ACTA LINGUISTICA ASIATICA

Univerza v Ljubljani
FILOZOFSKA FAKULTETA

ACTA
LINGUISTICA
ASIATICA

Volume 6, Issue 2, 2016

TABLE OF CONTENTS

Acta Linguistica Asiatica

Volume 6, Issue 2, 2016

ACTA LINGUISTICA ASIATICA
Volume 6, Issue 2, 2016

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. Branka Kalenić Ramšak, 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:
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

FOREWORD

With the year winding down, when people tend to reflect on their accomplishments in the past months, we dare offer miscellaneous to boast the broadness that ALA has been gaining. This number of the ALA journal thus includes different views on a language, a variety of different methodologies used and languages discussed, as well as several applicable fields the research outcomes take us to.

LI Wenchao in her work "Adjective Distribution in Modern Mongolian" investigates a scalar structure of adjective distribution in Mongolian to find out that, compared to Japanese, English, French, Mongolian only tolerates inherent resultatives.

Jasmina BAJRAMI's article "Speech level shift in Japanese and Slovene" is a part of her PhD thesis, which she has just submitted to University of Ljubljana. In it she discusses formality and politeness in general, as well as the extent and the ways they are executed in a Japanese and Slovene conversation.

"Interpretation of Daba Script: Gemu from Wujiao Villag" written by **XU Duoduo** is a result of a fieldwork research on the Daba script found in the Daba Calendar entitled Gemu, which origins in Wūjiǎo village, the province of Sichuan.

Nina GOLOB conducted a bidirectional perception experiment entitled "L1 Prosodic Interference: the Case of Slovene Students of Japanese", in which she evaluates L1 prosodic interference in recognizing (lexical) accent place in declaratives and interrogatives.

An interesting project report on the successful implementation of the e-learning system for Japanese, Korean, Hindi and Sanskirt comes from Croatia. Written by **Sara LIBRENJAK, Kristina KOCIJAN, and Marijana JANJIĆ** it is entitled "Improving Students' Language Performance Through Consistent Use of E-Learning: An Empirical Study in Japanese, Korean, Hindi and Sanskrit".

Last but not least is a technical note by **Qasier ABBAS** "Semi-Semantic Annotation: A Guideline for the URDU.KON-TB Treebank POS Annotation", in which the author proposes annotation guidelines of semi-semantic parts of speech for the URDU.KON-TB treebank.

Nina Golob

RESEARCH ARTICLES

ADJECTIVE DISTRIBUTION IN MODERN MONGOLIAN[1]

LI Wenchao
University of Zhejiang, China
widelia@zju.edu.cn

Abstract

This paper discusses adjective distribution in Mongolian based upon the mereological framework: scale structure. It investigates how adjectival complements are sensitive to the scalar structure of adjectival predicates (APs) in resultative constructions as well as direct perception expressions. The findings reveal that Mongolian only tolerates inherent resultatives; derived resultatives are ruled out. The acceptability of adjectival complements in inherent resultatives runs from 'Totally open-scale/Totally closed-scale' down to 'Lower closed/Upper closed-scale'. On the other hand, adjectival complements in direct perception expressions are of no diverse acceptability, i.e. all layers of APs are licensed. Furthermore, durative verbs are likely to yield open-scale APs whilst punctual verbs seem to favour closed-scale APs.

Keywords: Mongolian; adjective predicates; scale structure

Povzetek

Članek obravnava distribucijo pridevnika v mongolščini po mereološki teoriji skalarnih struktur. Avtor razišče občutljivost pridevniških dopolnil na skalarno strukturo pridevnikov v povedkovni rabi tako v posledičnih strukturah kot tudi v izrazih osebnega mnenja. Rezultati razkrivajo, da mongolščina dopušča samo inherentni vzročno-posledični odnos pridevnikov v povedkovni rabi, kjer pa sprejme pridevniška dopolnila z zelo raznolikim odnosom do pridevnika; od popolnoma odprtega/zaprtega do delno odprtega/zaprtega. Nasprotno je v izrazih osebnega mnenja, kjer so sprejemljiva bolj ali manj vsa pridevniška dopolnila, pri čemer nedovršni glagoli privlačijo odprti tip pridevnikov v povedkovni rabi, dovršni pa zaprti tip.

Ključne besede: mongolščina; pridevniki v povedkovni rabi; skalarna struktura

[1] This paper is based upon work supported by *National Foundation of Social Science* (15CYY002) China to Li Wenchao.

I would like to thank the anonymous reviewers and the editors for their comments, which have helped me sharpen and develop the manuscript a great deal. All remaining errors and shortcomings are entirely mine.

Acta Linguistica Asiatica, 6(2), 2016.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.6.2.9-22

1 Introduction

Mongolian, an Altaic language family member, is an exclusively suffixing agglutinative SOP language.[2] A salient feature of the language lies in that adjectival complements may directly precede the verbs, as seen in direct perception expression (1a) as well as in a resultative construction (1b).

(1) a. *Direct perception expression*

Нүцгэн	би	түүнийг	олж	харсан.	(well-formed)
naked	I	her.Acc	find	see-PAST	

'I found her naked.'

b. *Resultatives construction*

Тэр	ханаа	**улаан**/улаанаар	будсан.	(well-formed)
she	wall	red/red.Instr	paint-PAST	

'She painted the wall red.'

As far as (1b) is concerned, the adjective *улаан* 'red' is licensed in the resultative construction. However, not all adjectival complements appear to be welcome in Mongolian. The following adjective *хатуу* 'solid' is ruled out in the resultative construction, c.f. (2).

(2) *Нуур **хатуу_биет** хөлджээ. (ill-formed)
lake solid freeze-PAST
'The lake froze solid.'

The ungrammaticality of (2) lies in that the adjective *хатуу* 'solid' is a closed-scale AP whilst the adjective *улаан* 'red' in (1b) is an open-scale AP. It appears then that Mongolian resultatives seem to only license closed-scale APs. This linguistic phenomenon is the opposite of English resultatives. English only tolerates closed-scale APs as resultative complements, with open-scale APs being ruled out.[3] (3) provides the illustrations.

(3) a. Bill pounded the metal ***long**. (Open-scale AP: ill-formed)
b. Bill pounded the metal **flat**. (Closed-scale AP: well-formed)

The ungrammaticality of (3a) lies in that open-scale adjectives fail to describe certain culmination points (see Beavers, 2008; Wechsler, 2005; and Wyngaerd, 2001 for further discussion).

[2] There are three writing systems in Mongolian: Todo Bicig (Xinjiang area), Traditional Mongolian alphabet (Hudum) (Inner Mongolia) and Cyrillic Mongolian (Outer Mongolia). In this study, Cyrillic Mongolian was adopted. A list of Mongolian Cyrillic alphabet is provided at the end of the paper.

[3] A detailed explanation regarding *open-scale* and *closed-scale* APs will be given in Section 2.

Intriguingly, in terms of perception expression, on the contrary, English adjectives are of no diverse acceptability, i.e. both open and closed-scale APs are allowed, as illustrated by (4).

(4) a. Bill saw Mary **exhausted**. (Open-scale AP: well-formed)
b. Bill saw the dog **dead**. (Closed-scale AP: well-formed)

It is also salient to mention that German language, a relative of English, permits both open-scale and closed-scale APs in resultative constructions, as in (5). On the other hand, German has diverse acceptability of direct perception expressions: the distribution of APs runs from 'Lower closed/Upper closed-scale' down to 'Totally open-scale/Totally closed-scale', as in (6).

(5) *Resultative construction in German*

a. Mary hämmerte das Eisen **flach**. (Closed-scale AP: well-formed)
Mary hammer the metal flat
'Mary pounded the metal flat'

b. Bill hämmerte das Eisen **lang.** (Open-scale AP: well-formed)
Bill hammer the metal long
'Bill pounded the metal long'

(6) *Direct perception expression in German*

a. *Mary sah ihn müde. (Totally open-scale AP: ill-formed)
Mary see-PAST him tired
'Mary saw him tired.'

At this stage, then, it seems that the distribution of APs in German direct perception expression somehow resembles Modern Mongolian.

Resultative construction has been studied intensively in linguistic typological work. Various frameworks have been raised, intending to achieve a more thorough analysis on the subject. Below is a map of major theories that have contributed to this subject, based on Chomsky's (1965) 'Aspects Model', Levin and Rappaport Hovav's (1995) 'Projection Approach', Goldberg's (1995) 'Construction Grammar Approach', Jackendoff's (1997) 'Adjunct Analysis', Rappaport Hovav and Levin's (1998) 'Event Structure Template', Boas's (2003) 'Dynamic Usage-Based Model', and Kennedy, Christopher and Louise McNally's (2005) 'Scale Structure'.

Aspect Model
↓
Lexical-Functional Grammar
↓
Projection Approach
↓
Construction Grammar Approach, Adjunct Analysis, Event Structure Template
↓
A Dynamic Usage-Based Model
↓
Scale Structure

The distribution of APs in Mongolian seems to receive far less attention compared to those of European languages. This study tentatively explores how APs show sensitivity to the scalar structure in resultative as well as direct perception constructions.

The paper is mapped out as follows: section 2 sheds light on the framework that is adopted in the analysis, i.e. scalar structure. With this in place, section 3 first examines the scalar properties of Mongolian APs and then delves into APs' distribution in resultative constructions. Section 4 moves on to explore APs in direct perception expressions. Section 5 highlights the results and concludes the paper.

The Mongolian Cyrillic alphabet is adopted in this study and hand-made examples are used. Native speakers checked the examples. Due to numerous dialects in the country, native speakers from the Inner and Outer parts of Mongolia have both been asked to provide judgements. Moreover, a multilingual parallel electronic dictionary is employed: http://asuult.net/dic/.

2 Framework: scalar structure

A *scale* is constituted of a set of degrees (points or intervals indicating measurement values) on a particular dimension (e.g. cost, depth, height, temperature), with an ordering relation. The dimension represents an attribute of an entity, with the degrees indicating the possible values of this attribute (Kennedy and McNally, 2005). *Scale* measures the change undergone by the incremental theme, patient or figure participant. The following variations have been established:

(a) Totally open-scale: a scale may have neither a minimal nor maximal element
(b) Lower closed-scale: a scale may have a minimal but no maximal element
(c) Upper closed-scale: a scale may have a maximal but no minimal element
(d) Totally closed-scale: a scale may have both maximal and minimal elements

Kennedy and McNally (2005)

Scale structure may apply to adjectives[4]:

(7) a. Totally open-scale: tired, *müde*, nervous, *nervös*, sad, *traurig* , deep, *tief*, long, *lang*
b. Lower closed-scale: drunk, *betrunken*, shave, *rasiert*, pale, *blass*, flat, *flach*
c. Upper closed-scale: pure, *rein*, naked, *nackt*, unshaved, *unrasiert*
d. Totally closed-scale: dead, *tot,* empty, *leer*

Scale structure may apply to verbs:

(8) a. Totally open-scale: shift, *utsuru*, ply, *kayou*
b. Lower closed-scale: approach, *yoru*, leave, *hanareru*, flit, *kasumeru*
c. Upper closed-scale: return, *modoru*, cross, *wataru*, return, *kaeru*, pass, *heru*
d. Totally closed-scale: arrive, *tsuku*, reach, *todoku*, transcend, *koeru*, descend, *oriru*, exit
e. Non-scale change: tumble, *korogaru*, swim, *oyogu*, run, *hashiru*, crawl, *hau*, dance, *odoru*

Scale structure may apply to prepositions/postpositions:

(9) a. Totally open-scale: toward, *e*
b. Lower closed-scale: from, *kara*; *yori*
c. Upper closed-scale: *e-to*[5]
d. Totally closed-scale: until, *made*; along, *ni sotte*

The scalar properties of APs, verbs and (pre) postpositions (PPs) further link to the combinatorial possibilities with verbs. Wechsler (2005, p. 264) indicates that resultative constructions with maximal endpoint adjectives often have durative verbs, e.g. *wipe, pull*. On the other hand, non-gradable adjectives are more likely to occur with punctual verbs, e.g. *shoot, kick.* This view is supported by Beavers (2008), who proposes that punctual verbs tend to yield non-gradable de-verbal adjectives while durative verbs tend to yield gradable de-verbal adjectives.

As such, *scale structure* should be a great help in detecting the syntactic distribution of APs in Mongolian.

3 The distribution of adjectives in resultative constructions

With the classifications of adjective highlighted, this section proceeds by looking into: (a) the scalar property of Mongolian APs; and (b) their distributions in resultative constructions. (10) provides a list of the most-used adjectives.

[4] In (7)-(9), English lexicons are in normal font; the corresponding lexicons of other languages are in italic. (7) are German adjectives; (8) and (9) are Japanese adjectives.

[5] *e-to* is between direction *e* and delimitation *made*.

(10) эцсэн (tired) / гүн (deep) / урт (long) / хатуу (hard) / нүцгэн (naked) / үхсэн (dead) / өвчин (sick) / чийг (wet) / хавтгай (flat) / сэрїїн (awake) / хоосон (empty)

An examination in terms of the modifiers *жинхэнэ* 'very' and *хагас* 'half' is carried out. Three Mongolian native speakers provided their judgements. The results are given in (11).

(11) a. ***жинхэнэ* (very)**

жинхэнэ *чийг*	(very wet)	[natural]
жинхэнэ *эцсэн*	(very tied)	[natural]
жинхэнэ *гүн*	(very deep)	[natural]
жинхэнэ *урт*	(very long)	[natural]
жинхэнэ *хатуу*	(very hard)	[natural]
* жинхэнэ *өвчин*	(very sick)	[unacceptable]
? жинхэнэ *хавтгай*	(very flat)	[unnatural]
? жинхэнэ *нүцгэн*	(very naked)	[unnatural]
жинхэнэ *үхсэн*	(very dead)	[natural]
жинхэнэ *сэрүүн*	(very awake)	[natural]
жинхэнэ *хоосон*	(very empty)	[natural]

b. ***хагас* (half)**

хагас **чийг**	(half wet)	[unnatural]
? хагас *эцсэн*	(half tied)	[unnatural]
хагас *гүн*	(half deep)	[natural]
? хагас *урт*	(half long)	[unnatural]
? хагас *хатуу*	(half hard)	[unnatural]
? хагас *өвчин*	(half sick)	[unnatural]
? хагас *хавтгай*	(half flat)	[unnatural]
? хагас *нүцгэн*	(half naked)	[unnatural]
? хагас *үхсэн*	(half dead)	[unnatural]
хагас *сэрүүн*	(half awake)	[natural]
? хагас *хоосон*	(half empty)	[unnatural]

Based upon the test, we may arrive at the following classification of Mongolian APs:

(12) a. Totally open-scale: *эцсэн* (tied), *урт* (long), *гүн* (deep)
b. Lower closed-scale: *чийг* (wet), *өвчин* (sick), өлөссэнийг (hungry), хусугсанийг (shaved)
c. Upper closed-scale: *хатуу* (hard), *сэрүүн* (awake), чайсанийг (pale)
d. Totally closed-scale: *нүцгэн* (naked), *үхсэн* (dead), *хоосон* (empty), *хавтгай* (flat)

In light of the classification, we move onto inquiring as to how APs are distributed in resultative constructions. Tests along with the four different scalar structures of APs are provided in (13) to (16). Native speakers again assessed the examples.

(13) *Totally open-scale AP*

a. жан сан: төмөр таяг дабдажъ **урт** болгаб. (natural)
Zhang san metal stick pound long PAST
'Zhang san pounded the metal long.'

b. жан сан: өрөөн өсөрч **ядараб**. (natural)
Zhang san himself dance tired-PAST
'Zhang san danced himself tired.'

(14) *Lower closed-scale AP*

a. жан сан: өрөөн иниесээр байжъ **Өбчитай** болоб. (natural)
Zhang san himself laugh PROG sick PAST
'Zhang san laughed himself sick.'

b. ?жан сан: өбөст газар усулжъ **чийг** болгаб. (unnatural)
Zhang san garden water wet PAST
'Zhang san watered the garden wet.'

(15) *Upper-closed scale AP*

a. *жан сан: ус хӨ[6] лдээжъ **хатуу** болгаб. (ill-formed)
Zhang san water Ba freeze solid PAST
'Zhang san froze the water solid.'

b. жан сан: лъ сийг сажилажъ **сэргээб**. (natural)
Zhang san Li si shake awake PAST
'Zhang san shook Li si awake.'

(16) *Totally closed-scale AP*

a. ?жан сан: төмөр таяг дабдажъ **хабтагай** болгаб. (natural)
Zhang san metal stick pound flat PAST
'Zhang san pounded the metal flat.'

b. *Нуур **хатуу_биет** хөлджээ. (ill-formed)
lake solid freeze-PAST
'The lake froze solid.'

We now have several layers illustrating the acceptability thresholds of Mongolian APs in resultatives, running from 'Totally open-scale AP' down to 'Lower closed-scale AP, Upper-closed scale AP, Totally closed-scale AP'.

Note that (13) are inherenet resultatives.[7] It seems that Mongolian lacks derived resultative constructions. All layers of APs appear to be ruled out.

[6] хӨ is a co-verb.

[7] The terminology used to describe the two types of resultative constructions varies: Kageyama (1996) labels them as 'inherent resultatives' vs. 'derived resultatives'; Washio (1997) refers to them as 'strong resultatives' vs. 'weak resultatives'; Iwata (2006) uses the terms 'argument resultatives' vs. 'adjunct resultatives'; and in Levin & Rappaport Hovav (1995) and Kennedy's (1999) works,

(17) *Mongolian derived resultatives* (unnatural)

? жан сан: арих уужъ хогослаба ба.
Zhang san pub drink empty PAST
'Zhang san drank the pub empty.'

Another Altaic language, Japanese, also lacks derived resultative constructions, as shown in (18):

(18) a. *Japanese inherent resultatives* (well-formed)

Taro wa kabe o siroku nutta.
Taroo TOP wall ACC white paint-PAST
'Taro painted the wall white.'

b. *Japanese derived resultatives* (ill-formed)

*Kanojo wa sakana o zerii joo ni tataita.
She TOP fish ACC jelly into pound-PAST
'She pounded the fish into a jelly.'

We might contend that Altaic languages are likely to miss derived resultative constructions. This feature is shared with Romance languages; Italian and French do not tolerate derived resultative constructions, as seen in (19) and (20).

(19) a. *Italian inherent resultatives* (well-formed)

Ho taglito la carne in piccolo pezzi.
have.1stSg cut.PPT the meat in small pieces
'I cut the meat into small pieces.'

b. *Italian derived resultatives* (ill-formed)
*Gianni ha martellato il metallo piatto.
Gianni has hammer-PPT the metal flat
'Gianni hammered the metal flat.' (Napoli, 1992, p. 65)

(20) a. *French inherent resultatives* (well-formed)

Je coupe la viande **en** morceaux.
I cut the meat PREP pieces
'I cut the meat into pieces.'

'control resultatives' vs. 'exceptional case-marking resultatives' is used. Moreover, Dimitrova-Vulchanova (2002) employs 'connected resultatives' vs. 'disconnected resultatives' to describe resultatives. All these terms differ slightly but ultimately refer to the same thing. The current paper follows Washio (1997).

b. *French derived resultatives* (ill-formed)

?Elle a battu le poisson en gelée.
She pounded the fish PREP jelly.'
'She pounded the fish into jelly.'

Moreover, another character of French resultatives is worth highlighting, namely, that only prepositional complements are permitted to denote the result, c.f. (21):

(21) a. J'ai peint le mur **en** rouge. (well-formed)
b. Jean a cassé le vase **en** morceaux. (well-formed)
c.f. c. *J'ai peint le mur *rouge*. (ill-formed)

Germanic languages, on the other hand, license both inherent and derived resultative constructions.

(22) a. *German inherent resultatives*

Bill fror das Wasser hart.
Bill freez-PAST the water hard
'Bill froze the water hard.'

b. *German derived resultatives*

Bill trank die Kneipe leer.
Bill drink-PAST the pub empty
'Bill drank the pub empty.'

(23) a. *English inherent resultatives*

Mary wiped the table clean.

b. *English derived resultatives*

She pounded the fish into a jelly.

4 The distribution of adjectives in direct perception expressions

Having drawn a picture of the sensitivity of APs in resultatives, we are in a better position to engage in the analysis of direct perception expression. Tests along with the four various scalar structures of Mongolian APs are carried out. Once more, native speakers provided the judgements.

(24) *Totally open-scale AP*

жан сан лъ сийн **ядарсанийг** үзэб. (natural)
Zhang san Li si tired see-PAST
'Zhang san saw Li si tired.'

(25) *Totally closed-scale AP*

a. жан сан нохойн **үхүргэнийг** үзэв. (natural)
Zhang san dog dead see-PAST
'Zhang san saw the dog dead.'

b. жан сан лъ сийн **нүцгэн нь олжъ** үзэв. (natural)
Zhang san Li si naked see-PAST
'Zhang san saw Li si naked.'

(26) *Upper closed-scale AP*

жан сан лъ сийн **чайсанийг** үзэв. (natural)
Zhang san Li si pale see-PAST
'Zhang san saw Li si pale.'

(27) *Lower closed-scale AP*

a. жан сан лъ сийн **өлөссэнийг** үзэв. (natural)
Zhang san Li si hungry see-PAST
'Zhang san saw Li si hungry.'

?b. жан сан лъ сийн **хусугсанийг** үзэв. (ill-formed)
Zhang san Li si shaved see-PAST
'Zhang san saw Li si shaved.'

The oddness of хусугсанийг 'shaved' (c.f. 27b) might have revealed that a Mongolian perception verb only denotes a direct perception report. It does not contribute to a potential indirect perception, describing the observer's conceptualisation of the perceived event. Nonetheless, Mongolian seems to welcome all layers of adjectival complements in direct perception expressions.

5 Conclusion

This paper has delved into the adjective distribution in resultative constructions as well as direct perception expressions. The findings show that Mongolian only tolerates inherent resultatives; derived resultatives are ruled out. The acceptability of adjectival complements in inherent resultatives runs from 'Totally open-scale/Totally closed-scale' down to 'Lower closed/Upper closed-scale'. On the other hand, adjectival complements in direct perception expressions are of no diverse acceptability, i.e. all layers of APs appear licensed. The foregoing discussion is summarised in Table 1.

Table 1: The distribution of Mongolian adjectives in resultative and direct perception constructions

Scalar property	Inherent resultatives	Derived resultatives	Direct perception expression
Totally open scale	*high*	all ruled out	all licensed
Lower closed scale			
Upper closed scale			
Totally closed scale	*low*		

The scalar properties of APs further link to the combinatorial possibilities with verbs. Durative verbs are likely to yield open-scale APs whilst punctual verbs seem to favour closed-scale APs. Moreover, the scalar properties can be a syntactic diagnostic for split intransitivity, i.e. unergative verbs are likely to yield closed-scale postpositions while unaccusative verbs tend to occur with open-scale postpositions.

Table 2: Scalar properties of APs along with the combinatorial possibilities with verbs and postpositions

Durative verbs $\Rightarrow$ Open-scale AP	Open-scale postpositions $\Leftarrow$ Unaccusative verb
Punctual verbs $\Rightarrow$ Closed-scale AP	Closed-scale postpositions $\Leftarrow$ Unergative verbs

Mongolian Cyrillic alphabet

а *a*	б *b*	в *v*	г *g*	д *d*	е *ye*	ё *yo*	ж *j*	з *dz*	и *i*	й *y*	к *k*
л *l*	м *m*	н *n*	о *o*	ө *ö*	п *p*	р *r*	с *s*	т *t*	у *u*	ү *ü*	ф *f*
х *kh*	ц *ts*	ч *ch*	ш *sh*	щ *shch*	ъ "	ы *i*	ь '	э *e*	ю *yu*	я *ya*	

References

Asbury, A., Gehrke, B., van Riemsdijk, H., & Zwarts, J. (2008). Introduction: Syntax and semantics of spatial P. In A. Asbury, J. Dotlačil, B. Gehrke, & R. Nouwen (Eds.), *Syntax and Semantics of Spatial P* (pp. 1–32). Linguistik Aktuell / Linguistics Today 120, Amsterdam: John Benjamins.

Beavers, J. (2008). Scalar Complexity and the Structure of Events. In J. Dolling, T. Heyde-Zybatow & M. Schäfer (Eds.), *Event Structures in Linguistic Form and Interpretation* (pp. 245–265). Berlin: Mouton de Gruyter.

Beavers, J., Levin B., & Tham, S.W. (2010). A Morphosyntactic Basis for Variation in the Encoding of Motion Events. *Journal of Linguistics* 46, 331–377.

Boas, H. C. (2003). *A Constructional Approach to Resultatives.* Stanford: CSLI Publications.

Boas, H. C., (2000). *Resultative constructions in English and German.* Doctoral Dissertation. Chapel Hill: University of North Carolina.

Bolinger, D. (1972). *Degree Words*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Croft, W., Barddal J., Hollmann W. B., Sotirova V., & Taoka C. (2001). *Discriminating verb meanings: the case of transfer verbs* (Autumn Meeting of the Linguistic Association Great Britain, Reading).

Dimitrova-Vulchanova, M. (2002). *On two types of results: resultatives revisited*. Ms NTNU Trondheim.

Folli, R., & Ramchand, G. (2005). Prepositions and results in Italian and English: An analysis from vent decomposition (pp. 1–20). In H. Verkuyl, A. van Hout & H. de Swart (Eds.), *Perspective on Aspects.* Dordrecht: Kluwer.

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Goldberg, A., & Jackendoff, R. (2004). The English Resultative as a Family of Constructions. *Language 80,* 532–568.

Hay, J., Kennedy C., & Levin B. (1999). Scalar structure underlies telicity in 'degree achievements'. *Semantics and Linguistic Theory* 9, 127–144.

Iwata, S. (2006). Argument Resltatives and Adjunct Resultatives in a Lexical Constructional Account. *Language Sciences* 28: 449–496.

Jackendoff, R. S. (1997). *The Architecture of the Language Faculty*. Cambridge, MA: MIT Press.

Kageyama, T, (1999). Word formation. In N. Tsujumura (Ed.), *The Handbook of Chinese Linguistics.* Malden, MA: Blackwell Publishers.

Kageyama, T. (1996). *Doushiimiron-gengo to ninchi no setten*. Kuroshio.

Kageyama, T. (2001). *Nichiei Taishoodooshi no Imi to Koobun*. Taishuukan Publisher.

Kennedy, C. & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language 81*(2), 345–381.

Kennedy, C. (1999). *Projecting the Adjectives: The Syntax and Semantics of Gradability and Comparison*. New York: Garland.

Kennedy, C., & Levin, B. (2008). Measure of Change: The Adjectival Core of Degree Achievements. In L. McNally & C. Kennedy (Ed.), *Adjectives and Adverbs: Syntax, Semantics and Discourse* (pp. 156–182). Oxford, UK: Oxford University Press.

Kitahara, H. (2009). Dooshinogoigainenkouzoo to ni ku no sukeirukoozoo, toogokoozoonimotoduita, chakutennkoobun to kekkakoobun no heikoosei (pp. 315–364). In N. Ono (Ed.), *Kekkakoobun no taiporojii*.

Levin, B. & Rappaport Hovav, M. (1995). *Unaccusativity*. Linguistic Inquiry Monograph 26, MIT Press, Cambridge, MA.

Levin, B. & Rappaport Hovav, M. (2010). *Lexicalized scales and verbs of scalar change*. Paper presented at the CLS 46.

Levin, B. (2013). *Resultatives Revisited*. Paper presented at Secondary Predication in Formal Frameworks. Utrecht.

Levinson, S. C., & Wilkins, D. P. (2006). The background to the study of the language of space. In Stephen. C. Levinson & David. P. Wilkins (Eds.), *Grammars of space: Explorations in cognitive Diversity* (pp. 1–23). Cambridge: Cambridge University Press.

Li, W. (2011). Event Argument in Adjectival Perception Complements in German: Comparison with English. *Ars Linguistica*, *18*, 217–230.

Ma, Z., & Lu, J. (1997a). Study on adjectives as resultatives I. *Study of Chinese Language*, (1), 3–7.

Melka, F. (2003). Verbes de mouvement et verbes resultatifs en langues romanes et Germaniques. In F. Sanchez Miret (Ed.), *Actas del XXIII Congreso Internacional de Linguiy Filologia Romanica 2/2* (pp. 55–63), Niemeyer, Tubingen.

Mihara, K. (2009). Sukeirukoozoo kara miru kekkakoobun. In N. Ono (Ed.), *Kekkakoobun no taiporojii*, 141–170.

Ono N. (2010). A Lexical Resource View of Motion and Resultative Constructions. Morphology and Lexicon Forum, oral presentation, National Institute for Japanese language and linguistics.

Rappaport Hovav, M., & Levin B. (1998). Building Verb Meanings. In M. Butt & W. Geuder (Ed.), *The Projection of Arguments: Lexical and Compositional Factors* (pp. 97–134), CSLI Publications, Stanford, CA.

Sapir, E. (1944). Grading: a study in semantics. *Philosophy of Science*, (11), 93–116.

Simpson, J. (1983). Resultatives. In L. Levin, M. Rappaport & A. Zaenen (Eds.), *Papers in Lexical-Functional Grammar* (pp. 143–157). Indiana University Linguistics Club, Bloomington.

