitic; affix; agreement; prosody; Santali

Povzetek

Članek proučuje skladenjsko konfiguracijo zaimenskega označevanja števila v santaliju. Skladenjske, oblikoslovne in prozodične omejitve kažejo, da imajo zaimenski označevalci števila lastnosti tako pon (tj. pripon in predpon) kot naslonk. Označevalec deluje kot pona zato, ker ne more stopiti v navezovalno razmerje z drugimi argumenti, kot naslonka pa deluje zato, ker je del prozodično najmočnejšega sestavnika. Argumenti, ki sprožijo ujemanje s predmetom, ne izkazujejo nekega določenega sklona, temveč nakazujejo na razhajanje med sklonom in ujemanjem s predmetom. Po drugi strani pa argumenti, ki sprožijo ujemanje z osebkom, izkazujejo izključno imenovalnik, kar nakazuje na zvezo med imenovalnikom in ujemanjem z osebkom. Tako ujemanje z osebkom kot predmetom je občutljivo na sklon, ki izkazuje lastnost pone. Z vednostjo o distribuciji zaimenskih označevalcev števila v članku analiziramo preverjanje oznak dveh parametrov, ujemanja in sklona v santaliju.

Ključne besede: naslonka; pripona; ujemanje; prozodija; santali

Acta Linguistica Asiatica, 10(1), 2020.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.10.1.9-33

1 Introduction

This paper presents an analysis of pronominal number marking in Santali, which has in previous studies[1] been discussed either as a clitic (Kidwai, 2005; Anderson, 2007; Hock, 2013, among others) or as an affix/agreement (Subbarao, 2012). We prove that it is the same marker with a combination of affix and clitic features. A typological analysis and application of diagnostic tests show that the pronominal number markers exhibit properties of both clitics as well as affixes. These markers do not depict as two separate morphemes with two different functions, namely a clitic and an affix. Following the application of diagnostic tests, we analyse syntactic configuration of markers, keeping in view the properties of a clitic and an affix. Diagnostic tests reported in Kidwai (2005) show that the subject marker is a clitic in Santali. Additionally, we show that the subject and the object markers also function as affixes. A typological analysis of Khasi, Santali, Hindi-Urdu and Telugu[2] strengthen our argument that the preverbal position of a subject marker is the locus of a clitic, and that the post-verbal position is the locus of an affix. We notice that the preverbal position of the subject marker in Santali resembles Khasi's preverbal position of the subject marker. The post-verbal position of the subject marker resembles the post-verbal marker in Hindi-Urdu and Telugu. One significant property of affix is its inability to function as a pronoun by showing binding relations with other arguments as mentioned in Baker and Kramer (2016). Such binding relations are absent in Santali. Based on the results obtained from the application of diagnostic tests as in Suñer (1988), Mavrogiorgos (2010), Kramer (2014) and Baker and Kramer (2016), we claim that the subject and object markers syntactically behave like affixes along with their properties of a clitic. Evidence also comes from Hock's (2013) analysis of the subject marker in Santali as a Wackernagel element, where the subject marker attaches to prosodically strongest position within a sentence. This is the preverbal position. Hock's data also mentions the post-verbal occurrence of the subject marker, which is prosodically the weakest position. We consider the marker at the weakest prosodic prominence may be a syntactic locus of an affix based on the evidence obtained from the typological analysis and the application of the diagnostic tests.

The typological analysis of Khasi, Santali, Hindi-Urdu and Telugu shows a diss(association) of agreement and case in Santali. That is, nominative case marked argument obligatorily exhibit subject marking, indicating association. Case apart from

[1] Baker (2015, p. 60) citing *WALS* (Comrie, 2005; Siewierska, 2005b) mentions the status of accusative case marking and object agreement in 188 languages, where 13 out of 188 languages are having both person agreement with the P argument and accusative case marking: Cahuilla, Comanche, Greek, Guarani, Hungarian, Kanuri, Koasati, Kunama, Mangarayi, Miwok, Persian, Quechua, and Spanish. Among the languages mentioned above, the object agreement marker is probably a clitic and not true agreement as in Spanish and Greek.

[2] The data in Telugu is provided by the author of this reasearch, who is a native speaker of Telugu. The sentences are grammatical and acceptable to the best of the author's knowledge.

nominative, typically those which remain vP internally may or may not co-occur with object marking, indicating dissociation. Both subject and object agreement are sensitive to case. Keeping in view the results obtained from the investigation, we analyse the affix-clitic phenomenon of Santali by adopting Woolford's (2006) case system; feature interaction as in Davison (2004); movement of noun to higher D^0 to handle the (in)animates as in Kidwai (2005); Bhatt's (2005) AGREE to analyze the (diss-)association of case and agreement; finally, the Prosodic Inversion (Halpern 1992, Taylor 1995) to analyze the markers at prosodically strongest position.

The organization of the paper is as follows: section 2 discusses the basic description of pronominal number marking in Santali, and a typological analysis of select South Asian languages with a special focus on Santali; section 3 demonstrates the results of diagnostic tests, which show the role of a pronominal number markers as a clitic as well as an affix; section 4 elaborates the phrase structure in Santali, keeping in view the clitic-affix properties of subject and object markers; and finally, section 4 makes an overall conclusion.

2 Morphological description of the pronominal number markers

In this section, we demonstrate the morphological distribution of pronominal number markers in Santali (see Appendix A for the list of markers). Firstly, we provide a basic description of a subject marker (SM) and an object marker (OM) that correspond to various arguments such as subject, direct object (DO), indirect object (IO), object indicating possessor, object indicating beneficiary and non-nominative subject (NNS). We compare the pronominal number markers among Santali, Khasi, Hindi-Urdu and Telugu in order to identify their structural interaction with other syntactic operations such as case and agreement. Finally, the results of the diagnostic tests show that SM and OM have properties of an affix as well as a clitic. We also show how the evidence obtained from the typological analysis is parallel to the results obtained in the diagnostic tests.

2.1 Canonical structure of agreement in Santali

In this section, we discuss the basic structure of pronominal number markers in Santali. The SM –*y* occurs to the right and the OM –*e* occurs to the left of the finiteness marker –*a* in (1) and (2). Alternatively, SM is attached to the preverbal constituent *gidrə* as in (2b)[3]. An OM is absent from an intransitive verb as in (2a).

[3] See example (8) in section 3.1 for sentences with subject marking.

(1) *arel*$_i$ *uni*$_j$ *ɲel-ked-**e**$_j$-(y)a-**y**$_i$* (Santali)
Arel him see-PST:A-**OM**-FIN**–SM**
‘Arel saw him.’

(2a) *gidrə*$_i$ *udgɔrɔʔ-kan-a-**y**$_i$* (Santali)
child sweat-COP.PST-FIN**-SM**
‘The child was sweating.’

(2b) *gidrə-**y**$_i$ udgɔrɔʔ-kan-a* (Santali)
child**-SM** sweat-COP.PST-FIN
‘The child was sweating.’

Arguments such as object indicating possessor, object indicating beneficiary and NNS exhibit OM within the verb as in (3) (Neukom, 2001, pp. 111–112). The OM of the object indicating possessor is preceded by *–t,* a possessor incorporation as in (3a). The OM of the object indicating beneficiary is preceded by the applicative *–a-* as in (3b). The NNS in a psych-predicate exhibits OM as in (3c).

(3a) *baha **iɲ-ak** dal**-t-iɲ**-a-e* (Santali)
Baha **I-GEN** strike-**POSS-OM-**FIN-SM
‘Baha will strike mine.’

(3b) *baha **iɲ-renak** dal**-a-iɲ-**a-e* (Santali)
Baha **I-for** strike-**APP-OM**-FIN-SM
‘Baha will strike for me.’

(3c) ***baha***$_i$ *rabaŋ ɲam-akad-**e(y)**$_i$-a* (Santali)
Baha.NNS cold.NOM OBJ have-PRF:A-**OM**-FIN
‘Baha caught cold.’

In (4a), *puthi* ‘book’ is an inanimate noun, and therefore the corresponding OM (**ø**) does not realise. In a ditransitive sentence the DO takes precedence over the IO as in (4b) when both DO and IO are animate.[4] If one among DO or IO is animate, then the OM of the object with animate features realizes on the verb as in (4c).

[4] Baker (2012, p. 257) and also Leslau (1995, p. 186) show example from Amharic where the higher object (goal) among the two objects agrees within the ditransitive sentence. This property of the goal is a feature of object agreement and not cliticization. Such property of the goal does not exist in Santali. Instead, among the patient and the goal, it is the patient which realizes as the object marking in Santali.

(4a) *arel$_i$ puthi ɲel-ked-**ø**-a-y$_i$* (Santali)
Arel book see-PST:A-[-**OM]**-FIN-SM
'Arel saw the book.'

(4b) *baha$_i$ arel* ***gidra-kin$_j$*** *ema-t-**kin**$_j$-a-y$_i$* (Santali)
Baha.SUBJ Arel IO **child-dual DO** give-PST:A-**OM**-FIN-SM
'Baha gave the (two) children to Arel.'

(4c) *baha$_i$* ***gidra-kin$_j$*** *pɔtɔb ema-t-**kin**$_j$-a-y$_i$* (Santali)
Baha.SUBJ **child-dual.IO** book.DO give-PST:A-**OM-**FIN-SM
'Baha gave the book to the (two) children.'

The patient in a transitive sentence is accusative case marked and it possesses object marking as in (5a). If the same sentence is passivized, the patient is nominative case marked and it possesses subject marking as in (5b). Therefore, subject/object marking is sensitive to case as well as thematic roles. The markers being sensitive to other grammatical operations shows that we have a requirement to observe other grammatical operations in order to identify the true nature of the markers. Therefore, in the following section, we observe a typological analysis of the pronominal number markers among Khasi, Santali, Telugu and Hindi-Urdu.

(5a) *arel$_i$* ***uni$_j$*** *ɲel-ked-**e**$_j$-(y)a-**y**$_i$* (Santali)
Arel **him.patient** see-PST:A-**OM**-FIN**–SM**
'Arel saw him.'

(5b) ***uni$_i$*** *arel-ṭhen ɲel-oco-en-a-**y**$_i$* (Santali)
he Arel-by see-PASS-FIN-**SM**
'He was seen by Arel.'

In the following section, we provide a comparison of subject/object marking in Santali with the agreement structure of selected South Asian languages.

2.2 Variations and commonalities of agreement in Santali, Telugu, Hindi-Urdu and Khasi

Recall that the previous studies mention pronominal number marker as either a clitic or an affix in Santali. However, it was never considered as a marker with both of the features. The typological analysis of pronominal number markers shows that Santali contains features of Austro-Asiatic as well as non-Austro-Asiatic languages. In other words, the markers are combinations of affix-like and clitic-like properties. The typological analysis is elaborated below. One parallel among Khasi and Santali is the preverbal placement of SM. That is, *u* in Khasi occurs after *la wan* ‘came’ as in (6a) and *–y* in Santali occurs after *ɲelkedeya* ‘saw’ as in (6b). The variation between Khasi and Santali is that the SM in Khasi never occurs post-verbally. In contrast, the SM in Santali occurs post-verbally as in (6c). The object marking precedes the finite marker *–a* as in (6b) and (6c). In contrast, object marking is absent in Khasi, Hindi-Urdu and Telugu[5]. The analogy between Telugu, Hindi-Urdu and Santali is the post-verbal occurrence of SM as in (6c) to (6e). That is, the SM *–y* in Santali, *-i* in Hindi-Urdu and *-di* in Telugu attach to the right of the verb. In Section 3, we provide evidence to show that the pre-verbal occurrence of SM is the locus of a clitic, and the post-verbal occurrence is an affix.

(6a) ***u*** *briew* ***u*** *la* *wan* (Khasi)
M. human **M.** PST come
‘A man came.’ (Nagaraja, 1993)

(6b) *arel*$_j$ *uni*$_j$*-**y*** *ɲel-ked-**e**$_i$-(y)a* (Santali)
Arel him-SM see-PST:A-**OM**-FIN
‘Arel saw him.’

(6c) *arel*$_j$ *uni*$_j$ *ɲel-ked-**e**$_i$-(y)a-**y*** (Santali)
Arel him see-PST:A-**OM**-FIN-SM
‘Arel saw him.’

(6d) *rāmuḍu sīta-ni cūs-ā**-ḍu*** (Telugu)
Ram Sita-ACC see-PST**-3.SG.M**
‘Ram saw Sita.’

(6e) *umā kamīz lā**ī*** (Hindi-Urdu)
Uma.NOM shirt bring.PST**.F**
‘Uma brought a shirt.’ (Koul, 2008, p. 38)

The experiencer *pallavi* has overt dative case marking: *-ki* in Telugu and *ko* in Hindi-Urdu as in (7a) and (7b). In contrast, the OM *e* of the experiencer attaches the verb as in (7d). In the three languages mentioned above, the logical subject *pallavi* in Hindi-

[5] Irrespective of Santali, Hindi-Urdu and Telugu, and belonging to three different language families, they show commonalities. Such commonalities can besides be found in Hindi-Urdu and Icelandic, where they have quirky subjects/non-nominative arguments possessing all the properties of a subject except with the verbal agreement. Similarly, the nominative object possesses all the object properties except the verbal agreement (Thaínsson, 1979, p. 466). According to Subbarao (2012), the non-nominative experiencers have less probability to trigger verbal agreement.

Urdu/Telugu and *baha* in Santali exhibit object-like property. However, the mechanism exhibiting the object-like property in Hindi-Urdu/Telugu is different from Santali. That is, the NNS exhibits overt case marking in Hindi-Urdu/Telugu and object agreement in Santali. We consider the overt case marking and object agreement as object-like properties since they also occur on arguments such as direct and indirect objects. These are the arguments that do not move out of the verbal complex v/VP for case. The point that we emphasize is that the underlying syntactic interactions of NNS are same in all the three languages irrespective of the variations at the morphological level. Since, the underlying structure is similar to Hindi-Urdu/Telugu, we assume that the operations in Santali that are similar to Hindi-Urdu/Telugu may determine the affix-like properties of OM. In section 3, the application of the diagnostic tests shows that OM has some properties of an affix, apart from its properties of a clitic.

In (7a) and (7b), another parallel among Hindi-Urdu and Telugu is the nominative case and verbal agreement of theme *khušī* (Hindi-Urdu) and *jvaram* (Telugu) in psych-predicate. We assume that the theme has nominative case in Santali also. In South Asian languages as in (7), the subject agreement and nominative case obligatorily co-occur. Therefore, we consider such co-occurrence as a diagnostic test to determine whether or not the argument in a nominative case is marked. However, there is no evidence to show that the theme is not marked as a nominative case since the argument is inanimate, and as a result, SM is absent as in (7c).

Khasi[6] does not have constructions with experiencer subjects (Subbarao 2012: 12) and therefore we cannot show a comparison with Santali.

(7a) *pallavi-ki*$_i$ *jvaram*$_j$ *vacc-in-**di***$_j$ (Telugu)
Pallavi-DAT cold.theme come-PST-**3.SG.M**
‘Pallavi got fever.’

(7b) *allavi*$_i$ *ko bahut khušī*$_j$ *hu**ī***$_{*i,j}$ (Hindi-Urdu)
Pallavi DAT very happiness.theme happened
‘Pallavi felt very happy.’ (Subbarao, 2012, p. 147)

[6] In Khasi the sentence where the experiencer subject does not show any object-like properties as in (i) is possible.

(i) u john u don jingshit
m john m has fever
‘John has a fever.’

(Personal communication with Marshall Kharumnuid, Indian Institute of Technology Madras)

(7c) *baha*$_i$ *rabaŋ* *ɲam-akad-**e**$_i$-a* (Santali)
Baha cold.theme have-PRF:MID-**OM**-FIN
'Baha caught cold.'

In the following section, we demonstrate the properties of a clitic and an affix in Santali based on the diagnostic tests.

3 Pronominal number marker: clitic vs. affix

In this section, we demonstrate that subject/object marker has both the properties of a clitic and an affix. We strengthen the fact mentioned above with the evidence obtained from the diagnostic tests.

3.1 Pronominal number markers depicting clitic-like properties

Kidwai (2005) provides evidence that the subject/object marker is a clitic. The evidence is the following:

1. Markers have a high degree of selection with respect to hosts, as they can attach to nouns, postpositions, negation, verbs and light verbs.

2. Markers do not have unexplained gaps. For example, in English, the kinship terms are *fatherly, daughterly, motherly,* but not **sonly.* Such gaps are not possible in Santali.

3. Idiosyncratic semantics does not exist with these markers. For example, *-er* in *revolver* vs. *dancer* in English.

4. Markers do not undergo stem allomorphy or other morphological idiosyncrasies. For example, the morphological idiosyncrasies as in *slap, give* and *see* in English.

5. A marker has the ability to move.

6. The clitics attach after the markers of functional categories. Based on the evidence obtained from the data showing that the marker is a clitic, Kidwai argues that the clitics are copies of number marking that fail to delete and they occur as Backernagel (P-2) analysis. In the analysis, the animate NPs are [+R] featured indicating that they are referential expressions possessing [PERSON] feature and as a result, they undergo the operation MOVE involving COPY +MERGE. Following MOVE, a copy of the pronoun is left in the N^0. The stranded [NUMBER] featured pronominal copy at the level of PF is not eligible to get deleted and as a result it realizes into a number morpheme. Inanimate nouns have a structure similar to Bhattacharya's (1999) analysis of deixis indicating location such as *that*$^{\text{DISTAL}}$,

that$^{\text{REMOTE}}$ and *this,* where they occur lower than D^0. As a result, the inanimate noun undergoes operations such as COPY+ MERGE.

Hock's (2013) analysis of SM provides evidence that markers have a combination of affix-like and clitic-like properties. According to Hock (2013, p. 70), the SM is a Wackernagel element (P-2). That is, subject markers occur on any constituent to the left of the verb if they are prosodically dominant, as in (8). They occur preverbally since they are in the FOCUS position. The preverbal marker is a unique feature in Santali, which differs from other non-Austro-Asiatic languages. Alternatively, P-2 occurs post-verbally as in (9). The post-verbal position has the weakest prominence in the utterance. Hock also mentions Osada's (2008) observation that the younger speakers mostly prefer the post-verbal position[7]. In our analysis, we consider the post-verbal occurrence of the SM as an affix and not cliticization since its occurrence is not based on prosodical prominence. One more indication that SM may be an affix is its resemblance with the post-verbal occurrence of SM in Hindi-Urdu and Telugu as in section 2.1.

(8a) *abu hola sinema ɲɛl-lagit=**bun** cəlaw-len-a* (Santali)
we.INCL. yesterday cinema see-APP**.=we.INCL.** go-PST-FIN
'We had gone to see the movie yesterday.' (Kidwai, 2005)

(8b) *abu hola=**bu** sinema ɲɛl-lagi[t] cəlaw-len-a* (Santali)
we.INCL. yesterday=**we.INCL**. cinema see-APPL. go.PST-FIN
'We had gone to see the movie yesterday.'

[7] Givón (1976) on the other hand, 'the NP-detachment hypothesis', where the pronouns are reanalysed as affixes in three stages such as the following.

Stage I. Analytic pronoun argument agrees with a dislocated NP.

(Marie$_i$) she$_i$ like strawberries.
ADJUNCT SUBJECT VERB

Stage II. Incorporated pronoun agrees with a dislocated NP.

(Marie$_i$) she$_i$-like strawberries.
ADJUNCT SUBJECT-VERB

Stage III. Grammatical agreement marker agrees with subject NP.

Marie 3SG.FEM-like strawberries.
SUBJECT AGR-VERB

If we compare the structures as in (8a) with San tali, we find structure as in stage III only, where the marker's phonetic form resembles partially to the pronouns in Santali (See appendix 1 for the list of pronouns). It is our assumption that the pronoun preceding the verb might have undergone a transformation from a pronoun as in stage II to a bound morpheme as in stage III.

(8c) *abu=**bu*** *sinema ɲɛl-lagit cəlaw-len-a* (Santali)
we.INCL=**we.INCL**. cinema see-APPL. go-PST-FIN
‘We had gone to see the movie.’ (Kidwai, 2005)

(9) *ɲɛl-gɔt’-ka-t’-ko-a=**ko*** (Santali)
see-EMPH-COMPLASP-TR.3PL-FIN=**3PL**
‘They saw them off.’ (Anderson, 2007, p . 245)

3.2 Pronominal number markers depicting affix-like properties

Applications of the diagnostic tests of Suñer (1988), Mavrogiorgos (2010), Kramer (2014)[8] and Baker and Kramer (2016) among others, show that the pronominal number markers are affixes.

Mavrogiorgos (2010, p. 98) states that a subject triggers agreement and an object cliticizes. One reason is that a subject refers to one case role only (nominative) while an object depicts more than one case role (genitive, accusative and nominative). That is, a clitic co-indexes an argument irrespective of its case-marking. In Santali, the subject depicts only one case (nominative) and therefore, SM as an affix. The OM at first glance may behave like a clitic, due to variation in case features of corresponding objects. If we observe the objects, the OM co-indexes only to those arguments which achieve case within the verbal complex and as a result, they are eligible to co-index with an OM. Therefore, OM is sensitive to case similar to an SM.

The absence of subject/object marker due to semantic features such as specificity, animacy or definiteness is a feature of a clitic (Sũner, 1988)[9]. According to Corbett (2006, pp. 14–15), optional marking is the nature of a clitic. In other words, the

[8] Kramer (2014, p. 5) mentions three approaches to analyse an OM such as i. The OM involves in feature checking; ii. OM is a morpheme that moves into the verbal complex from within the DP; iii. A combination of both the analyses mentioned above. Kramer’s analysis of clitic doubling involves the option (iii). That is, Agree relationship between v and doubled clitic, A-movement (Nevins, 2011; Harizanov, 2014) of DP/D to [spec, vP] and finally m-merger (Matushansky, 2006) with v. We cannot adopt this analysis for Santali since the method of A-movement is suitable if the subject/object markers show any properties of pronouns syntactically indicating that the markers are clitics. However, the diagnostic tests show that the markers do not have the properties of a pronoun.

[9] Suñer (1988) shows clitic doubling in Spanish where the doubling takes place only when the object is [+specific +animate] as in (i) and (ii). The optionality is shown as a property of a clitic.

(i) La oian a Paca/ a la niña/a la gata
Her 3.PL-listened to Paca/ the girl/the cat [+anim, +spec, (+def)]
‘They listened to Paca/the girl/the cat.’

(ii) *La compramos (a) esa novella.
It-F1.PL-bought that novel [-anim, +spec, (+def)]
‘We bought that novel.’

agreement marking is obligatory for all the DPs irrespective of the semantic features. In Santali too, subject/object markers are absent if the argument is inanimate as in (10a) indicating clitic-like property. However, OM is sensitive to case along with animacy in a ditransitive sentence with animate DO and animate IO as in (10b). In this context, the OM of the DO realizes within the verb and not an IO, irrespective of IO being animate. Therefore, the object marking being sensitive to case is a necessary condition along with animacy feature.

(10a)	*arel*$_i$	*puthi*$_j$	*ɲel-ked-**ø***$_j$*-a-y*$_i$	(Santali)
	Arel	book	see-PST:A-TR-**ø**-FIN-SM	

‘Arel saw the book.’

(10b)	*baha*$_i$	*arel*	***gidra-kin***$_j$	*ema-t-**kin***$_j$*-a-y*$_i$	(Santali)
	Baha.SUBJ	Arel IO	**child-dual DO**	give-TR-**OM**-FIN-SM	

‘Baha gave the (two) children to Arel.’

If OM is an affix, it is entirely absent in passive and/or reflexive verbs (Kramer 2014).[10] In Santali too, the nominal reflexive in (11a) and the IO in a passive sentence in (11b) do not exhibit OM attached to the verb.[11]

(11a)	*peṭhue-ko*$_i$	*akotege-ko*$_i$	*sarhao-en**-(*ko)-**a*	(Santali)
	student-PL	by themselves-SM	praise-PST:MID**-(*OM)-**FIN	

‘The students praised themselves.’

[10] Baker and Kramer (2016) provide a different analysis of reflexives, where the absence of clitic doubling of a reflexive anaphor is a diagnostic test of a clitic. We do notice the absence of clitic with a reflexive in Santali. However, we consider the absence of the clitic is not due to the crossover effects, but due to the detransitivizing property of a verb depicting anaphoric relationship. A similar property of detransitivizing can be observed in passive sentences, sentences with reciprocal anaphors, sentences with self-benefactive marking and intransitive sentences. If the absence of the object clitic is due to crossover effect, its effect would have been found in interrogative NPs and/or universally quantified NPs also, which is not the case in Santali. In other words, the PNS occurs with interrogative NPs and universally quantified NPs without crossover effects.

[11] The OM in Amharic shows a different structure from Santali, where the clitic doubling takes place for passives and reflexives as in (i) and (ii). Therefore, the marker in Amharic is a clitic and not an affixes.

(i)	älmaz	mäx’haf-u	tä-sät’t’-**at**
	Almaz.F	book-DEF.M	PASS-give-(3MS.S)-**3FS.O**

‘The book was given (to) Almaz.’ (Baker, 2014, (16b); Kramer, 2014, p. 16)

(ii)	ɨd3d3-wa-n	t-at’t’äb-äʃtʃ- **ɨw**
	hand.M-her-ACC	REFL-clean-3FS.S-**3MS.O**

‘She washed her hands.’ (Leslau, 1995, p. 464; Kramer, 2014, p. 16)

(11b) *lər̩həi-re asok hɔtete ad̩i hor̩e got-akan-**(*ko)-**a* (Santali)
war-in Ashok by many people kill-PRF:M**-(*OM)-**FIN
'Many people were killed by Ashok in the war.'

According to Baker and Kramer (2016)[12], an OM is a clitic, if it does not co-occur with anaphoric DPs, quantified DPs containing a bound variable and interrogative pronoun.[13] This is because the marker functions as an intervener preventing any movement operations. In Santali, the presence of subject/object markers does not interrupt the binding relation between the antecedent *baha* and the verbal reflexive –*n- /nije-lagit* as in (12) and (13). The absence of OM is not due to its function as an intervener, but due to the intransitive nature of the verb. Such intransitive nature of the verb and the non-existence of OM is also found in intransitive, self-benefactive and passive verb as in (14).

(12) *baha arsi-re ɲel-e**n**-a-**y*** (Santali)
Baha mirror-in see-PST.MID-FIN-**SM**
'Baha saw herself in the mirror.'

(13) *baha nije-lagit gud̩iya-ko kiriŋ-ket-**ko**-a-**y*** (Santali)
Baha self-for doll-PL buy-PST:A-**OM**-FIN-**SM**
'Baha bought dolls for herself.'

[12] The diagnostic text with an interrogative pronoun cannot be applied to Santali since the wh-element occurs in-situ as in (i). We cannot show the interaction of the pronominal number markers with the wh-element since wh-movement does not exist in Santali. In Amharic, the presence of the clitic prevents wh-movement as in the sentence (1), footnote (20).

(i) uni ayo jete hor̩ dular̩-**ko**-a-y (Santali)
his mother everyone love.HAB-OM-FIN-SM
'His mother loves everyone. '

[13] The following are the constructions (Baker & Kramer, 2016, p. 4) with interrogative NP, universally quantified NP and reflexive anaphor in Amharic and the OM cannot occur in these constructions.

(i) Mann- ɨn ayy- ɨʃ? (*ayy- ɨʃ-iw)
who.M-ACC see.PH-2FS.S see.PF-2FS.S-3MS.O
'Who did you (feminine) see?'

(ii) Lämma hullu-n-ɨmm säw ayy-ä. (*ayy-äw)
Lemma.M-ACC person see.PF-3MS.S see.PF(3MS.S)-3MS.O
'Lemma saw everyone.'

(iii) Lämma rasu-u-n gäddäl-ä. (*gäddälä-w)
Lemma.M self-his-ACC kill-3MS.S kill(3MS.S)-3MS.O
'Lemma killed himself.'

-n- playing a role of *intransitive marker.*

(14a) *baha arel sala ragi-le***n***-a* (Santali)
Baha Arel with anger-PST**:MID**-FIN
'Baha was angry with Arel.'

-n- playing a role of *self-benefactive marker.*

(14b) *baha ub' get'-e***n***-a-y* (Santali)
Baha hair cut-PST:**MID**-FIN-SM
'Baha cut her hair.'

-n- playing a role of a *passive marker.*

(14c) *kumbḍu polis ṭhen sap-oco-le***n***-a-y* (Santali)
Thief police with catch-PASS-PST**:MID-**FIN-SM
'The thief was caught by the police.'

Summing up the discussion, pronominal number markers in Santali realize animacy, person and number features but not grammatical gender. SM attaches to preverbal constituents; to any constituent to the left of the verb, or else it attaches post-verbally. OM attaches to the left of the finiteness marker of a verb. Pronominal number marking for inanimate nouns is absent. In the case of a ditransitive sentence, when both DO and IO are animate, it is the OM of DO that attaches to the verb. However, when one among the objects is inanimate, the animate object exhibits object marking. In psych-predicate, the NNS exhibits OM. The evidence obtained from a typological analysis, the application of diagnostic tests and Hock's (2013) analysis of SM show that the pronominal number markers have properties of both a clitic and an affix.

Santali shows the preverbal occurrence of SM similar to Khasi and post-verbal occurrence of SM as in Hindi-Urdu and Telugu. Kidwai's (2005) diagnostic tests show that they are clitics since they have a high degree of selection of hosts; unexplained gaps of SM are absent; idiosyncratic semantics of the markers do not exist; the markers do not undergo stem allomorphy; the SM has the ability to move; the marker attaches to the functional categories.

Apart from Kidwai's findings, another feature that reflects a clitic-like property is its nature to be sensitive to animacy. Hock's (2013) observation of SM provides us evidence that there is a possibility for SM to have properties of both affix and a clitic. According to Hock, SM is a Wackernagel element, where the marker attaches to the constituent with highest prosodical prominence. Alternatively, SM occurs post-verbally, which is prosodically the least prominent position. The post-verbal occurrence of the SM attaching the least prominent position might be the position of subject agreement.

Diagnostic tests as in 3.2 show the following results:

1. Pronominal number markers is not optionally dropped in the presence of DO, since pronominal number markers do not function as in intervener.

2. The occurrence of pronominal number markers is restricted to case properties. That is, SM occurs only with nominative arguments and OM occurs only with arguments that are case-marked within the verbal complex.

3. OM is absent for passives and reflexive verbs.

4. Pronominal number markers do not function as an intervener for anaphoric DPs.

5. The OM of animate IO does not realize since the OM is sensitive to case, where the DO takes precedence over the IO.

In the following section, we discuss the syntactic configuration of pronominal number markers, keeping in view properties of an affix and a clitic.

4 Syntactic analysis of pronominal number markers

Keeping in view the restrictions of the pronominal number markers, we demonstrate the syntactic configuration of these markers. The analysis involves (non-)structural case licensing as in Woolford (2006); feature checking of case and agreement as in Davison (2004); analysing (diss-)association of agreement and case by adopting Bhatt (2005); distinguishing inanimates from animates, where animates move to a higher position D^0 as in Kidwai (2005). Finally, Prosodic Inversion as in Halpern (1993) in order to analyse the SM occurring in the prosodically prominent position.

The syntactic configuration can be presented in four steps. The first three steps are at LF while step 4 is at PF.

a. Step 1 shows the (non-)movement of a noun to higher D^0 within DP, depending upon the animacy feature of a noun.

b. Step 2 is feature checking of case.

c. Step 3 is feature checking of agreement, where agreement depends upon the position of the argument in a tree structure. It occurs as a consequence of steps (1) and (2) since the presence or absence of agreement and the type of agreement (subject or object) depends on the movement of the animate noun to a higher position within DP and also the location of the argument within or outside the verbal complex.

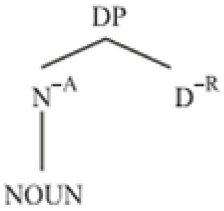
d. Step 4 is prosodic inversion.

4.1 Step 1

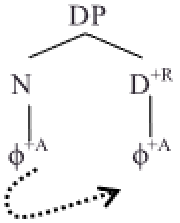
One of the properties of a clitic is the absence of pronominal number marker due to the inanimate feature of the argument. In contrast, the marker is present when the argument is animate as in (4a) in section 2.1, repeated in (15). The absence of the marker is indicated by ø. In line with Kidwai (2005: 203), the inanimates are not real pronominals because they do not refer to pronominals indicating person features. Instead, they refer to deixis of location such as *that*$^{\mathrm{DISTAL}}$, *that*$^{\mathrm{REMOTE}}$ and *this.* Hence, Kidwai shows a representation as in (16), where an inanimate noun originates at N^0 within NP and it does not have the ability to move to a higher position D^0 as in (16a). As a result, an inanimate noun is ineligible to participate in agreement due to its embeddedness within the DP. In contrast, an animate noun originates at N^0 within NP and it moves to D^0 as in (16b).

(15) *arel*$_i$ ***puthi*** *ɲel-ked-**ø**-a-y*$_i$ (Santali)
Arel **book** see-PST:A-[-**OM]**-FIN-SM
‘Arel saw the book.’

(16a)



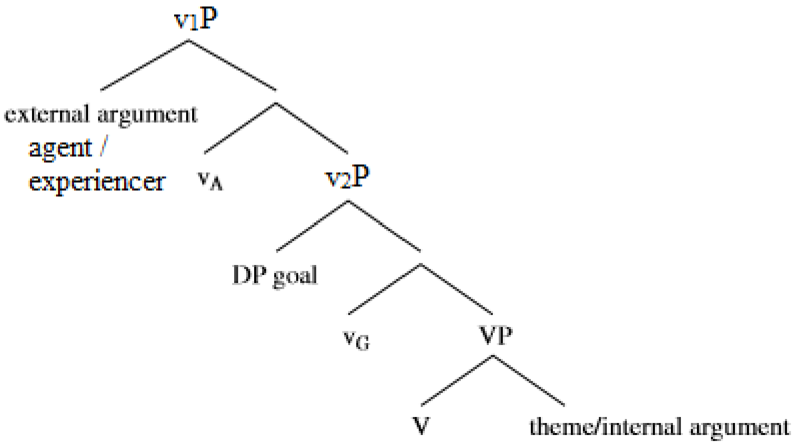
(16b)



4.2 Step 2

Adopting Woolford (2006)[14], verbal complex in Santali depicts the theta discharge of external argument (an agent or an experiencer[15]) at the higher [spec, vP], DP goal/recipient[16] at lower [spec, vP] and the theme/internal argument as the complement of V as in (17). We label the higher vP as v_1P and the lower vP as v_2P for the sake of convenience. The theta-discharge takes place right where they are base-generated. Consequently, the agent moves to [spec, TP] to check nominative case. In psych-predicates, a theme moves to [spec, TP] to check nominative case.

(17)



Tense and V function as goals with interpretable case features. The interpretable features of tense include nominative case, subject agreement and EPP. Interpretable features of V include structural accusative case and object agreement. The

[14] Woolford classifies non-structural case into lexical and inherent case. Lexical case is idiosyncratic; lexically selected by an individual verb or preposition. It is associated with the internal argument such as theme/internal arguments, case-marked arguments such as dative on ditransitive goal and not on (shifted) DP arguments. These arguments are checked by V in the VP proper.

Inherent case is associated with arguments external to the VP such as agent/external arguments (ergative case) and on (shifted) DP arguments, which are the positions with higher ɵ-roles. These arguments are checked by little/light v within vP. The inherent case is associated with specific ɵ-roles.

[15] Davison's concept of equidistance to TENSE is not suitable for Santali since the equidistance creates equal rights for the experiencer as well as the theme to achieve case, agreement and EPP of TENSE. However, the features of TENSE are unequally distributed among the arguments. That is, the experiencer receives only EPP feature from T and the rest of the features such as the nominative case and agreement are provided to the theme. However, note that the theme is ineligible to receive agreement features due to inanimate feature.

[16] Davison (2004) provides a v/VP structure where the IO/goal occurs within VP. In Santali, IO within VP is not required since the lower V has no role in providing lexical case to IO. Instead, the argument is checked lexical dative case by its null P in Santali.

interpretable feature of v is lexical case. Arguments function as probes with uninterpretable unvalued case features. Feature checking of case is elaborated below.

In intransitive, transitive and ditransitive sentences as in (2a), (5a) and (4b) in section 2.1, repeated in (18) below, agent/subject originates at [spec, v_1P] (in italics) and moves to [spec, TP]. Consequent to movement, T^0 checks nominative case and EPP with the agent as in (19).

(18a) ***gidrə***$_i$ *udgɔrɔʔ-kan-a-y*$_i$ (Santali)
child.agent sweat-COP.PST-FIN-SM
'The child was sweating.'

(18b) ***arel***$_i$ *uni*$_j$ *ɲel-ked-**e**$_j$-(y)a-y*$_i$ (Santali)
Arel.agent him see-PST:A-OM-FIN–SM
'Arel saw him.'

(18c) ***baha***$_i$ ***arel*** ***gidra-kin***$_j$ *ema-t-kin*$_j$*-a-y*$_i$ (Santali)
Baha.agent **Arel.recipient** **child-dual.patient** give-PST:A-OM-FIN-SM
'Baha gave the (two) children to Arel.'

(19) Before: [$_{TP}$ agent.$_{PROBE}$ [$_{T'}$ T^0.$_{GOAL}$ [$_{v1P}$ ….

nominative case, EPP

After: [$_{TP}$ agent.$_{PROBE}$ [$_{T'}$ T^0.$_{GOAL}$ [$_{v1P}$ …..

In transitive and ditransitive sentences as in (18a) and (18b) above, patient/theme originates as complements of the head V. The head V checks structural accusative case to the patient/theme as in (20).

(20) Before: [$_{VP}$ [$_{V'}$ V.$_{GOAL}$ patient/theme.$_{PROBE}$]]

accusative case

After: [$_{VP}$ [$_{V'}$ V^0.$_{GOAL}$ patient/theme.$_{PROBE}$]]

In a ditransitive sentence as in (18c) above, the goal/recipient originates at lower [spec, v_2P] and v checks lexical case features the goal/recipient as in (21).

(21) Before: [$_{v2P}$ goal.$_{PROBE}$ [$_{v2'}$ v^{0}.$_{GOAL}$…]]

Lexical case

After: [$_{v2P}$ goal.$_{PROBE}$ [$_{v2'}$ v^{0}.$_{GOAL}$…]]

In a psych-predicate as in (3c) in section 2.1 repeated in (22), experiencer gets demoted to achieve object-like properties. In Santali, the object-like property is object agreement. See Section 2.2 for the typological analysis of object-like property. Hence, it originates at [spec, v$_{1}$P] and v checks lexical case features to the experiencer as in (23a). On the other hand, the theme originates as a complement of V and moves to [spec, TP] as in (23b). Consequent to movement, T^{0} checks nominative case to the theme as in (18).

(22)	***baha$_{i}$***	*rabaŋ*	*ɲam-akad-**e(y)**$_{i}$-a*	(Santali)
	Baha.experiencer	cold.theme	have-PRF:A-**OM**-FIN	

‘Baha caught cold.’

(23a) Before: [$_{TP}$ [$_{T'}$ T^{0}.$_{GOAL}$… [$_{v1P}$ experiencer.$_{PROBE}$ [$_{v1'}$ v^{0}$_{GOAL}$ [$_{v2P}$ … [$_{VP}$ …]]]]]

EPP Lexical case

After: [$_{TP}$ [$_{T'}$ T^{0}.$_{GOAL}$ … [$_{v1P}$ experiencer.$_{PROBE}$ [$_{v1'}$ v^{0}$_{GOAL}$ [$_{v2P}$ … [$_{VP}$ …]]]]]

(23b) Before: [$_{TP}$ theme.$_{PROBE}$ [$_{T'}$ T^{0}.$_{GOAL}$ [$_{v1P}$ … [$_{v2P}$ … [$_{VP}$ [$_{V'}$ V^{0} …]]]]]]

Nominative case, EPP

After: [$_{TP}$ theme.$_{PROBE}$ [$_{T'}$ T^{0}.$_{GOAL}$ [$_{v1P}$ … [$_{v2P}$ … [$_{VP}$ [$_{V'}$ V^{0} …]]]]]]

4.3 Step 3

Arguments that trigger object agreement do not possess one type of case. They vary with respect to case. Such dissociation of agreement and case is due to two reasons. Firstly, arguments occupy different positions within VP. And secondly, the head agrees with an argument and the argument does not receive case-features from the same head. In contrast, arguments with subject agreement obligatorily have nominative case indicating association of agreement and case. This happens because the head that participates in subject agreement also participates in nominative case. Thus, there is both association as well as a dissociation of agreement and case in Santali. Keeping in view the (diss-)association, we adopt Bhatt’s (2005) AGREE which is an extension of

Chomsky's (1998, 1999, 2000) Agree[17]. Below, we elaborate feature checking of agreement in Santali.

In subject agreement as in (2a) in section 2.1 and repeated in (24) below, tense functions as a probe with unvalued, uninterpretable phi-features, and it searches for a DP which is a goal with interpretable phi-features. Agent in intransitive, transitive and ditransitive sentences, is the goal. Tense checks the phi-features with the agent located at [spec, TP] to trigger subject agreement as in (25).

(24) ***gidrə***$_i$ *udgɔrɔʔ-kan-a-y*$_i$ (Santali)
child.agent sweat-COP.PST-FIN-SM
'The child was sweating.'

(25) Before: [$_{TP}$ agent.$_{GOAL}$ [$_{T'}$ T^0.$_{PROBE}$ [$_{v1P}$ .....]]]
subject agreement
After: [$_{TP}$ agent.$_{GOAL}$ [$_{T'}$ T^0. $_{PROBE}$ [$_{v1P}$ .....]]]

In object agreement of a patient, a goal/recipient and an experiencer as in (1), (4b) and (3c) in section 2.1 repeated in (26), V functions as probe with uninterpretable unvalued phi-features. Patient and goal/recipient as in (26b) and the experiencer as in (26c) function as goalss with interpretable object agreement features. V searches for its nearest DP within the verbal complex and checks object agreement as in (27).

(26a) *arel*$_i$ ***uni***$_j$ *ɲel-ked-e*$_j$*-(y)a-y*$_i$ (Santali)
Arel **him.patient** see-PST:A-OM-FIN–SM
'Arel saw him.'

(26b) *baha*$_i$ ***arel*** ***gidra-kin***$_j$ *ema-t-kin*$_j$*-a-y*$_i$ (Santali)
Baha.SUBJ **Arel.goal** **child-dual.patient** give-PST:A-OM-FIN-SM
'Baha gave the (two) children to Arel.'

[17] Chomsky's Agree states that, the head T^0 functions as an unvalued, uninterpretable φ-PROBE and searches for the closest GOAL, such as a DP possessing its own interpretable features, in order to enter into an Agree relationship. Once the head T^0 finds a DP, T^0 gets a value from the DP, as a result the uninterpretable features of T^0 are erased. Simultaneously, case is checked in a similar manner. However, in case checking, DP functions as a PROBE with unvalued, uninterpretable case features and searches for a functional verbal head, a GOAL with interpretable case features. The feature interaction of agreement and case take place simultaneously. Once a DP is valued by case-features, the DP becomes inactive. An inactive GOAL is ineligible to value its φ features of agreement to a PROBE. Now, the difference between Agree and AGREE is that the GOAL is not inactive in AGREE. That is, a head agrees with an argument and the argument does not receive case-features from the same head.

(26c) ***baha**$_i$* *rabaŋ* *ɲam-akad-e$_i$-a* (Santali)
Baha.experiencer cold.theme have-PRF:MID-OM-FIN
'Baha caught cold.'

(27a) Before: [$_{VP}$ [$_{V'}$ V^0.$_{PROBE}$]] patient.$_{GOAL}$]]

object agreement

After: [$_{VP}$ [$_{V'}$ V^0$_{.PROBE}$ patient.$_{GOAL}$]]

(27b) Before: [$_{v2P}$ goal.$_{PROBE}$ [$_{v2'}$ v^0$_{.GOAL}$…]]

object agreement

After: [$_{v2P}$ goal.$_{PROBE}$ [$_{v2'}$ v0$_{.GOAL}$…]]

(27c) Before: [$_{v1P}$ experiencer.$_{PROBE}$ [$_{v1'}$ v.$_{GOAL}$ [$_{v2P}$ … [$_{VP}$ …]]]]

object agreement

After: [$_{v1P}$ experiencer.$_{PROBE}$ [$_{v1'}$ v.$_{GOAL}$ [$_{v2P}$ … [$_{VP}$ …]]]]

4.4 Step 4

Since SM is a Wackernagel element, it gets dislocated from the post-verbal position to the prosodically most prominent constituent through Prosodic Inversion (Halpern, 1992; Taylor, 1995) at PF. The Prosodic Inversion does not apply to OM since the OM does not possess a mechanism of movement similar to SM. Therefore, it remains in the same position where it originates. At LF, SM originates as an affix post-verbally. Due to prosodic inversion, SM moves to a constituent to the left of the verb as in (28a). In case SM occurs more than once within the utterance, it realizes as an affix post-verbally and consequently, it moves to leave a copy, where ever SM occurs within the utterance as in (28b). XP in (28) indicates any constituent in a preverbal position.

(28a) Before: [$_{XP}$ X^0 [$_{VP}$ V^0**=SM**]]
After: [$_{XP}$ X^0**=SM** [$_{VP}$ V^0]]

(28b) Before: [$_{XP}$ X^0 [$_{VP}$ V^0**=SM**]]]
After: [$_{XP}$ X^0**=SM** [$_{VP}$ V^0**=SM**]]

5 Conclusion

In this paper, we show that the pronominal number markers that indicate arguments of a sentence in Santali have properties of a clitic as well as an affix. Most of the properties of a clitic have already been identified in Kidwai (2005). In this paper, we identified the properties of an affix with the help of a typological analysis and the application of diagnostic tests. Kidwai's analysis that show properties of a clitic are the following.

1.The SM has a high degree of selection of hosts such as nouns, postpositions, negation, light verb etc.

2. Unexplained gaps corresponding the SM or OM are absent. Idiosyncratic semantics do not exist.

3. Allomorphy of the markers is absent.

4. The SM can move.

5. The marker attaches the morphemes indicating functional categories.

Additionally, we stated one more property of a clitic that the markers do not occur when the nouns are inanimates.

A typological observations of Khasi, Santali, Hindi-Urdu and Telugu show that some properties of a clitic resemble Khasi, where Khasi has a preverbal SM similar to Santali. Some properties of an affix resemble Hindi-Urdu and Telugu such as the post-verbal occurrence of SM. Observation of DO, IO and NNS show that the object-like properties of these arguments display morphological variations among these languages. However, they have a common underlying configuration that enables arguments to possess object-like properties. Another commonality determining the property of an affix is the obligatory co-occurrence of a nominative case and subject marking in Santali, Hindi-Urdu and Telugu. Looking at common operations, we assumed that there can be some operations in Santali similar to non-Austro-Asiatic languages, which allow a marker to be an affix. The obligatorily co-occurrence of nominative case and subject marking indicates association of agreement and case. In contrast, OM and case show dissociation since the OM does not correspond to one particular case.

An application of diagnostic tests show the following properties of an affix.

1. The subject marker and the object marker are sensitive to case similar to an affix.

2. In other words, SM occurs only with nominative case marked argument and OM occurs only with arguments such as DO, IO and NNS, which are the arguments that possess the properties of an object.

3. The OM of an IO in a ditransitive sentence is not realized irrespective of its animate feature, since the OM is sensitive to case along with animacy.

4. The pronominal number markers do not function as potential interveners preventing binding relations. This in turn indicates that the pronominal number markers do not function like pronouns.

Keeping in view the distribution of pronominal number markers, we analysed the operations of agreement and case in four steps. Step 1 showed that the inanimates exist lower in the DP and animates move from lower N^0 to D^0 for further process of agreement. Step 2 showed case checking of the arguments. Step 3 presented feature checking of subject and object agreement. Step 4 dealt with Prosodic Inversion depicting dislocation of SM.

Appendix

Below are the pronouns and the corresponding pronominal number markers.

The personal pronouns (Full/Free forms) of Santali:

Person	singular	Dual	Plural		
		INCL	EXCL	INCL	EXCL
First	*iɲ*	*alaŋ*	*əliɲ*	*abo*	*alɛ*
Second	***am***		*aben*		*apɛ*
Third	*ac’*		*əkin*		*ako*
	uni		*unkin*		*onko*

(Ghosh, 2008, p. 41)

Short/Bound forms of personal pronouns:

	Singular	Dual		Plural	
		INCL	EXCL		
First	*-ɲ(iɲ)*	*-laŋ*	*-liɲ*	*-bon*	*-lɛ*
Second	***-m***		*-ben*		*-pɛ*
Third	*-e*		*-kin*		*-ko*

(Ghosh, 2008, p. 54)

Abbreviations

3	: Third person
A	: Active
ACC	: Accusative
AGR	: Agreement
APP	: Applicative
APP	: Applicative
COP	: Copula
DAT	: Dative
DEF	: Definite
DO	: Direct Object
F/FEM	: Feminine
FIN	: Finite
GEN	: Genitive
HAB	: Habitual
IO	: Indirect Object
M	: Masculine
MID	: Middle
NNS	: Non-Nominative Subject
NOM	: Nominative
NOM OBJ	: Nominative Object
OM	: Object Marker
PRF	: Perfect
PL	: Plural
PRES	: Present
PROG	: Progressive
PST	: Past
REFL	: Reflexive
SG	: Singular
SM	: Subject Marker
TAM	: Tense Aspect Mood
VREF	: Verbal reflexive

References

Anderson, G. D. S. (2007). *The Munda Verb: Typological Perspectives*. Berlin–New York: Mouton De Gruyter.

Baker, M. (2015). *Case* (No. 146). Cambridge: Cambridge University Press.

Baker, M., & Kramer, R. (2016). Doubling clitics are pronouns: Reduce and interpret (Manuscript). *Rutgers University and Georgetown University*.

Bhatt, R. (2005). Long-distance agreement in Hindi-Urdu. *Natural Language and Linguistic Theory,* 23, 757–807. Retrieved from at https://people.umass.edu/bhatt/papers/bhatt-lda-nllt05.pdf.

Bhattacharya, T. (1999). *The structure of the Bangla DP. Department of Linguistics* (Doctoral Dissertation). University College London. Accessed at http://people.du.ac.in/~tanmoy/papers/Tanmoy_UCL_PHD.pdf

Chomsky, N. (1998). Minimalist Inquiries: The Framework, No. 15. In *MIT Occasional Papers in Linguistics*, Massachusetts: MITWPL.

Chomsky, N. (1999). Derivation by Phase, No. 18. In *MIT Occasional Papers in Linguistics*, Cambridge, Massachusetts: MITWPL.

Chomsky, N. (2000). Minimalist inquiries, the framework. In R. Martin, D. Michaels & J. Uriagereka (Eds.), *In honour of Howard Lasnik. Step by step: Essays on minimalist syntax* (89–155). Massachusetts: MIT Press.

Comrie, B. (2005). Alignment of case marking of full noun phrases. In M. Haspelmath, M. Dryer, D. Gil and B. Comrie (Eds.), *The world at language structures* (398-403). New York: Oxford University Press.

Corbett, G. G. (2006). *Agreement*. Cambridge: Cambridge University Press.

Davison, A. (2004). Structural case, lexical case and the verbal projection. In *Clause structure in South Asian languages,* 199-225. Springer, Dordrecht. Retrieved from https://link.springer.com/chapter/10.1007/978-1-4020-2719-2_7

Ghosh, A. (2008). Santali. In Anderson (ed.), *The Munda Languages.* 434-507. New York: Routledge.

Givón, T. 1976. Topic, pronoun and grammatical agreement. In Charles Li (ed.), *Subject and Topic* (149−188). New York: Academic Press.

Halpern, A. L. (1992). *Topics in the placement and morphology of clitics* (Doctoral dissertation). Stanford University.

Harizanov, B. (2014). Clitic doubling at the syntax-morphophonology interface. *Natural Language & Linguistic Theory, 32*(4), 1033-1088.

Hock, H. H. (2013). Backernagel is Wackernagel Lite. On the "P-Minus 2" Clitics of Santali. *Lingua posnaniensis, 55*(2), 67-75.

Kidwai, A. (2005). Santali 'Backernagel' Clitics: Distributing Clitic Doubling. In R. Singh (ed.), *The Yearbook of South Asian Languages and Linguistics* (189–207). Berlin: Mouton De Gruyter.

Kramer, R. (2014). Clitic doubling or object agreement: the view from Amharic. *Natural Language & Linguistic Theory, 32*(2), 593-634. Retrieved from http://faculty.georgetown.edu/rtk8/Clitic%20Doubling%20or%20Object%20Agreement.pdf

Leslau, W. (1995). *Reference grammar of Amharic*. Wiesbaden: Otto Harrassowitz Verlag.

Matushansky, O. (2006). Head movement in linguistic theory. *Linguistic Inquiry,* 37: 69–109.

Mavrogiorgos, M. (2010). *Clitics in Greek: A minimalist account of proclisis and enclisis*. Amsterdam/Philadelphia: John Benjamins Publishing.

Nagaraja, K. S. (1993). Agreement in Khasi and Munda languages. *Bulletin of the Deccan College Research Institute,* 53, 271-276. Accessed at https://www.jstor.org/stable/42936449.

Neukom, L. (2001). *Santali.* Muenchen: Lincom Europa.

Nevins, A. (2011). Multiple agree with clitics: Person complementarity vs. omnivorous number. *Natural Language & Linguistic Theory,* 29, 939–971.

Osada T. (2008). Mundari. In Anderson (ed.), *The Munda Languages.* 99–164. New York: Routledge.

Siewierska, A. (2005). Alignment of verbal person marking. In M. Haspelmath, M. Dryer, D. Gil, and B. Comrie (eds.), *The world atlas of language structures* (406-409). New York: Oxford University Press.

Subbarao, K. V. (2012). *South Asian languages: A syntactic typology*. Cambridge: Cambridge University Press.

Suñer, M. (1988). The role of agreement in clitic-doubled constructions. *Natural Language & Linguistic Theory*, 6: 391–434.

Taylor, A. (1995). The distribution of object clitics in Koine Greek. *University of Pennsylvania Working Papers in Linguistics, 2*(1), 7.

Thráinsson, H. (1979). *On Complementation in Icelandic.* New York: Garland.

Woolford, E. (2006). Lexical case, inherent case, and argument structure. *Linguistic inquiry, 37*(1), 111-130.