Slobin, D. (1996). Two ways to travel: Verbs of motion in English and Spanish. In M. Shibatani & S. A. Thompson (Eds.), *Grammatical constructions: Their form and meaning* (195–219). Oxford: Oxford University Press.

Slobin, D. (1997). Mind, code, and text. In J. Bybee, J. Haiman & Sandra A. Thompson (Eds.), *Essays on language function and language type* (pp. 437–467). Amsterdam: John Benjamins.

Slobin, D. (2000). Verbalized events – a dynamic approach to linguistic relativity and determinism. In S. Niemeier & R. Dirven (Eds.), *Evidence for linguistic relativity*. John Benjamins, Amsterdam.

Talmy, L. (1985). Lexicalization patterns: semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description, vol. III: Grammatical categories and the lexicon* (pp. 57–149). Cambridge: Cambridge University Press.

Talmy, L. (2000a). *Toward a cognitive semantics, vol.1: concept-structuring systems.* Cambridge, Mass: MIT Press.

Talmy, L. (2000b). *Toward a cognitive semantics, vol.2: typology and process in concept structuring*. Cambridge, Mass.: MIT Press.

Tsujimura, N. (2001). A constructional approach to stativity in Chinese. *Studies in Language, 25*(3), 525–553.

Uegaki, W. (2009). *A degree-based semantics for Chinese resultatives*. Paper presented at the University of Tokyo.

Vanden Wyngraerd, G. (2001). Measuring events. *Language, 77*, 61–90.

Vendler, Z. (1967). *Linguistics in Philology*. Cornell University Press.

Washio, R. (1997). Resultatives, compositionality, and language variation. *Journal of East Asian Linguistics, 6,* 1–49.

Wechsler, S. (2000). An Analysis of English Resultatives Under the Event-Argument Homomorphism Model of Telicity. In *Proceedings of the 3rd Workshop on Text Structure*, University of Texas, Austin.

Wechsler, S. (2005). Resultatives under the 'event-argument homomorphism' model of telicity (pp. 255–274). In N. Erteschik-Shir & T. Rapaport (Eds.), *The syntax of aspect*. Oxford: Oxford University Press.

SPEECH LEVEL SHIFT IN JAPANESE AND SLOVENE

Jasmina BAJRAMI
University of Ljubljana, Slovenia
jasmina81@gmail.com

Abstract

In verbal communication, we always aim to establish and maintain harmonious relations with others. Proper use of expressions and the choice of the way we speak are closely connected with politeness. In Japanese speech level is a level of formality or politeness in conversation, which is expressed by the use of linguistic forms (formal vs. informal) within and at the end of an utterance and the use of honorific expressions. In Slovene the level of formality or politeness in conversation is mainly expressed by the use of formal language and general colloquial language. Speech level shift is a shift from one speech level to another – e.g. from a formal style to an informal, etc. According to previous research, these shifts express speaker's psychological distance and a change of attitude towards a hearer. In this paper I will first briefly present the theoretical framework of politeness and an outline of speech levels in Japanese and Slovene. I will then present the data and the method used in this study. Finally, I will present and discuss the results of the analysis of both Japanese and Slovene conversation.

Keywords: speech level; speech level shift; politeness; formal style; informal style

Povzetek

V govorni komunikaciji vselej stremimo k vzpostavljanju in ohranjanju skladnih medsebojnih odnosov. Primerna raba izrazov ter izbira načina govora sta tesno povezani z vljudnostjo. V japonskem jeziku stopnja govora pomeni stopnjo formalnosti ali vljudnosti v pogovoru, ki se izraža z jezikovnimi oblikami (formalnimi oz. neformalnimi) znotraj ali na koncu stavka ter s spoštljivimi izrazi. V slovenskem jeziku se stopnja formalnosti ali vljudnosti v pogovoru večinoma izraža z rabo knjižnega jezika in splošnega pogovornega jezika. Menjava stopnje govora je menjava iz ene stopnje govora v drugo – tj. iz formalnega načina govora v neformalnega. Predhodne raziskave kažejo, da takšne menjave stopnje govora odražajo govorčevo psihološko razdaljo in spremembo drže do sogovorca. V članku je najprej na kratko predstavljen teoretični okvir vljudnosti in podan opis stopenj govora v japonskem in slovenskem jeziku. Nato so predstavljeni gradivo in metode, ki so bile uporabljene v tej raziskavi. Na koncu so podani rezultati analize japonskega in slovenskega pogovora.

Ključne besede: stopnja govora; menjava stopnje govora; vljudnost; formalni stil; neformalni stil

Acta Linguistica Asiatica, 6(2), 2016.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.6.2.23-51

1 Introduction

Politeness is a linguistic behavior used to ensure the harmony of human relations and plays a crucial role in interpersonal communication.

This paper analyzes the use of speech level and speech level shift in Japanese and Slovene formal conversation from the perspective of Brown and Levinson's (1987) theory of politeness and aims to clarify (1) the conditions which determine the choice of a speech level, (2) the factors that cause speakers to change the speech level, and (3) the kind of functions that these shifts carry in the conversation..

2 Previous research

2.1 Politeness (Brown & Levinson, 1987)

Brown and Levinson's theory of politeness is a very prominent early work on politeness among politeness research in pragmatics and in the research of discourse politeness. According to the theory, politeness does not represent courteous manners – it is a linguistic strategy used to establish and maintain harmonious human relations.

The theory is based on the notion of "face", defined as "the public self-image that every member wants to claim for himself", which consists of two related aspects, a negative face and a positive face (Brown & Levinson, 1987, p. 61). Authors also point out that there are certain kinds of acts that intrinsically threaten the face and call them "face-threatening acts" (FTAs).

The main components of Brown and Levinson's Politeness theory are the notion of "face", formula for estimating the degree of seriousness or weightiness of FTAs, concrete descriptions of politeness strategies, and factors influencing the choice of strategies. According to the theory, politeness strategies are used depending on the weightiness of a face threat (W), which is regulated by three factors (social variables): social distance (D) between speaker and hearer, relative power (P) of speaker and hearer, absolute ranking of impositions in the particular culture (R). The greater the face threat, the more polite strategy is used (Brown & Levinson, 1987, pp. 61–84).

The theory has widely been criticized by numerous scholars from its beginnings. Critics claim that the theory is not appropriate for describing those Asian cultures, in which the use of certain language forms used to express social distance is obligatory. Matsumoto (1988, 1989), for instance, believes that Brown and Levinson's claim that politeness strategies are used to mitigate speech acts which threaten face does not apply to Japanese because in Japanese politeness strategies are often used even when there is no face-threatening act. On the other hand, some scholars believe that the theory holds an all-embracing approach and that the concept of negative politeness does apply to Asian languages. Even more, they believe that its universal characteristics can account for a better understanding of a discourse in Japanese (Usami, 2001;

Pizziconi, 2003). Usami (2002) claims that Brown and Levinson's theory is universal and appropriate for comparisons of politeness systems in different cultures, if it is used at the level on discourse, that is, if we do not focus merely on utterances or sequences of short conversations.

It seems that the common problem of the critique and interpretation of Brown and Levinson's theory from the Japanese language perspective (such as Ide, 1989, and Matsumoto, 1988) lies in the fact, that it seems to only focus on the Japanese honorifics and their analysis, and therefore explanations tend to be overly detailed. In dealing with universality of politeness, it is important to look for common points and characteristics of different languages and not only point out the characteristics and exceptions in Japanese. For this reason, in order to analyze the dynamics of the use of politeness strategies, this study employs Brown and Levinson's theoretical framework and tries to determine, how the three factors – social distance, social power, absolute ranking of imposition – are interrelated and how they influence the use of politeness strategies in disagreements.

2.2 Previous research in Japanese

2.2.1 Speech level shift in Japanese

In Japanese speech level is a level of formality or politeness in conversation, which is expressed by the use of linguistic forms within and at the end of an utterance: by polite/non-polite copula *desu/da* and polite/non-polite verb suffix *-masu/-dearu* within and at the end of an utterance, and by the use of honorific expressions *keigo*.

Speech level shift in Japanese indicates a shift from formal *(desu*, *-masu*) to a plain form (*da*, *-dearu*) and vice versa, and a shift from using polite/formal expressions of *keigo* to using plain expressions and vice versa. The research of speech level shift in Japanese is relatively extensive: Ikuta and Ide (1983), Mimaki (1993, 2001), Usami (1995, 1999, 2001, 2002) etc. These researchers claim that speech level shift is a discourse strategy used to express speaker's psychological distance and a change of attitude towards the hearer. Moreover, they point out that speech level is influenced by three factors: social context (age, sex, degree of intimacy, educational background, etc.), psychological distance between the speakers and discourse unit (syntactic condition). Their analysis clarify that down shifts (a shift from a polite form to a plain form) as well as up shifts have the purpose of regulating a psychological distance between the speakers. Down shifts are seen mainly when the speaker expresses affection, support or empathy to the hearer. In such cases the shift has the role of shortening the psychological distance between the speakers. On the other hand, the up shifts are seen in more formal situations, for example, when the speaker tries not to invade the hearer's private territory. In such case, by lengthening the psychological distance, the conversation could continue without bumping into an obstacle and without the speaker in some way hurting the hearer. Moreover, the up shift is seen, when the speaker is nervous or there is tension in a conversation. Some of the functions speech level shifts carry in regard to

a discourse unit are stressing the importance of the topic or content of the conversation and adding an explanation or an example to a previous statement.

2.2.2 *Keigo*

Keigo is a honorific system in Japanese language. It is a special type of honorific speech and polite behavior, determined by relative social status of interlocutors or people who are the subject of the conversation. As a verbal behavior, *keigo* is a specific linguistic means for expressing the same referential meaning in different ways – i.e. on different levels of respect and politeness.

Keigo is a part of even more extensive system called *taigūhyōgen* (attitudinal expressions), linguistic expressions which are semantically the same, but differ depending on the addressee, content of the conversation, and different situations. These differences depend on how the speaker sees the addressee (e.g. if the addressee is hierarchically superior to the speaker, speaker will speak with them politely), or how the speaker behaves in certain situations. As a result of this "treatment", in addition to honorific expressions, there are also various expressions of negative meaning, love expressions, etc.[1] Among all these numerous attitudinal expressions, *keigo* represents only a part of specific expressions used for expressing respect and politeness.

In describing the characteristics of *keigo*, Minami (1987) points out *koryo* (consideration), evaluative attitude (*hyōkateki taido*) and a proper use of expressions (*hyōgen no tsukaiwake*). Depending on the subject of the speaker's consideration, Minami points out four different types of consideration. The first one is consideration oriented towards the hierarchical relations; depending on speaker's relationship (superior, subordinate, close, distant, etc.) with the hearer or a third person, the speaker chooses expressions of *keigo*. The second is consideration to the content of the conversation, where the speaker chooses appropriate expressions according to who or what the content is related to. The third type is consideration to situation, where a speaker chooses their expressions depending on the setting of the conversation. Fourth type is consideration to a medium. This means that the speaker might not use *keigo* when talking to someone in person, but would use it in a letter to this same person. According to Minami, speaker's evaluative attitude is always accompanied by these four types of consideration. In other words, the speaker will always make a choice of using certain expressions only after putting into consideration who the hearer is, what the setting of the conversation is, what the topic is etc. Moreover, this decision is mainly the result of three standpoints: hierarchical relations, level of closeness, and level of formality of the setting.

In a broad sense we can classify *keigo* into three main groups: *sonkeigo* (honorific language), *kenjōgo* (humble language) and *teineigo* (polite language). Furthermore, *teineigo* has a subcategory *bikago* (refined language).

[1] See Kikuchi 1997: p. 29-42 and Minami 1987: p. 12-16.

Sonkeigo is a group of honorific expressions referring to the addressee or the third person, their actions, and all things belonging to them, only when they are superior to the speaker. The speaker expresses respect by using honorific expressions for everything concerning that person or thing, by which he puts them higher than himself.

Kenjōgo is a group of humble expressions, referring to the speaker and speaker's actions, which have some kind of effect on the addressee or the third person (who are superior to the speaker). With *kenjōgo* the speaker shows respect by humbling himself and his actions, which puts the addressee or the third person above the speaker.

In contrast with the first two groups presented above, *teineigo* is not used to express respect, but to express formality and politeness of the speaker towards the addressee according to their relationship or the situation. *Teineigo* is formed with copula *desu* and a polite verb form *masu*. Moreover, it can also be formed with copula *degozaimasu*, which is even more polite than the first two and is usually used in very formal or specific situations (i.e. answering a phone at a work place). *Desu* is a formal form of the auxiliary verb *da* and is attached to nouns and adjectives, whereas *masu* is a formal form of verbs.

Speaker's choice of the way of speaking, e.g. use/non use of honorific expressions, proper use of the speech level, the use of certain vocabulary, etc., is closely related to Brown and Levinson's politeness theory. According to their theory, in interpersonal communication speakers use politeness strategies in order to preserve harmonious relationships. As mentioned above, Minami (1987) points out that in Japanese the use of *keigo* depends on human relations (i.e. the level of intimacy and hierarchy), content of the conversation, situation, the setting (formal/informal), etc. For example, the use of formal and informal style in Japanese (*teineitai/futsūtai*) depends not only on the level of intimacy, but also on the setting or the situation of the conversation. This means that people who are close to each other normally use an informal language in an every-day conversation. However, in a formal setting they would usually use a formal language. Looking from the viewpoint of Brown and Levinson's politeness theory, by using *keigo*, the speaker is using a negative politeness strategy, intended to satisfy the hearer's negative face. On the other hand, by using an informal language the speaker is attending to the hearer's positive face, which means he is using a positive politeness strategy.

2.3 Previous research on Slovene

2.3.1 "Formal" and "general colloquial" Slovene language

Toporišič (1973, 2000) divides Slovene language into two main categories – *social* and *functional* category. Furthermore, he divides the social category into literary language and non-literary language. Non-literary language is the spoken language with various dialects and regional colloquialisms, whereas literary language is the common language everywhere in Slovenia and is further divided into formal language and general colloquial language.

People who have received at least a secondary education are able to use the formal language. For example, formal language is used by writers, poets, reporters, and others for writing technical and scientific papers, newspaper and magazine articles, official documents etc. Moreover, formal language is used in everyday life as well – for writing journals, letters, emails, etc. It is also used for reading texts written for theater, television, radio, schools, meetings, governmental institutions etc., or speaking when these texts are learned by heart. Originally, formal language is used for writing, however, it is also widely used for speaking. For example, at educational institutions it is spoken by teachers, professors and students, and it is also spoken on television, radio, meetings etc. In other words, when there is some form of an audience, especially when there is a difference in education, dialects or social status between people, it is expected that speakers will use a formal language. Formal language is an old language with entirely determined orthography, pronunciation, accent and literary style and therefore Slovenian speakers know it well. However, its correct use depends mainly on the education and social environment of the speakers.

According to Toporišič (1973, 2000), general colloquial language is a freely spoken language, whose rules of grammar and pronunciation are not so rigid. Compared to formal language, general colloquial language differs mostly in pronunciation, however there are differences in functions, syntax and vocabulary. Regarding the pronunciation, perhaps the most obvious difference is the loss of *-i* at the end of a word in general colloquial language: *pisati* (to write, formal), *pisat* (to write, general colloquial). General colloquial language is mainly spoken in informal settings of a smaller number of people, where the social and psychological distance is often smaller, interlocutors can see each other's faces and the conversation is often accompanied by gestures and movements of arms and body. Because all this is especially connected to syntax, syntax of general colloquial language is a little different than the syntax of the formal language. Dialects have a strong influence on general colloquial language, but people, who have received at least secondary education, are able to separate themselves from their dialects and use a language which is closer to the formal language.

The base of formal language as well as general colloquial language is the language, spoken in Slovenian geographical, political and cultural center, the capital Ljubljana. And because in this paper we examine conversation in television debate show, which is a public and formal setting, the formal language and general colloquial language in this debate show is the subject of our analysis.

2.3.1.1 Differences between "formal" and "general colloquial" language

Toporišič (2000) points out the following categories where the most prominent differences are seen between formal language and general colloquial language: pronunciation, vocabulary, the way of writing, accent, linguistic forms, syntax structure. For example, very obvious differences between the two languages are: a) different pronunciation at the end of verb ending in male, singular form: *rekel* ((he)

said) is pronounced as *rekəu* in formal language and *reku* in general colloquial; b) the loss of short infinitive *-i*, c) difference in accent: *nosíti* (formal, to carry), *nôsit* (general colloquial, to carry), etc. Moreover, the pronunciation of general colloquial language is softer compared to the pronunciation of the formal language: short vowels *i ・u ・a ・ ə* are not pronounced as clearly in general colloquial language.

Another big difference, that we will look at more closely, is a difference in vocabulary. We can say that politeness level in Slovene differs according to the stylistic value of the vocabulary. According to their stylistic value, Toporišič (2000) divides vocabulary into stylistically unmarked words and stylistically marked words. The former are words which express the essential meaning of the word and do not express emotions, time, an origin (foreign) or a dialect, etc. They are used in both formal and general colloquial language, in all settings (formal and casual), and are most commonly known, which is why they are used more frequently than any other words. On the other hand, stylistically marked words are influenced by emotions, time, area, etc. and are therefore mostly used in colloquial language. For example stylistically unmarked word *glava* (head) is used as such in formal and general colloquial language, whereas its stylistically marked version *betica* is used in colloquial language and has a negative connotation.

Most words of the general colloquial Slovene language are stylistically unmarked and are frequently used in an everyday and casual conversation, therefore we can say that speakers are very accustomed to using them. This is why we can assume that these words would be commonly used in a formal setting as well. From the politeness point of view we can say that stylistically unmarked words have the highest level of politeness and are used in formal settings, whereas the politeness level, while depending on the type of the word, of stylistically marked words is lower and they are mainly used in a casual and private settings.

Pogorelec (1965) points out that in a formal discussion speakers use the short infinitive (for example, instead of formal *morali se bomo zavedati* (we must be aware), speakers say *moral se bomo zavedat* (general colloquial), which is common in colloquial language. Moreover, she found out that the loss of vowels in conjunction words and adverbs is very common. For example, *tud* (*tudi*) (also), *al* (*ali*) (or), *zlo* (*zelo*) (very), etc. According to Pogorelec, the differences in formal and general colloquial language mentioned above depend on the purpose and setting of the conversation. Whether or not someone can properly use formal language in a formal setting depends on their education, environment, social status, etc., therefore it is expected that, due to these factors, some people are not able to properly use formal language and only use general colloquial language or the mixture of the two.

Tivadar (2003) believes that in order to speak on a public television talk show, one must have a good command of formal language and says that the difference in using formal and general colloquial language is connected to the speaker's social and personal characteristics, as well as to their dialect. Moreover, he points out that in such public setting, shifts between a formal and general colloquial language are very frequent.

All the above researchers share a similar viewpoint regarding characteristics and the use of formal and general colloquial language. Speaker's background and education play an important role in the ability of having a good command of the formal language. Furthermore, according to these studies, differences in pronunciation and vocabulary are the most obvious differences between the formal and general colloquial language.

From all the above mentioned we consider pronunciation and vocabulary to be used deliberately at least to a certain degree, and are the main concern in our analysis about the speech level shift in Slovene.

3 Methodology

3.1 Data

As data for analysis in this study we used two conversations from two television debate shows – "Sunday Debate" (*Nichiyō tōron*) for Japanese language and "Friction" (*Trenja*) for Slovene language. Table 1 shows the basic information about the two shows.

Table 1: Basic data information

	Sunday Debate (*Nichiyou touron*)	**Friction (*Trenja*)**	
Time	60 minutes	60 minutes	
Topic of the debate	"What are policies regarding nuclear power plants and nuclear energy" 「どうする原発・エネルギー政策」 (*Dō suru genpatsu enerugii seisaku*)	"Marriage: yes or no?" ("*Poroka: da ali ne*?")	
Interlocutors	Host, male, 53 years old	Host, male, 39 years old	
	Participants: 4 men	12 participants	
		7 men:	5 women:
	JM01 (47) JM02 (66) JM03 (61) JM04 (43)	SM01 (35) SM02 (34) SM03 (34) SM04 (43) SM05 (36) SM06 (68) SM07 (35)	SF01 (66) SF02 (34) SF03 (45) SF04 (49) SF05 (64)
Number of utterances	550	724	

"Sunday Debate" (*Nichiyō tōron*)

Sunday Debate is a Sunday morning talk show in Japan broadcasted by national public broadcasting organization NHK. It is in a form of a debate, moderated and run by the host, where usually current political issues, domestic and world state of affairs, are being discussed.

In this particular episode there are five speakers altogether, four guests and a host, all male. The whole debate takes a form of a well moderated calm conversation, without any altercations and explicit conflicts. The verbal exchange is exclusively between the host and other participants; the participants speak only after being asked a question or are being invited to state their opinion by the host. The topic of the show are nuclear power plants – should the government keep them running or should they be shut down. The show was recorded in September 2012, when due to earthquake and nuclear disaster in 2011 the nuclear power plants and nuclear energy were a very hot topic. Regarding the number of running nuclear plants in the future, three scenarios were being discussed in the show: "zero scenario" – as proposed by the government (supported by JM01 and JM04), "15% scenario" – this means that 15% of the nuclear power plants would be left running (supported by JM03), and "20-25% scenario", where 20-25% of the plants would be left running (supported by JM02).

"Friction" (*Trenja*)

Friction was an evening talk show broadcasted on Slovenian commercial television channel POP TV, which dealt with various topics from politics to every-day matters concerning people and their every-day life. Friction had a similar format as Sunday Debate, with a host moderating the show and leading the conversation.

The analyzed episode has 15 participants, both male and female, and a male host. Primarily, the show has a similar course as Sunday Debate, meaning that all interaction is expected to be an exchange between the host and the participants, where the participants should speak only after being given the permission to speak or after being asked a question by the host. However, the speakers soon start to interrupt each other and start speaking without being given the permission to speak. In a few instances the host even intervenes in order to calm the speakers down or to change the topic of conversation in order to prevent further conflict.

The main topic of this show are the possible causes of the fact that Slovenia has the lowest rate of marriages in Europe. Moreover, a big part of the show is also a discussion about same sex marriages and what are the rights of homosexual people regarding marriages and children. Regarding this, there are speakers who think marriage (between a man and a woman) is extremely important, as only such union can provide a proper environment to raise children in. Furthermore, these same speakers are against legalizing same-sex marriages and refuse making them legally equal to heterosexual marriages. Consequently they think two people in same-sex partnerships should not be able to have and raise children, as in their opinion, this is against the

nature. Such speakers are SM01, SM04, SM07, SF04 and partially SF03. On the other hand, all the other speakers argue that all people, regardless of their sexual orientation, should have same rights and privileges. Most of these speakers also think that getting married will not necessarily ensure a lifelong partnership and that children are the ones that are really important. Moreover, they also think that children can have a good environment and be raised well in any form of the family, as long as they are loved.

Although both shows have a similar format and a similar purpose, i.e. discussing an important topic in order to persuade the audience, the dynamics is quite different. Japanese speakers discuss the topic in a very calm manner and present their opinions in a more subtle way, without any conflict. Whereas Slovene speakers tend to be much more direct, explicit, interruptive and even causing a conflict in a few instances.

3.2 Data analysis

3.2.1 Conversation analysis: Usami (1999): Socio-psycholinguistic Approach

As a methodology for research on interpersonal communication, Usami (1999) proposes a "Socio-psycholinguistic approach for natural conversation analysis". The approach with an empirical methodology based on quantitative analysis from traditional psychology deals with various social factors such as age of the speakers, their social status, gender, etc. The approach aims to not only explain the discourse structure, but also to clarify the principles of the language use and human behavior through the analysis of conversation as human interaction. Socio-psycholinguistic approach follows specific steps in analyzing the data, and this is also the course taken in this study: after gathering the appropriate data, all the conversations were transcribed based on "Basic Transcription System for Japanese" (see Usami, 1999), which has been developed to suit the approach as a useful tool for explaining human mentality and the mechanism of interpersonal communication. Next was the coding of analyzed items, which were then used for quantitative analysis. Finally, those items which could not be included in the coding process were examined by qualitative analysis.

3.2.2 Coding items

3.2.2.1 Coding items in Japanese

In order to examine factors connected with politeness in communication, we first analyzed the conversation on the level of linguistic forms. To do this, we set three types of coding items at the sentence level: speech level at the end of the sentence, speech level of the whole sentence and speech level of the vocabulary. After this, in order to comprehensively understand the relation between speech level and politeness, we looked at speech level shifts from the dynamic side of the discourse level. Table 2 shows coding items for speech level in Japanese and Table 3 shows coding items for speech level shift in Japanese.

Table 2: Coding items for speech level in Japanese

Coding items	Speech level at the end of the sentence	Speech level of the whole sentence	Vocabulary speech level
S: Super-polite form	/	Utterances which contain *sonkeigo*, *kenjōgo*, *bikago*	/
P: Polite form	Sentences which contain a formal style of *desu*/*masu*	Sentences which contain a formal style of *desu*/*masu*	Sentences which contain a honorific title (i.e. name + *san*), formal conjunctions (i.e. ですから *desukara*, because), etc.
N: Non-polite form	Sentences which contain an informal style of *da*/*dearu*	Sentences which contain an informal style of *da*/*dearu*	Sentences which contain only a name without a honorific title, sentences which contain informal conjunctions (i.e. だから、だけど *dakara, dakedo*, because)
NM: No-marker	Sentences without a marker showing level of politeness	Sentences without a marker showing level of politeness	Sentences without a marker showing level of politeness: sentences which do not contain any of the markers classified as "P" and "N"

Table 3: Coding items for speech level shift in Japanese

Coding items		Shift
D	DI: Down-shift from Interlocutor	Shift from interlocutor's preceding utterance: shift from formal style of *desu*/*masu* (P) to informal style of *da*/*dearu* (N)
	DS: Down-shif from Self	Shift from speaker's own preceding utterance: shift from formal style of *desu*/*masu* (P) to informal style of *da*/*dearu* (N)
U	UI: Up-shift from Interlocutor	Shift from interlocutor's preceding utterance: shift from informal style of *da*/*dearu* (N) to formal style of *desu*/*masu* (P)
	US: Up-shift from Self	Shift from own preceding utterance: shift from informal style of *da*/*dearu* (N) to formal style of *desu*/*masu* (P)
Ns	NI: No-shift from Interlocutor	No shift from interlocutor's preceding utterance (both utterances contain the same speech level)
	NS: No-shift from Self	No shift from speaker's own preceding utterance (both utterances contain the same speech level)

3.2.2.2 Coding items in Slovene

In any language, speech level is determined by the setting of a conversation, relationship between speakers, etc. and according to these factors speakers choose to speak formally or casually. In Slovene these factors are closely related to the use of second person pronouns, pronunciation, vocabulary, etc. As the use of second person pronouns does not usually change throughout the conversation (especially in a formal one), they were not subject of our research. Speech level dynamics in Slovene conversation is best explained by the change in pronunciation and the use of vocabulary, therefore, in order to examine speech level shift in Slovene we have focused on the difference between formal and general colloquial language and analyzed the pronunciation and the use of vocabulary. We have set two coding items for speech level in Slovene: speech level of pronunciation and speech level of vocabulary. Table 4 and Table 5 show speech level and speech level shift coding items in Slovene conversation.

Table 4: Coding items for speech level in Slovene

Coding items	Vocabulary speech level	Pronunciation speech level
FL: Formal language	Sentences which do not contain any of the words classified as "GCL", "PT", or "NT" (sentences which only contain neutral words of a formal language)	Sentences which contain a formal pronunciation
GCL: General colloquial language	Sentences which contain words of general colloquial language	Sentences which contain a pronunciation of general colloquial language
PT: Polite title	Sentences which contain honorific titles, such as *gospod, gospa* (Mr, Mrs)	/
NT: No polite title	Sentences which contain only a name without a honorific title	/
NP: Neutral pronunciation	/	Sentences where only a formal pronunciation is possible

Table 5: Coding items for speech level shift in Slovene

Coding items		Shift
D	DI: Down-shift from Interlocutor	Shift from interlocutor's preceding utterance: shift from formal language (FL) to general colloquial language (GCL)
	DS: Down-shift from Self	Shift from speaker's own preceding utterance: shift from formal language (FL) to general colloquial language (GCL)
U	UI: Up-shift from Interlocutor	Shift from interlocutor's preceding utterance: shift from general colloquial language (GCL) to formal language (FL)
	US: Up-shift from Self	Shift from speaker's own preceding utterance: shift from general colloquial language (GCL) to formal language (FL)
Ns	NI: No-shift from Interlocutor	No shift from interlocutor's preceding utterance
	NS: No-shift from Self	No shift from speaker's own preceding utterance

4 Results of analysis and discussion

4.1 Basic information about the data

Basic information about the Japanese and Slovene data is presented in Tables 6 and 7.

Table 6: Basic information about the conversation data

Language	Duration of conversation	Number of utterances	
		Number of all utterances	Number of utterances, which were subject of analysis[2]
Japanese	60 minutes	550	542
Slovene	60 minutes	724	722

Table 7: Basic information about the speakers

Speaker (age): profession	
Japanese	**Slovenian**
Host (53)	Host (39)
JM01 (47): Minister of State for National Policy (0% scenario)	SM01 (35): priest

[2] Utterances, which do not contain any of the coding items and utterances, which were not possible to entirely transcribe (due to an indistinct sound or pronunciation), were not a subject of analysis.

Speaker (age): profession	
Japanese	**Slovenian**
JM02 (66): specially appointed professor at Tokyo Institute of Technology (20-25% scenario)	SM02 (34): actor
JM03 (61): professor at Graduate school of Hitotsubashi University (15% scenario)	SM03 (34): journalist, human rights activist
JM04 (43): senior researcher at Fujitsu Research Institute 0% scenario	SM04 (43): priest
	SM05 (36): athlete
	SM06 (68): jurisconsult, journalist
	SM07 (35): high school teacher, family therapist
/	SF01 (66): actress
	SF02 (34): athlete
	SF03 (45): psychologist, singer
	SF04 (49): politician
	SF05 (64): singer
Legend: JM: Japanese Male; SM: Slovene Male; SF: Slovene Female	

As we can see in Table 6, out of 550 utterances produced in Japanese conversation, 542 utterances were a subject of analysis. This means that the sentences which did not include any of the coding items or which were not possible to be transcribed entirely (due to an indistinct sound or pronunciation) were not analyzed. In the Slovene conversation there were 724 utterances produced, of which 722 were analyzed.

Table 8 below shows frequencies and percentages of all utterances for every speaker.

Table 8: Frequency and the percentage of all utterances for every speaker in Japanese and Slovene

Japanese			Slovene		
Speaker	**No. of Utterances**	**Percentage (%)**	**Speaker**	**No. of Utterances**	**Percentage (%)**
Host	144	26,2	Host	140	19,3
JM01	127	23,1	SM01	31	4,3
JM02	106	19,2	SM02	75	10,4
JM03	102	18,6	SM03	130	17,9
JM04	71	12,9	SM04	45	6,2
			SM05	18	2,5

Japanese			Slovene		
Speaker	**No. of Utterances**	**Percentage (%)**	**Speaker**	**No. of Utterances**	**Percentage (%)**
			SM06	38	5,2
			SM07	17	2,4
			SF01	47	6,5
			SF02	14	1,9
			SF03	73	10,1
			SF04	71	9,8
			SF05	25	3,5
Total	**550**	**100,0**	**Total**	**724**	**100,0**

As we can see from the table above, the highest number of utterances in both languages belongs to the host, with the Japanese host taking up 26,2% of all utterances in Japanese conversation, and the Slovenian host taking up 19,3% of all utterances in Slovenian conversation. We can say that these highest percentages in both languages, as Komiya (1991) has pointed out, are due to the fact that the discussion is taking place under the guidance of the host. In other words, the speakers normally speak only after being given the permission or asked a question by the host. It is important to know how much a certain speaker speaks, as this consequently effects the number of shifts of each speaker – if the speaker produces more utterances, he has more opportunities to switch between speech levels.

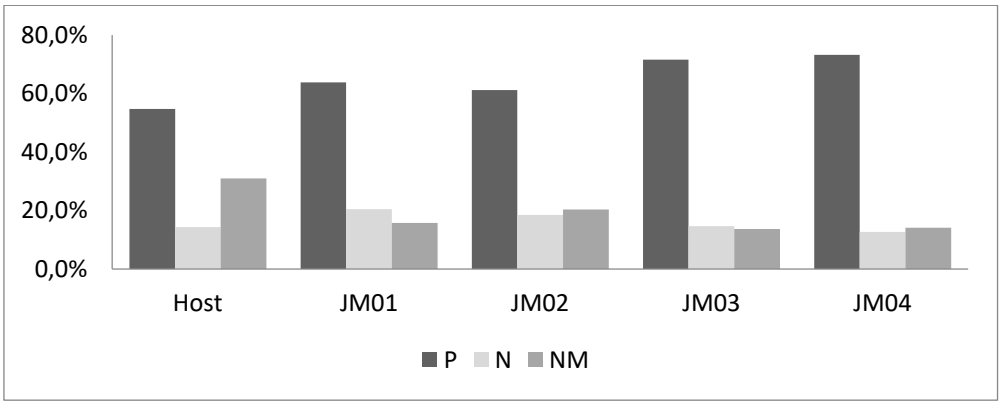
4.2 Japanese language: results of analysis and discussion

In analyzing Japanese conversation, we have analyzed every speech level (speech level at the end of the sentence, speech level of the whole sentence and vocabulary speech level) from four different point of views: a) the use of each speech level in regard to the number of all utterances in the conversation, b) the use of each speech level in regard to all utterances for every individual speaker, c) the use of each speech level only in utterances with politeness markers (utterances classified as S, P and N), d) the use of each speech level only in utterances with politeness markers for every individual speaker. Furthermore, we have looked at the speech level shift from the global (quantitative analysis) and local (qualitative analysis) point of view. Some of the main findings are presented below.

4.2.1 Speech level in Japanese

4.2.1.1 Speech level at the end of the sentence in Japanese

Figure 1 below presents the use of each speech level at the end of the sentence for each individual speaker in Japanese.



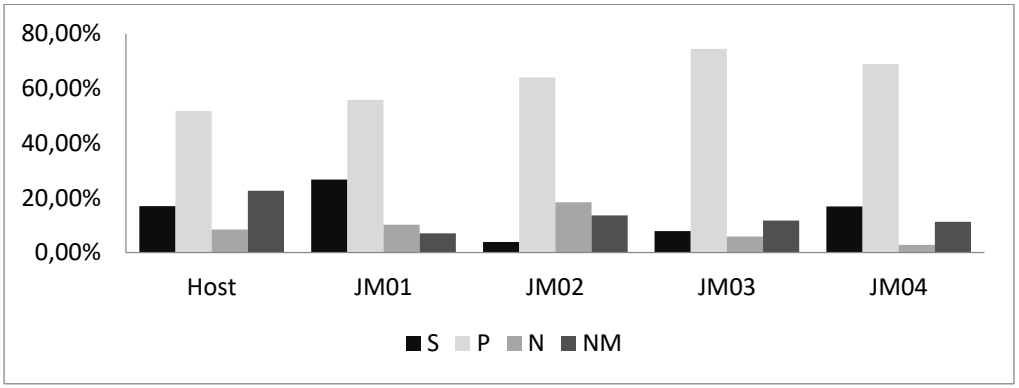
Legend: P = polite; N = non-polite; NM = no-marker

Figure 1: The use of each speech level by individual speakers

As we can see from Figure 1, formal style is most commonly used among all the speakers, which we can say is appropriate use of the speech level in a formal setting, while the use of informal style and utterances without politeness markers was not as frequent. If we take a look at all the participants (JM01, JM02, JM03, JM04), we can see that there is not much difference in speech level use between them. However, looking at the host, we can notice that he produced a rather high number of utterances without politeness markers. According to Usami (2001), the use of utterances without politeness markers functions as a strategy for avoiding acknowledgement of hierarchical relationship between the speaker and the addressee. When the host is speaking to all four participants he often uses utterances without politeness markers in order to stay neutral, however, when he addresses each participant respectively, he uses tends to use a formal style.

4.2.1.2 Speech level of the whole sentence in Japanese

To see the use of the speech level of the whole sentence in Japanese, in addition to the use of formal and informal style within the utterance, we have also looked at the use of *keigo* which we have labeled as "S" (super polite). Figure 2 presents the use of speech level of the whole sentence for each individual speaker.



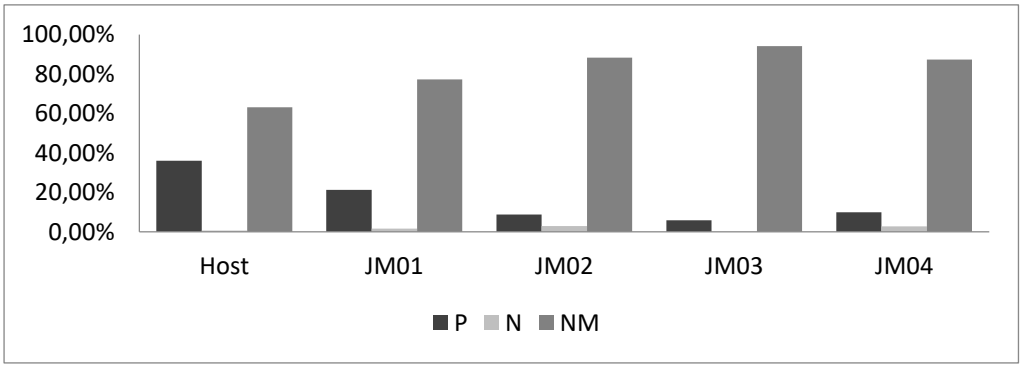
Legend: S = super polite; P = polite; N = non-polite; NM = no-marker

Figure 2: The use of each speech level of the whole sentence by individual speakers

From Figure 2 we can see that the use of the polite form (formal style) is the most frequent among all the speakers, which is a basic speech level for a formal setting such as a television talk show. A fact which should be pointed out is the use of *keigo* (S), which is relatively low. The reason for this is perhaps in the course of the conversation. *Keigo* is known to be mostly used to express the relations between the speakers, however, this conversation is not the kind of setting where speakers would be addressing each other as most of the interaction is between a host and a participant (participants do not speak directly to each other). Moreover, the topic of the conversation are not of a personal matter and experience, but rather of a more general nature (issues in the country), which may also be considered as a reason for such a low use of *keigo*.

4.2.1.3 Speech level of the vocabulary in Japanese

Figure 3 presents the use of the vocabulary speech level in Japanese, for each speaker individually.



Legend: P = polite; N = non-polite; NM = no-marker

Figure 3: The use of the vocabulary speech level in Japanese by individual speakers

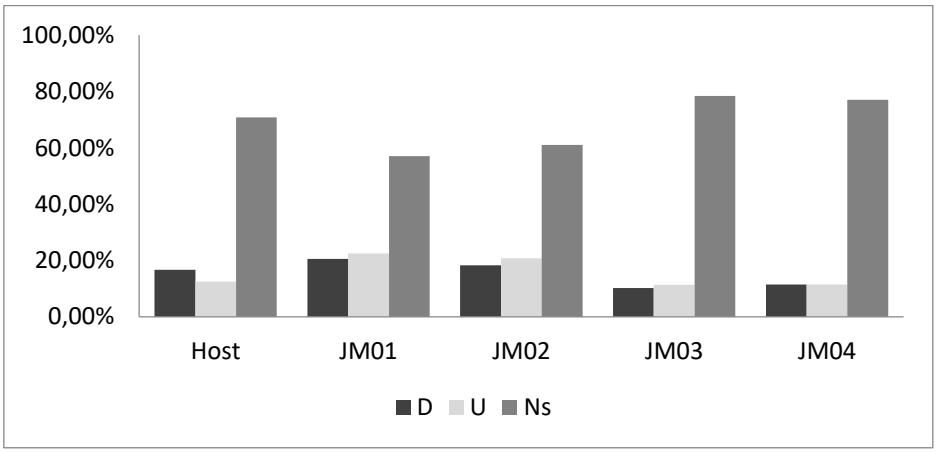
As we can see from Figure 3, the host uses polite vocabulary (i.e. polite titles) most frequently as he is the one mostly addresses of all the participants. Moreover, we can see that the use of non-polite vocabulary is very low, and all of the non-polite use was limited to the plain form of conjunction word *dakara, dakedo*, meaning there was no occurence of a non-use of polite titles.

4.2.2 Speech level shift in Japanese

In Japanese, speech level shift is the shift of a speech level at the end of a sentence. Analyzing speech level shifts will allow us to better understand the function of different elements related to politeness. As the speech level shift is a dynamic phenomenon within the conversation, we have looked at it from the global and local point of view. At the global level we have looked at the frequency of speech level shifts in all utterances of the conversation and at the frequency of speech level shifts by individual speakers. At the local level we have looked at the factors influencing the occurrence of the shifts and the functions that these shifts carry in the conversation.

4.2.2.1 Analysis of the speech level shifts from the global point of view

Figure 4 presents the use of speech level shifts by individual speakers.



D: down shift; U; up shift; Ns: no shift

Figure 4: The use of speech level shifts by individual speakers

If we take a look at the Figure 4, we can see that the use of Ns (no shift) was the highest, while the use of D (down shift) and U (up shift) was almost the same among all speakers. We can explain these results if we take into consideration the fact that the use of a formal style (P), which is the basic speech level in this conversation, was the highest among all the speakers and the use of informal style (N) was very low. This means that all the speakers tend to preserve the basic speech level (P) and therefore very rarely conduct the shifts. When they do shift from a formal to informal style, they

soon shift back up to the formal style – hence the relatively same use of up and down shifts.

4.2.2.2 Analysis of the speech level shifts from the local point of view

By analyzing the shifts from the local point of view, we are able to see what kind functions these shifts carry in the conversation. Because the conversation in a television talk show is of a formal nature, the number of shifts which carry the function of regulating the psychological distance between the speakers is very small. On the other hand, the number of shifts carrying the functions connected with the development of the discourse, is relatively high. Some of these shifts had the following functions: further clarifying the preceding utterance, pointing out or stressing the importance of the conversation content, preserving the basic speech level, changing the topic of conversation, etc.

According to Usami (1995), up-shift in a formal conversation is often conducted due to a sudden and temporary down-shift in order to return and preserve the polite level considered as appropriate (and basic) in such situation.

Below is an example of two shifts: the first one carries a function of explaining the previous utterance and the second one has the function of returning to a basic (polite) speech level.

Example (1)

Line	Speaker	Conversation content	SL	Shift	F
1	JM01	ですから、やはり、大事なことは政府として大きな方向性というものをですね、やはり共有していく、でしかも私たちが作っていこうとして新しいエネルギー社会は、国民の皆さん一人一人の参加が必要なんです。 *Desukara, yahari, daijina koto wa seifu toshite ōkina hōkōsei toiu mono wo desune, yahari kyōyūshiteiku, de shikamo watashitachi ga tsukutteikō toshite atarashii enerugii shakai wa, kokumin no minasan hitori hitori no sanka ga hitsuyōnandesu.* Therefore, of course it is important that we share this great course with government and we must all participate in creating a new energy-efficient society.	P	NS	

2	JM01	あの、これまでのように、あの電力会社がですね、集中的に大きな電力を作って供給するという仕組みから、一人一人が電気を作ったり、節約したりという、そいう分散ネットワーク型のエネルギー社会にしていく。 *Ano, koremadenoyōni, ano denryoku kaisha desune, shūchūtekini ōkina denryoku wo tsukutte kyōkyūsuru toiu shikumi kara, hitori hitori ga denki wo tsukuttari, setsuyakushitari toiu, sōiu bunsan nettowaakukata no enerugii shakainishiteiku.* Well, until now power companies have been creating and supplying us with a great amount of electric power and now we must change this by saving and so creating a dispersed network of energy-efficient society.	N	DS	Explanation of the previous utterance
3	JM01	ま、そのために国民の皆さん方の協力が必要で、そういうやっぱり国民の皆がですね、意識を共有してやっていく、ま、そういう意味では大きな方向性を示すということは非常に大事なことだというふうに思っています。 *Ma, sonotameni kokumin no minasangata no kyōryoku ga hitsuyōde, sōiu yappari kokumin no minna ga desune, ishiki wo kyōyūshiteyatteiku, ma, sōiu imi de wa ōkina hōkōsei wo shimesu toiu koto wa hijōni daijina kotoda toiufūni omotteimasu.* Well, for this all the people's cooperation is needed, everyone should be aware, well, I think in this sense it's important to follow the same great course.	P	US	Return to the basic speech level

Legend:
SL: Speech level at the end of the sentence
F: Factors influencing the occurrence of shifts

In this example JM01 is using a formal style in presenting his opinion, but in line 2 he suddenly conducts a down-shift to an informal style in order to give a further explanation. Then in line 3 he shifts back to the formal style in order to return to the basic level of the conversation.

Moreover, we have found out that when speakers conducted a shift, they conducted an up-shift to the formal style to assert themselves (speak about their own opinion) and a down-shift to explaining or point out general matters. Example (1) above shows these shifts: in line 2 JM01 explains something of a general knowledge and therefore conducts a down-shift. After this in line 3 he expresses his own opinion (asserts himself) and to do so he conducts an up-shift to a formal level.

Looking at this from Brown and Levinson's politeness theory point of view, when a speaker uses a formal *desu/masu* style, they show that they do not want to interfere with the hearer's personal space. In such case, according to the politeness theory, the speaker is expressing a negative politeness. An act of expressing one's own opinion is considered to be more face threatening than presenting general matters, because there is a possibility that a hearer will not share the same opinion with a speaker, and speaker's opinion may consequesntly be perceived as imposing. Therefore in order not to cause any damage to a hearer's negative face, a speaker uses formal style when expressing their own opinion.

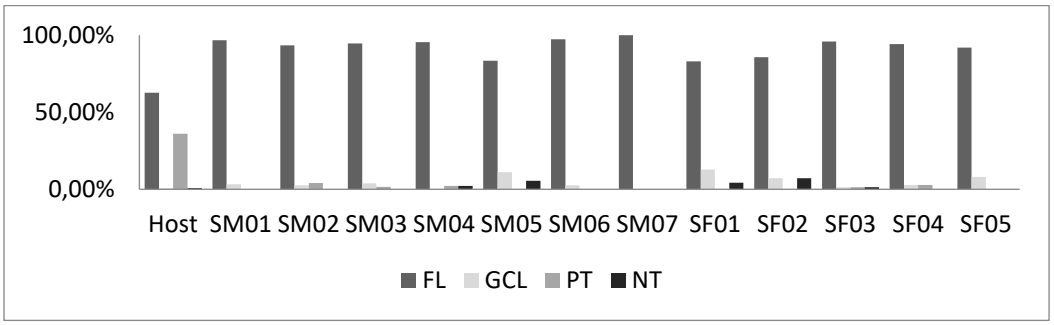
4.3 Results of analysis and discussion in Slovene language

4.3.1 Speech level in Slovene

To look at the speech level in Slovene, we have analyzed the speech level of the vocabulary and pronunciation.

4.3.1.1 Speech level of the vocabulary in Slovene

Figure 5 below presents the use of speech level of the vocabulary for each individual speaker in Slovene.



FL: formal language; GCL: general colloquial language; PT: polite title; NT: no polite title

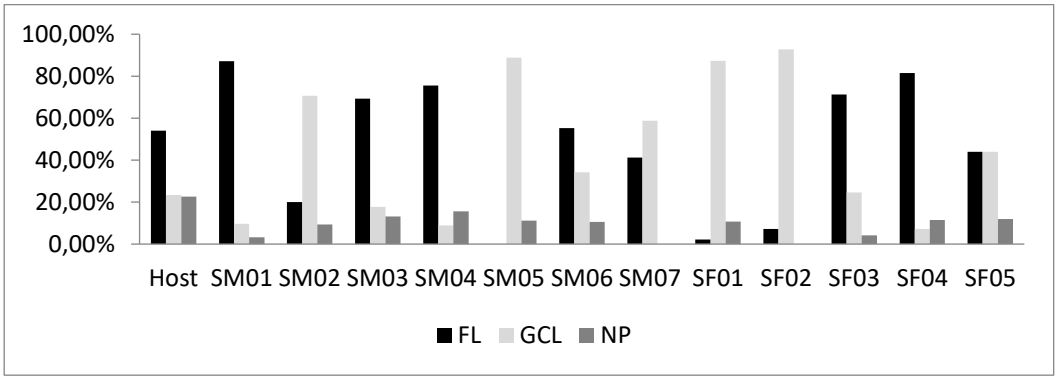
Figure 5: The use of the speech level of the vocabulary in Slovene by individual speakers

Looking at the results of speech level of the vocabulary, the frequency of formal language words (FL) is the highest. Such results were expected, because – as we have said before – most words in Slovene are formal or neutral (unmarked). Furthermore, we can see that the use of the general colloquial language (GCL) words is very rare, which means that most speakers choose the level of vocabulary according to the formal setting of a conversation. The use of the few GCL words may thus be attributed to the speaker's character and background. In other words, someone who frequently uses formal language in their everyday life (i.e. people with professions, where they are expected to use a formal language: politicians, professors, etc.) will most likely use it in a formal setting, such as a television talk show, as well. People who do not use formal language as much in their private lives due to the nature of their profession and other

social background, will, on the other hand, tend to use GCL words more frequently. For example, if we take a look at SM07: he only used the vocabulary of the formal language, which may be due to the fact that he is a teacher and is therefore used to using a formal language in his everyday life.

4.3.1.2 Speech level of the pronunciation in Slovene

We have looked at the use of pronunciation (formal or general colloquial) of words, which reflects the speaker's perception of a linguistic behavior and the situation of a conversation. Figure 6 presents the use of pronunciation by individual speakers.



FL: formal language; GCL: general colloquial language; NP: neutral pronunciation

Figure 6: The use of pronunciation by individual speakers in Slovene

As we can see from the Figure 6, the use of formal and general colloquial pronunciation varies greatly. While this is also due to the difference in the number of utterances by individual speakers, it is, even more importantly, also due to the factors like speaker's character and background. As we have seen with the use of vocabulary, the use of pronunciation depends on whether the speaker uses formal language in his everyday life or not – it therefore depends on the speaker's profession, education, life style, etc. For example, speakers SM01 (priest), SM03 (journalist), SM04 (priest), SM06 (jurisconsult, publicist), SF03 (psychologist), SF04 (politician) all use the formal pronunciation much more than the general colloquial, which may be attributed to their professions.

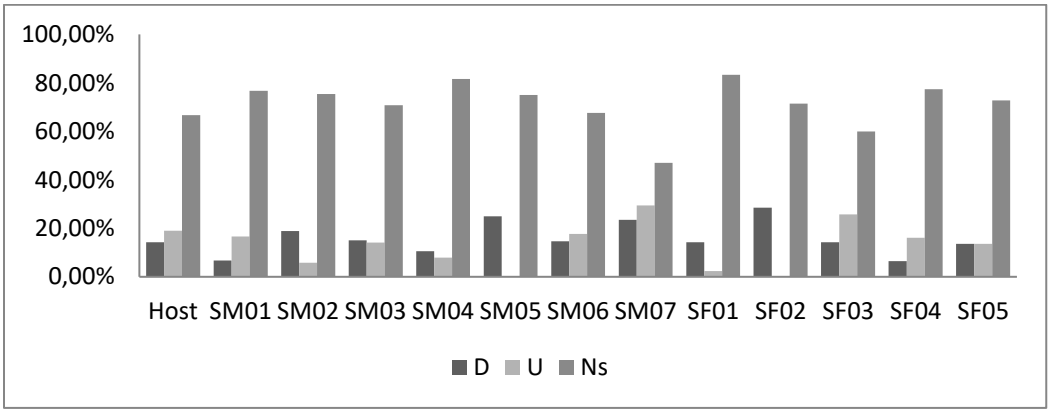
4.3.2 Speech level shift in Slovene

Speech level shift in Slovene is a shift in the level of pronunciation. For example, a change from using formal pronunciation, such as *delati, videl, žal* (work, saw, unfortunately), to using a general colloquial pronunciation, such as *delat, vidu, žou*, was considered a down-shift.

Same as in Japanese, speech level shift in Slovene was also analyzed from the global and local point of view.

4.3.2.1 Analysis of the speech level shifts from the global point of view

Figure 7 presents the use of speech level shifts by individual speakers.



D: down shift; U: up shift; Ns: no shift

Figure 7: The use of speech level shifts by individual speakers in Slovene

Results of the analysis of the use of speech level shifts show that all speakers tend not to shift speech levels. If we compare these results with the results of the use of pronunciation speech level we can see that most of the speakers, who generally use formal pronunciation (host, SM01, SM06, SF03, SF04), conduct more up-shifts than down-shifts. The reason for it is that they want to preserve the formal speech level, which is the basic speech level of the conversation. This means that when they suddenly conduct a down-shift by using a general colloquial pronunciation, or when the preceding speaker uses a general colloquial pronunciation, they perceive this as a deviation from the basic speech level and in order to preserve it, the conduct an up-shift (from GCL to FL).

4.3.2.2 Analysis of the speech level shifts from the local point of view

Analyzing the shifts from the local point of view we have found that most shifts carry the function of developing the discourse, among which more than half of shifts occurred when speakers continued to further explain the preceding utterance, and approximately 16% of shift occurred when the speaker changed the topic.

The example below shows speech level shifts carrying the function of presenting a new topic and further explaining the preceding utterance.

Example (2)

Line	Speaker	Conversation content	SL	Shift	F
1	SM04	*Potem pa gospodu ⌈SM03⌋ (mhm), replika.* And then a reply to SM03.	FL	NS	
2	SM04	*Seveda, a, cerkev ni vedno imela, e, zakonske pastorale tako visoko kot jo ima danes - danes ima zelo visoke standarde.* Of course, the church's pastoral of the marriage hasn't always been as high as it is today – today it has high standards.	FL	NS	
3	SM04	*V srednjem veku vemo, da je to področje blo precej razpuščeno, ne.* In the Middle Ages this field was quite unregulated.	GCL	DS	New topic
4	SM04	*Poroka se je začela uveljavljati pravzaprav z novim vekom pa ta, to insistiranje na, na zakonskem življenju.* Marriage and married life actually started to gain importance with the Modern history.	FL	US	Explanation of the previous utterance

Legend:
SL: Speech level at the end of the sentence
F: Factors influencing the occurrence of shifts

Before the example presented above, SM04 was talking about the importance of marriage as a ritual, and then in line 1 he said he wanted to add a reply to SM03. First in line 2 he said that the pastoral of marriage was very important (similar to what he discussed earlier in the conversation), but then explained that it had not always been so. In line 3 he then started a new topic – the historical background of the marriage, and conducted a down-shift. Then in line 4 he shifted back to the formal speech level in order to explain his preceding utterance and at the same time returned back to the basic (formal) speech level.

4.4 Comparison of the speech level shifts in Japanese and Slovene

Figure 8 presents the total use of speech level shifts in Japanese and Slovene.

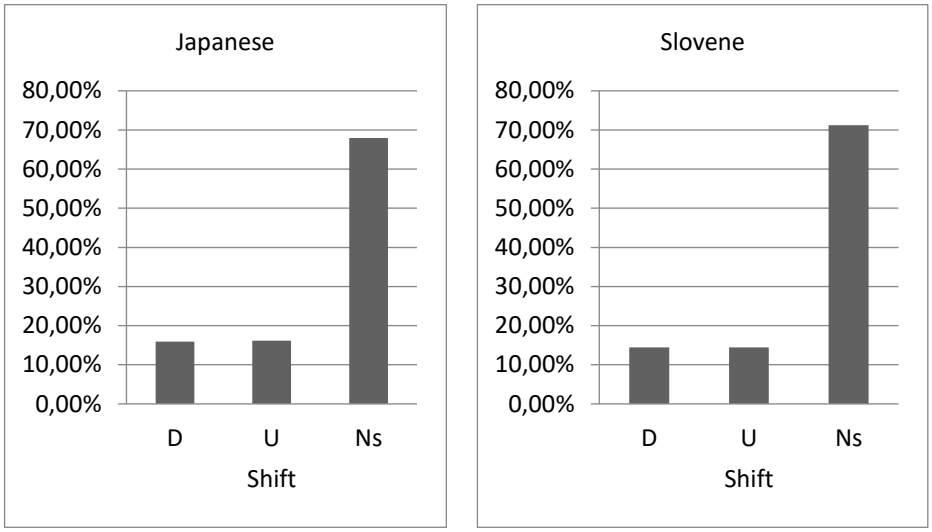


Figure 8: The use of speech level shifts in Japanese and Slovene

The frequency of the use of shifts (D=down, U=up or Ns=no-shift) was very similar in both languages: in both languages the use of down and up-shifts respectively was only around 15%, which means that no-shifts amounted to around 70%. This shows that in both languages speakers tend to preserve the basic speech level (the speech level that is expected to be used in a certain conversation) and that speakers usually shift back to the speech level they mostly use in this conversation (this also means that, for example, the Slovene speaker, who mostly uses the general colloquial pronunciation shifts down from the formal pronunciation, if the preceding speaker was using the formal speech level). Another reason why there are so many no-shifts in these two conversations is the formal setting of the conversation. Namely, according to the previous research of the speech level shift in Japanese, the shifts often occur due to the regulation of the psychological distance between the speakers. But because this setting is not the type of setting where human relations are being formed, the use of no-shifts is the highest.

Furthermore, we have conducted a local analysis in order to see what kind of functions the shifts in both languages carry in the conversation. Table 9 shows the percentage of shifts carrying different functions in Japanese and Slovene.