CONTEXTUAL FACTORS AND LANGUAGE: AN ANALYSIS OF ORDER PLACEMENTS FROM A JAPANESE CROWDSOURCING WEBSITE

Andrej BEKEŠ
University of Ljubljana, Slovenia
Andrej.bekes@ff.uni-lj.si

Abstract

Business documents, like other communications, are created in a specific social context to achieve various social goals. This study examines relationships between linguistic characteristics of order placements on Japanese dedicated crowdsourcing website and their context of situation, focusing on the power relations between the orderer and the subcontractor. As for the relationship between the orderer and the subcontractor, qualitative data analysis shows that it is the orderer who is overwhelmingly powerful in this relationship. This imbalance seems to be reflected in the linguistic characteristics of order placements, such as choices made in the system of grammar, and in the quality of information in the sense of Grice's maxims of conversation.

Keywords: crowdsourcing; order placement; orderer; subcontractor; Japanese

Povzetek

Poslovni dokumenti so tako kot vsako sporočanje ustvarjeni v specifičnem družbenem kontekstu z namenom doseganja različnih družbenih ciljev. Študija preučuje povezavo med jezikovnimi značilnostmi naročil na japonskem spletnem mestu za množično naročanje (crowdsourcing) in njihovim kontekstom situacije, s poudarkom na razmerjih moči med naročnikom in podizvajalcem. Kar zadeva razmerje med naročnikom in podizvajalcem, kvalitativna analiza podatkov nakazuje, da je naročnik v tem razmerju izjemno močan. Zdi se, da se to neravnovesje odraža v jezikovnih značilnostih kot so vrsta izbire v slovničnem sistemu in v kakovosti informacij v smislu Gricevih konverzacijskih načel.

Ključne besede: množično zunanje izvajanje; naročilo; naročnik; podizvajalec; japonščina

Acta Linguistica Asiatica, 10(1), 2020.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.10.1.35-47

1 Introduction

1.1 Social context of crowdsourcing

The advent of internet enabled a relatively new form of employment, i.e., people working under a short-term, ephemeral contracts or in self-employed capacity, known alternatively as crowdsourcing or gig economy, to begin its fast expansion globally. The best-known representatives of this phenomenon are companies such as Uber technologies, an international ride-hailing company, active on a global scale. This type of employment, while its potential is being hailed as a new promising form of employment, has also spawned a plethora of social problems accompanying it (cf. New York Times Sept. 15, 2019, Asahi Shinbun Nov. 17, 2019, Tokyo Shinbun Dec. 6, 2019, among others). It is thus attracting enough attention to be dealt with not only in a range of media all over the developed world but also in the film ("Sorry we missed you", the last film made by a veteran British filmmaker Ken Loach).

Yet, due to its recency as a phenomenon, crowdsourcing has not yet received much attention in academia. One early forerunner is a research group at the National Institute of Japanese Language and Linguistics (NINJAL) in Tokyo, led by Professor Ishiguro, which is at present working to determine linguistic features of Japanese crowdsourcing texts, using a large database of more than 100.000 individual order placements from a dedicated Japanese site, CrowdWorks, run by the eponymous company. We had the privilege to work with the group for a year, from the end of 2018 to the end of 2019.

With the potential for a stable long-term employment increasingly diminishing, crowdsourcing is, among others, taking advantage from the void in labor law framework that is supposed to protect workers (Tamura, 2014). Chomsky (1993), writing about the situation in the United States, has in his analysis of trends similar to crowdsourcing arrived at the conclusion that the employment terms are changing in the direction increasingly advantageous to employers, resulting in impoverishment of the social strata providing the labour force. Certainly, according to labor regulations in OECD countries, full-time employment implies various aspects of social security for workers such as minimum wage, paid leave, health insurance and pension insurance, while on the other hand it seems that there is no such social security provided in crowdsourcing environment. The reason is that the so-called 'worker' employed through crowdsourcing is not legally a worker. This is because the relationship between an orderer (also ordering party, in Japanese *kyūjinsha* 求人者, recruiters) and an order receiver (here referred to as subcontractor, in Japanese also *kyūshokusha* 求職者, job seeker) is in principle not a labor-management relationship, but a contract, where a job seeker is a subcontractor rather than a worker directly employed by an orderer (see Tamura, 2014; Mizuno, 2015). In case of crowdsourcing, the result of such relationship between an orderer and a subcontractor is that a subcontractor sells his or her work to the orderer purely as a commodity without any protection by the legal framework. In other words, the work performed by a subcontractor (or, more precisely, of those individuals who, in some cases, are providing the work at the end of the subcontracting chain) is increasingly reduced to being just the goods on the market, as Chomsky (1993) points out. Such a situation, not unlike that of the 19th century during the dawn of the Industrial Revolution, is expected to affect linguistic features found in the order placements.

In this paper, we focus on linguistic features found in the corpus of about 100.000 order placement data generously provided by CrowdWorks (https://crowdworks.jp/public/) from their eponymous dedicated crowdsourcing website. Relying on the framework proposed by Halliday (1978) and Hasan (2009), we try to interpret linguistic characteristics of these order placements within the context of situation in which they were issued. More specifically, we try to interpret these order placements in terms of the power relationship between an orderer and their subcontractor.

1.2 Crowdsourcing: its features and structure

At the CrowdWorks crowdsourcing website, the following flow of work can be discerned.

- Orderer (client) creates and publishes order placement (publishing);
- Applicant (worker) searches for a suitable order placement and sends application (application).
- Orderer judges the application, decides on the hiring and proceeds to contract (contract).
- Orderer and subcontractor communicate exclusively via the Internet (accomplishment).
- Subcontractor submits deliverables in time to meet the delivery date [delivery].
- Orderer checks the quality of the deliverables and pays for them [payment].

Crowdsourcing, at least in our case, is a system in which an orderer (client) and a subcontractor (worker) can complete their work only through online communication. At the CrowdWorks website there are four distinguishable business document types, i.e., order placements, application documents, exchanged documents, and deliverables. In this chapter, we limit our research to order placements. From the aforementioned newspaper articles in New York Times and Japanese media it can be deduced that from the viewpoint of subcontractor there are three advantages of crowdsourcing: (i) work from home, (ii) work with people from all over the world, and (iii) possibility using one's free time for additional income.

But the above benefits for the subcontractor also benefit the orderer. The fact that work can be done at order receiver's/subcontractor's home means that the orderer benefits from the zero cost for office space and equipment since both are provided by the subcontractor. The fact that people from all over the world can become subcontractors means that subcontractors can be recruited under the cheapest conditions. From the viewpoint of subcontractors based in Japan this is obviously a demerit. Needless to say, in the case of a subcontractor who has no regular job and therefore no economic power, the fact that crowdsourcing work can be done at home, even if offered at low wage, is perceived as a merit, because there are no transportation cost involved. But we must not forget that in the case of a regular job, transportation cost would be born by the employer.

The usual discussions about crowdsourcing also seem to lack the following important aspect, i.e., legal liability of parties arising from the orderer- subcontractor relationship. As already mentioned, order placements studied in this paper are order placements advertised on the site called CrowdWorks, run by the company of the same name. In these documents, there is usually a description of the content of work and how to order and apply, but in none of the examined randomly chosen 100 documents from the database of 100.000 documents was there a description of the orderer and the subcontractor nor the legal liability binding the two. Keeping legal issues ambiguous besides seems to favour the orderer over the subcontractor.

Based on the above assumptions, Section 2 introduces the framework of the analysis, the Section 3 analyzes the specifics of the context of the situation in which CrowdWorks order placements, and its relation with the linguistic aspects of order placements. Section 4 summarises the results.

2 Analysis framework: Discourse (text) and context of situation

In decades of research seeking to determine the link between discourse and social situation, various approaches have been attempted. With sociology as his starting point, Bourdieu (1991) offers interesting insights. As for approaches, rooted in discourse, Critical discourse analysis (CDA, see Fairclough 1989, 2012, Hart 2016) merits close attention. Like most CDA studies, the present study relies on the framework proposed by the Systemic functional linguistics (SFL), in particular the SFL's notion of the context of situation.

In the continuation of this section, we introduce Halliday and Hasan's notion of 'context of situation' as the framework of description and analysis used to investigate linguistic characteristics of order placements in relation to their social context.

In linguistic exchange language is realized as a text. According to Hasan (2009), based on the work of Halliday (1978) and others, this realization takes place in the ‘context of a situation’. The context of a specific situation in which language exchange takes place is, in a nutshell, who interacts with whom, in what situation, using what kind of language. The context situation conceived in this way consists of the following three elements and Following Halliday (2009, pp. 57-58) and Hasan (2009, p. 178) and its aspects could be subsumed as follows:

(i) Field, i.e., social activities of which specific discourse is a constituent part, and other relevant aspects of the situation where a particular discourse takes place.

(ii) Tenor, i.e., social relationship between parties involved in social activities of which discourse is a part.

(iii) Mode, i.e., the way language is used in a particular discourse, including the way how discourse participants are in contact (channel).

The above three elements are intrinsically related to discourse and stand out from various other elements included in the context where particular discourse is taking place. The above three elements thus make a reasonable description of the context of situation possible. Actually, Hasan (2009) addresses the connection between the context of situation and language in a much broader cultural and social perspective, but this goes beyond the scope of the present study.

3 Characteristics of the context of situation and linguistic expressions seen in CrowdWorks order placements

In this section, we explore how contextual factors relate to characteristics of linguistic expressions in CrowdWorks order placements.

3.1 Features of context of situation in CrowdWorks order placements

First, we present the analysis of the context of situation. Most of the information that can be obtained from CrowdWorks's website mainly focuses on the companies acting as orderers. Information concerning the subcontractors is limited to the number of applicants who applied for a particular job and the number of applicants that were accepted. The only additional information about subcontractors provided on the website is the subcontractors' profiles, which are required by the orderer for a specific task.

In order to identify elements in the context of situation, i.e., in the field, tenor, mode associated with particular order placements that may affect the linguistic expression therein, 1000 randomly extracted examples of order placements, and additionally, the CrowdWorks website, were examined. These elements are summarised in Table 1 below. The elements in Table 1 are treated as variables, and the values they can assume are given in bold print in parentheses.

Table 1: Relevant elements of the context of situation of CrowdWorks order placements.

Context elements	Detailed elements (where possible expressed as variables and their values)
Field	• Conditions for ordering work by an orderer → Properties related to deliverables and their production: *Added value* (of deliverables)*:* (**high, low**) *Period period of employment:* (**long, short, single job**) → Order (employment) conditions / payment method: *Form of work:* (**project, competition, task**) *Reward system:* (**fixed remuneration, hourly unit price, competition**) *Reward* (converted to hourly wage)*:* (**high, medium, low**) → Job description *Required skill level:* (**high, low**) → Legal framework *Description of legal and social responsibility:* (**yes, no**) → Place of work *Location:* (**at home, commuting**) <In case of **commuting**: *Transportation expenses:* (**yes, no**)> • Elements related to subcontractor → *Number of applicants:* (**1, more than 1**) → *Number of subcontractors:* (**0, more than 0**) → *Estimated self esteem performing the ordered work:* (**yes, no**)
Tenor	• *Orderer:* (**organization, individual**) orders work from *subcontractor:* (**individual**); • Relative bargaining power and relationship between orderer and subcontractor: → *Orderer:* (**strong**), → *Subcontractor:* (**weak**) → *Orderer* has the right to employ *order receiver/subcontractor* → *Subcontractor* has fewer options of action available than *orderer*; • Social distance between *orderer* and *order receiver/subcontractor* estimated as maximal
Mode	• Orderer openly recruiting via the dedicated website: → No direct visual or auditory contact → Handwritten exchange on the dedicated website → There is a conflict of interest

In Table 1, each element in the field is related to order placement seen in a discourse, but the relevance is not direct in the sense of Hasan (2009). Rather, these elements are considered as relevant factors because they may influence the power relationship between an orderer and a subcontractor. Most of these elements, except for *number of applicants* and *number of subcontractors*, can be regarded as qualitative variables.

As for tenor, i.e., the relationship between an orderer and an order receiver-subcontractor, the social distance between the two is considered to be the largest possible. In the power relationship, the orderer is considered to be strong while the subcontractor is considered to be weak. In the skill profile that is required in particular order placement, the availability of such a profile, i.e., whether such skill profile is very common or rare, may, in the case, when such a profile is rare, influence the power relationship to the advantage of the subcontractor. Furthermore, from the point of view of the subcontractor, the degree of pride and satisfaction one gets by having performed the required order, may affect the frequency of subcontractors' application.

As for mode, the recruitment is open to public thorough the dedicated CrowdWorks website, and there is no direct contact between the two parties. All the contact is limited to the written exchange online.

3.2 Relevance of the added value for the advertised job

Mode and tenor are being fixed permanently and therefore do not vary, though the indirect communication though the website and social distance are also influencing the power relation between the orderer and the subcontractor to the advantage of the orderer, who holds all the reins.

On the other hand, there are several elements in the field that are not fixed, on of them being the added value of the deliverables. Contents of ordered work range from high value-added tasks that require high skills, such as software development and training, to simple low value-added items such as 'copy and paste' jobs. Such added value of the deliverables can for each case be inferred from the proclaimed amount of remuneration for the deliverable, and additionally also from the skill level required of the subcontractor in order to perform the task in question. In fields where highly skilled human resources are required, the number of qualified subcontractors is limited, so their bargaining power is high, and it seems that they are for this reason relatively strong in relation to orderers. On the other hand, in jobs that do not require high skills, there are many qualified subcontractors that can potentially apply, so their bargaining power is weak, and it can be concluded that their relative power in relation to the orderer is weak. 'Relatively' in the previous sentence refers to the difference in bargaining power between a highly skilled subcontractor and a low-skilled subcontractor. Yet, needless to say, it is the orderer that has the absolute advantage in any of the relationships implied in crowdsourcing jobs. Otherwise, the orderer would seek the required labor in the form of full-time employment, rather than through crowdsourcing (cf. Muthoo, 1999; Tamura, 2014).

3.3 Qualitative analysis of the relationship between the characteristics of linguistic expressions in CrowdWorks data and the context of situation

For the purpose of this section we performed a qualitative analysis of the relationship between order placement and its context of situation. In order to obtain information from a subcontractor such as the number of applicants and the number of subscribers, which are not included in the order placement material of the randomly chosen subsample of 1000 order placements, we directly accessed the CrowdWorks website, which contains such information. However, since the recruitment period has not yet expired for some of the raw data taken from CrowdWorks's website, the number of applicants and contractors may not be final. We qualitatively analysed in detail 10 CrowdWorks website order placements obtained in this way, and 10 randomly selected from the subsample of 1000 order placements.

3.3.1 Linguistic features of order placements

Ishiguro (2018), summarising the goals of the research of crowdsourcing order placement data conducted by his group, proposes the analysis of linguistic features of order placements from the following four perspectives. These are (i) 'expression' (employment of grammatical items), (ii) 'psychological attitudes' (politeness), (iii) 'information' (conveying of content), and (iv) 'conditional aspects' (non-verbal factors).

In the present study of order placements we focused on the expression side, more specifically, the quality of linguistic expression and the content side, more specifically, the quality of the information provided. The basis for the evaluation of the quality of linguistic expression was quality of employment of grammatical items in order placements. The basis for the evaluation of the quality of information was based on Grice's (1975) maxims, in particular on the Maxim of quantity. Both aspects of order placements were judged independently by two evaluators, who arrived at mutually compatible conclusions. The combined result for each order placement was subsumed under the variable *comprehensive impression*, with values high, medium, or low, i.e., the *comprehensive impression* of each order placement could be judged as high, medium, or low. *Comprehensive impression* judged as high was not attested, examples of two order placements, one judged as medium-high and the other as low, are given in the appendix.

3.3.2 *Comprehensive impression* of order placements and added value

Elements of the context that were analysed are variables related to the field aspect of the context of situation, presented in Table 1 in italic. These elements are the only specific data that could be obtained directly from the analysed material. The overall interpretation also took into account the aforementioned characteristics of tenor and mode aspects of the context of situation.

For about 70% of order placements their *comprehensive impression* was judged as low; in such cases the amount and the quality of information were problematic without exception. Such order placement cases also had *rewards* in the range of low and required skills, and by implication, added value, were in the range of medium or low. Further, the highest evaluation of the *comprehensive impression* of order placement cases did not go beyond medium.

Cases where *comprehensive impression* of linguistic expression was evaluated as medium were found in order placements that offer relatively high rewards. Order placements with relatively higher *rewards* also have higher *added value*, and orderers had higher motivation to secure subcontractors meeting the appropriate skill levels. In other words, in such cases, with relatively few highly skilled personnel available on the labour market, power relationship of both parties was also relatively advantageous to subcontractors (see Muthoo, 1999). Yet even in this case, it is still the orderers who keep overall control in their hands. On the other hand, the subcontractors, though in a more advantageous position than those of lower *added value* jobs, requiring lower level of skills, are still at a disadvantage as they have less security in a one-off employment relationships offered by the CrowdWorks website. It can be concluded that in the case of relatively high value-added projects, the orderers seem to be motivated to pay more attention to the content and wording of their order placements, to secure the labour force that will perform such work.

On the other hand, order placements that offer low *rewards*, tend to require mechanic work, and are judged to have low *added value*, typically do not require high *skills*. For the large group of low-skilled subcontractors competition for the jobs is increasing and their relative strength in relation to orderers is thus becoming weaker. It is easy to see that in such cases, even if the orderer does not pay much attention to linguistic expression of order placements, expect sufficient number of applicants for the advertised jobs can be expected. Indeed, it seems that this is in fact the reason for the relatively low quality of such order placements. There might also be another factor that lowers the quality of order placements for projects with low *added value*. Namely, in case of low value-added projects, there must be a large number of work done (and consequentially, a large number of order placements issued) in order to make business profitable. If the number of order placements is large, this alone means that the effort that could be devoted to each order placement is necessarily limited.

4 Conclusions

In this paper, we examined the relationship between the context of a situation and the quality of order placements in the CrowdWorks order placement corpus. Quality, based on the relevance of provided information and on the quality of linguistic expression, has been judged independently by two evaluators. Since these order placements did not include detailed information about the situation of the orderer and the subcontractor apart from what was discernible from its language related aspects, only a limited number of factors related to the context of the situation could be identified. To make up for this, this pilot analysis was, besides the aforementioned order placement corpus data provided by the CrowdWorks, partially based also on data taken in random order directly from CrowdWorks dedicated website, where such personal details were still displayed. Results point in the direction of a possibility of negative correlation between the level of *added value* of ordering requirements, influencing the negotiating power of the orderer towards the subcontractor, and the quality of linguistic expression, which could also be conceived in terms of the Grice's (1975) maxims of conversation.

However, the above analysis is based on limited data. In that sense, it only provides grounds for a hypothesis about the orderer and subcontractor related factors that govern the quality of CrowdWorks order placements. After examining the results through a more detailed multifaceted quantitative analysis, based on ampler data, more accurate conclusions can be reached. A lot is expected from the results of Ishiguro's research group, which are to be made public throughout 2020.

Acknowledgements

The author would like to thank the National Institute of Japanese Language and Linguistics (NINJAL) in Tokyo, for being given the chance to conduct research there during the one-year period from November 30, 2018 to November 29, 2019, to Professor Kei Ishiguro, for the privilege to be able to join his research group during this period, to Dr Yuko Aoki, for her patient help with analysis of the order placements, and to all the members of the research group for sharing their vigour and enthusiasm.

References

Bourdieu, P. (1991). *Language and symbolic power*. In J. B. Thompson (Ed.), Gino Raymond and Matthew Adamson (transl.). Cambridge: Polity Press.

Chomsky, N. (1993). *Year 501: The conquest continues*. London: Verso.

CrowdWorks crowdsourcing. https://crowdworks.jp/public/, last accessed 30 December 2019.

Fairclough, N. (1989). *Language and power*. London: Longman.

Fairclough, N. (2003). *Analysing discourse: Textual analysis for social research*. London: Routledge.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. Morgan (Eds.), *Syntax and semantics 3: Speech acts*, pp. 41-58. New York: Academic Press.

Halliday, M.A.K. (2009). *Essential Halliday*. London: Continuum.

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. London: Edward Arnold.

Hart, C. (2016). *Discourse, grammar and ideology: functional and cognitive perspectives*. London: Bloomsbury.

Hasan, R. (2009). The place of context in a systemic functional model. In M.A.K. Halliday and J. J. Webster (Eds.), *Continuum Companion to Systemic Functional Linguistics*, pp. 166-189. New York: Continuum.

Ishiguro, K. (2017). Bunshō to wa nani ka - nihongo no hyōgen-men kara mita yoi bunshō [What is a sentence: A good sentence from the viewpoint of Japanese expression]. In Lee J.H. (Ed.), *Bunshō o kagakusuru* [Science of Writing] pp.14-37. Tokyo: Hituzi Syobo,.

Ishiguro, K. et al. (2018). Kuraudosōshingu o mochiita bijinesu hatchū bunsho no shitsuteki bunseki shiron [An Essay on Qualitative Analysis of Business Ordering Placemets Using Crowdsourcing]. *Proceedings of the 2018 Autumn Meeting of the Japanese Language Education Society*, pp. 29-38.

Mizuno T. (2015). Kuraudosōshingu wa shitsugyōmondai o kaiketsu dekiru ka [Can crowdsourcing solve unemployment?]. *National Institute of Informatics - NII Today*, No. 70, pp.12. Retreived December 30, 2019, from https://www.nii.ac.jp/about/publication/today/70-5.html

Muthoo, A. (1999). *Bargaining theory with applications*. Cambridge: Cambridge University Press.

Tamura S. (2014). *Kuraudosōshingu no hōritsu mondai (3) Mi wo mamoru tame ni [Legal problems with crowdsourcing (3) In order to protect oneself]*. Retreived June 17, 2019, from http://ashitaba-tam.hatenablog.com/entry/2014/04/26/013829

Newspaper Articles

Anonymous. (2019, Nov. 17). Netto tanpatsu rōdō jiyū to fuan to [One-off internet work: freedom and anxiety]. *Asahi Shinbun morning edition*, p.4.

Anonymous. (2019, Dec. 6). “Hataraki-kata kaikaku no shikaku”: kyodai IT ni ‘ko’ kusen - Ūbā-īitsu haitatsu-in dankō monzenbarai [“Blind Spot of Work Style Reform”: Difficult struggle of ‘individuals’ against giant IT - collective bargaining turned away]. *Tokyo Shinbun morning edition*. Retrieved Jan. 3, 2020, from https://www.tokyo-np.co.jp/article/economics/list/201912/CK2019120602000124.html

Irwin, N. (2019, Sep. 15). Maybe We’re Not All Going to Be Gig Economy Workers After All: Companies like Uber are hitting the turbulence of government regulation, worker resistance and labor market reality. *New York Times*. Retrieved Jan. 3, 2020, from, https://www.nytimes.com/2019/09/15/upshot/gig-economy-limits-labor-market-uber-california.html

Appendix: Examples of relatively good and bad order placements

Example CW ID:315154 (*comprehensive impression:* medium-high)

時間単価制	3,000円～4,000円
稼働時間／週	5時間／週
期間	1週間以内

仕事の詳細

Vagrant, Chef の基本、余裕があれば応用までマンツーマンでお教えいただける方を募集します。カリキユラムですとか、資料ですとか、その手の準備は不要です。

概念を教えてもらうというよりはその場で動作を試し試し習得していくような形式を望みます。

こちらから五月雨に質問していきますので逐次答えていただきたいです。

当方の環境はWindows 7 Home Premium。VirtualBox上にCentOS 6.6 64bitが入ってます。土日祝日の都合の合う時間に5時間程度、都内(可能であれば新宿近辺)の電源付きのカフェまたはコワキングスペス等での打ち合わせ(お仕事)を希望します。

▽ 使用するプログラミング言語/ツール/特殊技術(テクノロジー)

・Vagrant/Chef/その他有用なものがあれば

▽ 重要視する点・開発経験

・Vagrant/Chef/周りの浅い質問に対してある程度サクサク答えられること(その場でグクってもらって結構です)

▽ その他コメント

※実際のお仕事の進め方・別途詳細は、クラウドワークスのメッセージでやりとりして決められればと思います。多数のウェブエンジニア・プログラマの方からのご連絡・ご応募お待ちしております。

Example CW ID:1541997 (*comprehensive impression:* low)

タスク	8円／件
募集件数	120件

仕事の詳細

PCでの検索結果のI順位を報告して頂くお仕事です!

【こちらで指定する検索順位は日々変わりますので、ご注意ください。】

※この仕事は全国の各地域サーバーによって、どれだけ順位に差があるのかを調査する為、募集させて頂いております。サイトへの誘導・宣伝及び特定のワードで検索をして頂く事自体が目的ではありません。

ORTHOEPIC COMPETENCE DESCRIPTORS IN JAPANESE LANGUAGE EDUCATION: CEFR LEVELS B1 TO C2

ITO Hideaki
University of Tsukuba, Japan
ito.hideaki.gb@u.tsukuba.ac.jp

Abstract

The Council of Europe's 2001 Common European Framework of Reference for Languages (CEFR) has shown rapid global adoption, and now includes Japanese language education though it primarily aimed at alphabetically transcribed languages. It basically acknowledges that orthoepic competence relates to comprehension of characters yet does not indicate descriptors. Descriptors examining A1 and A2 levels, using altered techniques, have already been set. In this paper, I re-examine descriptors for levels B1 to C2, which have not yet been attempted, and combine them with the results for levels A1 and A2 to present descriptors for levels A1 to C2 in overall.

Keywords: orthoepic competence; letter knowledge; CEFR; co-occurrence relations

Povzetek

Skupni evropski referenčni okvir Sveta za jezike iz leta 2001 je bil precej hitro sprejet globalno, tudi na prodročju poučevanja japonskega jezika, čeprav je bil primarno namenjen latiničnim jezikom. Okvir predvideva, da se ortopska kompetenca nanaša na razumevanje znakov, vendar ne navaja konkretnih deskriptorjev. Deskriptorji, ki preučujejo ravni A1 in A2 z uporabo spremenjenih tehnik, so že nastavljeni. V tem prispevku preučim deskriptorje na ravni med B1 in C2, ki še niso bili preskušeni, ter ugotovitve združim s tistimi na ravneh A1 in A2. Na koncu v celoti predstavim deskriptorje za vse ravni od A1 do C2.

Ključne besede: ortopska kompetenca; poznavanje črk; SEJO (CEFR); sočasni odnosi

1 Introduction

The Common European Framework of Reference for Languages (CEFR hereafter), announced in 2001 by the Council of Europe, has spread quickly throughout the world's language education community. Today, it is in the process of becoming a global language standard. As a European framework, the CEFR has particularly been

Acta Linguistica Asiatica, 10(1), 2020.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.10.1.49-66

influential within European language education. Regardless of the stage of education, various European educational organizations now utilize CEFR criteria to determine course levels and degree of language ability a student should achieve upon graduating. Furthermore, the CEFR is getting accepted within Asia (including Japan) as a language education standard (Cheng, 2017). In case of Japanese language education, the expansion of its adoption has centered on the JF Standard for Japanese-Language Education (developed by the Japan Foundation) and draws upon CEFR benchmarks, which are now becoming commonplace. One example is the use of six common reference levels to describe language proficiency (Majima, 2018; Ito, 2019a). Nevertheless, the works of Meyer (2010) and North (2014) raise the problem of character comprehension when using the CEFR in Japanese language education.[1] North argues:

> In the context of current pedagogy for Japanese and Chinese it is not possible for a learner at A2 or B1 or B2 to read the types of text that appear in CEFR descriptors for the levels concerned, simply because they do not know enough signs. (p. 45)

North (2014) proceeds to suggest that:

> …using the CEFR for such languages implies either *profiling* proficiency, admitting that such learners are a higher level for listening and speaking than they are for reading and writing – which the CEFR scales will facilitate describing – or alternatively, developing completely new descriptors for reading and writing. (p. 45)

That is to say, in North’s view, in particular cases there is a need to formulate entirely new descriptors for reading and writing. Länsisalmi (2012) and Shigemori Bučar, Ryu, Moritoki Škof, & Hmeljak Sangawa (2014) also point out the character/kanji issue as one of the difficulties in using the CEFR framework in Japanese teaching. Furthermore, when conducting task-based testing within Japanese language education, insufficient writing ability can cause problems with respect to completing and evaluating tasks (Kumano, Ito, & Hachisuka, 2013). If the CEFR is to be used within Japanese language education, then it is necessary to resolve issues regarding the CEFR and orthoepic competence. According to Ito (2017), the CEFR acknowledges the existence of orthoepic competence but does not indicate relevant descriptors. As a means of engaging with this problem, Ito drafted some orthoepic competency descriptors that are graded according to proficiency. However, the initial methodology for creating this draft proposal relied upon a subjective evaluation of characteristic words; Ito (2019a) thus re-examined the draft plan for the A1 and A2 orthoepic

[1] With respect to the Chinese Language Proficiency Test (HSK) being redesigned in 2010 on the basis of the CEFR, Meyer (2010) raises an issue regarding Chinese language education. He argues that the required number of words is too low. North (2014) subsequently expands on this point to advance a similar view in terms of Japanese language education.

competency descriptors. In this paper, I present the results of my own attempt at re-exploring levels B1 to C2, which Ito (2019a) did not investigate. Furthermore, by combining these findings with those of Ito (2019a), I show how I have been able to form orthoepic competence descriptors for levels A1 to C2.

The structure of this paper is as follows. Section 2 outlines the six stages of the CEFR's common reference levels and defines orthoepic competence. Section 3 further clarifies the research goals after reviewing prior studies related to orthoepic competence in the CEFR. Subsequently, Section 4 explains the method of analysis and its results, while Section 5 proposes some considerations drawn from those findings. Finally, Section 6 summarizes this paper and points out topics for future research.

2 Common Reference Levels and Orthoepic Competence

2.1 The Six Stages of Common Reference Levels

The Council of Europe (2001) presents six common reference levels (Figure 1), stating that: "It seems that an outline framework of six broad levels gives an adequate coverage of the learning space relevant to European language learners for these purposes" (p. 23). These common reference levels are arranged in tiers from A1 (low language proficiency) to C2 (high language proficiency). A1 or A2 indicates a "Basic User"; B1 or B2 represents an "Independent User"; and C1 and C2 signal a "Proficient User" (Ito, 2019a). However, depending on the context of its application, the CEFR also recognizes smaller divisions of B1 such as B1.1 and B1.2. As Ito (2019b) points out, there are changes in the breadth of ability expected and the abilities that are emphasized for each level. From the above, we can see that A1, A2, B1, B2, and C1, C2 do not actually express sharply delineated stages. Rather, it is more appropriate to view them as rendering the level-like nature of language proficiency into a more easily understandable form.

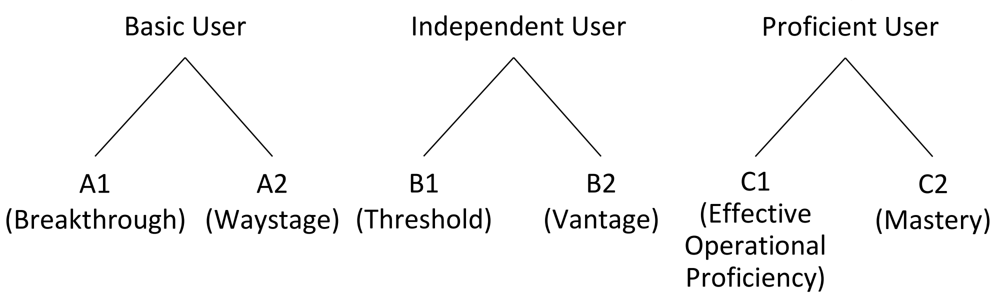


Figure 1: The Common Reference Levels (Council of Europe, 2001, p. 23)

2.2 Orthoepic Competence

The CEFR defines communicative language skills as “those which empower a person to act using specifically linguistic means” (Council of Europe, 2001, p. 9). It divides these abilities into three subgroups: (1) linguistic, (2) sociolinguistic, and (3) pragmatic. The first subgroup is composed of general linguistic range, vocabulary range, vocabulary control, grammatical competence, phonological competence, orthographic competence, semantic competence, and orthoepic competence.

With respect to orthoepic competence (the focus of this paper), the CEFR states that: “users required to read aloud a prepared text, or to use in speech words first encountered in their written form, need to be able to produce an accurate pronunciation from the written form” (Council of Europe, 2001, p. 117). Bellassen and Zhang (2008) define this skill slightly differently, as “the ability of the language user to accurately read aloud and pronounce a text or speech in a loud voice” (语言使用者在大声朗读文章或演讲稿时面对书写形式而体现出的朗读和发音能力) (p. 68).

After confirming the CEFR and the observations of Bellassen and Zhang (2008) with regard to orthoepic competence, Ito (2019a) advances the definition by describing it as “the skill of reading text or characters…aloud.” Furthermore, if we read Ito (2017), we find that the author’s understanding of “reading” as orthoepic competence is “the ability to pronounce vocabulary or characters within a text, and to understand the meaning of words or characters within context, and with respect to the function of how they are written” (p. 56). Ito (2019a) recognizes this sense, writing that “if we take into account the existence of kanji within Japanese, then it makes sense to consider not only pronunciation but also the meaning of words” (p. 75). However, Ito (2019a) considers that “being able to read” implies the existence of an ambiguous continuity in the space between pronunciation and the understanding of a word’s meaning. On this basis, Ito (2019a) proposes a more detailed definition, asserting that “being able to read” is a “state” or “condition” whereby “an individual can pronounce characters or vocabulary, and is at least partially capable of understanding the meaning of characters or words within context, as well as the function of how they are written” (p. 76). Ito (2019a) states that this definition is not in conflict with the former one (2017). Therefore, in this paper, I use Ito’s (2019a) definition of orthoepic competence as “being capable of reading.”

3 Previous Studies and Research Goals

A large number of studies on the CEFR have been written for a range of languages. Of these, many have focused on the ability to accomplish tasks, which is a typical CEFR skill (Ito, 2019a). However, previous investigations that have examined the relationship between the characters of non-alphabetic languages and CEFR orthoepic competence are, to the best of my knowledge, limited to Bellassen and Zhang (2008), Ito (2017), and Ito (2019a).

Bellassen and Zhang (2008) investigated the application of the CEFR within Chinese language education in France. They point out that, as with Japanese, there is hardly any relationship between the pronunciation of Chinese characters and the way in which they are written. For this reason, it is nearly impossible to read an unknown Chinese character, furthermore, a learner may even be led astray by their mistaken interpretation for a similarly looking characters. Bellassen and Zhang conclude that characters present a problem in terms of applying the CEFR within Chinese language education. They propose a method for measuring the ability to write Chinese characters, which they refer to as a "literacy threshold" (识字门槛). A literacy threshold selects Chinese characters based on the frequency with which they are written, their prevalence in everyday conversation, the incidence of their use in conversation, and the number of words that can be created when they are joined with other Chinese characters. The selected Chinese characters are then divided into different language proficiency groupings with the ability to recognize and write them, forming the basis for evaluating which level of language proficiency a person has reached (Table 1). The work of Bellassen and Zhang (2008) holds some significance with respect to the problem of the relationship between Chinese character and the CEFR. However, apart from noting that frequency of appearance informs the reason for their selection, authors do not offer any details on their selection of Chinese characters, which raises doubt on whether the set literacy threshold is appropriate. Ito (2019a) points out that the basis for proficiency evaluation is not a concrete descriptor but rather an individual kanji, and thus the design of the literacy threshold does not align with the thinking of the CEFR, according to which language users are understood as "social agents" that strengthen and improve their abilities by carrying out tasks within certain set environments.

Table 1: Literacy Threshold Levels and the Number of Chinese characters according to Bellassen & Zhang, 2008, p. 69 (author's translation)

Level	Number of Chinese characters (approximately)
C2	Over 3,000
C1	2,200
B2	1,500
B1	800
A2	500
A1	250

Furthermore, Ito (2017) suggests that the CEFR should not provide a concrete descriptor for orthoepic competence because the CEFR is not something complete and static, but is rather distinguished by an orientation toward the possibility of continual expansion and refinement. Ito (2017) indicates the Council of Europe's own language in this regard by pointing to the statement that: "the framework should be open and flexible, so that it can be applied, with such adaptations as prove necessary, to particular situations" (Council of Europe, 2001, p. 7). Ito (2017) further asserts that the creation of a descriptor for orthoepic competence is an urgent task, such that it is necessary to provide a concrete descriptor now, even if an outcome is in a form that leaves room for future debate. Using this understanding, the author advances his own proposal for an orthoepic competence descriptor for Japanese.

The work of Ito (2017) is significant for having provided a concrete orthoepic competence descriptor. However, it would be difficult to claim that the proposed descriptor is highly objective. The method that the author used to draft this descriptor is problematic because the method relies upon a subjective selection of characteristic words of the abovementioned six abilities (ranked by language proficiency), which have the following linguistic competence descriptors (excluding orthoepic and semantic competence): general linguistic range, vocabulary range, vocabulary control, grammatical competence, phonological competence, and orthographic competence.

Ito (2019a) states that "it is necessary to research [orthoepic competence] further in order to ensure that it becomes a more objective measure" (p. 78). Ito used a highly objective approach to choose words for each level of language proficiency, which he (2017) originally selected arbitrarily by employing the text mining technique, along re-considering the orthoepic competence descriptors for levels A1 and A2. Table 2 shows the results of this updated list of descriptors. Of the orthoepic descriptors originally given by Ito (2017), only the A1 and A2 descriptors have been replaced with his revisions (2019a).

Therefore, the purpose of this paper is to extend the reconsideration conducted at Ito (2019a) from the B1 level to the C2 level and to present new orthoepic competence descriptors in Japanese language education extracted in a highly objective.

Table 2: Orthoepic Competence Descriptors that Combine Ito (2017) and Ito (2019a)

Level	Orthoepic Competence Descriptors
C2	The learner can consistently read accurately and is even deeply familiar with difficult kanji outside of his/her field of expertise.
C1	The learner makes few mistakes and can read almost all characters without referring to a dictionary, including vocabulary related to his/her field of expertise.

Level	Orthoepic Competence Descriptors
B2	The learner misreads hardly any characters due to transference (interference) from his/her native language. While there are characters he/she cannot read due to a lack of orthoepic knowledge, he/she is familiar with enough characters to ensure that he/she does not struggle with communication in daily life.
B1	The learner makes mistakes in terms of mispronunciation and transference (interference) from his/her native language, but he/she is able to read the material with which he/she comes into contact in daily life.
A2	There are many cases where the learner may need to re-read a section of text or reads incorrectly; however, if he/she has the necessary basic vocabulary, then he/she is able to read material encountered in daily situations.
A1	The learner is able to read a section of text if he/she has studied the material and has a basic, concrete, and limited repertoire (words and expressions, etc.) that relates to his/her personal information.

4 Method and Results

4.1 Analytical Method

Similar to Ito (2019a) I used the linguistic competence descriptors indicated in the Japanese language version of the CEFR, including general linguistic range, vocabulary range, vocabulary control, grammatical competence, phonological competence, and orthographic competence. (There are 9 such descriptors for levels B1 and B2, 7 for C1, and 5 for C2.) I carried out the analysis using the text mining free software KH Coder 3.[2] On the first step, I identified words that most often emerged at various levels. For this purpose, I separated descriptors for different levels (from B1 to C2) and looked at the frequency rate of words that appeared at each level (Section 4.2). Next, I implemented the co-occurrence network analysis to determine co-occurrence relationships between words for each respective level (Section 4.3). Finally, I made use of the same Japanese language version of the CEFR as Ito (2019a), so that I could combine my findings with his, thus creating orthoepic competence descriptors for levels A1 to C2.

[2] KH Coder 3 is a text mining software created by Higuchi Koichi. It can carry out various forms of statistical analysis with respect to text-form data. As of July 2019, it has been used in over 2,800 research programs. For details on KH Coder 3, refer to Higuchi (2014).

4.2 Analysis of Word Frequency

In order to determine the frequency rate of the words used in the descriptors for levels B1 to C2, as mentioned previously, I conducted frequency analysis using KH Coder 3 for each proficiency level. Table 3 displays the highest occurring words from ranks 1 to 10 for each respective level. For B1, the first five words (in their order of highest frequency) are: "to state" (述べる), "vocabulary" (語彙), "to hold" (持つ), "state, condition" (状況), and "accurate" (正確). For B2, they are: "vocabulary", "high" (高い), "general" (一般), "to make a mistake" (間違う), and "to use" (駆使). For C1, we have: "expression" (表現), "vocabulary", "to say" (言う), "language" (言語), and "mistake" (誤り). Finally, for C2, the words are: "language", "vocabulary", "extremely" (非常), "expression", and "broad" (幅広い). However, since none of these words has a significantly high rate of occurrence, it is difficult to determine the characteristics for each level based on word frequency alone. Therefore, as we can see in the subsequent section, I used co-occurrence network analysis, which focuses on the words displayed in Table 3.

Table 3: Top Ten High Frequency Rankings for Words in Levels B1 to C2

	Level B1	Fr.	Level B2	Fr.	Level C1	Fr.	Level C2	Fr.
1	述べる (to state)	6	語彙 (vocabulary)	3	表現 (expression)	4	言語 (language)	2
2	語彙 (vocabulary)	4	高い (high)	3	語彙 (vocabulary)	3	語彙 (vocabulary)	2
3	持つ (to hold)	4	一般 (general)	2	言う (to say)	2	非常 (extremely)	2
4	状況 (state, condition)	4	間違う (to make a mistake)	2	言語 (language)	2	表現 (expression)	2
5	正確 (accurate)	4	駆使 (to use)	2	誤り (mistake)	2	幅広い (broad)	2
6	内容 (contents)	4	見る (to see)	2	些細 (unimportant)	2	あいまい (vague)	1
7	言語 (language)	2	言う (to say)	2	使用 (to use)	2	コノテーション (connotation)	1
8	考え (thought)	2	構造 (structure)	2	正確 (accurate)	2	モニター (monitor)	1
9	使う (use)	2	持つ (to hold)	2	イントネーション (intonation)	1	レパートリー (repertoire)	1
10	問題 (problem)	2	自分 (self)	2	ニュアンス (nuance)	1	意識 (consciousness)	1

4.3 Co-occurrence Network Analysis

Co-occurrence network analysis explores the relationship or strength of co-occurrence between words, which is then expressed as a network diagram. In a network diagram, each word is depicted as a circle, with the size of the circle representing the number of times the word appears. Furthermore, the existence (or lack) of lines connecting circles (words) together, as well as the thickness of those lines, embodies relationships between words and the strength of their co-occurrence. Co-occurrence network analysis is the strongest technique within the field of text mining (Ushizawa, 2018).

This section describes extracting the respective characteristic words of all levels from B1 to C2 using the co-occurrence network analysis. However, Ito (2019a) points out that:

> In order to accurately ascertain the characteristics of each level, it is desirable to focus on relationships and co-occurrence between high frequency words only. However, if we are only looking at a small number of words when conducting this analysis of high frequency words, then we may end up with an analysis of a region defined by extremely limited relationships and co-occurrence. (pp. 71–72)

According to the definition of Ito, the network diagrams consist of those constellations in which a minimum of three words had thick lines, expressing a strong relationship or co-occurrence. I followed this model. For example, in the case of B1, I only carried out co-occurrence network analysis for words with a frequency rate of six.

In cases where the network diagram did not fit this condition, I expanded the range to include words with a frequency rate that was 1 point lower. According to this approach, I determined the network diagrams used as the definition of analysis. With respect to the results of the co-occurrence network analysis for levels B1 to C2, each network diagram from B1 to C1 ended up containing words with a frequency rate as low as 2. On the other hand, level C2 only has a small number of descriptors (five), which was too few to create a co-occurrence network diagram. For this reason, I limited co-occurrence network analysis to levels B1 to C1. Regarding level C2, Ito (2019b) examined visual reception in the CEFR and commented on C2 level competence. I therefore referred to Ito (2019b) for this area while attempting to make my own slight revisions. In the next section, I present my findings from having utilized the word groupings attained through co-occurrence network diagrams (in terms of characteristic words for each level), and of having re-considered orthoepic competence descriptors for each level.

5 Considerations

In this section, I outline the findings of my re-consideration of orthoepic competence descriptors for levels B1 to C2, based on the outcomes of the co-occurrence network analysis and the results of Ito (2019b).

5.1 Orthoepic Competence Descriptors for Level B1

With respect to level B1, Ito (2017) proposes the following orthoepic competence descriptor: "The learner makes mistakes in terms of mispronunciation and transference (interference) from his/her native language, but he/she is able to read the material with which he/she comes into contact in daily life." Indeed, just as Ito (2017) states that the learner "is able to read the material with which he/she comes into contact in daily life," we can infer from the group in Figure 2 – comprised of "vocabulary" (語彙), "daily" (日常) and "topic" (話題) – that individuals at level B1 can "read *vocabulary* [associated with] *daily topics*".

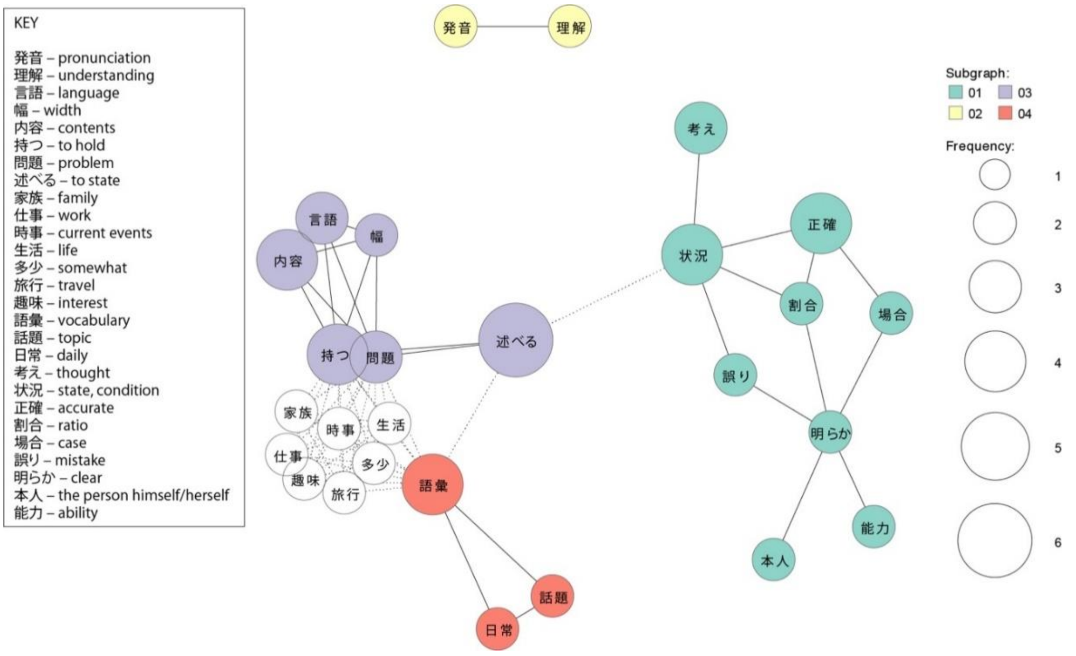


Figure 2: B1 Co-occurrence Network

In the case of the group centered on "to state" (述べる), "to state" does have a relationship with oral competence. However, "to hold" (持つ) has a relationship of co-occurrence with "to state," and "to hold" has a relationship of co-occurrence with "contents" (内容), "width" (幅), and "language" (言語). We can hence surmise that at this level, the individual can "state" or "speak" (述べる) about a "broad" (幅広い) range of "topics" (内容). Furthermore, if we look to the group centered on "state, condition" (状況) and "accurate" (正確), we find a relationship between "state, condition," "ratio" (割合), and "accurate." "State, condition," "ratio," and "mistake" (誤り) also have a co-occurrence relationship, while "ratio" and "mistake" have a co-occurrence relationship with "clear" (明らかな). Based on this analysis, I propose changing the B1 orthoepic competence descriptor to the following: "While there are cases where [learners] may make obvious (明らかな) mistakes (間違い), they are able to relatively accurately (割合正確に) read characters related to a broad range of material (幅広い内容), including everyday topics (日常的な話題)" (Table 4). The resulting change of the section "there are cases where [learners] may make obvious mistakes" (明らかな間違いを犯す場合もあるが) represents a revision toward a slightly higher degree of abstraction when compared to that of Ito (2017): "The learner makes mistakes in terms of mispronunciation and transference (interference) from his/her native language" (発音の間違えや母語の転移（干渉）などの読み間違えもあるが). If we consider this issue from the perspective of the CEFR's concept of "plurilingualism" – according to which all linguistic knowledge and experience contribute to the development of new communication – then Ito's (2017) comment on errors arising out of "transference (interference) from…native language" is actually due to a mistaken interpretation. Furthermore, as noted in Section 2.2, I use Ito's (2019a) definition of "being able to read." If we assume that mistakes are not restricted to "misreading," then we come closer to an orthoepic competence descriptor that better conforms to the reality of – or rather, a correct understanding of – the CEFR.

Table 4: Revisions to the Descriptor for B1 Orthoepic Competence

Level	Ito (2017)	Revisions based on the current analysis
B1	The learner makes mistakes in terms of mispronunciation and transference (interference) from his/her native language, but he/she is able to read the material with which he/she comes into contact in daily life.	While there are cases where [learners] may make obvious mistakes, they are able to relatively accurately read characters related to a broad range of material, including everyday topics.

5.2 Orthoepic Competence Descriptors for Level B2

Next, if we turn to level B2, Ito (2017) provides the following descriptor:

> The learner misreads hardly any characters due to transference (interference) from his/her native language. While there are characters he/she cannot read due to a lack of orthoepic knowledge, he/she is familiar with enough characters to ensure that he/she does not struggle with communication in daily life.

However, unlike Ito (2017) or the examination of level B1, Figure 3 does not portray a strong co-occurrence relationship between "mistake" (間違う) and other terms.

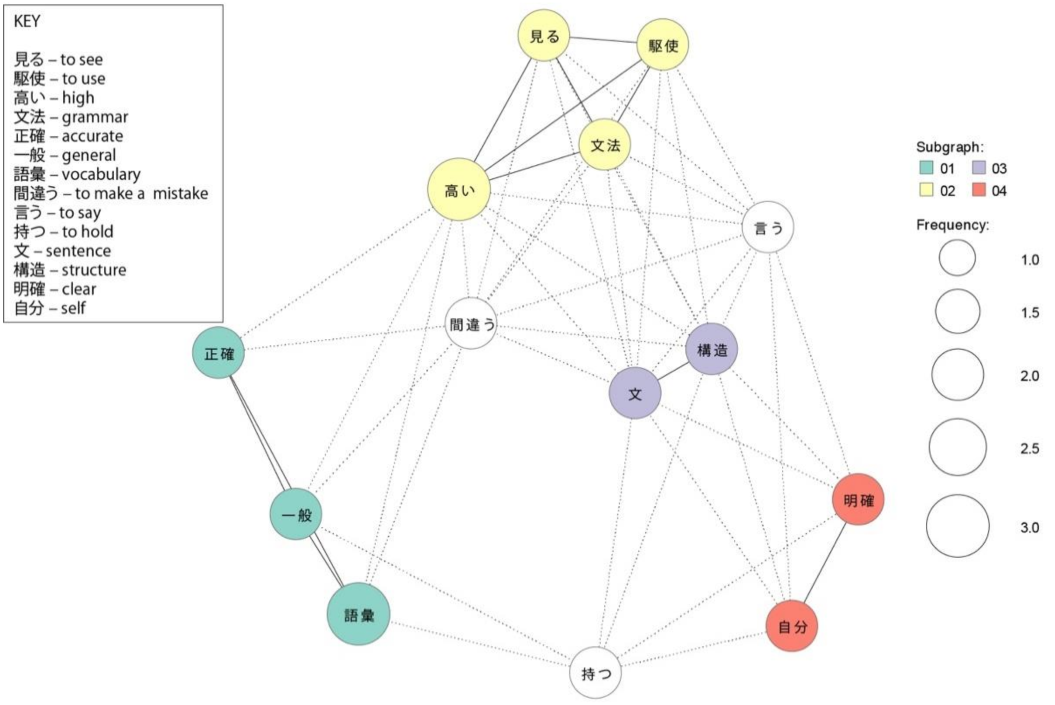


Figure 3: B2 Co-occurrence Network

Furthermore, with respect to "vocabulary" (語彙), in case of level B1, there is a co-occurrence relationship with "everyday topics" (日常的な話題), yet in case of level B2, there are instead co-occurrence relationships with "general" (一般) and "accurate" (正確). Thus, the breadth of vocabulary in level B2 seems slightly more extensive, making it a level where a language learner is "able to accurately use [用いる] words if they are common" (一般的な語について正確に用いることができる);[3] this segment represents another group in level B2 that displays a co-occurrence relationship. In this group, a strong, reciprocal co-occurrence relationship is found between "high" (高い), "to see" (見る), "use" (駆使), and "grammar" (文法). From this observation, we can infer that B2 users possess an extensive command of grammar. I propose that we combine this conclusion with the earlier statement that B2 users can "accurately use words if they are common." Thus, we attain a descriptor of orthoepic competence for this level, whereby the individual "has a high level of orthoepic competence, and is able to accurately read [読む] words if they are common" (高い読字能力を持っており、一般的な語であれば正確に読むことができる).

[3] Regarding this observation about the "expansion of vocabulary" at the B2 level, Ito (2019b), in his investigation of the abilities considered important for visual receptive activities, similarly notes that from B2 onward, "not only the type of reading or the method of reading, but also the depth of understanding of the content receives more emphasis at this stage, with a higher degree of competency expected" (p. 73). There may be a connection with my results.

Table 5: Revisions to the Descriptor for B2 Orthoepic Competence

Level	Ito (2017)	Revisions based on the current analysis
B2	The learner misreads hardly any characters due to transference (interference) from his/her native language. While there are characters he/she cannot read due to a lack of orthoepic knowledge, he/she is familiar with enough characters to ensure that he/she does not struggle with communication in daily life.	[The individual] has a high level of orthoepic competence, and is able to accurately read words if they are common.

5.3 Orthoepic Competence Descriptors for Level C1

Ito (2017) offers the following descriptor for level C1: “The learner makes few mistakes and can read almost all characters without referring to a dictionary, including vocabulary related to his/her field of expertise.” Figure 4 displays a group centered on “expression” (表現) and one centered on “mistake” (誤り), “unimportant” (些細), and “accurate” (正確). Of these, the “mistake,” “unimportant,” and “accurate” group shares a commonality with the section of Ito’s (2017) descriptor, which states that the individual “makes few mistakes” (literally “mistaken readings,” 読み間違え).