Table 9: Percentage of shifts carrying different functions in Japanese and Slovene

	Factors causing the shift	Function of the shift	Japanese		Slovene	
			D	U	D	U
Regulation of psychological distance	Expressing empathy to the addressee	Shortening of the psychological distance	1,5	0,0	0,0	0,0
	Joking	Shortening of the psychological distance	0,0	0,0	1,1	1,1
	Matching the addressee's level	Shortening of the psychological distance	0,0	1,4	0,0	0,0
Discourse development	Returning to the basic speech level	Preserving the basic speech level	0,0	28,2	0,0	0,0
	Moving on to the new topic	Indicating the change of the topic	7,3	2,8	15,4	15,4
	Explaining, giving an example, supplementing preceding utterance	Clarifying the preceding utterance	50,0	25,3	59,3	51,6
	Expressing or emphasizing the conclusion, intention, fact, argument, etc.	Pointing out the importance of the conversation content	29,4	32,4	5,5	5,5
	Question-answer session in order to confirm something	Deepening the understanding of the conversation/utterance content	4,4	4,2	2,2	7,7
	Asking a question, requesting a comment or information	Deepening the understanding of another speaker's opinion	7,3	5,6	5,5	5,5
	Requesting a right to speak	Indicating the wish to speak	0,0	0,0	4,4	4,4
	Leadership, addressing someone	Controlling/leading the course of conversation	0,0	0,0	6,6	8,8
	Total (%)		**100,0**	**100,0**	**100,0**	**100,0**

As we can see from the table above, in both languages the occurrence of shifts that carry a function of regulating the psychological distance between the speakers is very rare. This is, as we have said above, due to the formal setting of the conversation. On the other hand, in both languages most shifts carry the function of discourse development, however, there were some differences seen in this type of shifts. While there were 28% of shifts with the function of preserving the basic speech level in Japanese, there were no shifts with this function found in Slovene. The usage of all other shifts with the function of discourse development is very similar in both languages – there is only a noticeable difference seen in the use of shifts with the

function of indicating a wish to speak, and controlling/leading the conversation. The latter were the shifts conducted by the host in a situation, when speakers were interrupting each other, speaking over each other, or engaged themselves in a rather heated discussion. In order to calm them down, the host intervened. In Japanese there was no such instance, as the conversation was calm and steady throughout the whole show. This is also the reason why there was no instance of a speaker requesting to speak in the Japanese conversation, whereas in the Slovene conversation there were a few instances when speakers asked the host if they could speak. This also shows for a higher level of formality of the Japanese conversation and a slightly more relaxed nature of the Slovene conversation.

5 Conclusion

Results of the analysis have shown that the main factor determining the basic speech level of the conversation is the setting of a conversation. Most speakers will use appropriate speech level according to the setting, which in these two conversations was a formal style in Japanese and a formal language in Slovene. However, comparing the two languages from the politeness point of view, we have seen a difference in tendencies of the speech level use.

In order to understand how the relations between the speakers influence linguistic behavior, we have looked at three social factors (social distance (D), relative power (P) and absolute ranking of impositions (R)) to see what role they might play in linguistic behavior of the speakers. The proper use of speech level is determined according to hierarchical relationship between a speaker and a hearer, content of a conversation, context of a situation, etc. Minami (1987) proposes hierarchy and the level of closeness as the conditions which determine interpersonal relations. The level of closeness may be considered the same as the social distance in the politeness theory, while the hierarchy may be equated with the relative power. Within the three social factors proposed by Brown and Levinson, social distance and absolute ranking of impositions are considered to be fixed in this study, while the relative power differs among the speakers. In other words, as the speakers partake in the same talk show, the conversation content and formality of the situation are fixed (they stay the same throughout the show). However, as we do not have any information about the social distance between the speakers (how well they know each other), we cannot make any judgement on this. Therefore, due to the formal nature of the debate show and considering the attitude and linguistic behavior of the speakers, we may say that the social distance between the speakers is small. Furthermore, because all the speakers are speaking in the same context of situation about the same topic, the ranking of imposition is considered to stay the same throughout the conversation. Because the social distance and ranking of imposition do not change during the conversation, we wanted to see how the relative power (age, social status, etc.) influences the choice of the speech level and speech level shift in both languages.

We have come to a conclusion that, looking from the politeness theory point of view, there is a difference in the tendency of the use of speech level in both languages.

In Japanese, the informal style (i.e. lower speech level) was used mostly by the speakers who have a higher ranking of power[3]. Whereas in Slovene, speakers with a lower ranking of power were the ones mostly using the lower speech level (i.e. general colloquial language). This shows that the use of speech level in Japanese is influenced by hierarchical relations, whereas the use of speech level in Slovene reflects the speaker's background. Furthermore, we have seen that the Japanese speakers tend to use formal style when expressing their own opinion. By using a formal style a speaker creates certain distance between himself and a hearer, which allows a speaker to freely express their own (opposing) point of view. This corresponds to Brown and Levinson, who say that by using a negative politeness strategy, e.g. using a formal style, a speaker distances himself from a hearer and does not want to invade their private territory.

Moreover, speech level shifts in both languages are essentially not considered to be a deviation from the basic speech level of the conversation, but are rather understood as the means of conducting a smooth conversation.

References

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage.* Cambridge: Cambridge University Press.

Matsumoto, Y. (1988). Reexamination of the universality of Face: Politeness phenomena in Japanese. *Journal of Pragmatics*, *12*(4), 403–426.

Matsumoto, Y. (1989). Politeness and conversational universals – observations from Japanese. *Multilingua*, *9*(3), 207–221.

Pizziconi, B. (2003). Re-examining politeness, face and the Japanese language. *Journal of Pragmatics*, 35, 1471–1506.

Pogorelec, B. (1965). Vprašanje govorjenega jezika (The question of Spoken Language). In *Jezikovni pogovori* (pp. 132-156). Ljubljana: Cankarjeva založba.

Slovar slovenskega knjižnega jezika (Dictionary of Standard Slovenian) (1994). Ljubljana: DZS.

Tivadar, H. (2003). Podoba in funkcija govorjenega knjižnega jezika glede na neknjižne zvrsti. *Obdobja*, 22, 437–452.

Toporišič, J. (2000). *Slovenska slovnica* (Slovene Grammar). Maribor: Založba Obzorje.

Toporišič, J. (1973). *Slovenski knjižni jezik 4* (Slovene Literary Language 4). Maribor: Založba Obzorja.

[3] To determine the rank of social power among the speakers, we have taken into a consideration their age and profession: the higher the age the higher the ranking of power and speakers with professions for which more education is needed are considered as having a higher ranking of power.

Toporišič, J. et al. (Ed.) (1994). *Slovenski pravopis 1, Pravila* (Slovene Orthography 1, Rules). Ljubljana: DZS.

Usami, M. (2002). *Discourse politeness in Japanese conversation: Some implications for a universal theory of politeness*. Tokyo: Hituzi Syobo.

Wardhaugh, R. (1998). *An introduction to sociolinguistics*. Malden: Blackwell Publishers.

Ikuta S., & Ide, Y. (1983). *Shakai gengogaku ni okeru danwa kenkyū* (Discourse research in social linguistics). *Gekkan gengo*, 12 (12), 77–84.

Usami, M. (1995). Danwa reberu kara mita keigo shiyō: supiichi reberu shifuto seiki no jōken to kinō (Conditions for speech-level shift occurrence in Japanese discourse). *Gakuen*, 662, 27–42.

Usami, M. (1999). Danwa no teiryōteki bunseki – gengo shakai shinrigakuteki apurōchii (Quantitative discourse analysis – socio-psycholinguistic approach), *Nihongogaku*, 18 (10), 40–56.

Usami, M. (2001). Disukōsu poraitonesu toiu kanten kara mita keigo shiyō no kinō – keigo shiyō no atarashii toraekata ga poraitonesu no danwa riron ni shisasuru koto (Functions of the use from discourse politeness point of view), *Gogaku kenkyūjo ronshū* 6.

Usami, M. (Ed.) (2006). Shizen kaiwa bunseki eno gengo shakai shinrigakuteki apurōchii (Socio-psycholinguistic approach for natural conversation analysis). *Gengo jōhōgaku kenkyū hōkoku* 13.

Usami, M. (2006). Danwa ni okeru rōkaru bunseki to gurōbaru bunseki no igi (The meaning of local and global discourse analysis). *Gengo jōhōgaku kenkyū hōkoku* 13.

Usami, M. (2011). *Kihontekina mojika no gensoku* (Rules of basic transcription) (Basic Transcription System for Japanese: BTSJ).

Komiya, S. (1991). Tōronkai bamen no kaiwa sutoratejii (Strategies in debates), *Collection of essays on Japanese education of Center for foreign students education of Tsukuba University*, 145–165.

Kikuchi, Y. (1997). *Keigo*. Tokyo: Kōdansha.

Minami, F. (1987). *Keigo*. Tokyo: Iwanami.

Mimaki, Y. (1993). Danwa no tenkai yōshiki toshite no taigū reberu shifuto (Speech level shift as a mark of a discourse development). *Osaka Kyōiku University Repository – Social sciences*, *42*(1), 39–51.

Mimaki, Y. (2001): Taiwa ni okeru taigū reberu kanri no jisshōteki kenkyū (Empirical study on the management of speech level in conversat Equation 1on). *Heisei 9 – Heisei 12, Nendo kagaku kenkyū hojokin kenkyū seika hōkokusho*.

INTERPRETATION OF DABA SCRIPT: *GEMU* FROM WUJIAO VILLAGE

XU Duoduo
Nanyang Technological University, Singapore
duoduo001@e.ntu.edu.sg

Abstract

Daba calendars are the only written texts of Dabaism discovered so far. Some studies have been discontinuously conducted on this topic since the first report on this script in 1940. However, some unclear and obscure hermeneutic points of Daba script have still to be investigated in-depth. In this paper, I present an interpretation of one of the Daba calendars based on my original field work data. This Daba calendar, entitled "Gemu", origins in Wūjiǎo (屋脚) village. It is relatively old and, compared to others, shows an archaic version of the Daba calendars, according to the classification proposed in Song (2003). The present study aims to expand our knowledge of the Daba script and of the context in which it was and is still used. Further on, this paper can contribute to establishing a philological foundation for further research on the Daba script, and related comparative studies.

Keywords: Daba Script; *Gemu*; Wujiao Village; Lunar Mansion; Script Interpretation

Povzetek

Koledarji Daba so edina znana pisna besedila dabaizma. Od prvega poročila iz leta 1940 do danes je nastalo nekaj nepovezanih študij, ki niso povsem odgovorile na nekatera hermenevtična vprašanja v zvezi s pisavo Daba. Članek podaja razlago enega izmed koledarjev Daba in temelji na izsledkih terenske raziskave. Obravnavani koledar se imenuje "Gemu" in izhaja iz vasi Wūjiǎo (屋脚). Ta različica koledarja je starejša od primerljivih, ki so predstavljene in klasificirane v Song (2003). Pričujoča študija poglobljeno predstavi značilnosti pisave Daba in konteksta, v katerem je to besedilo nastalo, in se uporablja še dandanes. Članek lahko služi tudi kot filološka osnova za nadaljnje raziskave pisave Daba in s tem povezanih študij.

Ključne besede: pisava Daba; *Gemu*; vas Wujiao; lunarne lože; interpretacija pisave

Acta Linguistica Asiatica, 6(2), 2016.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.6.2.53-67

1 Introduction

Daba script used to write down calendars is the only written literature of Dabaism discovered so far. Dabaism is the local religion of Na People living on the border between Yunnan and Sichuan Provinces in South-West China. Dabaism shares the same origin with Dongbaism, another local religion of Naxi People living in Lìjiāng (丽江) area, Yunnan Province. Their different designations are due to dialectal differences between Na and Naxi. Dabaism derives from the local word for their priests, "Daba". Similarly, Dongbaism is derived from the word "Dongba", referred to their priests, in Naxi language.[1]

Dabaism, attested in Héngduàn (横断) Mountains, for a long time has been considered as an eastern branch of Dongbaism without writings (cf. Li, 1984, p. 38; He & Jiang, 1985, p. 117).[2] The earliest literature records of Daba script available now is by Zhuang Xueyou in 1940 on a Journal entitled *Liáng Yǒu* (良友) (*Good Friend*). However, it was just a bulletin reporting that Moso People have pictograms, without any detail about what kind of pictograms they are (cf. Song, 2003, p. 86). Inspired by that report, Song Zhaolin (宋兆麟) went to Moso area and collected one Daba calendar in 1963 in Yǒngníng (永宁) Township, Nínglàng (宁蒗) County, Yunnan Province. However, Song Zhaolin had not the opportunity to do research in depth until 2000 (cf. Song, 2003, p. 86). Basing on his field work data in 2000 in Lúgū (泸沽) Lake area, four versions of Daba calendars from four villages have been reported by him. Moreover, three among them are presented with a systematic interpretation for each symbol in Song (2003). Some other scholars have also considered the Daba calendar as the bearer of Daba symbols since 1980s. Yang (1994, pp. 32–35) wrote a brief report on this topic based on his first-hand field work data. In the paragraph of the book about a comparative study on Tibetan Buddhism among Tibetan People, Naxi People, and Pumi People, the author has shown a figure with 32 symbols, interpreted by him one-by-one. In the summer of 2010, Zhao Liming conducted a long field work following Song Zhaollin's suggestions. She has collected the latest information on the existing written texts of various ethnic groups living on the border of Yunnan and Sichuan Provinces. Later on, in the National Social Science Fund Project "Interpretation and Rescue of Endangered Scripts and Bibliographies in Southwest China" (10&ZD123) advised by her, I was in charge of the interpretation of Daba script. From January 2011 to January 2015, I have fulfilled several field work trips to the Daba and Dongba areas. The preliminary data presented in this paper have been collected during my two field work trips in 2011 from Wūjiǎo (屋脚) Village, Wūjiǎo (屋脚) Township, Mùlǐ (木里) County, Sichuan Province.

[1] Na is now listed as an independent language (ISO 639-3: nru), while Naxi is coded as ISO 639-3: nxq. In China, however, they are considered dialects of the same language: Na is the eastern branch of Naxi language.

[2] Distinguished from Dabaism, Dongbaism has a relatively mature writing system known as Dongba hieroglyphs (cf. Li et al., 1972).

Wujiao Village has a population of around 300 people, composed of Mongolian and Yi nationalities. The endonym of those Mongolian People living there is [nɑ˩hĩ˧]. Their language resembles to Naxi language. In other words, they are in fact Na People, the eastern branch of Naxi People, but not a branch of the Mongolian linguistic *milieu*.[3] Dabaism and Tibetan Buddhism coexist in Wujiao Village. There were six Daba priests and seven Lama priests (the Tibetan Buddhism monks) during my field work.

2 Literature Review

Yang (1994, pp. 32–35) presents an interpretation of 32 Daba glyphs[4] including: pronunciation of the glyph in local language transcribed by Chinese characters, the literal meaning of each glyph, and the divination meaning of the glyph. In this interpretation, nine glyphs are related to animals, seven are related to sex, four are related to stars, four are related to human body parts, two represents diseases, other two heterogeneous notions / items (No. 6: wealth/property and No. 25: poisonous food). Four other glyphs are not yet interpreted. The divination meanings are simply auspicious, ominous, or neutral, except for one glyph representing auspicious if there is a battle and ominous if it is a normal day. In his short conclusion, Yang has pointed out some characteristics of these Daba glyphs: 1) these glyphs are assigned to the days around one year without apparent links to the dates; 2) the pronunciations of the glyphs are mainly ancient Moso language, which is merely understood by current people; 3) the knowledge of these glyphs and the calendars are held only by a few elder Daba priests, not approachable by common people.

Song Zhaolin's interpretation of Daba script is a work based on each calendar. He presented the glyphs used for 28 days in the first month with 1) the pronunciation(s) of the glyph(s) in one grid of the calendar by Chinese characters, 2) the meaning(s) of the word(s), 3) the day is auspicious, ominous, or neutral, 4) the suitable activities and not suitable actions on that day. The three versions of Daba calendars interpreted in Song (2003) include two types: each day represented by either one glyph or two glyphs. The calendar with single glyphs for each day contains 28 different glyphs. The calendar with double glyphs for each day is a combination of these 28 glyphs with some other glyphs. The calendar written with the 28 single glyphs is considered the more aboriginal type. Applying the "Liù Shū" theory (六书)[5] on the analysis of Daba script, Song Zhaolin

[3] Na People have been recognized as Mongolian People during the official census of nationalities in China in 1950s, which has been a mistake. Naxi People, in their turn, have been named according to their endonym [nɑ˩ʂi˧] (He & Jiang, 1985, p. 2).

[4] In Yang (1994, pp. 32–35), Daba glyphs are defined as "picto-glyph" (túhuà wénzì 图画文字).

[5] "Liù Shū" theory (六书) is a traditional philological theoretical framework aimed at the analysis of the composition of Chinese characters proposed by Xǔ Shèn (许慎) (58–147 AD), a scholar from Han Dynasty (206 BC – 220 AD). According to this theory, Chinese characters can be divided into six categories based on the six manners they were created: xiàngxíngzì 象形字 (pictograms), zhǐshìzì 指事字 (simple indicatives), huìyìzì 会意字 (compound indicatives), xíngshēngzì 形声字 (phono-

has concluded that most of the glyphs are pictograms (xiàngxíng 象形) and simple indicatives (zhǐshì 指事). Moreover, he claims that the Daba glyphs are characters[6] according to three factors: 1) they have stable shapes; 2) these glyphs can write down the relatively complicate calendar, which has different contents for 360 days per year; 3) the glyphs in Daba calendars are widely recognized and used by Daba priests. Compared to Dongba writing, however, Daba script shows to be less mature as a writing system. The author listed three reasons: 1) pictograms are much more than simple indicatives; 2) variants of glyphs are spotted among different calendars; 3) these characters are used only to write down Daba calendars, not applicable to write down other Daba classics.

Yang (1994) and Song (2003) are two important references in the study of Daba script. Being pioneer researches on this topic, they have discovered the existence of Daba script and have provided a general idea of this pictographic writing. However, there is still need to work in-depth on the glyphs. For example, Chinese and Na have quite different phonemic systems. Therefore, the transcription of the pronunciations of the glyphs in Na by Chinese characters is not accurate enough as source. Moreover, due to inaccurate transcription, it is difficult to find out the relationship of the glyphs meaning with the daily vocabulary. This could lead to misinterpretation of these glyphs.[7] Further on, an unclear point in these two reports is the number of Daba characters. Song (2003) seems to count on the Daba calendars themselves, since he has provided scanned copies of Daba calendars. Conversely, the 32 Daba glyphs displayed in Yang (1994, p. 32) could be a combined elicitation from two versions of Daba calendars. In other words, the single glyph set used in Daba calendars contains 28 units. However, the number of the additional glyphs set was not clarified by Song (2003). Instead, Song has stated the total number of the glyphs appearing in double glyph calendars as it results from this sentence: "…it has several more characters, more than 30 pictographs in all" (Song, 2003, p. 90). According to his interpretation of data, the meanings of this additional set of glyphs include "throat", "sun", "moon", "eye", "palm", "throat", and "nose". The words for the two "throats" are different. Nevertheless, that was not explained in the paper.

Xu (2013) is a preliminary study on the philological issues of Daba script. Three versions of Daba calendars from three different villages have been systematically analyzed. Comparing forms, pronunciations, and meanings of the glyphs, the author established some relevant hermeneutic points: 1) Daba script should be configured as a kind of ancient conventional symbols (Istrin, 1987, p. 77), at the earliest stage of the local

semantic compound characters), jiǎjièzì 假借字 (phonetic loans), and zhuǎnzhùzì 转注字 (derived characters). Cf. Xu (2001, p. 314) and Boltz (1993, pp. 432–433).

[6] In Song (2003), Daba glyphs are defined as "pictographic characters" (xiàngxíng wénzì 象形文字).

[7] For example, the third symbol in the first calendar sounds as bàokuài (报快) (transcribed into Chinese pinyin) and is interpreted as "the penis of the goat" Song (2003, p. 86). However, according to my knowledge of Na language, the first syllable could be "frog" and the second is "mouth". Therefore, the compound word means the "the mouth of the frog".

writing; 2) the original Daba script contains 28 symbols representing the lunar mansions in Daba tradition; 3) some loan-symbols are likely to be derived from Tibetan Buddhism.

In order to clarify obscure points in the current available academic publications on Daba script, it is necessary to provide a more accurate description and interpretation of Daba writing, and of its bearer, Daba calendar, as well. In this paper, I will apply a linguistic perspective in order to try to understand the meanings of Daba symbols according to the local language.

3 Interpretation of *Gemu* from Wujiao Village

The specific Daba calendar interpreted in this paper is from Wūjiǎo (屋脚) Village, Wūjiǎo (屋脚) Township, Mùlǐ (木里) County, Sichuan Province. The calendar is about 40 cm in length and 18 cm in height. It consists of six pages, with the symbols written by calligraphic brush from the left to the right. It is a single symbol type of calendar (as stated by Song (2003)), where each day is written through a (only one) corresponding symbol.

The Daba priest preserving the calendar is called *Dawa*. *Dawa* was born in 1929, and passed away in 2012. According to the interview with him, the calendar is entitled [kɯ˧mɥ˧] "Gemu", literally means "the book of the stars". The first segment "ge" means "star" and the second segment "mu" means "book". It has been handed down from his ancestors. *Dawa* spent in Wujiao Village all his life.

This Daba calendar was firstly interpreted by me with the cooperation of Daba *Awo*, the son of Daba *Dawa*, during my field work in January 2011. Daba *Awo* is born in 1967. He is one of the most knowledgeable Daba priests in Wujiao Village. He started to learn Daba culture since he was 10 years old following Daba *Dawa*. When he was 21 years old, he started to conduct rituals by himself. The recorded data and interpretation have been checked with him again during my field work in July 2011 in Wujiao village.

Having noticed that the glyphs are repeated every 28 units, I have inferred they correspond to the 28 lunar mansions in Daba culture. I numbered the first cycle in the first month from 1 to 28 according to the star groups Daba *Awo* has divided during the interpretation work. Later on, in order to understand completely the constellation system, I have compared the Daba symbols with Dongba stars documented in academic publications. Since Dongba stars have been equated to Chinese and Western astronomical designations, I had the opportunity to connect these Daba mansions with those 28 lunar mansion systems from other main stream cultures, including Chinese, Tibetan, and Indian. The detailed comparison can be found in Xu (2015). In this paper, I annotate directly the correspondences of the 28 lunar mansions with the Chinese lunar mansions and the western designations for the stars or asterisms, if applicable, in order to provide a more comprehensive scenario of Daba lunar mansions represented by Daba glyphs.

The interpretation presented below lists: 1) Daba glyph; 2) IPA transcription of the glyph's pronunciation in the Na language; 3) Chinese translation of the name of the specific lunar mansion represented by the related symbol; 4) literal meaning of the lunar mansion; 5) the corresponding Chinese lunar mansion or star;[8] 6) the Western designation of the lunar mansion or star; 7) IPA transcription of the divination comments of the glyph;[9] 8) word-by-word gloss; 9) translation of the whole sentence; 10) my conclusive remarks on the suitable and non-suitable activities on that day.

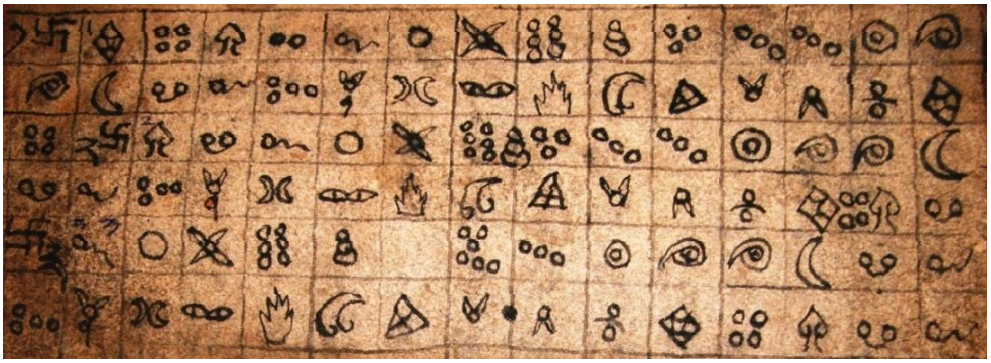


Figure 1: The First Page of Wujiao Daba Calendar

3.1 The "Human Being" Constellation

Script	28. 	1. 
Na	pʰæ˧mi˧	ŋ̩i˧ɖi˧
Chinese	Pàmǐ (帕米)	Nǐzhì (拟至)
Literal Meaning	---	---
Chinese Lunar Mansion	Hégǔ 2 (河鼓二)	
Western Designation	Altair (Aquila)	

IPA transcription of the divination comments of the glyphs:

(1) pʰæ˧ mi˧ mv̩˧ tɕʰi˧ ʑʌ˧, lɛ˩ tɕʰi˧ tʰɛ˧ tɕʰi˩.
the day of Pami funeral etc. to discard to go

Translation: (On the days of "Pami" and "Nizhi",) it is not allowed to celebrate funerals. If a funeral is celebrated, someone will die in the following day.

Remarks: The day when "Pami" or "Nizhi" is "on duty" is not good for funerals.

[8] I am providing the Chinese translations, since they are directly derived from field work notes.

[9] The transcriptions record the tones and pronunciations of the words in sentences in order to provide more comprehensive picture of Na language.

3.2 The "Horse" Constellation

Script	2. 
Na	ʐwæɬkɯɬ
Chinese	Mǎ Xīng (马星)
Literal Meaning	star of horse
Chinese Lunar Mansion	Húguā (瓠瓜)
Western Designation	Hugua (Delphinus)

IPA transcription of the divination comments of the glyph:

(2) ʐwæɬ tɕʰiɬ mʌɬ dʑʌȷ, lɛɬ χwɑɬ mʌɬ dʑʌȷ.
horse to sell not good to buy not good

Translation: (On the day of "Ma Xing",) the exchanges of horses are not auspicious.

Remarks: The day when Hugua is on duty is not good for horses' exchange.

3.3 The "Frog" Constellation

Script	3. 	4. 	5. 	6. 
Na	pʌɬkʰwʌɬ	pʌɬdʑɯɬ	dʑɯɬkɯȷ	pʌɬkɯȷpʰɯȷ
Chinese	Wāzuǐ Xīng (蛙嘴星)	Wāniào Xīng (蛙尿星)	Shuǐ Xīng (水星)	Báiwā Xīng (白蛙星)
Literal Meaning	the mouth of frog	the urine of frog	the star of water	a white star from the constellation "frog"
Chinese Lunar Mansion	Shì (室)	Bì (壁)	Kuí (奎)	Lóu (娄)
Western Designation	Markab and Scheat	Algenib and Sirrah	Legs	Bond

IPA transcription of the divination comments of the glyphs:

(3) gʮȷ boɬ χwɑɬ bjɛɬ dʑʌɻ, gʮɬ pʰoɬ bɑȷ pʰoɬ bjɛɬ dʑʌȷ,
livestock to buy too good crop to sow crop to sow FUT. good

ʁɯɬ tɕʰiɬ mʌɬ dʑʌȷ, bɯɻ tɕʰiɬ mʌɬ dʑʌȷ, χwɑɬ bjɛɬ mʌɬ dʑʌȷ.
bull to sell not good yak to sell not good to buy FUT. not good

Translation: (On the days of "Wazui Xing", "Waniao Xing", "Shui Xing", and "Baiwa Xing",) the exchanges of livestock are good. It is also auspicious to work in the field. But the exchanges of bulls and yaks are not good.

Remarks: During the days when stars/asterisms from the "Frog" Constellation are on duty, the exchanges of livestock are auspicious, except bulls and yaks. These are good days to sow and to work in the field.

3.4 The "Six Stars" Constellation

Script	7. 	8. 
Na	qʰɻ˩tʂæ˩qʰɻ˧	qʰɻ˩tʂæ˩gv̩˧mi˧
Chinese	"Kezha" Jiǎo ("科扎"角)	"Kezha" Shēn ("科扎"身)
Literal Meaning	the horn of "Kezha"	The body of "Kezha"
Chinese Lunar Mansion	Mǎo (昴)	
Western Designation	Pleiades	

IPA transcription of the divination comments of the glyphs:

(4) gv̩˧ pʰo˧ bɑ˩ pʰo˧ mʌ˧ dʑʌ˩, gv̩˩ bo˧ sɛ˩ ʑɯ˧ qʰo˥ bjɛ˧
crop to sow crop to sow not good livestock blood to have to kill FUT.

mʌ˧ dʑʌ˩, dʑɯ˩ bi˥ hṽ̩˧ næ̃˩ di˩ dʑʌ˨, dʐi˩ ʂi˧
not good mountain area wild animals to chase good money to look for

ʁwʌ˧ ʂi˧ bi˧ dʑʌ˩, ʔɑ˧ ʁo˧ gv̩˩ bo˧ qʰo˥ bi˧
mountain to look for FUT. good home livestock to kill FUT.

lɑ˧ mʌ˧ dʑʌ˩, tɕʰi˧ bi˧ lɑ˧ mʌ˧ dʑʌ˩.
then not good to sell FUT. then not good

Translation: (On the days of the "Kezha",) it is not good to sow and work in the field. The livestock with blood (mammals) cannot be killed. It is allowed to hunt. All the animals hunted from the mountains are good. To kill the household livestock (bulls, horses, sheep, and pigs) is not good (to other livestock at home). To sell them is not good either.

Remarks: During the days when stars/asterisms from the "Six Stars" Constellation are on duty, it is not allowed to work in the field, either to kill or exchange livestock. They are good days for hunting.

3.5 The "Red Eye" Constellation

Script	9. 
Na	ŋ̩jæ˩hṽ̩˧
Chinese	Hóngyǎn Xīng (红眼星)
Literal Meaning	red eye
Chinese Lunar Mansion	Bì (壁)
Western Designation	Aldebaran

IPA transcription of the divination comments of the glyph:

(5) gʮ˩ bo˧ dʑɯ˩ bi˥ hwʌ˩ bi˩ mʌ˧ dʑʌ˩, ŋ̩jæ̃˩ hʮ̃˧
livestock mountain area to shepherd to go not good eye red

mʮ˧ kɯ˩ ʔɑ˧ ʁo˧ ʑɯ˩ qɛ˥ mʮ˧ ɻʮ˩ tʂʌ˩ ʁæʌ ʑiʌ.
fire star home house to burn fire forest fire pay attention

Translation: On the day of "Hongyan Xing", it is not allowed to bring livestock (bulls, horses, sheep, etc.) into the mountains; (people are not allowed to go into the mountains). On this day, people should pay attention to fire.

Remarks: The day when "Hongyan Xing" is on duty, is not suitable to go into the mountains. It is a day to avoid fire.

3.6 The "Three Stars" Constellation

Daba *Awo* explained this star group as the "sheep" constellation and did not give specific names of each star. The comparison with other Daba calendars and Dongba lunar mansions, from the perspectives of pronunciation and characters' shape, indicates the designation of the constellation as "three stars". Significant evidence is findable in the word pronunciation, since the first syllable in these Daba and Dongba lunar mansions means "three". In Naxi language, the word "three" is [sɯ˩] (Li et al., 1972, p. 118, No. 1537), while in Na language the word "three" is [so˧]. Moreover, the lunar mansions are depicted by three circles in Daba calendar. In Dongba writing, these lunar mansions are composed by three circles with additional symbols indicating the pronunciation of some syllables of the lunar mansions' names.

Script	10. 	11. 	12. 	13. 
Na	so˩tʰɑ˧ʁo˧	so˩tʰɑ˧lo˩	so˩tʰɑ˧tʂʰwʌ˧mi˧	so˩tʰɑ˧kɯ˧pʰɯ˧
Chinese	Sānxīng Tóu (三星头)	Sānxīng Shǒu (三星手)	Sānxīng "Chuōmī" (三星"戳咪")	Sānxīng Báixīng (三星白星)
Literal Meaning	the stars of "sheep"			
Chinese Lunar Mansion	Shēn (参)	Fá (伐)	---	Tiānláng Xīng (天狼星)
Western Designation	Orion's Belt	Orion's Broadsword	---	Sirius

IPA transcription of the divination comments of the glyphs:

(6) ʔ æ˧ tso˧ mʌ˧ ŋ̩i˩ lɛ˩ ɖi˧ dʑʌ˩, bi˧ dʑo˧ lɛ˩ qʰæ˩ mʌ˧ dʑʌ˩,
what not COP. to get good outside EXIST. to give not good

ʑɯ˩ tsʰɨ˥ ʔwɑ˩ tsʰɨ˥ dʑʌ˩, gʮ˧ pʰo˧ bɑ˩ pʰo˧ dʑʌ˩.
house to build foundation to build good crop to sow crop to sow good

ʑoɬ kɯɬ gᵑ̩ɻ boɬ ʔɑɬ ʁoɬ qɨɬ ʁoɬ dʑʌɻ, dʐiɬ tsʰɨɻ
sheep star livestock home to get back good money to come

ʁwʌɬ bᵑ̩ɻ ʔɑɬ ʁoɬ qɨɬ ʁoɬ dʑʌɻ, lɛɬ tɕʰiɬ mʌɬ dʑʌɻ. ʑoɬ
mountain many home to get back good to sell not good sheep

tɕʰiɬ mʌɻ dʑʌɻ.
to sell not good

Translation: (On the days of "Sanxing Tou", "Sanxing Shou", "Sanxing Chuomī", and "Sanxing Baixing",) it is good to get things and not good to give away things; it is good to build the houses and to establish the foundations of houses; it is good to sow and to work in the field. On the days of the constellation "sheep", it is good to buy livestock, to achieve big income; it is not good to sell sheep.

Remarks: During the days when the stars/asterisms from the "Three Stars" Constellation are on duty, it is auspicious to take in instead of giving out stuff. They are days suitable to build up houses and to work in the fields. It is not auspicious to sell sheep.

3.7 The "Pheasant" Constellation

Script	14. 
Na	hoɬkɯɬ
Chinese	Yějī Xīng (野鸡星)
Literal Meaning	star of pheasant
Chinese Lunar Mansion	Guǐ (鬼)
Western Designation	Beehive Cluster

IPA transcription of the divination comments of the glyph:

(7) ʁɯɬ gwʌɻ dʑʌɻ, ʐwæɬ gwʌɻ dʑʌɻ, gᵑ̩ɻ boɬ gwʌ˥ kɯɬ dʑʌɻ.
bull to teach good horse to teach good livestock to teach star good

Translation: (On the day of "Yeji Xing",) it is good to teach bulls and horses how to work. It is good to teach livestock that day.

Remarks: The day in which "Yeji Xing" is on duty is suitable to tame livestock.

3.8 The "Hawk" Constellation

Script	15. 
Na	kʌ˩kɯ˧
Chinese	Yīng Xīng (鷹星)
Literal Meaning	star of hawk
Chinese Lunar Mansion	Wǔdìzuò (五帝座)
Western Designation	Denebola

IPA transcription of the divination comments of the glyph:

(8) gʲ˧ pʰo˧ ba˩ pʰo˧ mʌ˧ dʑʌ˩, gʲ˩ bo˧ tɕʰi˧ mʌ˧ dʑʌ˩, χwa˧
crop to sow crop to sow not good livestock to sell not good to buy

mʌ˧ dʑʌ˩, kʰʲ˧ ʂæ˥ ɕi˩ di˩ dʑʌ˩.
not good dog to pull to fish good

Translation: (On the day of "Ying Xing",) it is not good to sow or to work in the field; it is not good to buy or to sell livestock. Only fishing or hunting is suitable for that day.

Remarks: The day in which "Ying Xing" is on duty, it is suitable to fish and to hunt. It is not allowed to grow crop, to work in the field, or to exchange livestock.

3.9 The "Pig" Constellation

Script	16. 	17. 	18. 
Na	bo˩kʰwʌ˧	bo˩dʑɯ˧	bo˩ma˧
Chinese	Zhūzuǐ Xīng (猪嘴星)	Zhūniào Xīng (猪尿星)	Zhūyóu Xīng (猪油星)
Literal Meaning	the mouth of pig	the urine of pig	the fat of pig
Chinese Lunar Mansion	Xuānyuán Shísì (轩辕十四)	Xuānyuán Shí'èr (轩辕十二)	Tàiwēiyòuyuán (太微右垣)
Western Designation	Regulus	Algieba	Two among σ Leo, ι Leo, θ Leo, and δ Leo

IPA transcription of the divination comments of the glyphs:

(9) bo˩ tɕʰi˧ mʌ˧ dʑʌ˩ qʰo˥ mʌ˧ dʑʌ˩, χwa˧ mʌ˧ dʑʌ˩, ʔæ˧ tso˧
pig to sell not good to kill not good to buy not good what

mʌ˧ ŋi˩ dʑʌ˩.
not COP. good

Translation: (On the days of "Zhuzui Xing", "Zhuniao Xing", and "Zhuyou Xing",) it is not good to sell pigs, to kill pigs, or to buy pigs. All the others things are good to be done.

Remarks: During the days when the stars/asterisms from the "Pig" Constellation are on duty, everything is allowed to be done, except to exchange or to kill pigs.

3.10 The "Mdzo" Constellation

Daba Awo did not give the explanation for syllable [zɨȷ]. I gloss it as "mdzo" according to the results of comparisons with my field work data from other villages and other studies on Dongba lunar mansions (cf. Xu, 2015, pp. 70–71).

Script	19. 	20. 	21. 	22. 	23. 
Na	zɨȷzɣ˦	zɨȷq^{h}ɻ˦	zɨȷʱi˦	zɨȷnʝjæ˦	zɨȷgɣ˦
Chinese	Piānniú Sì (犏牛四)	Piānniú Jiǎo (犏牛角)	Piānniú Ěr (犏牛耳)	Piānniú Yǎn (犏牛眼)	Piānniú Zhǎng (犏牛掌)
Literal Meaning	four sides of mdzo	the horn of mdzo	the ear of mdzo	the eye of mdzo	the foot of mdzo
Chinese Lunar Mansion	The area between Jiǎo (角) and Jī (箕)				
Western Designation	The area between the Horn mansion (Spica) to Winnowing Basket				

IPA transcription of the divination comments of the glyphs:

(10) ɖæ˦ ɻæ˦ dʑʌȷ, (zɨȷ qʰɻ˦) ʐæȷ tsʰɨ˥ kɯ˦ dʑʌɻ.
foundation flat good mdzo horn column to build star good

Translation: (On the days of "Pianniu Si", " Pianniu Jiao", " Pianniu Er", " Pianniu Yan", and " Pianniu Zhang"), it is good to make the foundation of houses flat. It is good to erect columns on the day of " Pianniu Jiao". (Nothing is forbidden during these days.)

Remarks: During the days when the stars/asterisms from the "Mdzo" Constellation are on duty, it is allowed to build the foundation of the houses. During the day in which "Piānniú Jiǎo" is on duty, it is good to erect the columns. Nothing is forbidden.

3.11 The Unknown Constellation (Hǔzuǐ Xīng (虎嘴星) and Ròushí Xīng (肉食星))

They are two local 'original' stars. It has not yet been possible to find reliable parallels in Dongba culture and other major cultures.

Script	24. 	25. 
Na	lɑɬhy̰ɬkʰwʌɬ	ʂʌlʤɻlʤy̰ɬ
Chinese	Hǔzuǐ Xīng (虎嘴星)	Ròushí Xīng (肉食星)
Literal Meaning	the star of tiger's mouth	the star of carnivore
Chinese Lunar Mansion	--	
Western Designation	--	

IPA transcription of the divination comments of the glyphs:

(11) ʔɑɬ tsoɬ mʌɬ ŋ̩jl mʌɬ ʤʌl, lol by̰ɬ ʤʌl, kʰy̰ɬ ʂæɭ ɕil dil
what not COP. not good incantation to chant good dog to pull to fish

ʤʌl. ʔɑɬ ʁoɬ gy̰l by̰ɬ sɛɬ ʑɯɬ tɕʰiɬ bjɛɬ mʌɬ ʤʌl, qʰoɭ bjɛɬ
good other livestock blood to have to sell FUT. not good to kill FUT.

mʌɬ ʤʌl, χwɑɬ bjɛɬ mʌɬ ʤʌl.
not good to buy too not good

Translation: (On the days of "Huzui Xing" and "Roushi Xing",) nothing is good to be done except chanting incantations in order to expel ghosts, and pulling dogs (to hunt) and fishing. It is not good to sell livestock, to buy livestock, or to kill livestock.

Remarks: During the days when "Huzui Xing" and "Roushi Xing" are on duty, it is allowed to chant spells to expel ghosts and to hunt. Nothing else is allowed, especially exchanging and/or killing livestock.

3.12 The Unknown Constellation (Tóu Xīng (头星) and Wěi Xīng (尾星))

Script	26. 	27. 
Na	ʂwæɬqʰwʌɬ	mælqʰwʌɬ
Chinese	Tóu Xīng (头星)	Wěi Xīng (尾星)
Literal Meaning	the head star	the tail star
Chinese Lunar Mansion	Jī (箕)	Dǒu (斗)
Western Designation	Winnowing Basket	Dipper

IPA transcription of the divination comments of the glyphs:

(12) ʑɑl ɬɑɬ ʤʌl, tsʰɯl tɕʰiɬ mʌɬ ʤʌl. tsʰɯl kɯɬ tɕʰiɬ bjɛɬ mʌɬ ʤʌl,
others good goat to sell not good goat star to sell go not good

χwɑɬ bjɛɬ mʌɬ ʤʌl, ʔɑɬ ʁoɬ lɛl poɬ ʑiɬ ʤʌl.
to buy to go not good home to bring back good

Translation: (On the day of "Tou Xing" and "Wei Xing",) there is nothing bad, except to sell goats or to buy goats. It is good to bring back home stuff.

Remarks: During the days when "Tou Xing" and "Wei Xing" are on duty, it is good to bring stuff into the house, not good to exchange goats. Nothing else is forbidden.

4 Conclusion

The present study aims to be a primary source for philological documentation and research on Daba script, with an accurate transcription and introduction of field word data. After having reviewed the available academic studies on Daba script, I pointed out the need of a detailed report describing Daba script.

In the third section of this work, I have presented the interpretation of Daba script from an old Daba calendar, *Gemu*, from Wujiao Village. This is a single-symbol Daba calendar, indeed the most original kind of calendar according to the classification established by Song (2003, p. 86).

Daba calendar is used by the priests in order to select auspicious days for rituals. Therefore, it can be classified into hemerology among different types of calendars. During the field work in which I have tried to decipher these glyphs, I noticed that Daba priests recite by heart the 28 lunar mansions names in a certain order, rather than reading them one-by-one. In other words, they depend more on their memory of the star system than on the written texts. The designations are relatively fixed according to the related positions among these 28 glyphs. This explains why the visual representation of the same lunar mansion could be different among the villages.

This is the first report introducing all-embracing information on Daba glyphs and Daba calendar (in its different versions). It provides an exact transcription of the glyphs' pronunciations and religious prescriptions in the local language, with Chinese and English glosses and translations. Moreover, it lists all the divination prescriptions related to the 28 lunar mansions in Daba context.

Daba culture has developed a specific and original methodology in choosing stars/asterisms in order to calculate the dates. Interpreting this traditional system, I have annotated the corresponding Chinese lunar mansions and their Western designations as well. This can help in identifying Daba script and the lunar mansions recorded by these ideograms in their whole cultural background, considering also their relationships with lunar mansions/stars in other (related and unrelated) cultures.

References

Boltz, W. (1993). Shuo wen chieh tzu [說文解字]. In M. Loewe (Ed.), *Early Chinese Texts: A Bibliographical Guide* (pp. 429-442). Early China Special Monograph Series No. 2. Berkeley: Society for the Study of Early China, and the Institute of East Asian Studies, University of California.

Istrin, V. (1987). *Vozniknoveniye i razvitiye pis'ma [The Emergence and Development of Writing]*. (S. Zuo [左少兴], Trans.). Beijing [北京]: Beijing Daxue Chubanshe [北京大学出版社].

He, J. [和即仁], & Jiang, Z. [姜竹仪] (1985). *Nàxīyǔ Jiǎnzhì* [纳西语简志] [A Brief Description of the Naxi Language]. Beijing [北京]: Minzu Chubanshe [民族出版社].

Li, L. [李霖灿], Zhang, K. [张琨], & He, C. [和才] (1972). *A Dictionary of Mo-So Hieroglyphics*. Taipei [台北]: Wenshizhe Chubanshe [文史哲出版社].

Li, L. [李霖灿] (1984). Mósuō Yánjiū Lùnwénjí [麽些研究论文集] [Collection of Papers on Moso Studies]. Taipei: National Palace Museum.

Song, Z. [宋兆麟] (2003). Mósuōrén de Xiàngxíng Wénzì 摩梭人的象形文字 [Hieroglyphic Writing of Moso People]. *Southeast Culture,* (4), 86–93.

Xu, D. [许多多] (2013). Dábā Lìshū Jiědú jí Fúhào Xìngzhì Chūtàn [达巴历书解读及符号性质初探] [An Interpretation of Ephemeris of Dabaism and Discussion of the Symbols' Nature]. *Yuyanxue Yanjiu* [语言学研究] [Linguistic Studies], *13*(1), 40–51.

Xu, D. [许多多] (2015). A Comparison of the Twenty-Eight Lunar Mansions between Dabaism and Dongbaism. *Archaeoastronomy and Ancient Technologies*, *3*(2), 61–81.

Xu, S. [许慎][Han Dynasty] (2001). *Shuō Wén Jiě Zì* [说文解字] [An Etymological Dictionary Explaining Graphemes and Analyzing Characters]. Nanjing [南京]: Jiangsu Guji Chubanshe [江苏古籍出版社].

Yang, X. [杨学政] (1994). *Zàngzú, Nàxīzú, Pǔmǐzú de Zàngchuán Fójiào: Dìyù Mínzú Zōngjiào Yánjiū* [藏族, 纳西族, 普米族的藏传佛教: 地域民族宗教研究] [The Tibetan Buddhism of Tibetan People, Naxi People, and Pumi People: A Study of Regional Ethnic Religion]. Kunming [昆明]: Yunnan Renmin Chubanshe [云南人民出版社].

L1 PROSODIC INTERFERENCE: THE CASE OF SLOVENE STUDENTS OF JAPANESE

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

Abstract

A bidirectional perception experiment was conducted on Japanese and Slovene subjects to evaluate the result of a full L1 prosodic interference in recognizing (lexical) accent place in declaratives and interrogatives. Perceptual hypercorrection into L1 prosody on the side of the listener was achieved by making the subjects think they were listening to their own language, and results show clear tendencies for errors, which in general agree with predictions. However, mapping from phonetic to phonological representations was found to be asymmetric, suggesting that subjects of the two languages rely on different phonetic cues, as well as that distinctive function of certain phonetic cues, such as duration, has different effects on perception of segmental structure.

Keywords: perception; lexical accent; L1 interference; Japanese; Slovene

Povzetek

Raziskava razkriva rezultate obojestranskega testa slušnega zaznavanja, ki ga je avtorica izvedla na slovenskih in japonskih slušateljih z namenom, da bi ocenila vpliv maternega jezika na prepoznavanje besednega naglasa v besedah povednega in vprašalnega naklona. Percepcijska hiperkorekcija v prozodijo maternega jezika je bila dosežena tako, da so slušatelji mislili, da poslušajo besede v svojem maternem jeziku. Rezultati kažejo jasne tendence napak, ki se v splošnem skladajo s predvidevanji. Vendar pa hkrati rezultati kažejo, da je bilo uvrščanje fonetičnih oblik v fonološke nesimetrično, zaradi česar lahko sklepamo, da se slušatelji obeh jezikov pri slušnem zaznavanju besednega naglasa zanašajo na različne fonetične lastnosti. Poleg tega lahko sklepamo tudi, da ima pomensko-razločevalna funkcija fonetičnih lastnosti, kot je na primer trajanje, različne vplive na percepcijo segmentne strukture.

Ključne besede: slušno zaznavanje; besedni naglas; vpliv maternega jezika; japonščina; slovenščina

1 Introduction

Second language (L2) learners tend to make mental associations or interlingual identifications between structures of the two languages (Weinreich, 1953; Odlin, 1989; Jarvis & Pavlenko, 2008), and first language (L1) interference is especially high when

Acta Linguistica Asiatica, 6(2), 2016.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.6.1.69-76

phonological differences between L1 and L2 are hidden behind the phonetic similarities (Eckman et al., 1989; Eckman, 2008).

In this paper, a bidirectional perceptual experiment was conducted to discuss how well a learner perceives L2 accent place under the supposition of a full L1 prosodic interference, or in other words, how well a native listener can recognize accent place from the learner's 'foreign accent' pronunciation.

Much has been done on phonetics and phonology of Japanese accentuation (e.g. Haraguchi, 1977; Shibatani, 1972; Sugito, 1972; Uwano, 2003; etc.), and some on Slovene accentuation (e.g. Bhaskararao & Golob, 2006; Srebot-Rejec, 1988, 1997; Šuštaršič, 1995, 2004; Toporišič, 1965, 2000; etc.) respectively, however, very little is known on the prosodic interference of the two languages[1]. A comparative study on acoustics of accentuation (Golob, 2005; Golob, 2011 compared Japanese pitch accent and Slovene stress accent) pointed out a different interaction of accentual and intonational tier in the two languages, and proposed a possible L1 interference as shown in Figure 1 (Golob, 2005).

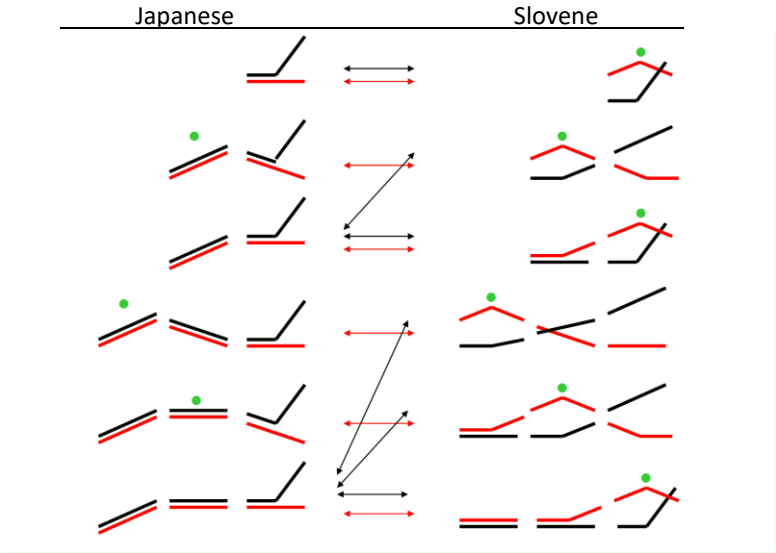


Figure 1: Hypothetical L1 interference between Japanese and Slovene based on pitch patterns. Declaratives are in red and interrogatives in black. Accent place is marked with a green dot.

Accent place in declaratives is expected to be perceived correctly since the similarities in pitch patterns stay within the so-called 'bilingual minimal pair'. This is not the case with interrogatives, where certain accent patterns in Japanese are expected to take more than one Slovene pattern while others show no matches.

[1] Prosodic influence was reported for Slovene speakers of Chinese, another language with tonal contrast (Petrovčič & Lin, 2015).

The present study aims to verify the above hypothesis and investigates on Japanese accent patterns that hypothetically show no matches. It further briefly examines the role of duration and intensity (which are ignored in the above hypothesis) in the misperception of an accent place.

2 Perception experiment

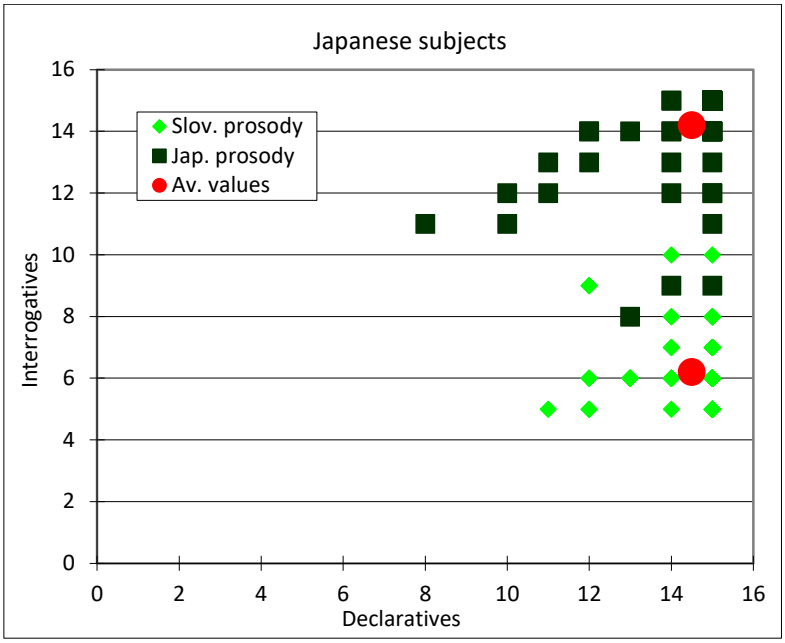
2.1 Methodology

Altogether 173 subjects, 106 Japanese and 67 Slovene native speakers, were asked to listen to their native language (reiterant speech, Larkey, 1983) and mark the accent place. The material, in total including 60 items, consisted of 2- and 3-syllabic words, half with Japanese and half with Slovene prosody, and chosen randomly from 10 Japanese and 10 Slovene native speakers. 'No accent place' answer option was given to Japanese subjects for the Japanese so-called *heibangata* accent pattern, cf. an accentless word, and was equated with the Slovene word accented on the final syllable.

Subjects scoring less than 50% in either declaratives or interrogatives of their native language were eliminated from further analysis (25 Japanese and 6 Slovene subjects).

2.2 Results

Figures 2a, 2b show results from 82 Japanese (top) and 61 Slovene subjects (bottom). Generally speaking, the difference between L1 and L2 perception was statistically significant for both groups of subjects ($p<0.001$), error rate being higher for L2 perception. Japanese subjects show a very high score for both intonational surroundings in L1 (96.3%, SD 9.0 and 94.7%, SD 10.2) as well as for declaratives in L2 (96.4%, SD 6.6), while L2 interrogatives, with only 41.6% in average (SD 6.9), show strong and clear error tendencies (see below). Slovene subjects also perceive accent places of L1 (89.0%, SD 12.5 and 82.1%, SD 14.1) better than those of L2 (76.7%, SD 15.9 and 57.7%, SD 17.9) but their error rate rises relatively evenly in both intonational surroundings, the tendency being typical for certain Japanese accent patterns (described in detail in 2.2.1).



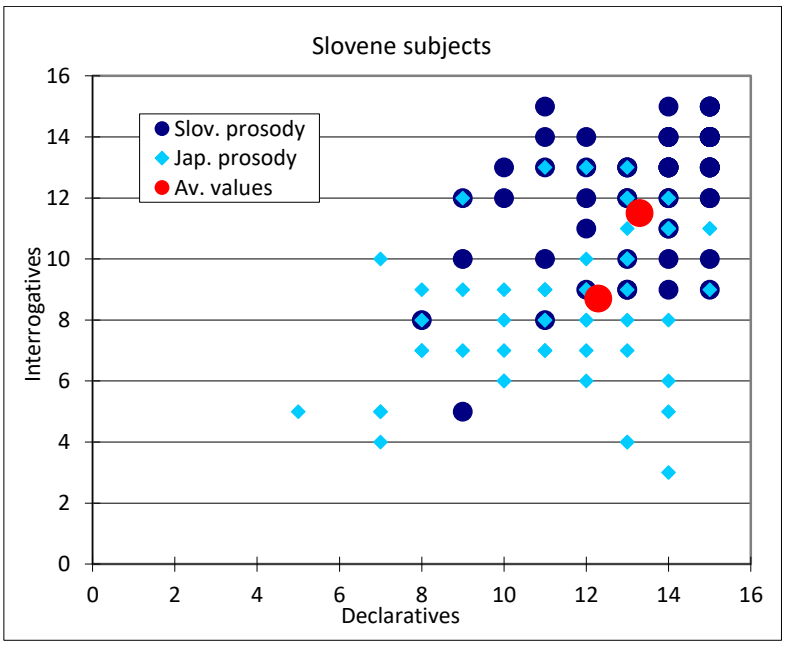


Figure 2a, 2b: Scattergrams showing correct answers in L1 and L2 perception by Japanese (top) and Slovene subjects (bottom).

2.2.1 Error analysis

Error rates were observed based on accent patterns and error tendencies were further evaluated to conclude on general L1 interference.

Results for Japanese subjects show that error rates in declaratives are negligible (Figure 3a), while in interrogatives (Figure 3b), there is a clear distinction between high and low error rates, depending on accent pattern. Validity is very low with words that are accented on a non-final syllable (average error rate is 92%, SD 4.1), of which 89.4% (SD 3.9) of cases were judged to have accent on the final syllable.

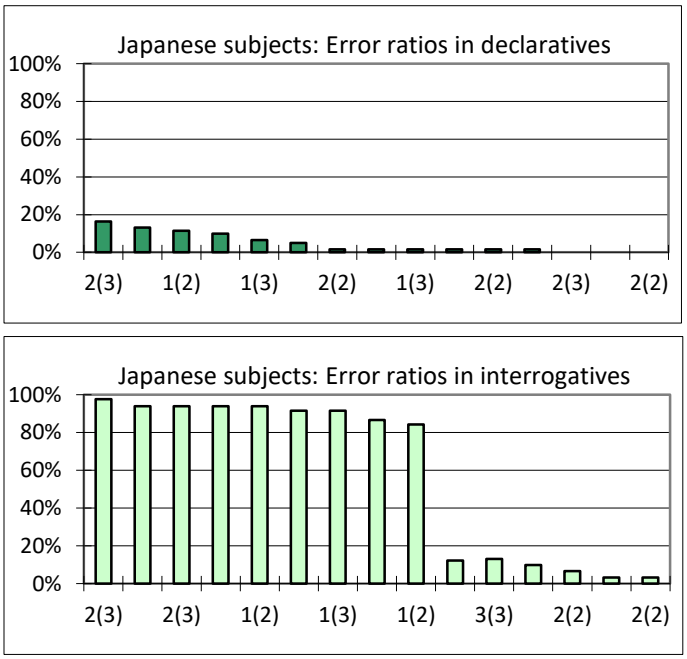


Figure 3a, 3b: Error rates for each word in declaratives (top) and interrogatives (bottom) by Japanese subjects. X-axis shows accent place and number of syllables in a word (in brackets).

Figures 4a and 4b show that, irrespective of intonational surrounding, Slovene subjects tend to make most errors in cases of Japanese non-accented words (28.8%, SD 16.5 and 50.2%, SD 19.6), placing the accent onto the first syllable (13.0%, SD 7.0 and 20.3%, SD 7.4) or penultimate syllable (11.7%, SD 6.5 and 18.7%, SD 12.4) in 3-syllabic words, and onto the first syllable in 2-syllabic words (10.1%, SD 3.6 and 20.3%, SD 5.5). Words accented on the first, penultimate, or first and penultimate syllable show slightly higher validity with the error ratios of 8.3%, SD 3.2 and 16.0%, SD 10.8, 16.7%, SD 4.7 and 27.0%, SD 7.5, 9.7%, SD 6.0 and 24.3%, SD 17.6, respectively. Their error tendencies seem to be random.