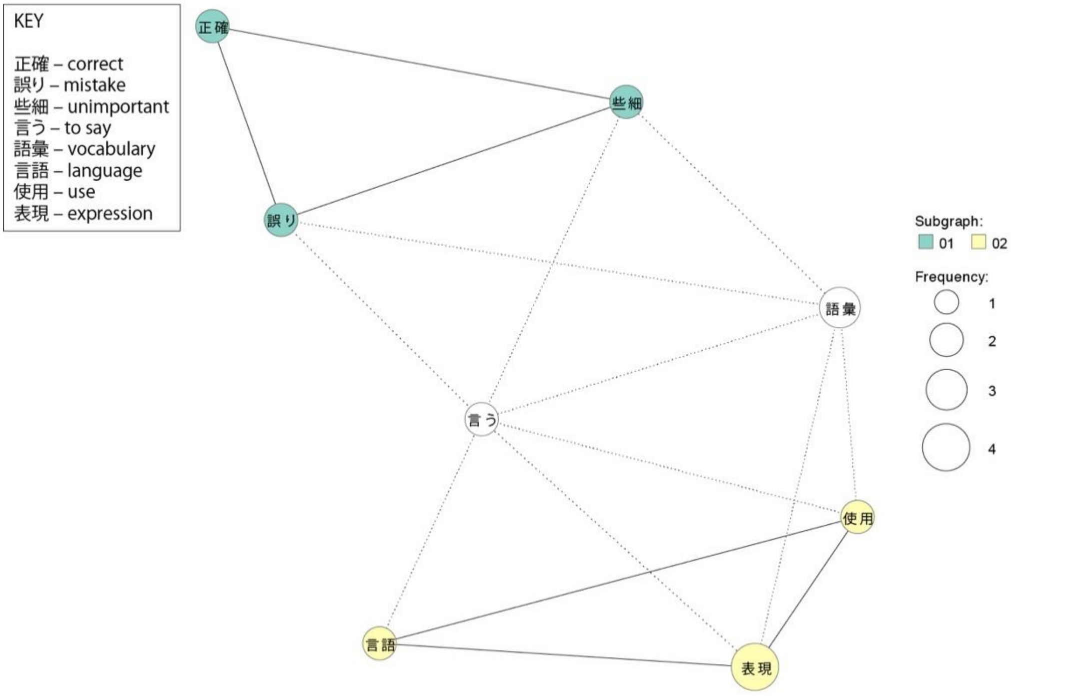


Figure 4: C1 Co-occurrence Network

Furthermore, “mistake,” “unimportant,” and “vocabulary” (語彙) are connected via a weak co-occurrence relationship. This connection points at users making “slight” (i.e., unimportant) “mistakes” at the level of “vocabulary.” On the other hand, as with our observation regarding level B1, “mistakes” are not limited to “mistaken readings.” For this reason, I suggest that Ito (2017)’s section of “makes few mistakes” be revised to “makes slight mistakes at the level of vocabulary, but can read accurately.” Regarding the group centered on “expression,” “expression,” “language” (言語), and “use” (使用) are connected through a strong co-occurrence relationship. Meanwhile, “to speak” (言う) is connected to the three abovementioned words, and “vocabulary” has a weak co-occurrence relationship with “expression” and “use.” We do not see keywords such as “field of expertise” or “dictionary” given by Ito (2017). However, in terms of C1 visual reception, Ito (2019b) notes that the framework emphasizes a language learner’s ability to have “a detailed understanding of opinions (etc.) regarding complex, difficult texts related to [his/her] own [field of] specialty” (自分の専門) (p. 74). We can surmise that “field of expertise” (専門分野) is an important part of C1. Hence, I altered this section to read: “can accurately read linguistic expressions and vocabulary such as [those] used in [his/her] field of expertise.” Combining this section with the above mentioned first part, my suggested reform of the orthoepic competence descriptor for C1 now reads as follows: “While [learners] may make slight mistakes at the level of vocabulary, they can accurately read linguistic expressions and vocabulary, such as [those] as used in their field of expertise” (語彙レベルの些細な誤りもあるが、専門分野などで使用される言語表現や語彙を正確に読むことできる).

Table 6: Revisions to the Descriptor for C1 Orthoepic Competence

Level	Ito (2017)	Revisions based on the current analysis
C1	The learner makes few mistakes and can read almost all characters without referring to a dictionary, including vocabulary related to his/her field of expertise.	While [learners] may make slight mistakes at the level of vocabulary, they can accurately read linguistic expressions and vocabulary, such as [those] as used in their field of expertise.

5.4 Orthoepic Competence Descriptors for Level C2

For the C2 level Ito (2017) offers the following descriptor: “A learner can consistently read accurately and is even deeply familiar with difficult kanji outside of his/her field of expertise.” Since I was unable to create a co-occurrence network, I attempted to use an alternative approach to revision. My strategy involves revealing information about C2 orthoepic competence by looking more closely at the capacity valued with respect to C2 visual reception, as outlined by Ito (2019b). According to Ito (2019b), the skill valued in terms of C2 visual reception is the language learner’s ability to “understand

and savor complex meanings in sources such as literature" (pp. 74–75). First, since "understanding complex meanings in sources such as literature" is expected, we can revise the section "outside of his/her field of expertise" by adding "such as literature" (文学など), which would be more suitable. Being able to read "difficult kanji" is not necessarily required for "understanding and savoring complex meanings in sources such as literature." However, since "difficult kanji" (literally "kanji difficult to read" or 難読漢字) is indicated, and because I was unable to conduct a detailed analysis into the C2 level, I deemed it inappropriate to make large changes. Therefore, regarding a modification to C2 orthoepic competence, I propose the following: "The learner can consistently read accurately, and can besides read difficult kanji such as those found in literature" (一貫して正しい読みをし、文学などで使われる難読漢字なども読むことができる).

Table 7: Revisions to the Descriptor for C2 Orthoepic Competence

Level	Ito (2017)	Revisions based on the current analysis
C2	The learner can consistently read accurately and is even deeply familiar with difficult kanji outside of his/her field of expertise.	The learner can consistently read accurately, and can besides read difficult kanji such as those found in literature.

5.5 Revised Orthoepic Competence Descriptors

This concludes the overview of my revisions using co-occurrence network analysis regarding the orthoepic competence descriptors for levels B1 to C2, as provided by Ito (2017). Table 8 lists the revised orthoepic competence descriptors for levels B1 and C2 as outlined in this paper, alongside the re-considered orthoepic competence descriptors for levels A1 and A2 advanced by Ito (2019a).

Table 8: Revised Orthoepic Competence Based on Current Research (B1-C2) and Ito (2019a).

Level	Revised Orthoepic Competence
C2	The learner can consistently read accurately, and can besides read difficult kanji such as those found in literature.
C1	While [learners] may make slight mistakes at the level of vocabulary, they can accurately read linguistic expressions and vocabulary, such as [those] as used in their field of expertise.
B2	[The individual] has a high level of orthoepic competence, and is able to accurately read words when common.
B1	While there are cases where [learners] may make obvious mistakes, they are able to relatively accurately read characters related to a broad range of material, including everyday topics.

Level	Revised Orthoepic Competence
A2	There are many cases where the learner may need to re-read a section of text or reads incorrectly; however, if he/she has the necessary basic vocabulary, then he/she is able to read material encountered in daily situations.
A1	The learner is able to read a section of text if he/she has studied the material and has a basic, concrete, and limited repertoire (words and expressions, etc.) that relates to his/her personal information.

6 Conclusion and Future Tasks

Building upon the earlier work of Ito (2017, 2019a) – that is, the creation of suggested orthoepic competence descriptors for the CEFR, as well as a re-investigation of levels A1 and A2 – I re-examined the descriptors for levels B1 to C2. My findings did not return characteristic words determined on the basis of frequency of appearance. However, by utilizing co-occurrence network analysis, I was able to extract co-occurrence networks for levels B1 to C1. By re-scrutinizing Ito's (2017) proposal for orthoepic competence descriptors, I was able to identify a more objective method than that utilized by Ito (2017). Employing this technique, I outlined revised orthoepic competence descriptors for levels B1 to C1. Regarding C2, a small number of descriptors for this level meant that co-occurrence network analysis was unfeasible. For this reason, I only made out light changes grounded in inferences drawn from the competences valued for visual reception, as shown by Ito (2019b). By combining my outcomes with those of the re-examination conducted by Ito (2019a) into levels A1 and A2, I was able to present orthoepic competence descriptors that are more objective than those initially put forth by Ito (2017).

Nevertheless, a number of issues remain that should be addressed in future research. First, as noted above, since co-occurrence network analysis was not possible for the C2 descriptors, it is similar to a tentative plan. Furthermore, the B1 level descriptor was altered from "The learner makes mistakes in terms of mispronunciation and transference (interference) from his/her native language" to "there are cases where [learners] may make obvious mistakes." This modification represents a higher degree of abstraction, and may be considered an orthoepic competence descriptor that more closely aligns with the reality of – or, rather, a correct understanding of – the CEFR. Nevertheless, at the same time, we cannot deny the possibility that for B2, the resulting descriptor – "[The individual] has a high level of orthoepic competence, and is able to accurately read words if they are common" – may in fact be too abstract. Hence, some further consideration as to whether this descriptor can be used in its current formn is needed. Going forward, relying on the analysis of descriptors alone may not be appropriate for further investigating problems such as the above (that is, the inability to conduct analysis for C2 due to a small number of descriptors, and the feeling that the descriptor for B2 is too abstract). Concerning the orthoepic competence descriptors presented in this paper, it will be necessary to focus on administering a Japanese language test and self-evaluation questionnaire to Japanese language learners in order to thoroughly comprehend the relationship between orthoepic competence descriptors and Japanese language ability. At present I do not have a concrete research plan; I can only offer these orthoepic competence descriptors. Future studies should explore the relationship between Japanese language ability and the self-evaluation of orthoepic competence.

References

Bellassen, J., & L. Zhang. (2008). Ou zhou yu yan gong tong can kao kuang jia xin li nian duihan yu jiao xue de qi shi yu tui dong– chu yu jue ze guan tou de han yu jiao xue [The CEFR: New concept and its implications and impetus – Chinese language teaching at its critical moment]. *Shi jie han yu jiao xue [Chinese Teaching in the World]*, *85*, 58–73.

Chéng, Y. (2017). *Chūka sekai ni okeru CEFR no juyō to bunmyakuka [The introduction and contextualization of CEFR in the Chinese speaking world]*. Tokyo, Japan: Coco Publishing.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge, UK: Cambridge University Press.

Higuchi, K. (2014). *Shakaichōsa no tameno keiryō tekisuto bunseki—naiyō bunseki no keishō to hatten o mezashite [Quantitative textual analysis for social surveys: Aiming for the succession and development of content analysis]*. Sapporo, Japan: Nakanishiya Publishing.

Ito, H. (2017). Kakuchō seichika no tameno dokuji nōryoku kijutsubun shian sakusei -CEFR/JFS no gengo kōzōteki nōryoku o sankō ni- [Devising tentative descriptors of orthoepic competence for extension and refinement: With reference to CEFR/JFS linguistic competence]. *Nihongo kyōiku [Journal of Japanese Language Teaching]*, *168*, 55–62.

Ito, H. (2019a). Dokuji nōryoku no hyōka shakudo no saikō- “kiso dankai no gengo shiyōsha” ni chūmoku shite- [Reconsidering can-do descriptors for orthoepic competence with a focus on the CEFR’s “Basic Language User]. *Kiso kyōiku hoshōgaku kenkyū [The Journal of the JASBEL], 3*, 72–86.

Ito, H. (2019b). CEFR no juyōteki katsudō dewa naniga jûshi sareteirunoka- shikakutekina juyōteki katsudō no kaku reberu no bunseki- [What is important in receptive activities of CEFR? An analysis of visual receptive activities for different levels]. *Nihongo kyōiku [Journal of Japanese Language Teaching], 173*, 69–76.

Kumano, N., Ito, H., & Hachisuka, M. (2013). JFS/CEFR ni motozuku JFS nihongo kōza reberu nintei shiken (A1) no kaihatsu [Level certification test (A1) development for JFS Japanese-language courses based on JFS/CEFR]. *Kokusai kōryū kikin nihongo kyōiku kiyō* [*The Japan Foundation Japanese-Language Education Bulletin*], *9*, 73–88.

Länsisalmi, R. (2012). Conquering the world with “cutting-edge curricula”: Global Citizens learning East Asian Languages. In D. Smakman & L. Willemsen (Ed.), *Proceedings of the 2012 “Van Schools tot Scriptie” Colloquium* (pp. 99-111). Leiden, Netherlands: University Library, Leiden University.

Majima, J. (2018). CEFR no kokunaigai no nihongo kyōiku eno inpakuto [The domestic and international impact of CEFR on Japanese language education]. In H. Sensui (Ed.), *Kotoba o oshieru kotoba o manabu fukugengo fukubunka yōroppa gengo kyōtsū sanshōwaku [Teaching language, learning language: Language education and the Common Framework of Reference for Languages for a multi-lingual, multi-cultural Europe (CEFR)]* (pp. 249–274). Shiga, Japan: Kōrosha.

Meyer, F. (2010, August). *Some considerations on the new HSK. Does it help spreading Chinese as a foreign language (CFL)?* Paper presented at the 16th Meeting on Chinese as a Foreign Language. Zürich, Switzerland: University of Zürich.

North, B. (2014). *The CEFR in Practice.* Cambridge, UK: Cambridge University Press.

Shigemori Bučar, C., Ryu, H., Moritoki Škof, N., Hmeljak Sangawa, K. (2014). The CEFR and teaching Japanese as a foreign language. *Linguistica, 54*(1), 455-469. https://doi.org/10.4312/linguistica.54.1.455-469

Ushizawa, K. (2018). *Yatte miyō tekisuto mainingu jiyū kaitō ankēto no bunseki ni chōsen! [Let's try text mining: Learn free answer questionnaire analysis!]* Fukuoka, Japan: Asakura Shoten.

DEVELOPMENT AND OPERATION RESULTS OF JAPANESE ACCENT PERCEPTION TEST BASED ON E-LEARNING SYSTEM

KIM Yu Young
Dongduk Women's University, South Korea
yuiyu1004@dongduk.ac.kr

Abstract

Main purpose of this study is to find and build a model of Japanese accent perception practice at Japanese lessons in regular classrooms and those based on utilization of network and multimedia. Until now, Japanese phonetic researchers and teachers had to spend a disproportionate amount of time and effort to conduct experiments or test, and to develop the means to process resulting data of the experiments conducted. With the proposed [AJ-phonetic Test] system, accent tests are conducted online. In this respect, Japanese learners can take part in the phonetics experiment in a time and location convenient to them. Also, researchers and teachers can work on the obtained data by using the database available. Moreover, AJ-phonetic test feedbacks provide not only test results but also a more comprehensive data analysis. Results of the 12-year operation of AJ-phonetic test in Korea proved to have a positive effect on students as well as teachers. This paper present general guidelines necessary to conduct the AJ-phonetic test. Firstly, in case of Korean learners of Japanese, a separate set of words accentless pitch and those with accent pitch is needed because the two sets present different difficulty levels for a learner. Secondly, as the number of moras in a word affects the difficulty, we introduced dummy words, which proved successful in controling the level of difficulty and increasing learning effect. Besides, dummy words worked well as a substitute for low tone (frequency) long words as well for long words in general. Thirdly, since there are only few cases with a distinctive long-short opposition in Korean, Korean learners of Japanese have difficulties recognizing accents in words with special moras. Such words therefore need special attention. Finally, the ability of Korean learners of Japanese to recognize accent in Japanese words differs based of the learners' native dialects. Best results were obtained by learners from the Jeonla-do region, followed by those from Gyeongsang-do region, while learners from Seoul had most difficulties in recognizing the accent. From all the above findings we conclude that Japanese accent education is highly necessary, and that in the process of education, both Japanese accent characteristics as well as learners' native dialects should be considered.

Keywords: e-learning; online test; Tokyo word accent; Japanese word accent exercises; Japanese word accent teaching

Acta Linguistica Asiatica, 10(1), 2020.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.10.1.67-86

Povzetek

Glavni namen študije je oblikovati model za utrjevanje zaznavanja japonskega besednega naglasa pri vajah japonščine v klasičnih učilnicah kot tudi v učilnicah, ki temeljijo na uporabi spleta in multimedijev. Do zdaj so morali raziskovalci in učitelji japonske fonetike porabiti nesorazmerno veliko časa in truda za izvedbo in analizo takšnih raziskav ter da so razvili programe za obdelavo rezultatov. S predlaganim sistemom, t.i. AJ-fonetičnim testom, je naglasne teste mogoče izvajati preko spleta. V tem pogledu lahko japonski učenci sodelujejo v fonetičnem poskusu v času in na kraju, ki jim ustreza. Prav tako lahko raziskovalci in učitelji dostopajo do pridobljene baze podatkov. Povratne informacije iz baze AJ-fonetičnega testa ne vsebujejo samo golih rezultatov temveč tudi celovito analizo podatkov. Rezultati 12-letnega izvajanja AJ-fonetičnega testa v Koreji imajo pozitivni učinek tako na študente kot tudi učitelje. V tem prispevku so predstavljene splošne smernice, potrebne za izvedbo AJ-fonetičnega preizkusa. Kot prvo je za korejske učence japonskega jezika potrebna ločena obravnava besed brez naglasnega mesta in tistih z naglasnim mestom, saj ta dva sklopa predstavljata različne stopnje zahtevnosti učenca. Drugič, ker dolžina moraičnih enot vpliva na težavnost prepoznavanja besednega naglasa, smo uvedli tudi brezpomenske besede, ki so se izkazale za uspešne pri nadzorovanju stopnje zahtevnosti in povečanju učnega učinka. Poleg tega so brezpomenske besede delovale kot nadomestek za dolge besede izrazito nizkih tonov (frekvenc) ter za besede s strukturno zahtevnejšimi moraičnimi enotami. Tretjič, ker je v korejščini le nekaj izrazitih primerov z opozicijo dolg-kratek glas, imajo korejski učenci japonščine težave pri prepoznavanju naglasa v besedah, ki vsebujejo posebne moraične enote. Takšne besede zato potrebujejo posebno pozornost. Nenazadnje se sposobnost korejskih japonskih učencev, da prepoznajo naglas v japonskih besedah, razlikuje tudi glede njihovo materno narečje. Najboljše rezultate so dosegali učenci iz regije Jeonla-do, sledijo jim tisti iz regije Gyeongsang-do, medtem ko so imeli učenci iz Seula največ težav pri prepoznavanju japonskega besednega naglasa. Iz vseh zgornjih ugotovitev sklepamo, da je potreba po utrjevanju znanja o japosnekm besednem naglasu zelo visoka ter da je v procesu izobraževanja potrebno upoštevati tako japonske naglasne značilnosti kot tudi materno narečje učencev.

Ključne besede: e-učenje; spletni test; tokijski besedni naglas; vaje japonskega besednega naglasa; poučevanje japonskega besednega naglasa

1 Introduction

Rosenberg (2000) defined e-learning by proposing three important and fundamental criteria of e-learning. First, e-learning should be able to update, store, distribute, and share training or information through the network immediately to give access to the system for the students' benefit and education. Second, it should be delivered to end-users through computers using HTML based standard Internet technologies readily available to the learners. Third, e-learning should focus on understanding and interpreting learning concepts as learning solutions that go beyond the paradigm of traditional learning methods. Also, Rosenberg (2000) set standards for accessibility that enables learners to access education and information remotely, instructor's immediate feedback to learners, and learning solutions that go beyond the traditional paradigm of learning. However, many present studies, including Rosenberg (2000), focus on the criteria of e-learning which aim the development of theories catering learner's perspective. However, we believe that the instructor’s perspective is equally important.

Until now, teachers and researchers of Japanese word accents had to spend a large amount of time and effort to conduct Japanese accent tests and processing the result. Above all, instructors who wanted to conduct such practice or tests have had to gather learners in one place so far, but it was a complicated effort to gather learners in one place at one time. Besides, Japanese accent education has limitations on multimedia equipment and test location, which makes it challenging to train with defined and identified limited resources that were meant to be used in a limited schedule of time.

Studies have already shown that for instructors, e-learning should have a comprehensive academic management system that would manage information such as learner's learning history as well as their progress online (e.g., LMS; Learning Management System). This system enables the instructor to give learners appropriate feedback according to their level and previous achievements, while also improving the content of the textbook and adjusting the difficulty according to the learners’ understandings and reactions. This system works best by building an automated online test system, whereby instructors can effectively invest much of their time spent on non-education, such as exam submission and grading, collecting and classifying result data. This personalized educational tool enables the instructor to maintain learners’ interest in learning, and to reduce the probability of learners becoming frustrated and in a worst scenario dropping out of the program. For all the above reasons, this study developed *The Online Tokyo Accent test Program* [AJ-phonetic Test].

With the AJ-phonetic test we also wanted to meet the requirements of e-learning from the learners’ point of view. Learners who participate in e-learning need feedback and evaluation gauges on the quality and progress, in other words, a device that can maintain a sense of reality about students’ learning in general and individually. Much resembling traditional teaching-learning, a learner is not alone when using e-learning platforms. Within the AJ-phonetic Test system, they can always communicate with an instructor about learning contents, and this can be done through communication means such as bulletin boards (Q and A, Freeboard) or e-mail, and likewise, exchange opinions with virtual classmates learning together or working on a same project. Through this virtual community, a learner can eliminate distance from the instructor even within the e-learning system, and experience a sense of peer-to-peer. Thus it can be said that the system is similar to the effect of learning in a traditional teaching-learning system. While learners can overcome spatial and temporal limitations by using the network, where participating in such an education can be adjusted to their schedule, learners can also follow their progress and achievements, which are available in the system anytime. Learners can besides accurately identify their current knowledge and performance, and set learning goals through an objective evaluation system and analysis function, which raises learning motivation and consequently boosts learning at all stages.

On the other hand, according to the Japanese Education Association (1991) surveying the learning needs of Japanese learners, learners were more interested in speaking skills than anything else. According to the survey results, 59.0% of the learners said that among the four functions of writing, reading, listening, and speaking, the speaking function was the most necessary, and about 38.1% of the learners wanted to be able to speak with natural pronunciation and intonation in the future. This is the second highest response rate, which indicates that many learners strive for fluent and correct pronunciation when learning Japanese.

Among studies on listening ability of Japanese accents, Study Group 3, *Comprehensive Study on Japanese Language in the International Society* (Osamu Mizutani), which was launched in 1994, was designed to examine the reality of Japanese (Tokyo dialect) accent listening. Ayusawa (1997) found that there was a difference in perceiving Japanese accents according to the student learner's native language differences. Besides, the results of the longitudinal study of Ayusawa, Nishinuma and Kawatsu (2000) *Tokyo Accent Listening Test* reported that with 24 subjects, including 13 Koreans, the subjects were divided into three groups in the order of the percent correct. In this study, three Korean learners among eight in the upper group, four Korean learners among eight middle-level group, and six Korean learners among eight subgroups, and the longitudinal results showed that most Korean learners placed in the lower middle level group. This represents the necessity and importance of Japanese accent practice such as this study for Korean Japanese learners.

Ogawara (1997) examined the relationship between Korean learner's pronunciation and listening in the study of the student's accent listening ability. As a result, students with better listening ability tended to have a valid listening standard, and that tendency led to the improvement of pronunciation in those learners.

Choi (2006, pp. 200-201) conducted an experiment on how Korean Japanese learners (Seoul and Busan dialect speakers) and Japanese native speakers responded to each other according to the change in the accent position of each word in Japanese. As a result, learners from Seoul who had no accents tended to find similarities in Korean rhyme, rather than feeling Japanese accents in areas where there was a sharp drop in voice. However, the learner from Busan, with an accent, confirmed that he responded similarly to the Japanese native speaker. In other words, there is a difference in the cognitive ability of foreign languages depending on whether there is an accent in the dialect of the place of origin. Moreover, according to Jeong (2009, p. 234), because of interference of the mother tongue Japanese learners of Seoul dialects who do not have Mora, generate and perceive Japanese by syllable units. And most of all, they have a very weak point in their perception of special phonemes.

The above results suggest that listening tests of Japanese accents with Koreans who are particularly vulnerable to Japanese-accents, can be useful for improving their Japanese proficiency. Therefore, this study intended to develop and operate the Tokyo accent perception test based on the system that meets the criteria of e-learning mentioned above, and then analyze the results.

2 Research design

In this study, we considered concepts of e-learning in previous research, and reflect the instructor's perspective necessary for the development of appropriate student e-learning strategies. Then, based on the concept of e-learning, we developed an online Tokyo accent test, [The AJ-phonetic test].

It is noted that exchanging high and low tones shape Japanese accent. We designed the AJ-phonetic test by introducing e-learning technology with the aim that learners could practice Japanese accent in regular and multimedia classes, or online. We further analyzed its process and results for 12 years and two months, and revealed our findings on what is needed for future Japanese accent tests and education.

In the following sections, we present specifications of Japanese accent test [AJ-phonetic Test], details about the experiment design, and detailed features of the AJ-phonetic test.

2.1 Technical specifications

The AJ-phonetic Test (http://www.japanese.or.kr/phonetic_main.aspx) was designed using the Visual Basic.Net / ASP.Net 2.0 and MS-SQL. The basic structure of AJ-phonetic test is presented in Figure 1.

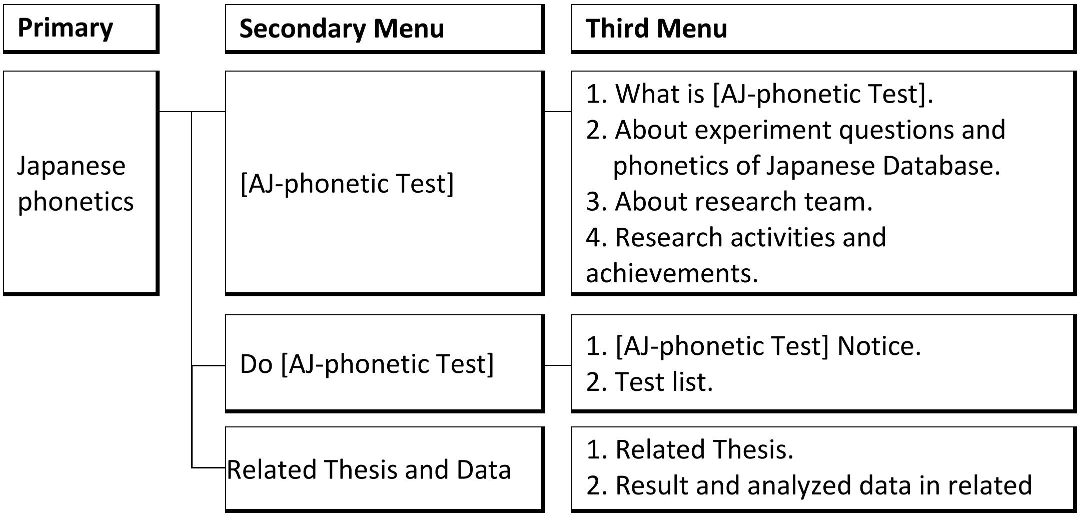

Primary
Secondary Menu
Third Menu
Japanese phonetics
[AJ-phonetic Test]
1. What is [AJ-phonetic Test].
2. About experiment questions and phonetics of Japanese Database.
3. About research team.
4. Research activities and achievements.
Do [AJ-phonetic Test]
1. [AJ-phonetic Test] Notice.
2. Test list.
Related Thesis and Data
1. Related Thesis.
2. Result and analyzed data in related

Figure 1: The basic structure of [AJ-phonetic test]

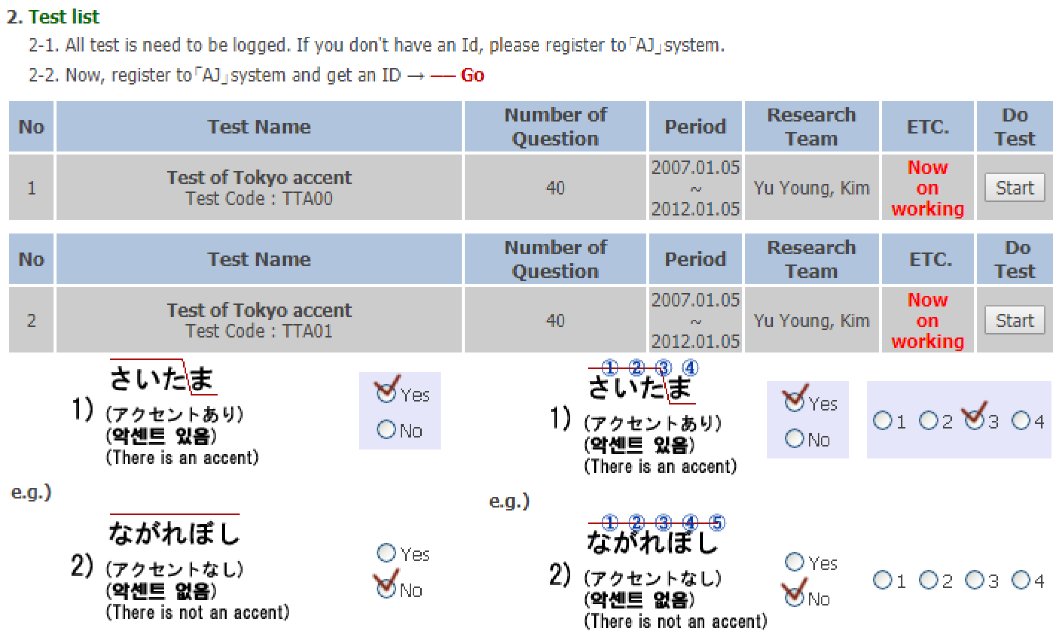
2.2 Research timeline and participants

This research started in October 2007 and lasted until December 2018. Throughout the period of 12 years experiments were taken 8094 times.