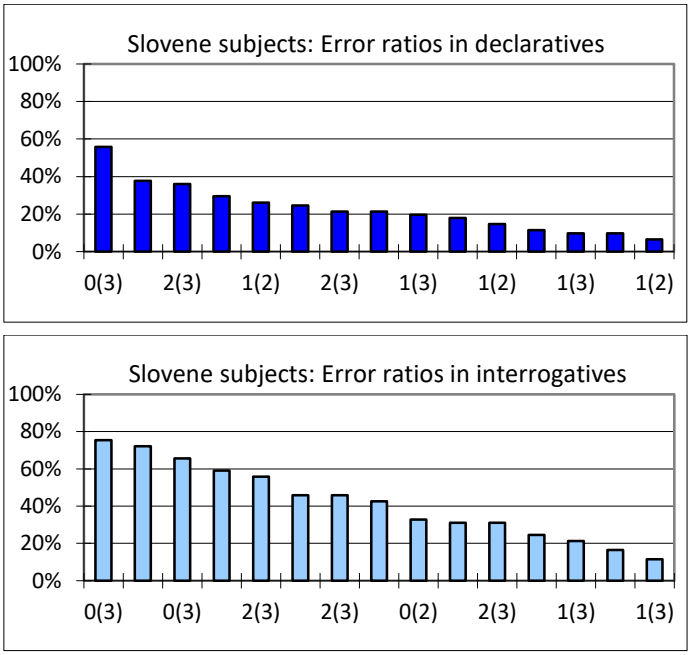


Figure 4a, 4b: Error rates for each word in declaratives (top) and interrogatives (bottom) by Slovene subjects. X-axis shows accent place and number of syllables in a word (in brackets).

3 General discussion and conclusions

A bidirectional study predicted symmetric mapping of phonetic information into L1 phonology. Predictions were confirmed for Japanese subjects, who extremely well perceived all but interrogatives with a non-final accent place, mistaking them for non-accented words. Similar results were obtained by Ayusawa (reviewed in Ayusawa, 2003) for several other language groups, which additionally shows that Japanese native listeners rely on the pitch fall in perception of accent place (Sugito, 1972; Uwano, 2003). Despite the low error rate, some Japanese subjects pointed out the possible incorrect transcription on the answer sheet, claiming a missing moraic vowel on the accented syllable, which points at L1 prosodic characteristics interfering with segmental organization. On the other hand, Slovene subjects show a weak tendency to place accent on a non-final syllable, following a common accent patterning in Slovene (Toporišič, 2000 [1976]). In general, however, errors were observed to occur randomly, suggesting that pitch information alone is not sufficient to correctly perceive accent place. Duration was shown to be an important phonetic cue in Slovene (Bhaskararao & Golob, 2006; Toporišič, 2000 [1976]), and further investigation is needed to show how much the lack of durational information (also those on intensity?) contributes to the misperception of accent place.

References

Ayusawa, T. [鮎澤孝子] (2003). Gaikokujingakushuushano nihongo akusentointoneeshon shuutoku [外国人学習者の日本語アクセント･イントネーション習得]. O*nseekenkyuu* [音声研究], *7*(2), 47–58.

Bhaskararao, P. & Golob, N. (2006). *What matters in Slovene accent? An acoustic comparion of stress and pitch accents.* 1st Slovene International Phonetic Conference, Ljubljana.

Eckman, F., Moravcsik, E. & Wirth, J. (1989). Implicational universals and interrogative structures in the interlanguage of ESL learners. *Language Learning, 39*(2), 173–205.

Eckman, F. (2008). Typological markedness and second language phonology. In J.G. Hansen-Edwards & M.L. Zampini (Eds), *Phonology and Second Language Acquisition* (pp. 95–115). Amsterdam: John Benjamins Publishing Company.

Golob, N. (2005). *Phonological approach to acoustic analysis of Japanese and Slovene prosody: accent and intonation*. 11th International Conference of EAJS, Vienna.

Golob, N. (2011). Acoustic Prosodic Parameters in Japanese and Slovene: Accent and Intonation. *Acta Linguistica Asiatica*, *1*(3), 25–44.

Haraguchi, S. (1977). *The Tone Pattern of Japanese: An Autosegmental Theory of Tonology. Tokyo: Kaitakusha.*

Larkey, L. S. (1983). Reiterant Speech: an acoustic and perceptual validation. *Journal of the acoustical Society of America*, *73*(4), 1337–1345.

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. Cambridge: Cambridge University Press.

Petrovčič, M. & Lin, M. (2015). *Sodobna kitajščina 1*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Srebot Rejec, T. (1988). *Word accent and vowel duration in standard Slovene: An acoustic and linguistic investigation*. München: Verlag O. Sagner.

Srebot Rejec, T. (1997). Nekaj o stavčni intonaciji v knjižni slovenščini. *Slavistična revija*, *45*(3–4), 429-455.

Šuštaršič, R. (1995). Naglas in poudarek v angleščini in slovenščini. *Slavistična revija*, *43*(2), 157–182.

Šuštaršič, R. (2004). A contrastive analysis of the vowel qualities of English and Slovene. In E. González-Álvarez & A. Rollings (Eds.), *Studies of contrastive linguistics* (pp. 601–608). Santiago de Compostela: Universidade de Santiago de Compostela.

Shibatani, M. (1972). The non-cyclic nature of Japanese accentuation. *Language*, *48*(3), 584–595.

Sugito, M. [杉藤美代子] (1972). 'Ososagari' koo: Dootai sokuteeni yoru nihongo akusentono kenkyuu [“おそ下り”考—動態測定による日本語アクセントの研究—]. In Tokugawa [徳川宗賢] (ed.). *Ronshuu nihongo kenkyuu 2 Akusento* [論集日本語研究 2 アクセント]. Tokyo [東京]: Yuseido [有精堂].

Toporišič, J. (1965). Naglasni tipi slovenskega knjižnega jezika. *Jezik in slovstvo*, *10*(2–3), 6–79.

Toporišič, J. (2000 [1976]). *Slovenska slovnica*. Maribor: Obzorja.

Uwano, Z. [上野善道] (2003). Akusentono taikeeto shikumi [アクセントの体系と仕組み]. *Asakura nihongo kouza* [朝倉日本語講座], *3*. Tokyo [東京]: Asakurashoten [朝倉書店].

Weinreich, U. (1953). *Languages in Contact: Findings and Problems*. The Hague: Mouton.

Acknowledgement

My warm thanks are due to Prof. Hiroshi Nakagawa (Tokyo University of Foreign Studies) and Prof. Chikako Shigemori-Bučar (University of Ljubljana) who generously created the conditions for this experiment, and also to the subjects for their substantial contributions.

RESEARCH ARTICLES (PROJECT REPORTS)

IMPROVING STUDENTS' LANGUAGE PERFORMANCE THROUGH CONSISTENT USE OF E-LEARNING: AN EMPIRICAL STUDY IN JAPANESE, KOREAN, HINDI AND SANSKRIT

Sara LIBRENJAK
University of Zagreb, Croatia
sara.librenjak@gmail.com

Kristina KOCIJAN
University of Zagreb, Croatia
krkocijan@ffzg.hr

Marijana JANJIĆ
University of Zagreb, Croatia
marijanajanjic@yahoo.com

Abstract

This paper describes the backing theories, methodology, and results of a two-semester long case study of the application of technology in teaching four Asian languages (Japanese, Korean, Hindi, and Sanskrit) to Croatian students. We have developed e-learning materials to follow the curriculum in Croatia and deployed them in Asian language classrooms. Students who agreed to participate in the study were tested before using the materials and after each semester, and their progress was surveyed. In the case of Japanese students (N=53), we have thoroughly monitored their usage and compared the progress of students who have diligently studied vocabulary and grammar using our materials on Memrise, and those who have neglected their studies. This was measured through their scores on the Memrise, which shows the user's activity. Also, their progress was measured using standardized tests that were designed in such a manner to resemble Japanese Language Proficiency Test. We have found that frequent users progressed averagely 20,3% after each semester, while non-frequent users have progressed only 11,6%, proving this method to be related to stable and constant use of e-materials.

Keywords: e-learning; second language teaching; Japanese; Korean; Hindi; Sanskrit; Memrise; Anki; Quizlet

Povzetek

Članek obravnava dva semestra trajajočo študijo primera o uporabi tehnologije pri poučevanju štirih azijskih jezikov hrvaškim študentom; japonščine, korejščine, hindijščine in sanskirta. Razvili smo elektronsko učno gradivo in ga dali v uporabo študijskim programom z omenjenimi štirimi azijskimi jeziki. Študentje, ki so bili pripravljeni sodelovati, so izvedli preizkus znanja jezika pred uporabo elektronskega učnega materiala ter po vsakem končanem semestru. Analiziran je bil njhov napredek. V primeru študentov japonščine – teh je bilo 53 –, smo se osredotočili na njihovo uporabo elektronskega učnega gradiva in primerjali študente, ki so pri učenju besedišča in slovnice redno uporabljali e-gradivo na spletni strani Memrise, s tistimi, ki niso uporabljali e-gradiva. Uporabili smo število dostopov na spletno stran, kar nakazuje na aktivnost učenja študentov. Njihov napredek smo

Acta Linguistica Asiatica, 6(2), 2016.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.6.2.79-94

merili s testi, ki so bili oblikovani na osnovi standardnih testov japonščine JLPT. Ugotovili smo, da so se študentje, ki so redno uporabljali e-gradivo, na testih izkazali bolje, saj so v primerjavi z rezultati na prejšnjih testih napredovali za 20,3%, medtem ko so tisti študentje, ki niso uporabljali e-gradiva, napredovali le za 11,6%. S tem smo dokazali koristnost redne in dolgoročne uporabe elektronskih učnih materialov.

Ključne besede: e-učenje; poučevanje drugega jezika; japonščina; korejščina; hindi; sanskrt; Memrise; Anki; Quizlet

1 Introduction

This paper presents the results of a year-long project, with the goal of producing systematic and sustainable materials for e-learning of Asian languages in Croatian language. We have covered four languages within the project: Japanese, Korean, Hindi and Sanskrit. The materials for each language were created by a language teacher and a team of collaborators, using the most common textbooks used in Croatia for the reference. Our primary goal was to create e-learning and m-learning materials which can be most beneficial to learners and for that reason the materials should follow the classroom curriculum as much as possible. A secondary goal of the project was the research of the correlation between e-learning using spaced repetition algorithms and learners' improvement of vocabulary, grammar, and writing in written language performance only, i.e. test speaking and listening skills were not included. This is due to the format of the materials and short-term length of the project, but it is certainly planned in the future should this method prove to be successful.

The first phase of the project consisted of conducting the comprehensive survey with the learners of Japanese, Korean, Hindi and Sanskrit throughout Croatia, and learning about difficulties in their studies, specific needs in the study materials, and their attitudes towards e-learning. The second phase was creating materials according to the survey results, on the e-learning and m-learning platforms Memrise, Quizlet and Anki, and their employment in the language classrooms as much as possible. Throughout the project, we have conducted language testing every three months, in order to assess the improvement in students' performance. The final phase included reviewing the results and using this input to improve the materials for the new generation of learners. In this article we will present some current studies in the similar fields, the methodology of our research (creation of materials and testing), our results and their commentary.

2 Languages and technology

2.1 Terminology issues: CALL, e-learning, m-learning and languages

In the *MemAzija* project, we dealt with using technology in language learning. It should be recognized that there is a lot of terminology relating to this field, some of which intersect. Computers and language learning have been a topic of research even before the broad spread of Internet and ever-present smart-phones. This project falls into the broad field of computer-assisted language learning, or CALL (Hanson-Smith, 2002). According to Warschauer and Healey (1998), CALL is present from the 1960s, but the presence of Internet and its myriad of opportunities 'can truly revolutionize it.'

Thus, this project fits into the field of e-learning, but also online learning as we are using the Internet as a medium. It is important to mention that mobile learning, or m-learning, is an important part of the project because all materials can be, and are, accessed through mobile devices such as smart-phones and tablets. Terms e-learning and m-learning are not limited to language studies. In fact, they are often not systematically applied to the field of language learning. According to Simkova, Tomaskova and Nemcova (2012), e-learning is limiting learners because it does not enable them to study at any place, while m-learning enables them not only to study anywhere but also to receive information and news faster. In language study, the continuous input is of a great importance, so m-learning is an ideal option for languages that are not often encountered in Croatia, such as our goal languages. We are aiming to use the technology for learners' benefit, in the long term and providing systematically organized content which follows the curriculum.

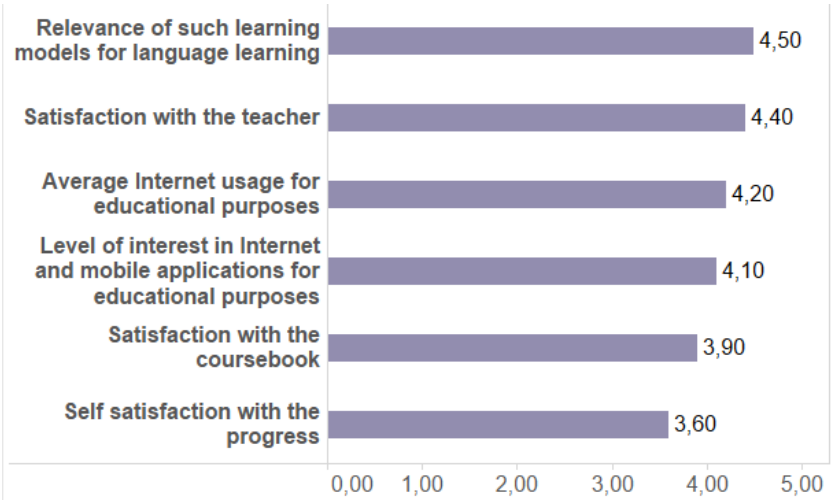
There are many benefits found in CALL and mobile/electronic language learning fields such as improving the results (Golonka et al., 2014) and also attitudes of students toward certain subjects (Yushau, 2006). However, we have discovered that in the case of Asian languages, long-term studies with a larger corpora of participants are very hard to find. One such study is about using Skritter software for Chinese character practice to improve proficiency (McLaren & Bettinson, 2016). With the lack of such studies for Asian languages, we hope that our research will contribute to the field in the area of technology usage in Asian language learning, especially considering promising results of our smaller-scale research, which we present in the next section.

2.2 Previous research

There is a plethora of research which deals with students' attitudes towards CALL (Ayres, 2002; Kuo, 2008; Rahimi & Yadollahi, 2011; Mashhadizadeh & Rezvani, 2015), without actually having tested materials in practice. In cases where some tests have been performed, they were either short-lived (e.g. one session with results of pre- and post- testing, as in De la Rouviere, 2013), or with a very small number of students (Tam & Huang, 2012). While these articles provide some insight into relationship between technology and language teaching and learning, we were interested in longer and more

systematic exposure to e-learning, especially that which is conducted as complementary to classroom curriculum.

Studies generally suggest favorable views on using technology in language classroom (Tam & Huang, 2012), as we have also found in our survey in the first phase of the project. Out of 203 surveyed students of Asian languages, most have given a favorable grade to the prospect of using technology in their studies, but they also reported mostly on not being previously exposed to it in the classrooms of Japanese, Korean, Hindi or Sanskrit. A small number of students used some e-learning websites on their own, while only 23.6% have had a chance to encounter e-learning and technology in language learning in the classroom. Graph 1 (adapted from Janjić et al., 2016) shows Croatian students' attitudes toward e-learning in Asian languages. This research was conducted on all age groups of learners of Asian languages, most of them being university students.[1]



Graph 1: Attitudes of Croatian students of Asian languages towards e-learning

Even though literature review encourages us to start developing technology-based materials for studying, we should be wary of overstating their usefulness. Wiebe and Kabata (2010) warn us that instructors tend to assess the usefulness of their computer-based materials higher than students. This does not necessarily signify they are not as useful as we have thought, but that we need to develop materials tailored to specific learners' needs. This can not be done without proper research and testing. This project aims to develop the first draft of materials, measure their usefulness in this state, but also to collect feedback in order to improve them for the next generation of learners.

[1] Detailed demographics and survey results are presented in Janjić, Librenjak and Kocijan (2016).

3 Methodology

In order to systematically conduct research with e-learning and m-learning materials amongst students, we needed to establish a relatively strict methodology. Figure 1 shows the draft of the method for this research.

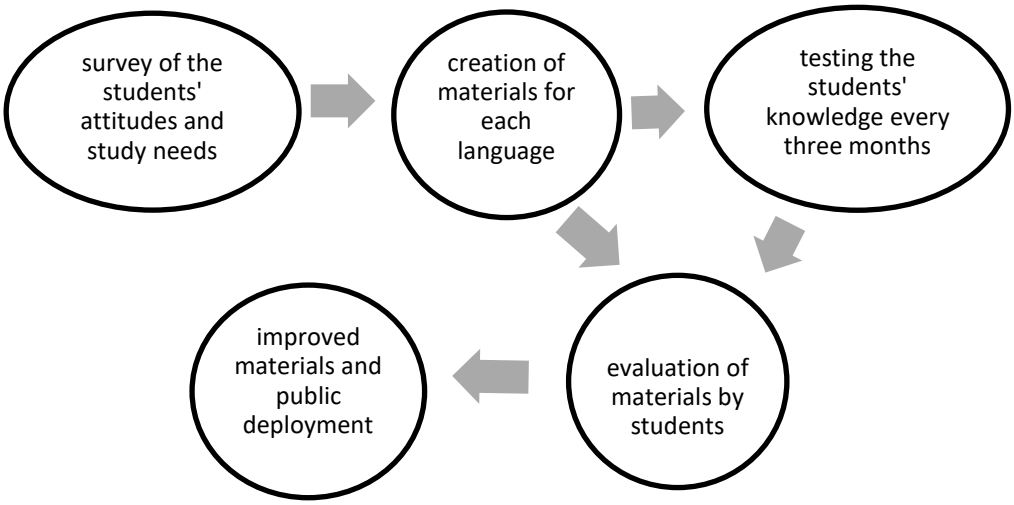


Figure 1: Schematic view of the methodology for this project

3.1 Student survey

Before creation of materials, the survey among 203 Croatian students of Asian languages (Japanese, Korean, Hindi and Sanskrit) was conducted, and specific information gathered (Janjić et al., 2016). We have found most difficult areas for each language because each was treated differently due to various differences: the script (e.g. Japanese uses most difficult writing system, as rated in the survey), the grammar (Indo-European versus non-Indo-European languages), pragmatic level (use of Sanskrit in a conversation is not very probable), etc. We also collected information about the most common textbooks used in Croatia, and reasons behind studying these languages. This information was used in order to determine how to create the best possible materials for the first phase, but having in mind that they would have to be further improved after receiving students' feedback.

3.2 Creation of e-materials on Memrise, Anki and Quizlet platforms

After the survey, we concluded that for the creation of e-materials we should use platforms that are available both as desktop and mobile version but are also easy to use from the teachers' and students' perspectives. For us, this meant that the platform has simple procedures for course creation as well as for course materials usage. Two of the chosen platforms (Memrise and Anki) use spaced repetition algorithm to improve students' memory, while the third one (Quizlet) uses games and text to speech

synthesis to improve students' listening skills and add more fun to learning. We have also divided the used platforms by language in order to respond to language specific problematic areas which were reported in the survey. The choice of a platform per language, as well as most common learning issues, are shown in Table 1.

Language	Most common learning issues	Platform(s) chosen	Reason
Japanese	script (kanji)	Anki	spaced repetition: the best choice for a large number of separate items to memorize
	speech production, grammar, vocabulary	Memrise	adaptable to various skills (grammar and vocabulary), best interface
Korean	speech understanding, writing and composition, grammar	Quizlet	automatic TTS (text to speech synthesis) ensures a lot of listening games to practice sentence formation
Hindi	speech production and understanding, memorizing vocabulary	Memrise	best choice for memorizing words in an attractive setting Quizlet provides no spaced repetition algorithm
Sanskrit	writing and composition, grammar, pronunciation	Memrise	adaptable to various language skills, popular with students

Table 1: Choices of platforms for creation of materials per language

It is important to note that these materials do not cover all necessary language skills. We have focused on improving vocabulary and script acquisition, and to some extent grammar. Platforms specialized for memorization, such as those used in our project, have their limitations. Thus, for specific grammar exercises some other e-learning materials should be used. A mobile and desktop applications for practicing grammar are being produced by authors for additional assistance in learning.

In this project, we have covered levels from A1 to B1, since there are not that many students who have so far reached level B1 in any of the languages. For the convenience of understanding and comparison it is important to keep in mind that names of these levels correspond to CEFR name levels. A1 designates a student beginner and material that should be covered by such student to reach level A2 which marks a learner who is a beginner with limited working proficiency. B1 level marks a student experienced with longer exposure to language learning, wider vocabulary than A1 and A2 level students, and is at beginning stages of intermediate level of his /her skills. We have extended the same classification over students of Sanskrit as well, which is like Latin considered a dead language even though it is claimed as the mother tongue by a small community

in India.[2] In this research we saw it fit to extended the division of levels to Sanskrit for one more reason: the research is focused on the study of students' acquisition of grammatical topics and vocabulary, and in that case whether a language is dead or not does not influence the distinction of beginner levels from intermediate levels, or in our case A levels and B levels. All the materials are published publicly under Creative Commons license (Attribution-NonCommercial-ShareAlike) and are available on the project website (http://memazija.ffzg.unizg.hr/), as well as on respective websites of each platform. They are not meant to be used only by the learners who actively participate in the project, but are free to be used by anyone who wants to learn either of the four languages in Croatian. Since the materials are easily adaptable and translatable, the raw materials are available upon request and could be localized for usage in another language.

3.3 Pre-usage and post-usage testing of students' knowledge

In order to assess the usefulness of materials, standardized tests were performed every three months. A number of learners who agreed to participate in the project for two semesters were monitored in their usage of materials through their user-names, and tested every three months for the improvement in their knowledge. In the case of Japanese, their test results were also compared to their individual usage of materials.

Students were asked to complete the test before we distributed the materials (pre-usage testing). After the first three months of using the materials, they were asked to complete another test of the equal format, but with a different set of questions. Another three months later, second post-usage testing was performed. All three tests were designed following the format of standardized tests for Japanese and Korean, and aligned with curriculum. However, for Hindi and Sanskrit such standardized tests were not available and therefore the tests followed topics covered by Croatian curriculum.[3] All tests included sections on vocabulary (with script), grammar and reading. As in the standardized tests published by the Ministries of Education of Japan or Korea, questions were multiple choice. The tests for Hindi and Sanskrit followed the same format. For the purpose of this project, the skills of writing, listening and speaking were not tested. Although they are certainly important language skills, they were not a part of this research project.

In the case of Japanese, we have used JLPT (Japanese Language Proficiency Test) format, dividing them into three levels: A1 approximately corresponding to N5, A2 approximately corresponding to N4 and B1 as a level approximately between N3 and

[2] See statistical data on mother tongues in India:
http://www.censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement1.aspx.

[3] Since standardized tests are an internationally objective way of measuring knowledge, we opted to use them for comparison of results whenever possible, while the materials themselves follow the curriculum.

N2. [4] Each test consisted of following sections: kanji reading and recognition, vocabulary, particles, grammar in context, dialogues, reading passages.

Korean tests were constructed on the basis of TOPIK (Test of Proficiency in Korean), published before the 2014 reform. Students who were taking A1 and A2 Korean class, were given a test similar to TOPIK I, and students who are taking B1 class took the test of the format similar to TOPIK II. The students were tested on vocabulary, particles, grammar, completing the paragraphs, ordering text and reading comprehension.

Hindi and Sanskrit tests were constructed following information from teachers on grammatical topics covered by students of each year and the textbooks used for study. Thus, the first year students of Hindi were given tests that correspond to A1 level, students of second year tests of A2 level and students of third year or higher tests that correspond to B1 level. They were tested on vocabulary, grammar, ordering text and reading comprehension.

In the case of Sanskrit, the main issue was that it is being taught primarily in a different way from other languages included in the study. As a language which is considered dead, and is thus comparable to Latin and ancient Greek, the dominant teaching method is concentrated on enabling students in the first year to independently recognize different grammatical units (verbal tenses, cases, sandhis, etc.). In collaboration with the teacher, these units were classified as levels which could correspond to A1, A2, etc. Vocabulary, deemed necessary for students to absorb, was extracted from texts students mostly encounter during their studies and was based on the frequency of lexemes' appearance. The students were tested on vocabulary, grammar and reading comprehension.

All tests are computer based and automatically graded, and could be used in any language since the questions are all in the target language, with the exception of cues and instructions. A student writes their user-name for the platform they are using when writing the test, and their result is stored in our database. In other words, we have not only monitored their grades on the standardized testing, but also its correlation to the frequency of their usage (i.e. scores which are awarded with the activity on Memrise). For the purpose of this study, we were able to perform this only for the students of Japanese, so the results pertaining to this comparison are not generalized for all four languages.

3.4 Evaluation of materials by students and further distribution

Finally, the materials needed to be graded by their users in order to improve them further. During the two semesters of their use, we accepted all reports of errors, bugs, and imperfections that could have been solved while not disrupting the availability of materials to users. Upon the completion of continuous two semester usage, students

[4] It should be noted that JLPT does not officially correspond to CEFR.

will also be asked to complete a comprehensive survey which will collect their attitudes towards this type of study, their grade of the materials in the whole as well as for each specific field and type. We also hope to be able to collect the suggestions to construct the lessons and tasks better, and re-issue faulty materials where necessary.

After we employ the needed improvements, we will be ready for the distribution and promotion of our materials to all types of users in Croatia including universities, language schools and private students. Although they have been publicly available since 2015, after additional improvement, a new wave of promotion throughout Croatia will be implemented. They will be promoted through workshops, school visits and the Internet.

4 Results and discussion

The materials were distributed to students throughout Croatia, and 65 of learners also agreed to participate in the two-semester monitoring of their progress while using the materials. All users were given a chance to participate in the testing in order to compare their results before and after using the materials. In the case of Japanese, their results were also compared with the frequency of using vocabulary and grammar exercises on Memrise. Other languages have not had a sufficient number of users to provide us with a detailed statistical account of the correlation between their Memrise scores and test scores, but we have recorded their test scores nevertheless.

4.1 An overview of offered courses and their users

Learners have had a number of courses to partake in. All of them follow most schools' curricula in the target language, so this should have been an additional motivational factor. It is important to note that instructors also play a significant role in motivating their students to use materials. According to Schmidt and Watanabe (2001), the role of a teacher as a motivator in language learning is significant in both generating, as well as in maintaining students' motivation. It was not possible to completely control and stimulate usage throughout two semesters in the various places of learning, so it was expected that course completion rate would not be 100%, i.e. that a number of students would stop using materials at some point. We expected that at lower levels, especially A1, would have most users since there are more students in e.g. the first year of Japanese or Hindi course than there are students in a second or third year. Table 2 shows courses created in the *MemAzija* project, as well as their current number of units, their main source and the current number of users.

Language and content		No. of units	Main source	Users	Audio + meme
Japanese	Vocabulary A1	768 words	Genki 1	85	yes
	Grammar and sentences A1	134 sentences		63	yes
	Vocabulary A2	751 word	Genki 2	28	yes
	Grammar and sentences A2	180 sentences		23	partially
	Vocabulary B1	772 words	Integrated approach to intermediate Japanese	11	not yet
	Grammar and sentences B1	266 sentences		10	not yet
	Vocabulary B2	85 words (currently in making)	Tobira: A gateway to advanced Japanese	N/A	not yet
Korean	Hangeul course	235 excercises	Active Korean 1	31	yes
	Vocabulary A1.1	402 words	Active Korean 1	25	yes
	Grammar and sentences A1.1	97 sentences			yes
	Vocabulary A1.2	413 words	Active Korean 2	8	yes
	Vocabulary A2.1	465 words	Active Korean 3	5	yes
	Vocabulary A2.2	387 words	Active Korean 4	4	yes
Hindi	Vocabulary A1	720 words	Complete Hindi: A Teach Yourself Guide	21	partially
	Grammar and sentences A1	501 sentence	Complete Hindi: A Teach Yourself Guide	22	yes
	Vocabulary A2	1148 words	Complete Hindi: A Teach Yourself Guide	4	partially
	Vocabulary A2.2 (synonymes, antonymes, expressions)	83 pairs	Complete Hindi: A Teach Yourself Guide	2	partially
	Grammar and sentences A2	453 sentences	Complete Hindi: A Teach Yourself Guide	3	yes
	Vocabulary B1	683 words	Living Language Hindi	2	partially
	Grammar and sentences B1	494 sentences	Living Language Hindi	2	not yet

Language and content		No. of units	Main source	Users	Audio + meme
Sanskrit	Vocabulary A1	297 words	Elements of Sanskrit Grammar, most frequent words from Mahabharata and Ramayana	21	partially
	Grammar and sentences A1.1	377 sentences	Elements of Sanskrit Grammar	20	not yet
	Grammar and sentences A1.2	219 sentences	Elements of Sanskrit Grammar	3	not yet
	Grammar and sentences A2.1	186 sentences	Elements of Sanskrit Grammar	2	not yet
	Grammar and sentences A2.2	73 sentences	Elements of Sanskrit Grammar	1	not yet
	Vocabulary A2	531 word	Elements of Sanskrit Grammar, most frequent words from Mahabharata and Ramayana	4	partially
	Vocabulary B1	522 words	most frequent words from epics, puranas and classical literature	4	partially
	Grammar and sentences B1	20 sentences (in making)	Elements of Sanskrit Grammar	N/A	not yet

Table 2: List of all MemAzija courses, their units, sources and users

Along with those courses, specifically for the acquisition of scripts (hiragana, katakana and kanji for Japanese, hangeul for Korean), Anki flashcard decks were constructed. As one can see, a number of users are significantly higher at lower levels, which is in accordance with the number of students in lower and higher school years, as well as the testing attendance throughout the academic year, which is discussed in the following paragraphs.

4.2 General testing results for all languages

Tests were conducted in three intervals: first test (pre-usage) in the beginning of the academic year 2015/2016, second test after the first semester (around February 2016) and the third test after the summer semester (June 2016). We invited all learners who agreed to partake in the study, but we were not always able to ensure their complete attendance, since it was voluntary, and the largest part of the participants are studying these languages as a hobby or a facultative subject, and not their major.

Still, we were able to collect enough information to compare their progress in each of the intervals. As Asian languages and their teaching methodology is a severely underdeveloped field in Croatia, at the point of writing this article there were not enough students to form control groups. Thus, the figures shown in Table 3 only serve as the record which should be compared with the future research. They cannot testify about the influence of e-learning on the general performance, as it is impossible to ascertain to which extent did e-learning and m-learning help in their improvement. The following chapter will discuss the case of Japanese, where we were able to perform more detailed analysis of the students' progress.

Table 3 summarizes the testing results per language per levels for each three testing intervals, as well as the number of attending students. Average scores are displayed in percentages of correct answers. Percentages in brackets mark improvement in correlation to the prior testing. As already mentioned, number of Hindi and Sanskrit students does not allow any statistically valid conclusions. The first year Hindi students (marked with * in Table 3) show higher mean score in pre-usage test than other Hindi students, whereas with Sanskrit it is opposite. One could speculate that it shows that first year students are more interested in living language, whereas at later stage their understanding of a language like Sanskrit becomes more consolidated, hence their pre-usage results are better. In general in all four languages, the positive surge in results after pre-usage test could point out to psychological effect test had on students, i.e. their motivation in having better results. Similarly, the second test mean score could point out to the decreased level of motivation or, on the other hand, to students experiencing more challenges in their further study and overview of topics.

Language	Pre-usage testing (Sep. 2015)		First testing (Feb. 2016)		Second testing (Jun. 2016)		Average score improvement
	Mean score	No. of students	Mean score	No. of students	Mean score	No. of students	
Japanese	29%	48	50% (+21%)	30	61% (+11%)	17	**17,89%**
Korean	63%	12	*n/a (not enough students stayed in the course)*[5]				
Hindi	20,42% (30%)*[6]	14 (+9)*	42,1% (+21,68 %)	6	53,28% (11,18 %)	4	**16,43%**
Sanskrit	44,27% (32%)*	11 (+16)*	51,11 % (6.84 %)	6	60,66% (9,55%)	5	**8,19%**

Table : Testing results for MemAzija users and their general improvement

[5] If the number of the Korean language students who stay in the classes after one semester reaches a statistically significant number, the research will be repeated. In this research, it regrettably had to be omitted because only 3 students stayed in the class.

[6] First year students (marked with *), tested as a reference for non-first year students learning Hindi and Sanskrit. First year students did not come forward to participate in first and second testing.

4.3 Correlation between Memrise usage and improvement in language performance: a case of Japanese

Additionally, we have specifically measured the activity of students of Japanese, since they were most active and most numerous. This is done in order to determine the connection between the usage of e-learning materials and language performance more precisely since the number of Asian language students in Croatia is not high enough to perform specific statistical analyses.

For this part of the research, 53 participants took part during the two semesters of research and study. Out of 53, 21 did not use Memrise frequently (category „analog"), 22 have used it regularly (category „digital"), and 10 have used it but not as diligently (category „mixed"). This was measured through Memrise activity scores: „analog" was <500.000, „mixed" was between 500.000 and 1.000.000, and „digital" counted the users who have scored more than 1.000.000 points on materials issued by MemAzija on the Memrise. All participants were invited to regular testings, and their results were compared according to the category. It should be noted that the „mixed" category also signified that a participant has also used Memrise a lot, but has taken a break at some point, so they are not as consistent in both semesters. Nevertheless, since they have certainly used the materials a lot, they are grouped together with the category „digital" for the overall results, in opposition to the category „analog".

Table 4 displays the results and average attendance rate for all groups. We can recognize two patterns: those who are not using Memrise a lot are not participating in the tests that much either; and those who are using Memrise more display better results in tests, having progressed 5,76% more after first, and 10,20% more after second testing, compared to the first group, i.e. showing mean progress of 11,6% for non-regular users and around 20% for users of Memrise.

Category	N	Average scores			Average change in scores		
		Pre-usage	After 1 semester	After 2 semesters	After 1 semester	After 2 semesters	Mean
Analog (non-frequent Memrise users)	21	28,7%	40,9%	51,9%	18,9%	4,3%	11,6%
Attended test:		18/21	6/21	4/21			
Mixed (used Memrise but with breaks)	10	29,2%	53,2%	61,0%	17,4%	22,9%	20,1%
Attended test:		10/10	7/10	2/10			

Category	N	Average scores			Average change in scores		
		Pre-usage	After 1 semester	After 2 semesters	After 1 semester	After 2 semesters	Mean
Digital (used Memrise consistently)	22	27,6%	52,0%	63,1%	27,8%	12,9%	20,3%
Attended test:		20/22	17/22	11/22			
Overall	53	28,1%	50,39%	60,21%	**+23,79%**	**+11,99%**	**+17,89%**
Attended test:		48/53	30/53	17/53			

Table 4: Testing results and average progress compared to various activity categories on Memrise

These scores point us to the existence of correlation between consistent and systematic learning using Memrise and improvement in standardized test scores. It is interesting to note that Memrise does not significantly lose its effect on a learner even if it is not used extremely consistently, as the users in the „mixed" category have had similar overall progress to those who have had a bigger score, i.e. have used Memrise more. However, we can conclude that using Memrise to study likely improves learner's performance on a Japanese standardized test over time. Unless we could perform complete surveillance over students' habits, it is impossible to completely discern all the factors contributing to their improvement. Although we could speculate that, for example, digital group students are more prone to technology, however, it is not possible to determine the influence of such factors in this research. For a possible topic of future research, we propose using the system consistently on a larger number of students to prove its usefulness more statistically significant. Lastly, vocabulary e-learning is certainly not sufficient for overall language proficiency, but we are inclined to conclude that it is an important addition and should be welcomed in language classrooms.

5 Conclusion

The conclusion that can be drawn from presented data can be summed up as following: students are interested in adding new learning materials to the already existing one. They have positive attitudes towards the use of technology. However, the data suggests that the consistency of students' usage of new materials mediated via technology is not ensured by the mere existence of such materials. Whether this should lead us to conclusion that e-learning is beneficial in case of any language learning is beyond the scope of this paper. It is, however, a question, that is interesting and surely important for science to answer. Hence, the general conclusion can be that further research in students' motivation, teachers' input and reflection on various aspects of created materials in language classrooms is very much in need, if not already overdue.

References

Anki – powerful, intelligent flashcards. (2016). *Ankisrs.net*. Retrieved 11 September 2016, from http://ankisrs.net/.

Ayres, R. (2002). Learner attitudes toward the use of CALL. *Computer Assisted Language learning 15*(3), 241–249.

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

Creative Commons. (2016). *Creative Commons*. Retrieved 11 September 2016, from https://creativecommons.org/.

De la Rouviere, J. (2013). *Chinese radicals in spaced repetition systems: a pilot study on the acquisition of Chinese characters by students learning Chinese as a foreign language*. M. Phil dissertation, Stellenbosch University.

Golonka, E.M., Bowles, A.R., Frank, V.M., Richardson, D.L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, *27*, 70–105.

Hanson-Smith, E. (2002). Computer-assisted language learning. In R. Carter & D. Nunan (Eds.), *The Cambridge guide to teaching English to speakers of other languages.* (pp. 107–113). UK: Cambridge University Press.

Janjić, M., Librenjak, S., & Kocijan, K. (2016). Croatian Students` Attitudes Towards Technology Usage in Teaching Asian Languages – a Field Research. *Mipro proceedings 2016.* Retrieved 11 September 2016, from http://docs.mipro-proceedings.com/ce/ce_34_3870.pdf.

JLPT Japanese-Language Proficiency Test. (2016). *Jlpt.jp*. Retrieved 11 September 2016, from http://www.jlpt.jp/e/index.cgi.

Kuo, M. (2008). *Learner to teacher: EFL student teachers' perceptions on internet-assisted language learning and teaching*. Online Submission. Education Resources Information Center.

Learning tools & flashcards, for free | Quizlet. (2016). *Quizlet.com*. Retrieved 11 September 2016, from https://quizlet.com/.

Mashhadizadeh D., & Rezvani, E. (2015). Iranian EFL learners' attitude towards the use of WBLL approach in writing. *International Journal of Research Studies in Language Learning* 5.

Materijali | MemAzija. (2016). *Memazija.ffzg.unizg.hr.* Retrieved 11 September 2016, from http://memazija.ffzg.unizg.hr/index.php/ankete/materijali/.

McLaren, A. E., & Bettinson, M. (2016). Digital Tools for Chinese Character Acquisition and Their Impact on Student Motivation. In R. Moloney & H. L. Xu (Eds.), *Exploring Innovative Pedagogy in the Teachingand Learning of Chinese as a Foreign Language* (pp. 235–251). Multilingual Education 15, DOI 10.1007/978-981-287-772-7_13.

Memrise – Learn something new every day. (2016). *Memrise*. Retrieved 11 September 2016, from http://www.memrise.com/.

Rahimi M., & Yadollahi S. (2011). Foreign language learning attitude as a predictor of attitudes towards computer-assisted language learning. *Procedia Computer Science*, 3, 167–174.

Schmidt, R. & Watanabe, Y. (2001). Motivation, strategy use and pedagogical preferences in foreign language learning. In Z. Dörnyei & R. Schmidt (Eds.), *Motivation and second*

language acquisition. Honolulu: University of Hawaii Second Language Teaching and Curriculum Center.

Simkova, M., Tomaskova, H., & Nemcova, Z. (2012). Mobile education in tools. *Procedia – Social and Behavioral Sciences*, 47, 10–13.

Skritter – Learn to Write Chinese and Japanese Characters. (2016). *Skritter.com*. Retrieved 11 September 2016, from https://www.skritter.com/.

Tam, V., & Huang, C. (2012). An intelligent e-learning software for learning to write correct Chinese characters on mobile devices. *Interactive Technology and Smart Education*, *9*(4), 191–203.

The official website for comparing UK higher education course data – Unistats. (2016). *Unistats.direct.gov.uk*. Retrieved 11 September 2016, from https://unistats.direct.gov.uk/ when entered data for Japanese course (direct link not available).

TOPIK 한국어능력시험 (2016). *Topik.go.kr*. Retrieved 11 September 2016, from http://www.topik.go.kr/usr/lang/index.do?home_seq=221.

Yushau, B. (2006). The Effects of Blended Elearning on Mathematics and Computer Attitudesi. *Pre-Calculus Algebra, The Montana Mathematics Enthusiast*, *3*(2), 176–183.

Warschauer, M., & Healey, D. (1998). Computers and language learning: An overview. *Language Teaching*, 31, 57–71.

Wiebe, G., & Kabata, K. (2010). Students' and instructors' attitudes toward the use of CALL in foreign language teaching and learning. *Computer Assisted Language LearningI, 23*(3), 221–234.

TECHNICAL NOTES

SEMI-SEMANTIC ANNOTATION: A GUIDELINE FOR THE URDU.KON-TB TREEBANK POS ANNOTATION

Qaiser ABBAS
University of Sargodha, Pakistan
qaiser.abbas@uos.edu.pk

Abstract

This work elaborates the semi-semantic part of speech annotation guidelines for the URDU.KON-TB treebank: an annotated corpus. A hierarchical annotation scheme was designed to label the part of speech and then applied on the corpus. This raw corpus was collected from the Urdu Wikipedia and the Jang newspaper and then annotated with the proposed semi-semantic part of speech labels. The corpus contains text of local & international news, social stories, sports, culture, finance, religion, traveling, etc. This exercise finally contributed a part of speech annotation to the URDU.KON-TB treebank. Twenty-two main part of speech categories are divided into subcategories, which conclude the morphological, and semantical information encoded in it. This article reports the annotation guidelines in major; however, it also briefs the development of the URDU.KON-TB treebank, which includes the raw corpus collection, designing & employment of annotation scheme and finally, its statistical evaluation and results. The guidelines presented will be useful for linguistic community to annotate sentences not only for the national language Urdu but for the other indigenous languages like Punjab, Sindhi, Pashto, etc. as well.

Keywords: semi-semantic part of speech; rich information; deep learning; parsing aid; linguistically motivated annotation; humanistic annotation

Povzetek

Rezultat tega dela so navodila za označevanje polsemantičnih besednih vrst v drevesnici URDU.KON-TB. Hierarhična označevalna shema je bila oblikovana z namenom, da razvrsti besedne vrste in jih kot take uporabi na korpusih. Tokratni koprus, ki je sestavljen iz strani Urdu Wikipedie in časopisa Jang, je bil označen s predlaganimi polsemantičnimi besednimi vrstami. Korpus vsebuje tekste lokalnih in mednarodnih novic, zgodbe s socialno temo, šport, kulturo, finance, vero, potovanja in druge teme. Uspešen poskusa označevanja je nadgradil drevesnico URDU.KON-TB. Dvaindvajset osnovnih besednih vrst je razdeljenih v podkategorije z morfološkimi in semantičnimi informacijami. Članek podaja jasne osnovne smernice označevanja. Hkrati ponuja kratek pregled razvoja drevesnice URDU.KON-TB, ki vsebuje zbirke surovih korpusov, oblikovanje in uporabo shem za označevanje ter nenazadnje tudi statistično oceno in rezultate. Predlagana navodila za označevanje so namenjena jezikovnim skupnostim, ki označujejo stavke tako v državnem jeziku Urdu kot tudi v drugih jezikih, kot so *Punjab*, *Sindhi* in drugih.

Ključne besede: polsemantična besedna vrsta; številne informacije; globoko učenje; pomoč pri razvrščanju; jezikoslovno utemeljeno označevanje; humanistično označevanje

Acta Linguistica Asiatica, 6(2), 2016.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.6.2.97-134

1 Introduction

A treebank or a parsed corpus is a text corpus of sentences annotated with a syntactic structure. Today, many natural language processing (NLP) and machine learning (ML) applications rely on treebanks. Treebanks are heavily used in corpus linguistics for investigating syntactic phenomena or in computational linguistics for training or testing parsers. The sentences in the treebank should be annotated according to a devised annotation scheme as presented in Figure 1 to 4 in our case. Annotation schemes can include the labeling to represent morphological forms, word classes, syntactic structures, semantics, grammatical arguments, co-references, etc. So, the corpus annotation is simply the addition of interpretative linguistic information to a corpus (Leech, 2005).

Annotation scheme that was used to develop the URDU.KON-TB treebank (Abbas, 2012, 2014a, 2014b) for the South Asian language Urdu is presented next with complete guidelines in Section 2. This annotation is actually encoded with the morphology, POS, syntactical and functional information including the handling of displaced constituents, empty categories, antecedents and anaphors, etc., but here only the case of semi-semantic part of speech (SSP) is discussed concisely. Such development of an annotation scheme is the fundamental step to build a treebank, for which the computational linguists then devise the annotation guidelines (Section 2), which is a compulsory part to build, and without which the annotation scheme has no worth at all. Annotation structure for the development of the URDU.KON-TB treebank has the combination of the PS (Phrase Structure) and the HDS (Hyper Dependency Structure) annotation detailed in Section 3.2. Annotation issues emerged during the development (Abbas, 2012) have been corrected in (Abbas, 2014a & 2014b) and the annotation guidelines presented in Section 2 is the most updated version. The corpus containing 1400 sentences[1] (discussed in Section 3.1) for the development of the URDU.KON-TB treebank was collected from the Urdu Wikipedia[2] and the Urdu Jang newspaper.[3]

[1] The size has been augmented recently to 2000 sentences by a student in her master's thesis "Annexing Treebank and the Urdu Parser"

[2] https://ur.wikipedia.org/wiki/اول_صفحۂ

[3] http://jang.com.pk/index.html

ADJ (Adjective)	.REL (Relative)	.ROOT (Root)
.DEG (Degree)	.DEM (Demons...)	.SUBTV (Subjunctive)
.ECO (Echo)	.PERS (Personal)	.PAST (Past)
.MNR (Manner)	POSTP (Postposition)	.PRES (Present)
.SPT (Spatial)	.CMP (Comparative)	.LIGHTV (Light Verb)
.TMP (Temporal)	.MNR (Manner)	.IMPERF (Imperfective)
ADV (Adverb)	.POSS (Possessive)	.INF (Infinite)
.DEG (Degree)	.REP (Repeat)	.PERF (Perfective)
.MNR (Manner)	.SPT (Spatial)	.ROOT (Root)
.NEG (Negative)	.TMP (Temporal)	.SUBTV (Subjunctive)
.SPT (Spatial)	PRAY (Pray)	.MOD (Modal)
.TMP (Temporal)	PREP (Preposition)	.IMPERF (Imperfective)
.REL (Relative)	.MNR (Manner)	.PERF (Perfective)
C (Conjunction)	.SPT (Spatial)	.SUBTV (Subjunctive)
.CAUS (Causative)	.TMP (Temporal)	.PERF (Perfective)
.CONS (Concessive)	PT (Particle)	.REP (Repeat)
.CORD (Coordinative)	.ADJ (Adjective)	.ROOT (Root)
.CORR (Co-relative)	.EMP (Emphatic)	.REP (Repeat)
.SBORD (Subordinating)	.INTF (Intensifier)	.SUBTV (Subjunctive)
.COND (Conditional)	.RESULT (Result)	.PAST (Past)
CM (Case Marker)	Q (Quantifier)	.PRES (Present)
DATE (Date)	.ADJ (Adjective)	VALA (Vala)
.D (Day)	.CARD (Cardinal)	VAUX (Verb Auxiliary)
.M (Month)	.FRAC (Fractional)	.IMPERF (Imperfective)
.Y (Year)	.ORD (Ordinal)	.INF (Infinite)
HADEES (Hadees)	QW (Question Word)	.MOD (Modal)
INT (Interjection)	.REP (Repeat)	.IMPERF (Imperfective)
M (Marker)	.TMP (Temporal)	.PERF (Perfective)
.P (Phrase)	.SPT (Spatial)	.SUBTV (Subjunctive)
.S (Sentence)	.MNR (Manner)	.PASS (Passive)
N (Noun)	SYM (Symbol)	.IMPERF (Imperfective)
.ADJ (Adjective)	TTL (Title)	.INF (Infinite)
.MNR (Manner)	.REG (Regard)	.PERF (Perfective)
.REP (Repeat)	U (Unit)	.ROOT (Root)
.PROP (Proper)	V (Verb)	.SUBTV (Subjunctive)
.SPT (Spatial)	.COP (Copula)	.PERF (Perfective)
.TMP (Temporal)	.IMPERF (Imperfective)	.PROG (Progressive)
.REP (Repeat)	.PERF (Perfective)	.ROOT (Root)
.SPT (Spatial)	.ROOT (Root)	.SUBTV (Subjunctive)
.REP (Repeat)	.SUBTV (Subjunctive)	.FUTR (Future)
.TMP (Temporal)	.PAST (Past)	.PAST (Past)
.REP (Repeat)	.PRES (Present)	.PRES (Present)
P (Pronoun)	.IMPERF (Imperfective)	
.DEM (Demonstrative)	.REP (Repeat)	
.INDF (Indefinite)	.INF (Infinite)	
.PERS (Personal)	.LIGHT (Light)	
.POSS (Possessive)	.IMPERF (Impe...)	
.REF (Reflexive)	.INF (Infinite)	
.REP (Repeat)	.PERF (Perfective)	
.REF (Reflexive)	.PROG (Progressive)	

Figure 1: A detailed version of the SSP tagset for the URDU.KON-TB treebank

The reliability of the treebank annotation or the annotation guidelines can be measured by calculating the agreement or the homogeneity among the annotators of the treebank. The reliability evaluation is a complex task for the treebank that contains rich information, but it is an essential part to play for the production of a quality treebank, so that the annotation can be readable. The annotation evaluation (Abbas,

2014a & 2014b) resolved most of our annotation issues except few. The guidelines of the URDU.KON-TB treebank are evaluated using a statistical measure known as the Krippendorf's α coefficient (Krippendorf, 2004). This can be used to evaluate the inter-annotator agreement (IAA). Randomly selected one hundred (100) sentences from the URDU.KON-TB treebank were given to five trained annotators for annotation. The annotated sentences then evaluated using the Krippendorf's α co-efficient. The α values of the IAA obtained for the part of speech (SSP) annotation is 0.964. The annotation guidelines were revised during and after this annotation evaluation. A little detailed presentation of evaluation is given in Section 4.

Section 2 describes the up to date annotation guidelines revised after the annotation evaluation (Abbas, 2014a & 2014b). Guidelines regarding the SSP annotation are detailed here in this article for easiness and simplicity along with their respective examples. To remain on track, the annotation tags are discussed according to the order of the SSP tags given in Figure 1. The discussion of the annotation guidelines is kept concise.

ADJ (Adjective)	PRAY (Specific statements of prayers)
ADV (Adverb)	PREP (Preposition)
C (Conjunction)	PT (Particle)
CM (Case marker)	Q (Quantifier)
DATE (Date)	QW (Question word)
HADEES (Narration of prophets deeds)	SYM (Symbol)
INT (Interjection)	TTL (Title)
M (Marker)	U (Unit)
N (Noun)	V (Verb)
P (Pronoun)	VALA (Special Word Vala)
POSTP (Postposition)	VAUX (Verb auxiliary)

Figure 2: The main POS-tag categories for the URDU.KON-TB treebank

PROG (Progressive form)	PROG (Progressive form)
PASS (Passive form)	PASS (Passive form)
FUTR (Future tense)	FUTR (Future tense)
PAST (Past tense)	PAST (Past tense)
PRES (Present tense)	PRES (Present tense)

Figure 3: Morphological tag set to annotate subcategories of verbs and auxiliaries

Semantic labels	
CMP (Comparative)	POSS (Possessive)
INST (Instrumental)	SPT (Spatial)
MNR (Manner)	TMP (Temporal)

Figure 4: Functional tag set for the URDU.KON-TB treebank

2 Semi-Semantic POS (SSP) Annotation

The term semi-semantic (partly or partially semantic) is used with the POS because some tags are encoded with semantics but not all e.g. N.SPT (a spatial noun) tag for a word *house*, ADJ.TMP (a temporal adjective) tag for a word *previous* in *previous year*, etc. There are twenty two (22) main POS tag categories, which are displayed in Figure 2. The description of the tags is given in the respective cells of the figure. These main categories are further divided into morphological and semantical subcategories according to the Figures 3 and 4, respectively. The final and detailed version of the SSP tag set is given in Figure 1. The dot "." is used to add the morphological or semantical features to the main category e.g. in V.PERF, a verb V is the main POS category like nouns, adjectives, etc., which has a perfective PERF morphology. The description of each category is as follows. It is to be noted that the Urdu script is written from right to left in coming examples. The row beneath the Urdu script is the transliteration of the sentence as proposed in Malik et al. (2010). Similarly, the row beneath the transliteration of the sentence contains the translated-word/POS-tag pair according to the SSP tag set given in Figure 1. At the end in examples next, a complete English translation of the Urdu sentence is presented. The complete guideline is going to be presented next, however, its employment procedure on the raw corpus to form the URDU.KON-TB treebank can be seen in Section 3.3. It is advised to skip this Section 2 for later reading and go to Section 3 to understand the flow of the article as this section concludes the deep design of the annotation guidelines.

2.1 Adjectives

Adjectives are used to modify a noun or pronoun (Aarts et al., 2014; Matthews, 2007; Miller et al., 1990; Stevenson, 2010). The first main category in Figure 1 is about ADJ (Adjective), which is divided into further five sub categories of tags included DEG (Degree), ECO (Echo), MNR (Manner), SPT (Spatial) and TMP (Temporal). The relevant POS annotations are provided in examples 1. Example 1(a) is the case of main POS category ADJ of adjective. There are some words like *tar* 'more' and *tarIn* 'most', which truly act as a degree adjective and not as degree adverb but there are some words which can play the role of a degree adjective or a degree adverb e.g. *ziyAdah* 'more/much', *bohat* 'more/much', etc, (Schmidt, 2013). Example 1(b) is the case of degree adjective ADJ.DEG. Example 1(c) is the case of reduplication[4] (Abbi, 1992; Boegel et al., 2007). Reduplication has two versions. First is discussed in a footnote below, while the other is the repetition of the original word e.g. *sAtH sAtH* 'with/along-with'. These two versions are named as echo reduplication and full word reduplication by Boegel et al. (2007), which are refurbished in our annotation as ECO (echo

[4] In Urdu like other South Asian languages, the reduplication of a content word is frequent. Its effect is only to strengthen the proceeding word or to expand the specific idea of a proceeding word into a general form e.g. *kAm THIk-THAk karnA* 'Do the work right' or *kOI kapRE-vapRE dE dO* 'Give me the clothes or something like those'

reduplication) and REP (full word reduplication/repetition) respectively. The echo words normally start with the letters *S* or *v* or *m*. The next examples from 1(d) to 1(f) are the cases of adjectives, which have the meaning of MNR, TMP and a SPT respectively. The addition of this MNR, TMP and SPT after the POS tag ADJ represents the semantics.

(1) a. اچھا لڑکا

acHA laRkA
good/ADJ boy/N
'Good Boy'

b. اہم ترین شخصیت

aham tarIn Saxs2iat
important/ADJ most/ADJ.DEG personality/N
'Most important personality'

c. برا ورا کام

burA vurA kAm
ugly/ADJ ADJ.ECO work/N
'Ugly work'

d. جابرانہ حکومت

jaberaanah hakUmat
cruel/ADJ.MNR government/N
'Cruel Government'

e. گزشتہ سال

guzaStah sAl
previous/ADJ.TMP year/N
'Preveious Year'

f. ملتانی کھسہ

mUltAnI kHUsah
multani/ADJ.SPT shoe/N
'Multani shoe'

2.2 Adverbs

Adverbs can modify verbs, adjectives or other adverbs. They can also modify phrases, clauses and sentences (Aarts et al., 2014; Matthews, 2007; Miller et al., 1990; Stevenson, 2010). Adverbs are mostly used as a qualifier of the verbs but they can also be used independently. They are subcategorized into six forms presented in Figure 1. The annotations are given in example 2. The main category of adverbs ADV is annotated in 2(a), which is further divided into five subcategories DEG (degree), MNR (manner),

NEG (negative), SPT (spatial) and TMP (temporal). The final TMP has another subcategory REL for relative temporal adverb. In 2(b), an adverb *bohat* 'very' is used before an adjective *acHI* 'good' and it is highlighting the adjective at a certain degree, hence annotated as ADV.DEG. In 2(c), *biltartIb* 'respectively' behaves as an adverb and advocates a manner of order as ADV.MNR. The word *nah* 'not' is a negative adverb negating the action in 2(d) and it is annotated with ADV.NEG relatively. A word *sAmnE* 'front/before' is a spatial adverb and annotated as ADV.SPT in 2(e). The case of temporal adverb is displayed in 2(f), where a word *ab* 'now' is annotated as ADV.TMP. This temporal adverb is divided into another hierarchy named relative-temporal adverb, which can be seen in the last example 2(g). A word *jab* 'when' is given a POS tag as ADV.TMP.REL as follows.