Altogether 605 participants were Korean native speakers and students of Japanese (1st to 4th graders) from Dongduk Women's University, Duksung Women's University, Korea University and Konkuk University. They have all gone through basic education on types and pronunciation of Japanese accents and are qualified for the Japanese Language Proficiency Test from N4 to N1. This research included the results to the point when the rate of correct answers reached 90% or above.

2.3 AJ-phonetic Test and its features

The AJ-phonetic Test is available at http://www.japanese.or.kr/phonetic_do.aspx. Figure 2 shows an excerpt from this page.



2. Test list

2-1. All test is need to be logged. If you don't have an Id, please register to「AJ」system.

2-2. Now, register to「AJ」system and get an ID → — Go

No	Test Name	Number of Question	Period	Research Team	ETC.	Do Test
1	Test of Tokyo accent Test Code : TTA00	40	2007.01.05 ~ 2012.01.05	Yu Young, Kim	Now on working	Start

No	Test Name	Number of Question	Period	Research Team	ETC.	Do Test
2	Test of Tokyo accent Test Code : TTA01	40	2007.01.05 ~ 2012.01.05	Yu Young, Kim	Now on working	Start

e.g.)

1) さいたま (アクセントあり) (악센트 있음) (There is an accent) ☑ Yes ○ No

2) ながれぼし (アクセントなし) (악센트 없음) (There is not an accent) ○ Yes ☑ No

e.g.)

1) ①②③④ さいたま (アクセントあり) (악센트 있음) (There is an accent) ☑ Yes ○ No ○1 ○2 ☑3 ○4

2) ①②③④⑤ ながれぼし (アクセントなし) (악센트 없음) (There is not an accent) ○ Yes ☑ No ○1 ○2 ○3 ○4

Figure 2: Web page of the AJ-phonetic test

There are four types of tests available on this page. They are presented in the following subsections.

2.3.1 TTA00 (Test of Tokyo Accent 00)

TTA00 contains 40 words of two or more moras (2-mora-words: 4 items, 3-mora-words: 8 items, 4-mora-words: 12 items/ 5-mora-words: 16 items; see Table 1 for details). Some words are accented and others accentless. Among them are dummy words (e.g., まま, ままま, etc). Participants should check [Yes] when they judge that a word contains accent (phonetically defines as a radical tone drop), or otherwise check the [No] button.

2.3.2 TTA01 (Test of Tokyo Accent 01)

TTA01 is structured similarly to TTA00 but with an additional task. Namely, when participants judge that there is an accent (a radical tone drop) and check the [Yes] button, they should also define the location of the accent place (defines as the last high mora before a tone drop).

Table 1: Word list of TTA00 and TTA01

No.	Word	No.	Word	No.	Word	No.	Word
1	となり	11	ままままま	21	たからもの	31	おしぼり
2	たかとび	12	まままま	22	わすれもの	32	まままま
3	おなまえ	13	まままま	23	ままままま	33	ままままま
4	ままままま	14	ままま	24	まま	34	ながれぼし
5	ねこ	15	あなた	25	みかいはつ	35	だいどころ
6	ままま	16	いす	26	おととい	36	ままままま
7	しごと	17	まままま	27	まま	37	さいたまし
8	まままま	18	おしいれ	28	ままま	38	まままま
9	ままままま	19	だいきらい	29	ままままま	39	ままま
10	みやげもの	20	ままままま	30	もしもし	40	なみだ

2.3.3 TTA02 (Test of Tokyo Accent 02)

TTA02 contains accented 20 words of the length of two or more moras (2-mora-words: 2 items, 3-mora-words: 4 items, 4-mora-words: 6 items/ 5-mora-words: 8 items; see Table 2 for details). Some of the words are dummies (e.g., まま, ままま, etc).

Table 2: Word list of TTA02

No.	Word	No.	Word	No.	Word	No.	Word
1	たかとび	6	ままままま	11	ままままま	16	おしぼり
2	ままままま	7	まままま	12	みかいはつ	17	ままままま
3	ねこ	8	あなた	13	まま	18	ながれぼし
4	ままま	9	まままま	14	ままま	19	さいたまし
5	まままま	10	だいきらい	15	もしもし	20	なみだ

2.3.4 TTA03 (Test of Tokyo Accent 03)

According to Onzuka (2011, p. 247) Korean learners of Japanese have difficulties recognizing the so-called special moras (also called special moraic phonemes in this article). The difficulty of its recognition depends on its word position; most difficulties present a prolonged sound and a syllabic nasal at the word-ending, followed by a mid-positioned geminate consonant and a prolonged sound, while a mid-positioned syllabic nasal is relatively easy to recognize.

Words in the TTA03 test were selected in consideration of this difficulty (Table 3). All 32 words contain more than 2 moras. The task, however, is the same as in the previous test. In this test participants checked [Yes] when they judged that there is an accent (a radical tone drop), and if clicking so, they further defined the location of the accent place (defines as the last high mora before a tone drop).

Table 3: Words list of TTA03

No.	Word	No.	Word	No.	Word	No.	Word
1	しんせつな	11	せんもんか	21	こうぎ	31	おじさん
2	おおぜい	12	へいきんち	22	えほん	32	ゆっくり
3	こっか	13	きんにく	23	きっぷ		
4	さんぽ	14	こんど	24	そうじ		
5	あさって	15	よかったら	25	おかあさん		
6	あんない	16	がっこう	26	どろぼう		
7	みなさん	17	すいようび	27	ひこうき		
8	けっこんしき	18	にほんじん	28	かっこく		
9	アイスティー	19	おめでとう	29	ふかんぜん		
10	きっさてん	20	さとう	30	たかっけい		

Table 4: Special moraic phonemes in TTA03

<table>
<tr><th></th><th>3 Moras</th><th>4 Moras</th><th colspan="3">5 Moras</th><th>6 Moras</th><th>Total</th></tr>
<tr><td>ン(撥音)
A syllabic nasal</td><td>3</td><td>4</td><td>4</td><td rowspan="2">2</td><td>-</td><td>-</td><td>13</td></tr>
<tr><td>ー(長音)
A prolonged sound</td><td>3</td><td>3</td><td>3</td><td rowspan="2">1</td><td>-</td><td>12</td></tr>
<tr><td>ッ(促音)
A geminate consonant</td><td>2</td><td>4</td><td>2</td><td></td><td>1</td><td>10</td></tr>
<tr><td>Total</td><td>8</td><td>11</td><td colspan="3">12</td><td>1</td><td>32</td></tr>
</table>

The following features are common to all of the tests:

- the AJ-phonetic test play voice file of each word twice;
- order of questions for all words is random;
- learners solve problems in a new word order each time;
- voice was recorded and saved at 22kHz, 16Bit;
- words were pronunced by a 30-year-old Tokyo native speaker and a university teacher of Japanese.

The words in each experiment were chosen so that learners can accurately judge and learn the accented listening ability without being disturbed by factors other than their listening ability. The detailed principle will be explained later.

3 Results

From the instructor's perspective, the flow of the AJ-phonetic test was designed as shown in Figure 3.

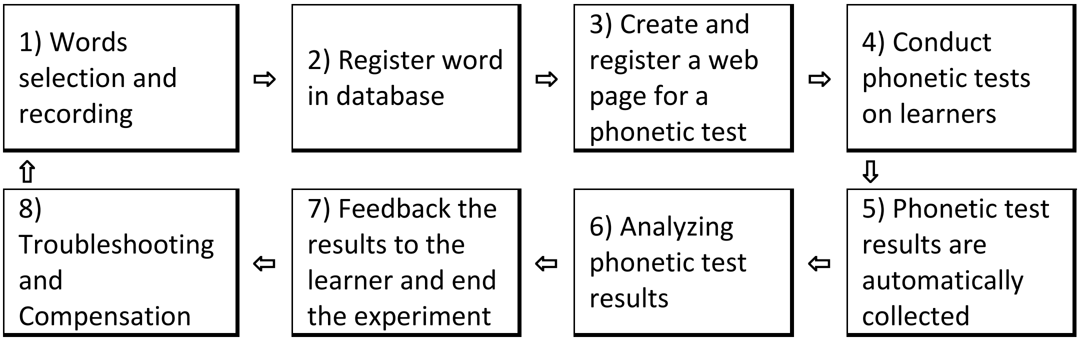
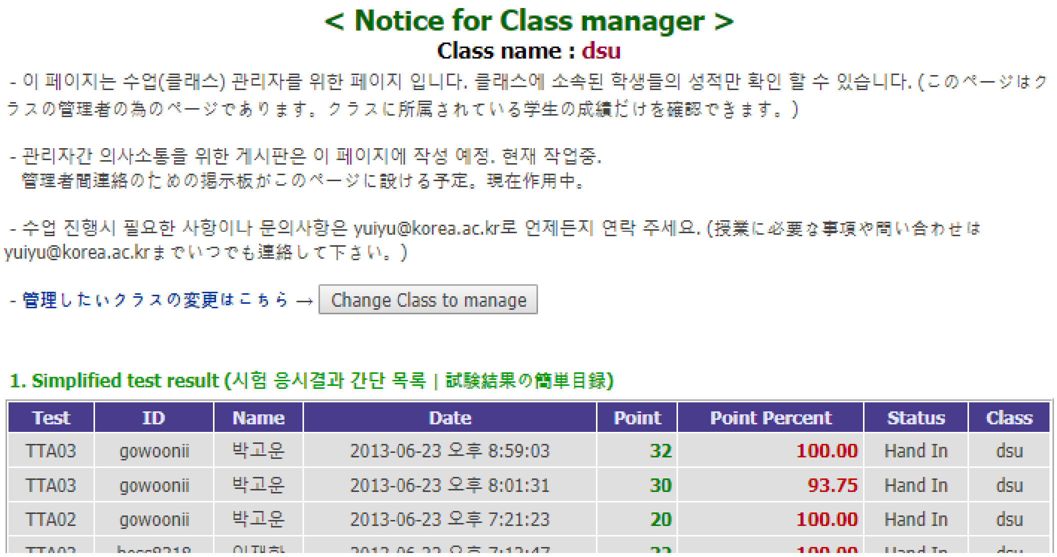

1) Words selection and recording
2) Register word in database
3) Create and register a web page for a phonetic test
4) Conduct phonetic tests on learners
5) Phonetic test results are automatically collected
6) Analyzing phonetic test results
7) Feedback the results to the learner and end the experiment
8) Troubleshooting and Compensation

Figure 3: The flow of the AJ-phonetic test from the instructor's perspective

Instructors have access to the results of students from their classes. In Figure 4, the class name is ‘dsu’ meaning Duksung Women’s University. Besides, instructors can also view students results for a specific period (and download this data as an Excel file) or access test statistics. The latter is shown in Figure 5.



< Notice for Class manager >

Class name : dsu

- 이 페이지는 수업(클래스) 관리자를 위한 페이지 입니다. 클래스에 소속된 학생들의 성적만 확인 할 수 있습니다. (このページはクラスの管理者の為のページであります。クラスに所属されている学生の成績だけを確認できます。)

- 관리자간 의사소통을 위한 게시판은 이 페이지에 작성 예정. 현재 작업중.
  管理者間連絡のための掲示板がこのページに設ける予定。現在作用中。

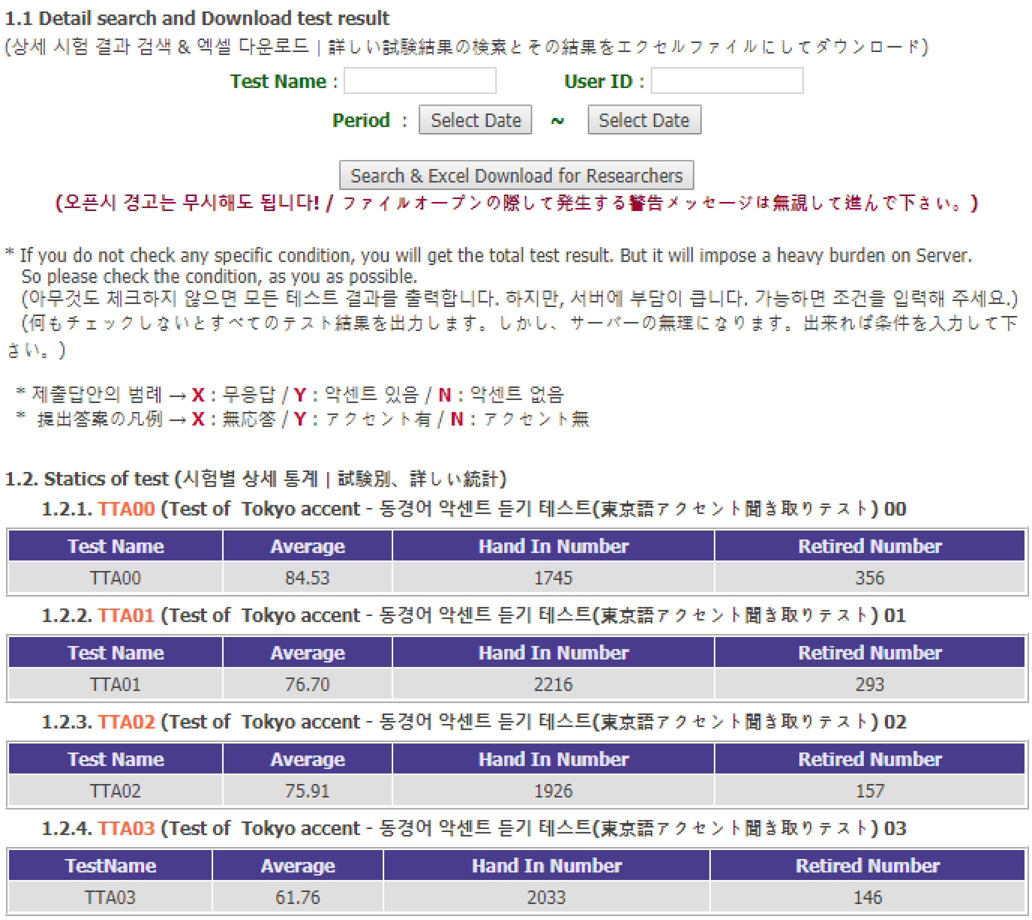
- 수업 진행시 필요한 사항이나 문의사항은 yuiyu@korea.ac.kr로 언제든지 연락 주세요. (授業に必要な事項や問い合わせはyuiyu@korea.ac.krまでいつでも連絡して下さい。)

- 管理したいクラスの変更はこちら → Change Class to manage

1. Simplified test result (시험 응시결과 간단 목록 | 試験結果の簡単目録)

Test	ID	Name	Date	Point	Point Percent	Status	Class
TTA03	gowoonii	박고운	2013-06-23 오후 8:59:03	32	100.00	Hand In	dsu
TTA03	gowoonii	박고운	2013-06-23 오후 8:01:31	30	93.75	Hand In	dsu
TTA02	gowoonii	박고운	2013-06-23 오후 7:21:23	20	100.00	Hand In	dsu

Figure 4: Example of test results of the ‘dsu’ class (administrator page for instructors)



1.1 Detail search and Download test result

(상세 시험 결과 검색 & 엑셀 다운로드 | 詳しい試験結果の検索とその結果をエクセルファイルにしてダウンロード)

Test Name : 　　User ID :

Period : Select Date ~ Select Date

Search & Excel Download for Researchers

(오픈시 경고는 무시해도 됩니다! / ファイルオープンの際して発生する警告メッセージは無視して進んで下さい。)

* If you do not check any specific condition, you will get the total test result. But it will impose a heavy burden on Server.
So please check the condition, as you as possible.
(아무것도 체크하지 않으면 모든 테스트 결과를 출력합니다. 하지만, 서버에 부담이 큽니다. 가능하면 조건을 입력해 주세요.)
(何もチェックしないとすべてのテスト結果を出力します。しかし、サーバーの無理になります。出来れば条件を入力して下さい。)

* 제출답안의 범례 → X : 무응답 / Y : 악센트 있음 / N : 악센트 없음
* 提出答案の凡例 → X : 無応答 / Y : アクセント有 / N : アクセント無

1.2. Statics of test (시험별 상세 통계 | 試験別、詳しい統計)

1.2.1. TTA00 (Test of Tokyo accent - 동경어 악센트 듣기 테스트(東京語アクセント聞き取りテスト) 00

Test Name	Average	Hand In Number	Retired Number
TTA00	84.53	1745	356

1.2.2. TTA01 (Test of Tokyo accent - 동경어 악센트 듣기 테스트(東京語アクセント聞き取りテスト) 01

Test Name	Average	Hand In Number	Retired Number
TTA01	76.70	2216	293

1.2.3. TTA02 (Test of Tokyo accent - 동경어 악센트 듣기 테스트(東京語アクセント聞き取りテスト) 02

Test Name	Average	Hand In Number	Retired Number
TTA02	75.91	1926	157

1.2.4. TTA03 (Test of Tokyo accent - 동경어 악센트 듣기 테스트(東京語アクセント聞き取りテスト) 03

TestName	Average	Hand In Number	Retired Number
TTA03	61.76	2033	146

Figure 5: Download of test results and test statistics (administrator page for instructor)

On the ther hand, the flow of the AJ-phonetic test from the test-taker's (learner’s) perspective was designed as shown in Figure 6.

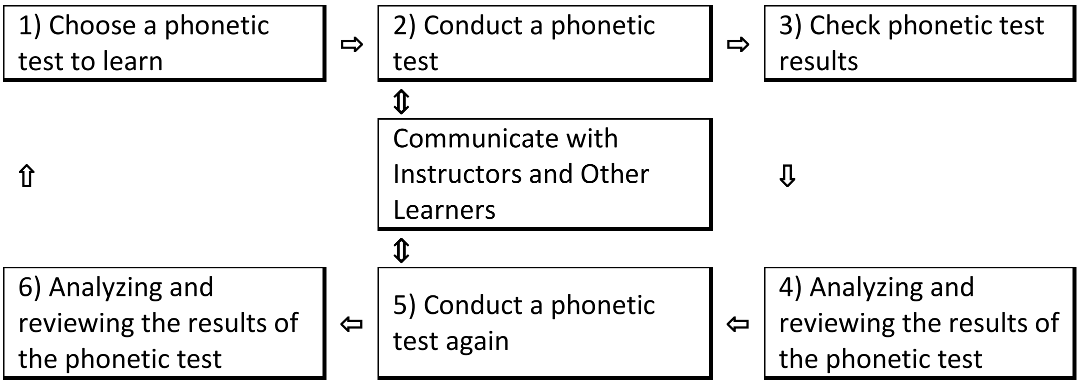
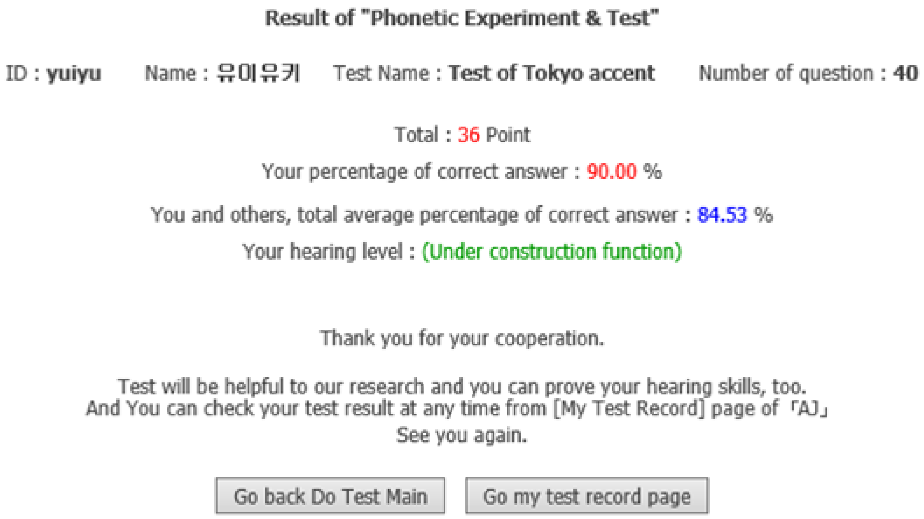

1) Choose a phonetic test to learn
2) Conduct a phonetic test
3) Check phonetic test results
Communicate with Instructors and Other Learners
4) Analyzing and reviewing the results of the phonetic test
5) Conduct a phonetic test again
6) Analyzing and reviewing the results of the phonetic test

Figure 6: The flow of the experiment from learner's perspective

The AJ-phonetic test makes it possible for a learner to check their results immediately after the test, an example of which is shown in Figure 7. The result page provides learners with their individual total scores and percentages, as well as average scores of all participants, in order that they can see their relative level rank and position in their class compared to the standard.



Result of "Phonetic Experiment & Test"

ID : **yuiyu** Name : **유이유키** Test Name : **Test of Tokyo accent** Number of question : **40**

Total : 36 Point

Your percentage of correct answer : 90.00 %

You and others, total average percentage of correct answer : 84.53 %

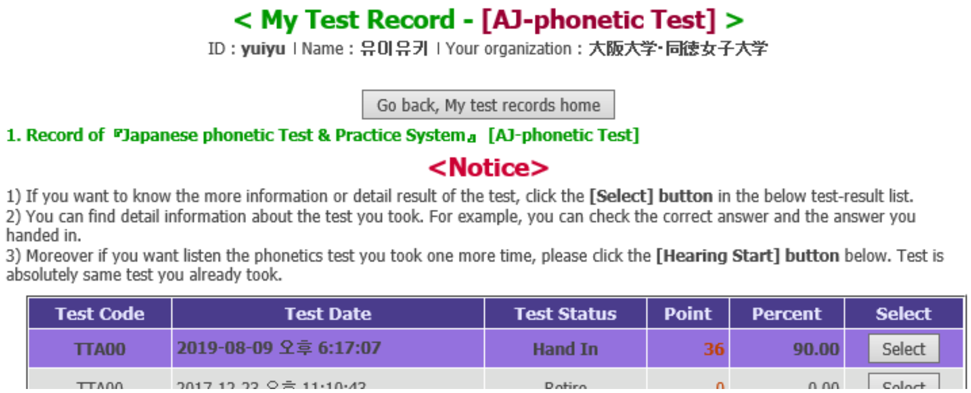
Your hearing level : (Under construction function)

Thank you for your cooperation.

Test will be helpful to our research and you can prove your hearing skills, too.
And You can check your test result at any time from [My Test Record] page of 「AJ」
See you again.

Figure 7: Example of accent test results for learners

Learners also have access to a detailed history of data collecting, such as their progress and test results on the so-called *My Test Record page* (Figure 8), and further can review the tests that they have already completed (Figure 9). All these information enables them to identify their weaknesses such as accents that they need to work on more comprehensively, for example (Figure 10).



< My Test Record - [AJ-phonetic Test] >

ID : **yuiyu** | Name : **유이유키** | Your organization : **大阪大学·同徳女子大学**

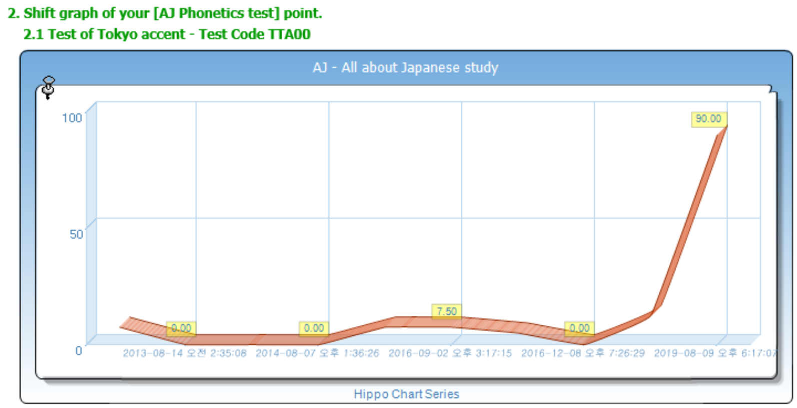
1. Record of 『Japanese phonetic Test & Practice System』 [AJ-phonetic Test]

<Notice>

1) If you want to know the more information or detail result of the test, click the **[Select] button** in the below test-result list.
2) You can find detail information about the test you took. For example, you can check the correct answer and the answer you handed in.
3) Moreover if you want listen the phonetics test you took one more time, please click the **[Hearing Start] button** below. Test is absolutely same test you already took.

Test Code	Test Date	Test Status	Point	Percent	Select
TTA00	2019-08-09 오후 6:17:07	Hand In	36	90.00	Select
TTA00	2017-12-22 오후 11:10:43	Retire	0	0.00	Select

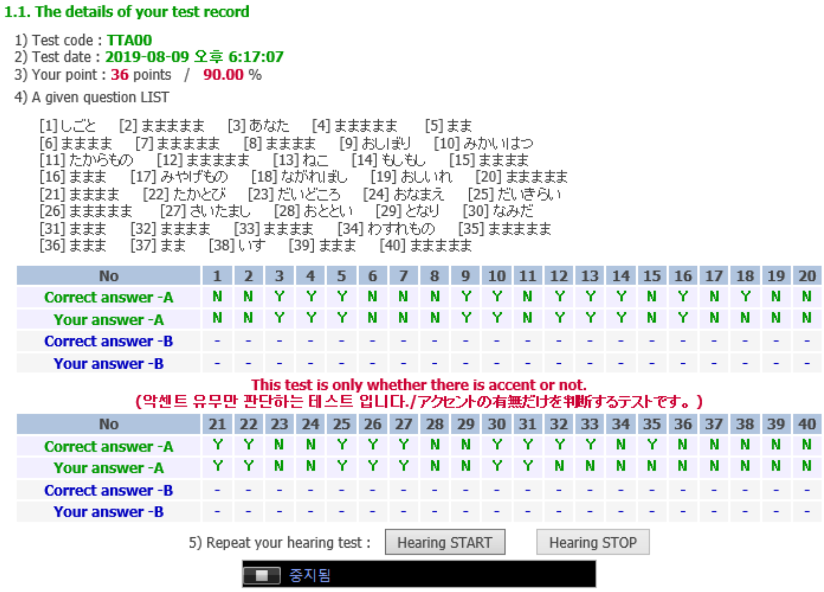
Figure 8: Tests history of learner



2. Shift graph of your [AJ Phonetics test] point.

2.1 Test of Tokyo accent - Test Code TTA00

Figure 9: Progress based on test results visualized



1.1. The details of your test record

1) Test code : TTA00
2) Test date : 2019-08-09 오후 6:17:07
3) Your point : 36 points / 90.00 %

4) A given question LIST

[1]しごと [2]まままままま [3]あなた [4]まままままま [5]まま
[6]まままま [7]まままままま [8]まままま [9]おしぼり [10]みかいはつ
[11]たからもの [12]まままままま [13]ねこ [14]もしもし [15]まままま
[16]ままま [17]みやげもの [18]ながれぼし [19]おしいれ [20]まままままま
[21]まままま [22]たかとび [23]だいどころ [24]おなまえ [25]だいきらい
[26]まままままま [27]さいたまし [28]おととい [29]となり [30]なみだ
[31]ままま [32]まままま [33]まままま [34]わすれもの [35]まままままま
[36]ままま [37]まま [38]いす [39]ままま [40]まままままま

No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Correct answer -A	N	N	Y	Y	Y	N	N	N	Y	Y	N	Y	Y	Y	N	Y	N	Y	N	N
Your answer -A	N	N	Y	Y	Y	N	N	N	Y	Y	N	Y	Y	Y	N	Y	N	N	N	N
Correct answer -B	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Your answer -B	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

This test is only whether there is accent or not.
(악센트 유무만 판단하는 테스트 입니다./アクセントの有無だけを判断するテストです。)

No	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Correct answer -A	Y	Y	N	N	Y	Y	Y	N	N	Y	Y	Y	Y	N	Y	N	N	N	N	N
Your answer -A	Y	Y	N	N	Y	Y	Y	N	N	Y	Y	N	N	N	N	N	N	N	N	N
Correct answer -B	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Your answer -B	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

5) Repeat your hearing test :

Figure 10: Details of test results

3.1 Obtaining statictical data

Functions of the administrator mode can be classified into three types. First, instructors can view the results and statistical data of individual learner’s accent listening tests, as made available in detailed Excel files and on web pages. The administrator can also view necessary information about learners who participated in the test. It is also important to state that because the AJ-phonetic test can set administrator ID for each educational institution and class, several instructors and educational institutions can simultaneously conduct AJ-phonetic test experiment or education individually (Figures 4 and 5).

3.2 AJ-phonetic test and its benefits

In Lee (1995), Lee et al. (1997), and Ayusawa (1997, 2000), Tokyo accent listening experiments were conducted for Japanese learners from various regions in Korea. The results showed that Japanese language learners differed in their ability to hear Japanese accents depending on their native language. However, this was an experiment in which the difficulty level of the test items was not adjusted in the experiment and analysis stages. In other words, an experimental design that divides the accentless pitch (the flat type) and the accent pitch (the relief type) is necessary for proper analysis.

3.2.1 Comparison of accented and accentless words

In case of the accentless words (the flat type), the correct answer rate is 50%. However, accented words (those with a tone drop) are estimated to have 50%, 33.3%, 25%, and 20% chances of being noted as correct, depending on the number of moras (2-5). Therefore, experimenting with accentless and accented words at the same time was avoided, as it may be confusing for learners. In this study, we added an experiment that clearly distinguished two types. (e.g. TTA00 vs TTA02)

3.2.2 Comparison between accent pitches

High numbers of moras in words are likely to increase the difficulty. At the same time it is known that longer words have a lower frequency of appearance. Therefore, in the AJ-phonetic test, dummy words were added to balance the number of moras, in order that learners can increase their learning effect. Because learners generally have a relatively small chance to encounter words with a large number of moras, they may develop a sense of accent by recognition and repetition exercises on dummy words.

Of course, dummy or meaningless words, such as 'まま' *mama*, in which a single note is listed consecutively have a disadvantage of being monotonous and easily accustomed to learners. However, the AJ Japanese Accent Test was designed with the primary objective of not only evaluating students' ability to recognize accents but also to develop a sense of accents in repeated practice. As emphasized in previous studies, Korean learners of Japanese with accentless native dialects, such as Seoul dialect, tend to listen to japanese through their native language system, which makes them ignore Japanese pitch.

As mentioned in Yang (2011, p. 71), accent education must be done before the wrong accent is fixed (early learning stage). And the experience of getting the right Japanese accent is more important than anything else. Repeated practice of consciously changing the height or listening to the pronunciation will be a useful learning strategy. Therefore, dummy words are suggested to be a good solution for improving this mora bias.

3.2.3 Intensive Test for Special phoneme

Korean and Japanese have many similarities, however, the number of phonemes and syllables in Japanese is smaller than that of Korean. This acts as a barrier to Japanese language learning. In particular, Japanese speech and listening education is easy to be neglected and is easy to be biased towards grammar and reading.

As mentioned earlier, Japanese is moraic language. It is well-known that the concept of mora represents difficulties for Korean learners of Japanese because there are no special moraic phonemes in Korean. Therefore, training of such special phonemes as well as a thorough understanding of Japanese accent are urgently needed.

In this study, we designed an accent test that consists of words containing such moraic phonemes, a syllabic nasal ‘ン’ (撥音 *hatsuon*), a prolonged vowel sound ‘一’ (長音 *choon*), and a geminate consonant ‘ッ’ (促音 *sokuon*), in order for learners to adapt to special moraic phonemes. This study selected Japanese accent test words in consideration of the difficulty to recognize special moraic phonemes by Korean learners of Japanese.

4 Analysis of the AJ-phonetic test results

Finally, we will analyze test results submitted by the learners. For reference, test results can be reviewed directly at:

http://www.japanese.or.kr/phonetic_related_thesis.aspx.

4.1 Answer ratio based on the number of moras in a word

Table 5 presents results related to the difficulty levels provoked by the number of moras in a word.

Table 5: Difficulty level and the number of moras

Moras	Level	Count
2	8.17	6
3	7.75	20
4	7.21	29
5	6.94	36
6	5	1
Average and sum	7.26	92

Results confirmed predictions, namely that words containing higher number of moras induce lower percentages of correct learners. (For reference, the AJ-phonetic test introduces a system that automatically changes the difficulty level depending on whether or not the word is correct. Also, it was found that the lower the level number, the higher the difficulty.)

We have thus decided to balance the number of moras in the accent test, and to add dummy words.

4.2 The percentage of correct answers per test

As expected, the TTA00 had the highest average score to determine the accent (see Figure 3).

Table 6: The percentage of correct answers per phonetic test

Test	TTA00	TTA01	TTA02	TTA03	All
Average	86.31	80.59	79.33	69.21	76.43

However, this study found some interesting results concerning the TTA01 and TTA02. First of all, TTA02 has accents on every word, and therefore it only is needed to find the accent location. However, in the case of TTA01, it is expected to be relatively difficult to solve because it is necessary to find the accent and the position of the accent at the same time. However, *learners scored relatively high on TTA01, which was expected to be more difficult than TTA02.* (Figure 3)

The reason for this result is that half of the words in TTA01 are words that have no accents, and therefore the correct probability of words that have no accents in TTA01 is evaluated at 50%. However, the TTA02 which must accurately locate the accent has a less than 50% chance of getting an answer. It is essential in this respect to understand this distinction. Because it proves that the hypothesis of this study which designing a test by dividing the accentless pitch, and the accent pitch words are fairer and more accurate as a representation of what the learners understand in this case.

Also, given the difficulty of these tests, the instructor may suggest that learners perform accent exercises in the order of TTA00 – TTA01 – TTA02 – TTA03.

Finally, the learner was found to have the most difficulty in TTA03, which consists of words containing special moraic phonemes. It indicates that instructors should spend much time practicing accents on words that contain special moraic phonemes. The following sections discuss which of the special moraic phonemes represents most difficulties for Korean learners of Japanese.

4.3 Difficulty according to special moraic phonemes

Words with either a syllabic nasal 'ン' (撥音 *hatsuon*), a prolonged vowel sound '一' (長音 *choon*), and a geminate consonant 'ッ' (促音 *sokuon*) are included into the test TTA03, which contains 13, 12, and 10 items, respectively. For this reason, the average difficulty of each special phoneme is calculated as follows (see Figure 6). That is, the higher the number of moras, the higher the difficulty (6.78 → 6.27 → 5.65 → 5).

It was acknowledged that words containing a geminate consonant represent most difficulties to learners. Instructors of accent education are thus suggested to pay special attention to words including a geminate consonant.

Table 7: Difficulty level and special moraic phonemes

<table>
<tr><th></th><th>3 M/L</th><th>4 M/L</th><th colspan="6">5 M/L</th><th>6 M/L</th><th>Total/L</th></tr>
<tr><td>ン(撥音)
Syllabic nasal</td><td>3/6.67</td><td>4/6.5</td><td>4/5.75</td><td rowspan="2">2/5.5</td><td>-</td><td></td><td rowspan="3">1/
6</td><td>7/5.71</td><td>-</td><td>13/6.29</td></tr>
<tr><td>一(長音)
Prolonged sound</td><td>3/6.67</td><td>3/6.3</td><td>3/5.67</td><td rowspan="2">1/6</td><td>-</td><td>6/5.67</td><td>-</td><td>12/6.21</td></tr>
<tr><td>ッ(促音)
Geminate consonant</td><td>2/7</td><td>4/6</td><td>1/5</td><td>-</td><td></td><td>3/5.67</td><td>1/5</td><td>10/5.91</td></tr>
<tr><td>Total/Level</td><td>8/6.78</td><td>11/6.27</td><td colspan="6">12/5.65</td><td>1/5</td><td>32/5.93</td></tr>
</table>

4.4 Test results by speakers of different Korean dialects

Lee (1995) and Lee et al. (1997) conducted experiments with Japanese learners from Seoul and Busan through the same experiments as were used with Ayusawa et al. (2000). They reported that their Japanese accent listening scores were lower than those from Busan. Learners from Seoul who did not have high or low accents were less able to perceive high or low accents in Japanese, as compared to those learners from Busan who used high and low accents to distinguish meanings. Learners from Busan who speak dialects where high and low accents are involved in discrimination of meanings and learners from Seoul who use dialects where high and low accents are not involved in discrimination of meaning are different in their ability to perceive high and low accents in Japanese. It was thus verified that the difference depends on the presence or absence of high and low accents on the mower.

In other words, Japanese learning could be influenced by local dialects, and by the presence or absence of a distinctive tone drop in them. This study verified this point and conducted a detailed comparison between Korean learners of Japanese from

Gyeongsang-do and Jeonla-do regions respectively. Both regions employ pitch accent dialects.

Table 8: Test results and their difference by region

Region	No. of participants	Average				
		TTA00	TTA01	TTA02	TTA03	ALL
Seoul	234	80.00	76.81	76.25	63.54	***74.16***
Gyeonggi-do	111	83.46	74.63	75.69	69.21	75.76
Chungcheong-do	11	86.11	78.63	61.50	51.44	64.07
Gyeongsang-do	24	85.41	80.55	75.88	69.35	***78.51***
Jeonla-do	6	86.96	91.18	83.93	73.27	***84.11***

*Note: Gangwon-do region was excluded from the comparison due to the lack of data (2 participants only).

Through the results of this study, it was confirmed that learners of Gyeongsang-do and Jeonla-do, in whose native dialects relatively strong tones or accents are detected, perform better on Japanese accent tests, which was already shown in previous studies. However, no comparative studies have yet been conducted on perception of Japanese accents among speakers of Gyeongsang-do and Jeonla-do regions. This longitudinal study gathered and analyzed a large amount of data to find that Koreans whose native dialect is that of Jeonla-do region performed better than those of Gyeongsang-do region dialect in all accent tests.

5 Conclusions

In this paper, we divided subjects participating in e-learning into instructors and learners, and conducted the AJ-phonetic test. Through the LMS system, instructors were able to access academic performance information of learners, their history and results obtained through the AJ-phonetic test. All these information enabled instructors to provide each learner with an appropriate feedback, which further helped learners to develop new strategies for learning. Since learners took the AJ-phonetic test online, results were immediately available to both instructors and learners. Such information allow learners to check their learning progress and achievements, and offers an objective analysis as to how well they learn Japanese language.

This study conducted the Tokyo accent test (the AJ-phonetic test) to verify the following hypothesis empirically.

First of all, when preparing accent tests, we needed to make a separate set of words because the accentless pitch and the accent pitch represent different levels of difficulty for korean learners of Japanese. Secondly, as the presence or absence of special moras affects the difficulty in the test, we added dummy words to control the level of difficulty and increase the learning effect of the test. Thirdly, since Korean learners of Japanese have difficulty recognizing accents of words containing special moras, it was necessary to pay additional attention to the practice of accents of such words. Finally, Korean learners of Japanese differ in their ability to recognize Japanese accents depending on their native dialect. All these findings bring us to conclusion that it is necessary to provide Japanese accent education based on thorough information about a learner (native language, language education etc.) as well as theory on Japanese accent and its learning and teaching. Last but not least, the AJ-phonetic test was designed as part of a computer-enabled Japanese class in a way to suit learner's individual schedule; it is available anytime and anyplace.

For further studies, we plan to add new accent tests with new experimental stimuli, which would include additional words with special moraic phonemes and different dummy words. However, prior to such work, we would like to further consolidate the analysis of the test contents currently available. We therefore plan to upgrade the system based on the advice received, such as those concerning types of data required by instructors and methods of analysis, as well as types of feedback material and methods of presentation to learners.

References

Ayusawa, T. (1997). Tokyogo akusento no kikitori tesuto ni tsuite [About Tokyo Accent Perception test], *Spoken Japanese language education: looking to the 21st century*, 189-20, The National Language Research Institute.

Ayusawa, T., Nishinuma, Y., & Kawatsu, M. (2000). Akusento shutoku no odankenkyu – sannenhan no chousakekka – [Longitudinal study of accent acquisition: Survey results for three and a half years-]. *Proceedings of the 14th Annual Conference of the Phonetic Society of Japan*, 177-182.

Choi, J. (2006). Ilbono accentheki wichibyonhwae kanhan ilgochal [Investigation on the acceptability of Japanese prosody -the effect of the position of accent nuclei-]. *Journal of Japanese Language and Literature, 58*(1), 187-204.

Choi, J., & Kim, Y. (2008). Iraninggu wo katsuyosita Nihongo akusento kikitori renshu – maruchimediajyugyo deno katsuyo wo chusin ni [Japanese Accent Perception Test based on e-learning: Using tool to improve the skill of accent perception based on multimedia technology]. *Journal of Japanese language and literature*, *64*, 263-279.

Jeong, S. (2009). Souru hougenwo bogotosuru nihongo gakushushano nihongo akusentono seisito chikaku [Generation and Perception of Japanese Accents of Japanese Learners with Seoul direct]. *Journal of Japanese language Association of Korea*, *26*, 217-234.

Lee, M. (1995). Kankokujin gakushusha no tokyogo akusento no chikaku – souru chiho no baai [Perception of Korean accent by Korean learners; In the case of Seoul region]. *Proceedings of the 1995 Japanese Language Education Society Fall Conference*, 159-164.

Lee, M., Ayusawa, T., & Kim, S. (1997). Kankokugowasha no "tokyogo akusento kikitori tesuto" no kekka – Souru, Pusan, Koushu hougenwasha no baai [Results of the "Tokyo Accent

Listening Test” by Korean native speakers: In the case of Seoul, Busan, and Gwangju dialect speakers]. *Toward 21st Century Japanese Speech Education*, 22-30.

Lee, M. (2002). *A Study on the development of web-based learning model and the effect of feedback type for Japanese language listening comprehension* (doctoral dissertation). Pusan University.

Rosenberg, M. J. (2001). *E-learning: Strategies for delivering knowledge in the digital age*. New York: McGraw-Hill.

Ogawara, Y. (1997). Hatsuon kousei bamen ni okeru gakushusha no hatsuon to kikitori nokankei ni tsuite [The relation between learners' pronunciation and listening in pronunciation correction situations]. *Journal of Japanese language teaching*, *92*, 10-52.

Onzuka, C. (2011). Kankokugo bogowashaniokeru tokushuonso no ninshiki – onin ninshiki tohyoki: riron to jikkenkara no apoorochi [The Phonological Awareness of the Japanese Special Phoneme by the Korean Native Speakers —The Phonological Awareness and The Notation: An Approach from a Theory and an Experiment—]. *Journal of Japanese Language and Literature, 77*(1).

Yang, N. (2011). Seoul banonhwajawa jeonnam bangonhwajarul desanguro han ilbono accent chonchi gyonghyang bigyo [Japanese accent hearing trend comparison having made to Seoul dialect speaker and Jeonnam dialect speaker]. *Journal of Japanese Language Education Association*, *56*, 49-72.

GRAMMAR ERRORS BY SLOVENIAN LEARNERS OF JAPANESE: CORPUS ANALYSIS OF WRITINGS ON BEGINNER AND INTERMEDIATE LEVELS

Miha PAVLOVIČ
University of Ljubljana, Slovenia
miha.pavlovic1@gmail.com

Abstract

This paper presents the construction of a corpus of writings by Slovene learners of Japanese as a foreign language at the beginner and intermediate levels and an analysis of the grammar errors contained within it, with the purpose of providing a simple and effective means of acquiring data on errors made by students of Japanese as a second language. Additionally, an error analysis of the grammar errors in the corpus and a comparison of the most common errors found on both levels, reveals the types of errors that carry over from the beginner to the intermediate level, negatively affecting the learning process. By compiling and analyzing a collection of 182 written texts written by Japanese learners, 492 cases of grammar misuse were observed on the beginner and 564 on the intermediate level. A comparative analysis of the most common types of grammar misuse on each level highlights the types of errors that seem to carry over from the beginner to the intermediate level. The findings can be useful to Japanese language learners as well as teachers. Furthermore, the learner's corpus created in the process marks the first step towards the creation of a larger, annotated and publicly accessible learner corpus of writings by Slovenian learners of Japanese to be used for further research in the field of second language acquisition.

Keywords: learner corpus; corpus construction; error analysis; grammar error; second language acquisition

Povzetek

Članek opisuje izgradnjo korpusa usvajanja jezika slovenskih študentov japonščine na osnovni in srednji ravni in analizo slovničnih napak v njem z namenom ustvarjenja orodja, ki bo uporabnikom omogočalo na enostaven in pregleden način pridobiti podatke o najpogostejših napakah v spisih slovenskih učencev japonščine in s pomočjo analize napak v le-tem ugotoviti, katere slovnične strukture povzročajo največ težav slovenskim učencem japonskega jezika na posamezni ravni ter s pomočjo primerjave rezultatov izpostaviti tipe napak, ki se prenašajo iz osnovne na srednjo raven. Korpus vsebuje 182 spisov, v katerih so označene in kategorizirane napake. Napak je 492 na osnovni in 564 na srednji ravni. S primerjavo najpogostejših napak na posamezni ravni so se bili izpostavljeni tipi napak, ki se prenašajo iz osnovne na srednjo raven. Te ugotovitve lahko koristijo tako učencem kot tudi učiteljem japonščine pri učnem procesu, hkrati pa je tako nastali korpus prvi korak k izgradnji obsežnega, označenega in javno dostopnega korpusa besedil slovenskih učencev japonščine za nadaljnje raziskave o učenju japonščine kot tujega jezika.

Ključne besede: korpus usvajanja jezika; gradnja korpusa; analiza napak; slovnične napake; usvajanje tujega jezika

Acta Linguistica Asiatica, 10(1), 2020.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.10.1.87-104

1 Introduction

The Slovenian and Japanese language are genealogically not related and thus differ on all levels of linguistic analysis: from script and phonology to grammar and syntax. At the syntactic level, the predicate in Slovene sentences mostly appears in second place, usually following a subject or adverbial, while in Japanese the predicate always appears at the end of a sentence or subordinate clause. On the grammatical level, there is a difference in the way cases are expressed; while in Slovene cases are expressed by noun declension, in Japanese particles (*kakujoshi* 格助詞) are attached to grammatical elements to mark their relation to the verb; while Japanese adjectives ending with an -i (*i-keiyōshi* イ形容詞) have past forms, Slovenian adjectives do not have different forms to express tense and a past form of the auxiliary verb is used, and there are many other subtler differences. It is therefore considerably more challenging and time-consuming for a Slovenian learner to learn Japanese than a more related language like English or German, which share grammatical similarities with the Slovenian language.

The occurrence of grammar errors is a natural part of the language acquisition process; thus it is only natural that learners make more errors when using the elements that are fundamentally different from those in their native language. The reason for the occurrence of such errors is usually attributed to the lack of knowledge about those elements. If such errors can be recognized and corrected, a strong foundation for further language acquisition may be guaranteed. Some types of errors disappear naturally, through exposure to the language. However, some errors, if not recognized and dealt with, persist and negatively influence the process of language acquisition. For these purposes, researchers in the field of second language acquisition (SLA) conduct so called “error analyses”, which, as the name suggests concern themselves with the quantitative and qualitative analysis of the errors produced by learners of a specific language. The tools used in such studies most commonly include databases or corpora containing examples of language use by students of a specific skill level (e.g. English learners on the intermediate level).

Due to the field of Japanese studies in Slovenia being fairly new, similar studies focused on the errors made by Slovenian learners of Japanese have been very few in number. Thus, there was a lack of and need for a tool that would allow users to easily access data on the types of errors Slovenian learners of Japanese tend to make in written compositions on a certain level. One of the aims of the present study is therefore, through the acquisition and digitalization of learners’ compositions, to create a corpus of errors by Slovenian learners of Japanese on both the beginner and intermediate level. Grammar errors on both levels were analyzed with the purpose of exposing the problematic grammatical elements that are prevalent on both levels. As mentioned previously, such types of errors, when unidentified, may hinder language acquisition. Expozing and consequently targeting them can have a positive effect on the learning process.

In short, the purpose of the study was to produce a resource in which teachers and SLA researchers can easily access data on the types of grammar errors Slovenian learners of Japanese tend to make, and by using the data expose the most problematic grammar error types.

Sections 2 contains a summary of previous research, used as reference. Sections 3 to 5 describe the creation of the corpus: section 3 the metadata added to the students' compositions, section 4 the process of data acquisition and digitalization, and section 5 the categorization of error types. The second part of the paper presents a first analysis of this corpus: section 6 describes the methodology used in the analysis, sections 7 and 8 the results of the analysis of grammar errors on the beginner and intermediate levels respectively, section 9 a comparison of the results on each level, followed by their discussion in section 10 and conclusions in section 11.

2 Previous research on errors in a second language acquisition

In the last decades, a number of error analyses targeting the grammar errors made by foreign students of Japanese (mostly native speakers of English, Chinese and Korean) have been conducted, mostly by Japanese linguists. Examples of such studies include: Teramura (1990), Ichikawa (1993), Kawaguchi (1995), Otsuka & Hayashi (2010), Harasawa (2012), Noda and Sakoda (2019) and others.

Present research is the first study to analyze a corpus of Slovenian Japanese learners, and as such seeks to verify whether the findings from previous studies are valid for native speakers of Slovene as well.

The following three surveys were primarily used as an important source of information and guidance for this analysis.

Kawaguchi (1995) analyzed writings of five students with different middle-level native languages. The compositions averaged around 400 characters, which caused 267 cases of errors. The most numerous types of errors involved particles, case particles in particular. The author concluded that such types of errors are often carried over to the advanced level. The comparison of the results for different levels of acquisition was taken as a model for the present research.

Han 2014 identified 2875 errors using quantitative analysis of 204 compositions. Grammar and semantic errors together accounted for almost 90 % of all errors, of which grammatical presented as much as 54.6 % while 33.8 % were semantic. The most common type of grammar errors (30.5% of all grammar errors) associated with the group of articles, of which case particles were found to be most problematic and represented 65% of all errors related to the use of particles. The most common mistakes were made in distinguishing between the use of *ga* が and *wa* は. Similar difficulties was also observed with the distinctions between: *ni* に, *de* で, *wo* を, *ga* が and no の. Methodology used in this research was a model for our research.

Finally, Online Dictionary of Errors in Japanese 2011 (*Onrain nihongo goyō jiten* オンライン日本語誤用辞典 2011), created at the University of Foreign Languages in Tokyo, is introduced not only for its error analysis but also as a tool designed to further conduct this type of research. The tool is based on a corpus containing more than 1000 entries of errors identified from 40 files, totaling more than 20,000 characters. The online dictionary is currently one of the few, if not the only, online corpora or dictionary that categorizes collected errors on multiple levels and allows the user to view them in a simple and transparent way. This online glossary is very important for the present research because the categorization used in building the corpus is based on the categorization of errors used in this corpus.

3 Slovenian learners of Japanese: corpus analysis of grammar errors

3.1 Methodology

3.1.1 Metadata structure and annotation

The Slovenian learners' written Japanese corpus consists of two sub-corpora: *Slovenian beginner learners' written Japanese corpus* and the *Slovenian intermediate learners' written Japanese corpus*.

The sub corpus of the beginner level consists of 142 shorter compositions, each with an average length of about 280 characters. The compositions were written by 29 first-year students of the Japanese studies program at the Department of Asian Studies in the Faculty of Arts, University of Ljubljana in the academic year 2016/2017. The compositions were not written in a test environment, but as homework at two of the Japanese language classes. The topics of the compositions cover a range of simple everyday topics (9 in total), such as descriptions of one's room, one's family, hobbies, a diary, a self-presentation and a reading diary.

The sun-corpus of the intermediate level consists of 40 longer compositions, each with an average length of about 500 characters. The compositions were written in 2017/2018 by 11 of the same 29 students (one year later than the first compositions). The compositions include 4 topics which require the use of more complex grammatical structures and vocabulary than the topics of the beginner corpus, and the students were asked to state and argue their opinion on the subject. These topics are: "telephone", "time", "world heritage" and "my country". These compositions were written as part of a mid-term exam, where dictionaries and grammar checkers were not allowed.

3.1.2 Acquisition and digitalization of the compositions

The compositions were submitted as homework or parts of mid-term exams. Each of the authors signed a waiver, allowing the inclusion of their compositions into the corpus and their use scientific purposes, under the condition that all personal data be anonymized.

The next step was digitization. The creation of the corpus required a tool for the annotation of grammar errors and search of both specific parts of the data (compositions), as well as the metadata (categories, data on the compositions, etc.), easy acquisition of statistical data and that would be portable on and compatible with different platforms (Mac, Windows, etc.). While several sets of open-source annotation software (such as "Slate", "WebAnno", "SketchEngine" and others) were available, none of these tools appeared to satisfy all of the required criteria. The tool that finally provided an almost surprisingly simple solution to the problem was Microsoft's Excel.

First, all of the texts were manually typed into a Microsoft Excel spreadsheet (each sentence in a separate row) verbatim as they appeared in the handwritten physical version; all errors, including orthographical errors, errors in the use of *kanji* 漢字, were transcribed as in the original.

This was done to enable the created corpus to be used for different types of error analysis in the future and to provide possible context for the occurrence of errors. All personal data was anonymized and replaced with a placeholder (*jinmei* [人名] for personal names, or *chōmei* [町名] in the case of town names).

Non-standard character forms were not annotated, because the inclusion of such errors would require a fairly different approach and toolset. Thus it seemed best to omit these types of errors.

The final step of the digitization process was error annotation. All error annotation from the original correction, done by the teacher in charge of the class, was carried over. Where annotations other than those made by the teacher were marked differently from the original annotations.

Finally, in a separate spreadsheet, a corrected version of each sentence containing an error was added in a column next to the original sentence and the corrected part marked with one of three colors, depending on the type of error: red for grammar errors, yellow for orthographical errors or errors connected to the use of Chinese characters and green for stylistic errors and errors in vocabulary choice.

3.1.3 Error categorization

While other types of errors were also included and annotated in the corpus, the analysis described in this paper focuses solely on grammar errors. Each of the grammar errors was categorized first into a main group, followed by a subgroup and finally within each subgroup according to the supposed cause for the error. However, when being categorized, the error was not categorized according to the grammar element that was mistakenly used, but according to the element that should have been used to form a grammatically correct sentence. The basis for this is the idea that, as mentioned in the first chapter, the cause for the occurrence of the error is a lack of knowledge about the element; in this case knowledge of the fact that this specific element needed.

Table 1: Examples of grammar errors due to a wrong choice

Grammatically incorrect sentence	Sentence as corrected by teacher	Error	Grammatical category	Sub-category	Cause of Error
ゲームを好きです。	ゲームが好きです。	が ⇔ を	格助詞	が	誤選択
Gēmu wo suki desu.	*Gēmu ga suki desu.*	*Ga ⇔ wo*	*kakujoshi*	*ga*	*gosentaku*
I like games.	I like games.		Case particle	Particle *ga*	Wrong choice

As seen in the above table, in the sentence "*Gēmu wo suki desu.*" the grammar error occurred due to the student using the particle *wo* instead of the particle *ga*, which this sentence structure calls for. The error would be classified as an error connected to the use of case particles, more precisely, the case particle *ga*, with the contributing cause being marked down as *wrong choice*.

Table 2: Examples of grammar errors due to lack of use

Grammatically incorrect sentence	Sentence as corrected by teacher	Error	Grammatical category	Sub-category	Cause of Error
ゲーム Ø 好きです。	ゲームが好きです。	が ⇔ Ø	格助詞	が	誤不足
Gēmu suki desu.	*Gēmu ga suki desu.*	*Ga ⇔ Ø*	*kakujoshi*	*ga*	*gofusoku*
I like games.	I like games.		Case particle	Particle *ga*	Lack of use

In the case of the sentence "*Gēmu suki desu.*" the grammar error occurred due to the student not using the particle *ga*; therefore, this type of error would again be categorized as an error connected to the use of the case particle *ga*, the difference here being that the contributing cause would be marked as "lack of use".

This categorization was adopted from the categorization used in a similar learner's corpus of Japanese learners' grammar errors, namely the *Online corpus of Japanese* *learners' errors* by Umino's et al. (2012, originally: *Onrain nihongo goyō jiten* オンライン日本語誤用辞典) published by the Tokyo University of Foreign Studies.

The reason for this choice is that the former corpus is one of the few corpora of Japanese learners in which errors are not only annotated, but also categorized in groups and subgroups according to their grammatical properties in a very similar manner as demonstrated in the above table. The reason an already existent classification was used was to make the data in these two corpora easily comparable, thus further increasing the number of possible uses for the assembled data in potential future studies.

Following below are three tables. The first contains all the main grammatical categories used. The second one contains the sub-categories of specific types of elements within each of the main grammatical categories. And the third table contains the five types of contributing causes that were determined for each error. The left column of each table contains the Japanese name of the category accompanied by its transcription and the right one an English translation by the author.

Table 3: Grammatical categories

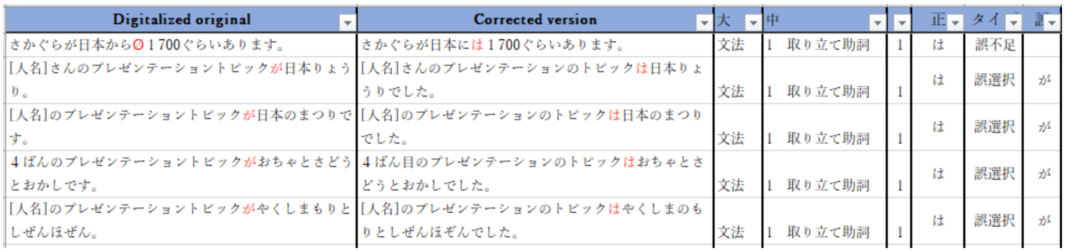
	Japanese original	Transcription	English translation
1-1	取り立て助詞	*toritatejoshi*	focus particles
1-2	格助詞	*kakujoshi*	case particles
1-3	終助詞	*shūjoshi*	final particles
1-4	複合辞	*fukugōji*	compound particles
1-5	ヴォイス	*voisu*	voice
1-6	テンス・アスペクト	*tensu-asupekuto*	tense and aspect
1-7	基本文型	*kihonbunkei*	basic sentence structure
1-8	表現文型	*hyōgenbunkei*	modal expressions
1-9	待遇表現	*taigūhyōgen*	polite expressions
1-10	形式名詞	*keishikimeishi*	formal nouns
1-11	指示詞	*shijishi*	demonstratives
1-12	疑問詞	*gimonshi*	interrogatives
1-13	2語の接続	*ni-go no setsuzoku*	word level conjunction
1-14	2文の接続	*ni-bun no setsuzoku*	sentence level conjunction
1-15	修飾	*shūshoku*	modifiers

Table 4: Error causes

Japanese original	*Transcription*	English translation
誤選択	*gosentaku*	wrong choice
誤不足	*gofusoku*	lack of use
誤形態	*gokeitai*	form error
誤付加	*gofuka*	redundance
誤位置	*goichi*	wrong position

In order to classify and annotate the errors, a framework needed to be created, so as to create space for the marks, enabling the different functions of MS Excel to work as intended.

As mentioned in the above paragraphs, the original text was placed in an excel spreadsheet, accompanied by the corrected version in the neighboring column. The column next to it (column C in the example bellow) contains the data on the type of error, ranging from "grammar", "style and vocabulary" to "orthography and script". The fourth column was created for data on the grammatical category and the one next to it for data on the grammatical sub-category (as explained in 4.1) to be inserted. The sixth column was made for data on the specific grammar element that was supposed to be used in the sentence where the error occurred (in some cases this data was the same as that in the fifth column, however in cases where the sub-category was an umbrella term, such as "temporal conjunctions" it served to further pinpoint the specific type of error). The seventh column was used to determine the cause of the error, while the eight one was used to mark which element was wrongly used instead of the right one. The final, ninth column was used to add numerical IDs to each of the sentences, making it possible to restore their original order within the whole framework after using different sorting options in Excel.



Digitalized original	Corrected version	大	中		正	タイ	誤
さかぐらが日本から∅1 700ぐらいあります。	さかぐらが日本には1 700ぐらいあります。	文法	1　取り立て助詞	1	は	誤不足	
[人名]さんのプレゼンテーショントピックが日本りょうり。	[人名]さんのプレゼンテーションのトピックは日本りょうりでした。	文法	1　取り立て助詞	1	は	誤選択	が
[人名]のプレゼンテーショントピックが日本のまつりです。	[人名]のプレゼンテーションのトピックは日本のまつりでした。	文法	1　取り立て助詞	1	は	誤選択	が
4ばんのプレゼンテーショントピックがおちゃとさどうとおかしです。	4ばん目のプレゼンテーションのトピックはおちゃとさどうとおかしでした。	文法	1　取り立て助詞	1	は	誤選択	が
[人名]のプレゼンテーショントピックがやくしまもりとしぜんほぜん。	[人名]のプレゼンテーションのトピックはやくしまのもりとしぜんほぞんでした。	文法	1　取り立て助詞	1	は	誤選択	が

Figure 1: Example of corpus

This design now enables the user to use Excel's sorting and search functions to e.g. search for all the instances of a specific error, to find all the cases in which a specific grammar element was used wrongly, sort the data according to each of the three categories (grammatical category, grammatical sub-category and error cause), search for specific terms used in either the original or the corrected data, easily acquire statistical data for cases of any of the above, and many more.

3.1.4 Data analysis

By using Microsoft Excel's sort and search functions the number of errors correlating to each group was counted for all the categories mentioned in Chapter 5.

The number of errors in each group and sub-group were then compared to the sum of all errors and were henceforth represented with percentages rather than actual numbers. This was also partially done to enable easier comparison of the results on each level in the second part of the analysis.

Next the grammatical categories and sub-categories with the highest amount of errors were determined alongside the most common causes for the occcurance of each type of error.

However, it must be said, that the percentages of errors described in the following sections are not a direct indicator of the relative difficulty of a particular morphological or syntactic category, only of the frequency of errors being made. To determine the relative difficulty of specific categories, a different approach would be necessary.

The number of errors related to categories that are more frequent (e.g. case particles) is necessarily larger than the number of errors related to categories that are less frequent in any text (e.g. final particles).

Originally, one of the goals was to identify the most numerous types of errors, and based on the ratio between the amount of correct and incorrect use of an element. However, because of the relatively small amount of data in each sub corpus and the uneven use of different grammatical elements within it, the calculated results were unreliable. In addition, in previous research, which served as the basis for this analysis, this step was also omitted.

Finally, if a learner were to misuse a rarely used grammar in 10 out of 10 cases, compared to a more common grammar being misused 200 out of 500 cases, the latter type of error would hinder communication between the author and the reader much more, simply because of its frequency. Additionally, the calculation itself would be too time consuming, in proportion to the unreliable results to be gained. Thus, only misuse frequency was determined.

This was done with both the Slovenian beginner learners' written Japanese corpus and the Slovenian intermediate learners' written Japanese corpus respectively. Thus the results on both levels were acquired and the elements the students struggle with the most were determined. The results are presented in the following sections.

The next step was the comparison of the results on both levels. As mentioned in chapter 1, this was done with the goal of exposing the types of grammar errors that appear on the beginner level and are still present on the intermediate level. The persistence of such errors means that they present a huge hurdle to the learner, which, if not overcome, would exert negative influence on the language acquisition process further on.

To expose these errors, the appearance rate of each of the error groups and sub-groups was observed and compared.

As a result, groups of problematic grammatical elements were successfully exposed and analyzed. A detailed summary of the results can be found in the following sections.

3.2 Results

3.2.1 Grammar errors on the beginner level

The beginner sub corpus includes 142 compositions in which 496 grammar errors were observed. The average length of the compositions is about 210 Japanese characters.

Table 5: Error data on the beginner level

Error category	Number of occurrences	Percentage of all errors
1-2 case particles	129	26,2 %
1-7 basic sentence structure	100	20,2 %
1-15 modifiers	74	15,0 %
1-6 tense and aspect	58	11,8 %
1-14 sentence level conjunction	40	8,1 %
1-1 focus particle	38	7,7 %
1-5 voice	12	2,4 %
1-10 formal nouns	12	2,4 %
1-4 composed particles	10	2, 0%
1-8 modal expressions	7	1,4 %
1-13 word level conjunction	6	1,2 %
1-3 final particles	3	0,6 %
1-11 demonstratives	3	0,6 %
1-12 interrogatives	1	0,2 %
SUM	**493**	**100,0 %**

The most common error categories, sorted from most to least common, are presented in table 5. Most errors were found in the group of case particles, amounting to 26 % of all errors found. The second most common were errors connected to the basic sentence structure which represent 20,2 % of all errors found; the third most common being the group of modifiers with 14,9 % of all errors. A considerable number of errors was also found in the category of sentence level conjunctions with a sum of 8,1 % and focus particles with 7,7 %.

Table 6: Error cuases on the beginner level

Type of error	Number of occurrences	Percentage of all errors
wrong choice	214	43,4 %
lack of use	164	33,3 %
form error	55	11,2 %
redundance	50	10,1 %
wrong position	10	2 %
SUM	**493**	**100,0 %**

As represented in the table above, the most common cause of errors was wrong choice with 43,4 %, followed by lack of use with 33,3 % of all cases. Other causes were much less common.

3.2.2 Grammar errors on the intermediate level

The subcorpus of Slovenian intermediate learners' written Japanese contains 40 compositions in which 564 grammar errors were observed. The average length of the compositions amounts to about 550 Japanese characters per composition.

Table 7: Errors on the intermediate level

Error category	Number of occurrences	Percentage of all errors
1-2 case particles	127	22,5 %
1-14 sentence level conjunction	99	17,6 %
1-1 focus particles	95	16,8 %
1-6 tense and aspect	51	9,0 %
1-7 basic sentence structure	42	7,4 %
1-8 modal expressions	36	6,4 %
1-15 modifiers	33	5,9 %
1-5 voice	29	5,1 %
1-10 formal nouns	23	4,0 %

Error category	Number of occurrences	Percentage of all errors
1-11 demonstratives	13	2,3 %
1-4 composed particles	9	1,6 %
1-13 word level conjunction	6	1,0 %
1-3 final particles	1	0,2 %
1-12 interrogatives	0	0,0 %
SUM	**564**	**100,0 %**

As seen in the table above, the most common errors were those related to the use of case particles with 22,5 % of all the errors observed. Also very common were errors from the categories of sentence level conjunction (17,6 %) and focus particles (16,8 %). Errors from the category tense and aspect (9 %), basic sentence structure (7,4 %) and modal expressions (6,4 %) were also common.

Table 8: Error cuases on the intermediate level

Type of error	Number of occurrences	Percentage of all errors
wrong choice	322	57,1 %
lack of use	133	23,6 %
redundance	64	11,3 %
form error	42	7,4 %
wrong position	3	0,5 %
SUM	**493**	**100,0 %**

As can be seen in the above table, the predominantly common cause of errors was wrong choice *gosentaku* with 57,1 %, followed by lack of use *gofusoku* with 23,6 % of all cases. Other causes were much less common.

3.2.3 Comparison of the results on both levels

After grammar analysis on each level was completed, a comparative analysis of the results on both levels was conducted. First, we will present comparison of the most common error categories, which will be followed by comparison of error causes across both levels.

The following table presents a comparison between the most common error categories on each level (as described in chapters 6 and 7). The categories in which a difference of more than 2 % was observed between the beginner and intermediate level are marked with blue if the percentage decreased, and red if the percentage increased. The threshold was first set to 5 %, but was later lowered down to 2 %, to accommodate for and include categories with differences between the two levels lower than than 5 %.

Table 9: Comparison of analysis results on both levels

Analysis of errors on the beginner level			Analysis of errors on the intermediate level		
1-2 case particles	129	26,2 %	1-2 case particles	127	22,5 %
1-7 basic sentence structure	100	20,2 %	1-14 sentence lev. conjunction	99	17,6 %
1-15 modifiers	74	15,0 %	1-1 focus particles	95	16,8 %
1-6 tense and aspect	58	11,8 %	1-6 tense and aspect	51	9,0 %
1-14 sentence lev. conjunction	40	8,1 %	1-7 basic sentence structure	42	7,4 %
1-1 focus particle	38	7,7 %	1-8 modal expressions	36	6,4 %
1-5 voice	12	2,4 %	1-15 modifiers	33	5,9 %
1-10 formal nouns	12	2,4 %	1-5 voice	29	5,1 %
1-4 composed particles	10	2,0 %	1-10 formal nouns	23	4,0 %
1-8 modal expressions	7	1,4 %	1-11 demonstratives	13	2,3 %
1-13 word level conjunction	6	1,2 %	1-4 composed particles	9	1,6 %
1-3 final particles	3	0,6 %	1-13 word level conjunction	6	1,0 %
1-11 demonstratives	3	0,6 %	1-3 final particles	1	0,2 %
1-12 interrogatives	1	0,2 %	1-12 interrogatives	0	0,0 %
SUM	493	100 %	SUM	564	100 %

By comparing the two tables, in 8 of the 14 categories changes in appearance percentage can be observed. At the transition from beginner to intermediate level a decrease of occurrence can be seen in errors connected to the use of:

- case particles (26,2 % → 22,5 %) – however still the most common error category;
- basic sentence structure (20,2 % → 7,4 %);
- modifiers (15 % → 5,9 %);
- tense and aspect (11,8 % → 9,0 %).

An increase in occurrence can be seen in errors connected to the use of:

- sentence level conjunction (8,1 % → 11,8 %);
- focus particles (7,7 % → 16,8 %);
- voice (2,4 % → 5,1 %9;
- modal expressions (1,4 % → 6,4 %).

Aditionally, by comparing the two tables a more equal spread of error percentage across all categories can be observed. This can be explained by the fact that the students on the intermediate level use a wider range of grammatical structures and grammar types from all groups, which causes a higher diversity in error types.

Below is a table comparing the supposed causes attributed to the errors on each level.

Table 10: Comparson of error causes on both levels

Analysis of errors on the beginner level			Analysis of errors on the intermediate level		
wrong choice *gosentaku*	214	43,40 %	wrong choice *gosentaku*	322	57,10 %
lack of use *gofusoku*	164	33,30 %	lack of use *gofusoku*	133	23,60 %
form error *gokeitai*	55	11,20 %	addition *gofuka*	64	11,30 %
addition *gofuka*	50	10,10 %	form error *gokeitai*	42	7,40 %
wrong position *goichi*	10	2 %	wrong position *goichi*	3	0,50 %
SUM	493	100 %	SUM	564	100 %

Through comparison of the results, the following conclusions can be drawn:

- the most common cause of errors on both levels is due to wrong choice;
- at the transition from beginner to intermediate level an increase in the errors caused by wrong choice can be observed;
- on both levels a considerable ammount of errors was also caused by lack of use – however the percentage decreased by almost 10 % when transitioning to the intermediate level;
- the errors caused by error in form decreases when transitioning to the intermediate level.

4 Overall discussion

The following subsections compare the results of the error analysis on the beginner and intermediate level.

4.1 Determining problematic errors

In the cases where a substantial reduction in the appearance rate of an error category was observed, it was interpreted as, depending on the degree of reduction, successfully alleviated; on the other hand, error groups in which a decrease in appearance rate was hardly present, non-existent or an increase of appearance rate was observed, were interpreted to be potentially problematic and were therefore marked and examined more carefully.

4.2 Errors concerning particles

Error types connected to particles (especially case particles) tend to carry over from the beginner level to the intermediate level, and are the most common type of errors on both levels.

Errors in the use of the case particle *ga* tend to carry over to the intermediate level most; while the most common cause for such mistakes is confusing its use with the focus particle *wa*.

Errors connected to the use of the focus particle *wa* present one of the most common error types on both levels. With the transition to the intermediate level an increase of such levels can be observed. This suggests that a further increase might be present in the transition to the advanced level as well. Most commonly the cause of these errors is due to confusing its use with the case particle *ga*.

While errors connected to the case particle *wo* do tend to carry over to the intermediate level, they appear less commonly.

Errors connected to the case particles *de* and *ni* are especially common and seem to carry over to the intermediate level. The predominant cause for these errors is due to learners confusing the use of one with the other.

Errors connected to the attributive particle *no* present the most common type of error on the beginner level. However, through the transition to the intermediate level these types of errors are far less common, which suggests that the learners seem to be growing accustomed to its use. A further decrease might appear at the transition to the advanced level.

4.3 Other error groups

On the beginner level learners seem to struggle with the use of the copula *da*/*desu*. Such errors are hardly present on the intermediate level.

Errors connected to verb and adjective conjugation are very rare on the intermediate level, in contrast to their prevalence on the beginner level, indicating that learners on the intermediate level are already fairly familiar with the conjugations and forms of the adjectives and verbs, thus most of the cases of misuse actually appear to be mistakes rather than errors. The difference between the two is that mistakes happen accidentally (typos, etc.), unlike errors, which happen due to a lack of knowledge (the student has incorrect information on the use of a specific grammatical element).

The same reduction can be observed with errors connected to the use of the past tense of adjectives and verbs.

Errors in the use of sentence level conjunctions are less common on the beginner level, where the learners are only familiar with a small amount of such grammatical structures. They were mostly observed in cases of enumeration and basic sentence conjunctions. On the intermediate level however, an increase in all of the subcategories was observed. This can be attributed to the fact that the learners on the intermediate level are familiar with a much wider range of different conjunctions, which makes for a higher chance of an incorrect one being used. Furthermore, in many cases the errors occur due to conjunctions being mistakenly used in the place of other conjunctions within the same subcategory (i.e. potential clauses).

4.4 Error causes

The types of errors that proved most persistent were those caused by wrong choice – errors where a grammatical element is used instead of another one.

Errors caused due to wrong form of a grammatical element are fairly common on the basic level, but tend to disappear when transitioning to the intermediate level.

4.5 Comparison to previous studies

When comparing the results of the analysis with those of preceding analysis' quite a few similarities can be observed. Similar to Ichikawa (1993) the ratio of errors due to misuse of conjunctions is fairly high. Similar to Kawaguchi (1995) and Yō (2014) the most common type of mistakes are mistakes connected to the use of particles, especially case particles.

5 Conclusions

Having conducted the present research, we have recognized several limits and will here introduce possibilities for their improvement.

The first point we would like to highlight is the scope of the corpus. It is currently comprised of 182 texts (142 shorter and 40 longer) written by students at both levels. Compared to other corpora, this number is quite low. For the purposes of future research, and in particular to increase the credibility of the results, both sub-corpora will need to be expanded and a corpus of advanced learners added.

Another point that should be improved is the categorization. Initially, the categorization was created to be used with a corpus, but given that it was not made specifically for this one, categorization, made specifically for this corpus should be made. Yō 2014 also highlights the lack of a generally established standard for categorizing grammar errors in the Japanese language as a common problem.

Usually, when annotating and categorizing errors in the creation of a learner's corpus, the work is done in groups, then the errors are determined according to the most commonly marked category. Because the categorization process has mostly been done individually a revision of the categorized errors will be needed. When the corpus is made publicly available, a system, that allows the users to submit suggestions or report errors will be set up, so that the corpus and the data within can constantly keep evolving and improving.

Another possibility for improvement is the optimization of software used as a corpus framework. As mentioned in 4.2, Microsoft Excel is currently used for the corpus framework. Although it currently meets all the needs of the corpus and has many positive features, with the growth of the corpus there will also be a need for a tool that makes it easier to add and annotate texts, analyze content and the like.

Last but not least, while findings obtained from both of the sub-corpora analyzes certainly provide useful data with a sufficient degree of credibility, due to the small size of the corpus, an adequate measure of criticality is also required when interpreting the results. As mentioned in the introduction, the purpose of the analysis was to provide students and teachers with an insight into the most common types of grammar errors and to, through the construction of the corpus, take the first step towards the final goal of an online corpus of Slovenian Japanese students. While further research is indeed required in this area, the goals set at the beginning of the analysis have been achieved.

References

Corder, S. P. (1967). The Significance of Learner's Errors. *IRAL* 1967 (5), pp. 161-170.

Harasawa, I. 原沢伊都夫. (2012). Nihongo sho chūkyū gakushū-sha no sakubun shidō: Gakushū-sha no goyō bunseki o moto ni [日本語初中級学習者の作文指導：学習者の誤用分析をもとに] (Composition learning for learners of Japanese on the basic and intermediate level, based on an analysis of learner errors). *Shizuokadaigaku kokusai kōryū sentā kiyō* 静岡大学国際交流センター紀要, 6, pp. 79-92. Accessed 2. 9. 2018. https://ci.nii.ac.jp/naid/110008917835

Ichikawa, Y. 市川保子. (1993). Chūkyūreberu gakushūsha no goyō to sono bunseki - fukubun kōzō shūtoku katei o chūshin ni [中級レベル学習者の誤用とその分析—複文構造習得過程を中心に—] (The errors of students on the intermediate level – with focus on the process of acquisitions of compound sentence structures). *Nihongo kyōiku* 日本語教育, 81, pp. 55-66.

Ichikawa, Y. 市川保子. (1997). *Nihongo goyō reibun shōjiten [日本語誤用例文小辞典] (Small dictionary of examples of misuse in Japanese)*. Tokyo: Bonjinsha.

Kawaguchi, R. 川口良. (1995). Chūjōkyū nihongo gakushūsha no sakubun ni miru goyō no ichirei [中上級日本語学習者の作文にみる誤用の一例] (Types of errors that appear in the compositions of learners of Japanese on the intermediate and advanced level). *Gengo bunka to nihongokyōiku 言語文化と日本語教育*, pp. 178-188.

National Institute for Japanese Language and Linguistics. (2016). Learner Corpus Study of Aquisiton of Japanese as a Second Language. *NINJAL*, http://lsaj.ninjal.ac.jp/, Accessed 10. 4. 2018.

Noda, H. 野田尚史, & Sakoda, K. 迫田久美子. (2019). *Gakushūsha kōpasu to nihongo kyōiku kenkyū 学習者コーパスと日本語教育研究 (Learners' Corpora and Japanese Language Education Research)*. Tokyo: Kurosio.

Otsuka, K. 大塚薫, & Masayoshi, H. 林翠芳. (2010). Chū jōkyū reberu no Nihon gogakushūsha no sakubun shidō — iken bun ni miru goi kanji shiyō oyobi goyō no bunseki kekka o fumaete — [中上級レベルの日本語学習者の作文指導—意見文にみる語彙・漢字使用及び誤用の分析結果を踏まえて—] (Teaching composition of Japanese language learners at middle and upper level-based on analysis of vocabulary, kanji use and misuse in opinion sentences). *Kōchidaigaku sōgō kyōiku sentā shūgaku ryūgakusei shien bumon kiyō 高知大学総合教育センター修学・留学生支援部門紀要,* 4, pp. 47-66. Accessed 2. 9. 2018. https://ci.nii.ac.jp/naid/120002187909

Suzuki, T. 鈴木智美. (2002). 2000-nendo chūkyū sakubun ni mirareru goi imi ni kakawaru goyō — sho chūkyū reberu ni okeru goi imi kyōiku no jūjitsu o mezashite [2000 年度中級作文に見られる語彙・意味に関わる誤用—初中級レベルにおける語彙・意味教育の充実を目指して—] (Misuse of vocabulary and semantics found in the composition of students on the intermediate level in the year 2000 - Aiming at enhancement of vocabulary and semantics education on the beginner and intermediate level -). *Tōkyōgaikokugodaigaku ryūgakusei nihongo kyōiku sentā ronshū 東京外国語大学留学生日本語教育センター論集*, 28, pp. 27-42. Accessed 2. 9. 2018. http://repository.tufs.ac.jp/bitstream/10108/20943/1/jlc028003.pdf

Teramura, H. 寺村秀夫. (1990). *Gaikokujingakushūsha no nihongo goyōreishū* [外国人日本語学習者の日本語誤用例集] (Collection of misuse of foreign Japanese learners). Teramuragoyōreishū database 寺村誤用例集データベース. Accessed 15. 1. 2018. http://teramuradb.ninjal.ac.jp/teramura.goyoureishu.pdf

Umino, T. et al. (2012). Learners' Language Corpus of Japanese. *Tokyo University of Foreign Studies*. Accessed 1. 9. 2018. http://cblle.tufs.ac.jp/llc/ja/index.php?menulang=en

Yō, H. 楊帆. (2014). Chūkyū Nihongo gakushūsha no sakubun ni okeru konnan-ten: Bun kōzō no koōkankei ni tsuite [中級日本語学習者の作文における困難点 : 文構造の呼応関係について] (Difficulties in the compositions of Japanese learners on the intermediate level: on correspondence of sentence structure). *Akitadaigaku kokusai kōryū sentā kiyō 秋田大学国際交流センター紀要*, 3, pp. 15-28. Accessed 2. 9. 2018. https://ci.nii.ac.jp/naid/110009768148/en/