(2) a. تقریبا ساری دنیا میں

taqrIban	sArI	dunIyA	mEN
almost/ADV	whole/Q	world/N.SPT	in/CM

'Almost in the whole world'

b. بہت اچھی لڑکی

bohat	acHI	laRkI
very/ADV.DEG	good/ADJ	girl/N

'Very good girl'

c. تعداد بالترتیب ۵ اور ۶ تھی

te2dAd	biltartIb	5	aor
quantity/N	respectively/ADV.MNR	5/Q.CARD	and/C.CORD
6	tHI		
6/Q.CARD	was/V.COP.PAST		

'The quantity was 5 and 6 respectively'

d. عمارت مکمل نہ ہو سکی

e2emArat	mukammal	nah	hO	sakI
building/N	complete/ADJ	not/ADV.NEG	be/V.LIGHT.ROOT	could/V.MOD.PERF

'The building could not be completed'

e. تفصیلات سامنے آئیں گی

tafs2IlAt	sAmnE	AyIN	gI
details/N	front/ADV.SPT	come/V.SUBTV	will/VAUX.FUTR

'The details will come out'

f. اب دیکھنا یہ ہے

ab	dEkHnA	yE	hE
now/ADV.TMP	to-see/V.INF	this/P.PERS	be/V.COP.PRES

'Now, this is to be seen'

g. جب یہاں کھیت ہوتے تھے

jab	yahAN	kHEt	
when/ADV.TMP.REL	here/ADV.SPT	crop-field/N.SPT	
hotE	tHE		
be/V.IMPERF	was/VAUX.PAST		

'When, there were crop fields here'

2.3 Conjunctions

Conjunctions are used to connect words, phrases, clauses or sentences (Aarts et al., 2014; Matthews, 2007; Miller et al., 1990; Stevenson, 2010). The main category of conjunction C is divided into five subcategories e.g., CAUS (causative), CONS (concessive), CORD (coordinative), CORR (correlative) and SBORD (subordinating). The last subcategory has another division of COND to represent conditional subordinate conjunction. The annotation of all divisions is presented in example 3. Words like *cUnkEh* 'since, because', *cUnAcEh* 'so, therefore', *kIUnkEh* 'because' are candidates for a causative conjunction in a clause. An example of causative conjunction is depicted in Example 3(a). The POS annotation examples of CONS and CORD are given in 3(b) and 3(c), respectively. The word *agarcEh* 'although' is acting as a concessive conjunction in the beginning of sentence in 3(b), while the other word *aor* 'and' is a coordinating conjunction in 3(c). The word *nah* 'not/neither' as a correlative conjunction is presented in 3(d), in which it is annotated with C.CORR tag. The subordinating conjunction C.SBORD is annotated in 3(e) for a word *kEh* 'that'. The C.SBORD is divided into another subcategory proposed as COND for conditional subordinating conjunction. Its annotation for a word *agar* 'if' is presented in 3(f).

(3) a.

SAyad	voh	akElA	tHA	kIUnkEh
perhaps/ADV	he/P.PERS	alone/ADJ	be/V.COP.PAST	because/C.CAUS
kHAnA	hOtEl	sE	kHAtA	tHA
meal/N	hotel/N.SPT	from,in/CM	eat/V.IMPERF	be/VAUX.PAST

'Perhaps, he was alone because he used to eat his meals in a hotel'

b.

agarcEh	AdmI	kam	tHE	magar
men/N	less/ADJ	were/V.COP.PAST	although/C.CONS	but/C.CORD
voh	pHir	bHI	jIt	gayE
they/P.PERS	then/ADV	too/PT.INTF	won/V.ROOT	V.LIGHTV.PERF

'Although the men were less but they had won either'

c.

te2dAd	biltartIb	5	aor
quantity/N	respectively/ADV.MNR	5/Q.CARD	and/C.CORD
6	tHI		
6/Q.CARD	was/V.COP.PAST		

'The quantity was 5 and 6 respectively'

d. nah tO tUm kHElE nah
neither/C.CORR PT.EMP you/P.PERS played/V.PERF nor/C.CORR
hI kHElnE diyA
PT.INTF play/V.INF gave/V.LIGHTV.PERF
'Neither you played yourself nor you allowed to play others'

e. nabI nE farmAyA kEh a2lI a2ilm
prophet/N CM said/V.PERF that/C.SBORD Ali/N.PROP knowledge/N
kA darvAzah hEN
of/CM door/N.SPT is/V.COP.PRES
'The prophet stated that Ali is the door to knowledge'

f. agar yEh mErA mAl hOtA
if/C.SBORD.COND it/P.PERS my/P.POSS property/N be/V.IMPERF
tO mEN xarc kartA
then/PT.RESULT I/P.PERS spend/V.ROOT do/V.LIGHTV.IMPERF
'If it would be my property then I will spend it'

2.4 Case markers

Case markers (CM) distinguish the grammatical functions of words, phrases, clauses, or sentences (Aarts et al., 2014; Matthews, 2007; Miller et al., 1990; Stevenson, 2010). Urdu case markers are syntactic clitics (Butt and Sadler, 2003) and divided into different forms by Butt and King (2004) e.g., ergative, accusative, dative, possessive, etc. All Urdu case markers are annotated with a simple CM tag at POS level. Four annotated examples can be seen in 3(a), 3(e) and 2(a) for instrumental case marker *sE* 'from', ergative case marker *nE*, possessive case marker *kA/kI/kE* 'of' and spatial case marker *mEN/par/tak* 'in/on/at'. The different forms of case markers play an important role in identification of argument structure like subject, object, etc. The effect of different forms and their related argument structure is discussed in Abbas (2014b).

2.5 Date

The DATE tag is used to represent dates of a month e.g. 14, 2, 31, etc. This tag is divided into three subcategories, which includes DATE.D, DATE.M and DATE.Y. Annotated examples can be seen in 4. The days of a week, month name and a year number are represented by DATE.D, DATE.M and DATE.Y, respectively.

(4) aetvAr 16 mayI 2004 kO
sunday/DATE.D 16/DATE May/DATE.M 2004/DATE.Y on/CM
'On Sunday, 16 May 2004'

2.6 Hadees

The Hadees is a report of deeds and saying of the prophet Muhammad (PBUH). These are tagged as HADEES in the URDU.KON-TB treebank. The Ahadees (plural of Hadees) in Arabic script in Urdu text are tagged only with this tag HADEES. The translated form of Ahadees in Urdu is annotated in a normal way. An example is depicted in 5 as follows. The Hadees with double quotes in the following sentence is in Arabic and hence tagged as HADEES.

(5) rasUl nE kahA “ h2UsyEno-minnI-va-anA-min-al-h2UsyEn ”
prophet/N CM said/V.PERF M.P HADEES M.P
'The prophet said, “Hussain is from me and I am from Hussain”'

2.7 Interjections

Interjections are the words or phrases used to exclaim, protest or command in a sentence. These are annotated with a tag INT. The example can be seen in 6 as follows.

(6) oE kHAnA kHAO
OE/INT food/N eat/V.SUBTV
'OE! eat the food'

2.8 Markers

The markers are used to identify the boundary of phrases, clauses, or sentences as marked by punctuation. The markers are divided into two subcategories e.g. phrase markers (M.P) and sentence markers (M.S). The punctuation within the sentence like single quotes, double quotes, colon, comma, etc., are annotated with M.P, however the boundary of the sentence like full stop and question mark is annotated with M.S. The annotated example can be seen in 7 as follows. The comma and period is marked by M.P and M.S respectively.

(7) in mEN bHakar , layIah
these/P.PERS in/CM Bhakkar/N.PROP.SPT comma/M.P Layyah/N.PROP.SPT
aOr lOdHrAN SAmil hEN .
and/C.CORD Lodhran/N.PROP.SPT include/N be/V.LIGHT.PRES full-stop/M.S
'Bhakkar, Layyah and Lodhran are included in these'

2.9 Nouns

The main noun tag N is divided into six subcategories, which includes adjectival noun (N.ADJ), noun having a manner (N.MNR), proper noun (N.PROP), repeated noun

(N.REP)[5], spatial noun (N.SPT) and temporal noun (N.TMP). The words *cHotE* 'younger' and *baRE* 'elder' are representing people having some property of young age and old age in 8(a), hence both are annotated with N.ADJ. In 8(b), the word *t2arah2* 'way, like, type' is first annotated with N.MNR but when the same word is repeated next then it gives the meaning of 'different types' and its repetition is annotated simply with N.MNR.REP. In 8(c), a subcategory N.PROP is annotated for a person name *marIyam* 'Maryam'. This subcategory is divided into two subcategories spatial and temporal, which are annotated as N.PROP.SPT and N.PROP.TMP for *panjAb* 'Punjab' and *a2Id-ul-fit2r* 'Eid festival', respectively. A common noun N is annotated in 8(b) for a word *taklIfEN* 'hardships'. There are some special common nouns, which can be repeated e.g. *kOrI kOrI* 'single penny'. When some noun is usually repeated then N.REP tag is used. So, this .REP along with the respective POS tag can be used to represent the presence of a repeated word. The annotation of N.SPT and N.TMP can be seen in 8(c) for *iz3lAa2* 'districts' and *din* 'day'. In both the subcategories, the repetition is possible for which the addition of REP with dot "." can be used accordingly.

(8) a.

cHOtE	baRE	sab	xUS	
younger/N.ADJ	elder/N.ADJ	all/Q.ADJ	happy/ADJ	
hOtE		hEN		
become/V.COP.IMPERF		be/VAUX.PRES		

'Younger and elder all become happy'

b.

UnhEN	t2arah2	t2arah2	kI	taklIfEN
they/P.PERS	type/N.MNR	type/N.MNR.REP	of/CM	hardships/N
dI	jAnE	lagIN		
give/V.PERF	go/VAUX.PASS.INF	start/VAUX.SUBTV		

'They were given hardships of different types'

c.

marIyam	panjAb	kE	ba2z	iz3lAa2
Maryam/N.PROP	Punjab/N.PROP.SPT	of/CM	some/Q	districts/N.SPT
mEN	a2Id-ul-fit2r	kE	din	gayI
into/CM	Eid-ul-Fitr/N.PROP.TMP	of/CM	day/N.TMP	went/V.PERF

'Maryam went into some districts of Punjab on the day of Eid-ul-Fitr'

2.10 Pronouns

The main category of pronoun P is divided into six subcategories P.DEM (demonstrative pronoun), P.INDF (indefinite pronoun), P.PERS (personal pronoun), P.POSS (possessive pronoun), P.REF (reflexive pronoun) and P.REL (relative pronoun). The first two subcategories P.DEM and P.INDF are annotated in 9(a) for words *yeh* 'this' and *kOI* 'any' respectively. The difference between P.PERS and P.DEM is this that when P.PERS refers to some person, place or thing, then this P.PERS behaves as a P.DEM like

[5] It lies in the category of full word reduplication as discussed in Section 2.1

in 9(a). The 3rd and 4th category P.PERS and P.POSS are annotated in 9(b) for words *mEN* 'I' and *tumhArA* 'your' respectively. P.POSS is further divided into P.POSS.REF, which is annotated for a word *apnA* 'own' in the same sentence. The repeated subcategory can be annotated after addition of .REP at the end. The fifth and sixth subcategory P.REF and P.REL are annotated in 9(c) for words *Apas* 'themselves' and *jO* 'which' respectively. The subcategory P.REL is further divided into P.REL.DEM and P.REL.PERS. These are annotated in 9(d) for words *jO* <u>*kUcH*</u> 'what ever' and *jIs* 'who' respectively.

(9) a. yeh meh2kama kOI kAm nahI kartA
this/P.DEM department/N any/P.INDF work/N not/ADV.NEG do/V.IMPERF
'This department does not do any work'

b. mEN tumhArA apnA bHAI hUN
I/P.PERS your/P.POSS own/P.POSS.REF brother/N be/V.COP.SUBTV
'I am your own brother'

c. jO Apas mEN moh2abat kI
which/P.REL themselves/P.REF among/CM love/N of/CM
mIs2Al hE
example/N be/V.COP.PRES
'Which is an example of love among themselves'

d. jIs kO jO kUcH milE
who/P.REL.PERS CM what/P.REL.DEM ever/P.INDF find/V.PERF
UTHA lEnA cAhIE
pick/V.PERF take/V.LIGHTV.INF should/VAUX.MOD.PERF
'Who finds what ever, should pick it up'

2.11 Postpositions

The postpositions are placed after a word to which it is grammatically related e.g. *sAtH* 'with' is a POSTP (postposition) in a postpositional phrase *Us kE sAtH* 'with him'. The postpositions are divided into six subcategories hierarchically as displayed in Figure 1. These include POSTP.CMP (comparative postposition), POSTP.MNR (postposition having a manner)[6], POSTP.POSS (possessive postposition), POSTP.REP (repetitive postposition), POST.SPT (spatial postposition) and POSTP.TMP (temporal post-position). The first two subcategories are annotated in 10(a) for postpositions *sE* 'than' and *t2arah2* 'like' respectively. In 10(b), the third and fourth subcategories are annotated for postpositions *pAs* 'have/has' and *sAtH* 'with' respectively. The last two

[6] The prepositions are divided into basic manner, manner by comparison and manner with a reference point by Saint-Dizier (2008) but I applied only manner in general to all related prepositions and postpositions for Urdu.

subcategories are annotated in 10(c) for postpositions *qarIb* 'near' and *ba2d* 'after' respectively.

(10) a. 25 sE z2yAdah laRkE
25/Q.CARD than/POSTP.CMP more/ADJ.DEG boys/N
aslam kI t2arah2 hEN
Aslam/N.PROP of/CM like/POSTP.MNR be/V.COP.PRES
'More than 25 boys are like Aslam'

b. xUrAk kE sAtH sAtH mErE
food/N of/CM with/POSTP with/POSTP.REP I/P.POSS
pAs pEsE bHI hEN
have/POSTP.POSS money/N also/PT.INTF be/V.COP.PRES
'I have also the money along with the food'

c. us kE qarIb h2amlE kE
him/P.PERS of/CM near/POSTP.SPT attack/N of/CM
ba2d bam pHatA
after/POSTP.TMP bomb/N exploded/V.PERF
'The bomb exploded near him after the attack'

2.12 Pray

The PRAY tag is used to annotate all types of prayers normally used in religious literature after the name of prophets, caliphs, and the righteous religious personalities e.g. the *alEh salAm* 'peace be upon him' is annotated with PRAY after the name of Jesus in 11(a) along with the other example as follows.

(11) a. h2az3rat a2IsA alEh-salAm allah
h2az3rat/TTL.REG Jesus/N.PROP AS/PRAY Allah/N.PROP
kE Ek a2z4Im nabI hEN
of/CM a/Q.CARD great/ADJ prophet/N be/V.COP.PRES
'Jesus (peach be upon him) is a great prophet of God'

b. h2az3rat mUhammad s3al-lalAhO-a2laehE-va-AlEhI-vasalam nE
TTL.REG Muhammad/N.PROP SAAWW/PRAY CM
h2az3rat a2lI kO apnA bHAI banAyA
TTL.REG Ali/N.PROP CM his/P.POSS.REF brother/N made/V.PERF
'Muhammad (peach be upon him and his descendant) made Ali his brother'

2.13 Prepositions

The prepositions are placed before a word to which it is grammatically related e.g., *bE* 'without' is a PREP (preposition) in a prepositional phrase *bE mUhAr Sutar* 'a camel without a hook). Prepositions are divided into three subcategories hierarchically as

displayed in Figure 1. These include PREP.MNR (preposition having a manner), PRET.SPT (spatial preposition) and PREP.TMP (temporal preposition). The first two subcategories are annotated in 12(a) for prepositions *bat2Or* 'as' and *andrUnE* 'in' respectively. The last subcategory is annotated in 12(b) for prepositions *dOrAnE* 'during'.

(12) a. us nE bat2Or DrAlvar andrUnE Sehar
he/P.PERS CM as/PREP.MNR driver/N in/PREP.SPT city/N.SPT
nOkrI kI
job/N do/V.PERF
'He did the job as a driver in the city'

b. voh yahAN dOrAnE taftIS
he/P.PERS here/ADV.SPT during/PREP.TMP investigation/N
A giA
come/V.ROOT go/V.LIGHTV.PERF
'He came here during the investigation'

2.14 Particles

The particles can appear after a word. These are divided into four subcategories, which include PT.ADJ (adjectival particles), PT.EMP (emphatic particles), PT.INTF (Intensifying particles) and PT.RESULT (resultant particles). All the subcategories are non-inflected except the PT.ADJ, which appears after adjective, adverb, noun or pronoun and agrees with the qualifier. The first and third subcategories are annotated in 13(a) for the particles *sA* 'like' and *bHI* 'too'. The annotation of PT.EMP is displayed in 13(b) for a word *tO*. The contrastive meaning is understood by default due to usage of PT.EMP in this sentence. In 13(c), the annotation of PT.RESULT is given for a word *tO* 'then'.

(13) a. voh Ek nAxUSgavAr sA bandah
he/P.PERS a/Q.CARD unpleasant/ADJ like/PT.ADJ man/N
bHI hE
too/PT.INTF be/V.COP.PRES
'He is like an unpleasant man too'

b. ab maslah falastIn tO
now/ADV.TMP problem/N Palestine/N.PROP.SPT PT.EMP
h2al hO gA
resolve/N be/V.LIGHT.ROOT will/VAUX.FUTR
'Now, the problem of Palestine will resolve (contrast: “the other problems will not” due to ‘tO’ effect)'

c. bAriS AyI tO mElah nahI
rain/N come/V.PERF then/PT.RESULT festival/N not/ADV.NEG
hO gA
be/V.ROOT will/VAUX.FUTR
'If the rain comes, then the festival will not hold'

2.15 Quantifiers

The quantifiers Q are used to show the amount of something. These are divided into four subcategories, which include Q.ADJ (adjectival quantifier), Q.CARD (cardinal quantifier), Q.FRAC (fractional quantifier) and Q.ORD (ordinal quantifier). In 14(a), the quantifiers *tamAm* 'all/whole', *har* 'every' and *dUsrA* 'second/other' are annotated with Q, Q.ADJ and Q.ORD, respectively. The remaining subcategories of quantifiers Q.CARD and Q.FRAC are annotated in 14(b) for words *Ek* 'one' and *cOtHAI* 'one 4th', respectively.

(14) a. tamAm mumAlik mEN har dUsrA
all/Q countries/N.SPT in/CM every/Q.ADJ second/Q.ORD
Saxs xUS hE
person/N happy/ADJ be/V.COP.PRES
'In all countries, every second person is happy'

b. mujHE Ek cOtHAI raqam dO
me/P.PERS one/Q.CARD fourth/Q.FRAC amount/N give/V.SUBTV
'Give me one fourth amount'

2.16 Questions Words

The question words QW identify a question in a sentence. These are divided into four subcategories, which include QW.REP (repeated question words), QW.TMP (temporal question words), QW.SPT (spatial question words) and QW.MNR (question words having a manner). The main category QW is depicted in 15(a) for a question word *kiyA* 'what'. If any question word is repeated then QW.REP can be used for annotation. The remaining three subcategories QW.TMP, QW.SPT and QW.MNR are annotated in a single sentence 15(b) for related question words *kab* 'when', *kidHar* 'where' and *kEsE* 'how', respectively.

(15) a. tumhArA nAm kiyA hE ?
your/P.POSS name/N what/QW be/V.COP.PRES ?/M.S
'What is your name?'

b. kab , kidHar aOr kEsE
when/QW.TMP ,/M.P where/QW.SPT and/C.CORD how/QW.MNR
jAO gE ?
go/V.SUBTV will/VAUX.FUTR ?/M. S
'When, where and how will you go?'

2.17 Symbols

The symbols SYM include brackets, parentheses, percent symbols, currency symbols, etc. All are dealt within a single category SYM as can be seen as follows.

(16) Gazvah fatah2 [fatah2e Makkah]
Gazvah/N.PROP fatah/N.PROP [/SYM fatahe/N Makkah/N.PROP.SPT]/SYM
: yeh ramz3An mEN hUI
:/M.P it/P.PERS Ramadan/DATE.M in/CM happen/V.PERF
'The Battle Conquest [conquest of Makkah]: It happened in Ramadan (a month name in Islamic Calendar)'

2.18 Titles

The titles are used to show respect or regard to personalities before addressing their names. At present it has only one subcategory TTL.REG (regard titles). Its annotation can be seen as follows.

(17) h2az3rat ImAm h2Ussyn a2lEh-salAm
his-highness/TTL.REG religious-head/N Hussain/N.PROP AS/PRAY
tIsrE ImAm hEN
third/Q.ORD religious-head/N be/V.COP.PRES
'His highness, the religious head, Hussain (PBUH) is the third religious head'

2.19 Units

The Unit U is used to represent different measuring units e.g., meter, liter, bar, grams, etc. The example of an annotation is given for the following sentence.

(18) qEsar nE 70 miliyan tan tEl
Qaiser/N.PROP CM 70/Q.CARD million/Q.CARD ton/U oil/N
darAmad kiyA
import/N do/V.LIGHT.PERF
'Qaiser imported 70 million ton oil.'

2.20 Verbs

The main verb V is divided into 11 subcategories, which are further divided into hierarchical subcategories discussed as follows. The hierarchical division of a special

word the VALA[7] and verb auxiliaries will be discussed in respective Sections 2.21 and 2.22.

2.20.1 Copula Verbs

The copula verb V.COP is used to connect the subject with the subject complement or the predicate link of a sentence (Aarts et al., 2014; Butt, 1995; Matthews, 2007; Miller et al., 1990; Stevenson, 2010). For example a sentence like *The weather is horrible* contains a subject *The weather* and a predicate link as an adjective *horrible*. The predicate of the sentence is the copula verb *is*. The copula verb connects the subject with the predicate link in this sentence. The V.COP (copula verb) is divided into six subcategories hierarchically as V.COP.IMPERF (a copula verb with imperfective morphology), V.COP.PERF (a copula verb with perfective morphology), V.COP.ROOT (a copula verb with root form), V.COP.SUBTV (a copula verb with subjunctive morphology), V.COP.PAST (copula verb with past tense) and V.COP.PRES (copula verb with present tense). The future form of copula verb itself is not possible in Urdu. In future construction, the copula verb 'be/become' always proceeds the future tense auxiliary *gA/gI/gE/gEN* 'shall/will' as can be seen in 19(d). The V.COP.IMPERF is annotated in 19(a) for a copula verb *hOtE* 'be/become' in imperfective form. Its perfective form is given in 19(b). The root form of another copula verb *ban* 'become' (Abbas and Raza, 2014; Raza, 2011) is presented in 19(c). The subjunctive form of copula verb *rahEN* 'remain' (Abbas and Raza, 2014; Raza, 2011) is annotated in 19(d). The copula verb with present and the past tense *hEN/tHE* 'are/were' is annotated in 19(e).

(19) a. SehrI parESAn nahI hOtE hEN
citizens/N worry/ADJ not/ADV.NEG be(become)/V.COP.IMPERF be/VAUX.PRES
'The citizens do not worry.'

b. xAhiS pUrI nah hUI
desire/N fulfill/ADJ not/ADV.NEG be(become)/V.COP.PERF
'The desire did not become fulfilled.'

c. sIrAj acHA ban giyA
Siraj/N.PROP good/ADJ become/V.COP.ROOT go/V.LIGHTV.PERF
'Siraj became good.'

d. Ith2AdI kAmyAb rahEN gE
allies/N successful/ADJ remain/V.COP.SUBTV will/VAUX.FUTR
'The allies will remain successful.'

e. aEsE lOg mOjOd hEN/tHE
such/ADJ people/N present/ADJ be/become/V.COP.PRES/.PAST
'Such people are/were present.'

[7] The VALA as a head word gives the phrase with adjectival property most likely and follows infinitive verb, nominal, adjectives, etc.

2.20.2 Imperfective Verbs

The imperfective verb tag V.IMPERF describes actions or states occurring generally or regularly (Schmidt, 2013). It can be identified by the inflected suffixes *tA, tI, tE, tEN* at the end of a verb. These suffixes co-exist after the root of a verb. At present, it has only one tag as V.IMPERF; however, if it is repeated then .REP can be used at the end. The example of V.IMPERF is given as follows.

(20) voh aks2ar cOrI kartE hEN
they/P.PERS often/ADV stealing/N do/V.IMPERF be/VAUX.PRES
'They often do stealing.'

2.20.3 Infinitive Verbs

An infinitive verb V.INF can be identified through its inflected suffixes *nA, nI, nE, nEN* concatenated to the root form of a verb. The annotation example is given as follows.

(21) h2akUmat kO kAm karnA hE
government/N CM work/N do/V.INF be/VAUX.PRES
'The government has to do work.'

2.20.4 Light Verbs I

A light verb V.LIGHT is a verb that contains a little semantic content of its own and it forms a predicate with some additional expression such as a noun or an adjective (Ahmed and Butt, 2011; Butt, 2003; Raza, 2011). It is subcategorized into eight different forms, which includes V.LIGHT.IMPERF (light verb with imperfective morphology), V.LIGHT.INF (light verb with infinitive morphology), V.LIGHT.PERF (light verb with perfective morphology), V.LIGHT.PROG (light verb with progressive morphology), V.LIGHT.ROOT (light verb with root form), V.LIGHT.SUBTV (light verb with subjunctive morphology), V.LIGHT.PAST (light verb with past tense) and V.LIGHT.PRES (light verb with present tense) as can be seen in 22(a)-(g) for light verbs *AtA* 'use to come', *AnA* 'to come', *AyA* 'came', *rAhA* 'remain', *A* 'come', *dORAEN* 'let to run' and *thA/hE* 'was/is' respectively. The respective subcategories can also be seen in Figure 1. All the light verbs presented in annotated sentences shared their semantic content with a preceding noun N or adjective ADJ.

(22) a. mUjE jin naz4ar AtA tHA
me/P.PERS ghost/N vision/N come/V.LIGHT.IMPERF be/VAUX.PAST
'I had used to sight the ghost.'

b. mUjE jin naz4ar AnA tHA
me/P.PERS ghost/N vision/N come/V.LIGHT.INF be/VAUX.PAST
'I had to sight the ghost.'

c. mUjE jin naz4ar AyA tHA
me/P.PERS ghost/N vision/N came/V.LIGHT.PERF be/VAUX.PAST
'I had sighted the ghost.'

d. a2li kO a2lambardArI kA a2uhdah h2Asil
Ali/N.PROP CM flag-bearer/N of/CM designation/N gain/N
rAhA
remain/V.LIGHT.PROG
'Ali had the designation of flag-bearer'

e. lOg taSadud bardASt nahI
people/N torture/N bear/N not/ADV.NEG
kar saktE
do/V.LIGHT.ROOT VAUX.MOD.IMPERF
'The people can not bear the torture'

f. ham naqSE par naz4ar dORAEN gE
we/P.PERS map/N on/CM vision/N run/V.LIGHT.SUBTV will/VAUX.FUTR
'We will look on the map'

g. SirAj pUrAnE AdmION mEN SAmil thA/hE
Siraj/N.PROP old/ADJ persons/N in/CM include/N was/is/V.LIGHT.PAST/PRES
'Siraj was/is included in old persons'

2.20.5 Light Verbs II

A light verb V.LIGHTV is a verb that contains a little semantic content of its own and it forms a predicate in the presence of an additional verb (Butt, 2003), hence is called as the verb-verb complex predicate. It is subcategorized into five different forms, which include V.LIGHTV.IMPERF, V.LIGHTV.INF, V.LIGHTV.PERF, V.LIGHTV.ROOT and V.LIGHTV.SUBTV as can be seen in 23(a)-(e) for light verbs *kartA* 'used to do', *dEnA* 'to give', *liyA* 'took', *hO* 'be/become' and *jAyE* 'should go' respectively. All the light verbs presented in the annotated sentences as follows shared their semantic content with a preceding verb V or a light verb V.LIGHT.

(23) a. voh patHar luRHkA dEtA tHA
he/P.PERS stone/N roll/V.PERF give/V.LIGHTV.IMPERF be/VAUX.PAST
'He used to roll the stone'

b. mUjHE jarmany cHOR dEnA
I/P.PERS Germany/N.PROP.SPT leave/V.ROOT give/V.LIGHTV.INF
cAhIyE
should/VAUX.MOD.PERF
'I should have to leave Germany'

c. mEN nE h2aj kar liyA
I/P.PERS CM pilgrimage/N do/V.ROOT take/V.LIGHTV.PERF
hE
be/VAUX.PRES
'I have performed the Hajj (pilgrimage)'

d. janral mUSaraf kO fOj muth2arik
general/N Musharraf/N.PROP CM army/N mobilization/N
karnA hO gI
to-do/V.LIGHT.INF be/become/V.LIGHTV.ROOT will/VAUX.FUTR
'General Musharraf will have to mobilize the army'

e. yeh saRak sitambar mEN mUkammal
this/P.DEM road/N.SPT September/DATE.M in/CM complete/ADJ
hO jAyE gI
be/become/V.LIGHT.ROOT should-go/V.LIGHTV.SUBTV VAUX.FUTR
'This road will be completed in September'

2.20.6 Modal Verbs

A modal verb V.MOD expresses a scale ranging from possibility to necessity (Abbas and Nabi Khan, 2009). It is subcategorized into three morphological forms, which includes V.MOD.IMPERF (modal verb with imperfective morphology), V.MOD.PERF (modal verb with perfective morphology) and V.MOD.SUBTV (modal verb with subjunctive morphology). This category of modal verbs is different from modal auxiliaries discussed in Section 2.22.3, in which the main verb (predicate) of the sentence is annotated with V and the modal auxiliaries are annotated with VAUX. The examples of *cAhnA* 'may want to' modified from Facchinetti et al. (2003) contain modal verb V.MOD acting as the predicate of the respective sentence and not as an auxiliary, and are presented as follows.

(24) a. voh kitAb paRHnA cAhtA hE
he/P.PERS book/N read/V.INF want/V.MOD.IMPERF be/VAUX.PRES
'He may wants to read the book.'

b. tUm nE intEqAm lEnA cAhA tHA
you/P.PERS CM revenge/N take/V.INF want/V.MOD.PERF be/VAUX.PAST
'You might have wanted to take the revenge'

c. voh intEqAm lEnA cAhEN gE
they/P.PERS revenge/N take/V.INF want/V.MOD.SUBTV will/VAUX.FUTR
'They will want to take a revenge'

2.20.7 Perfective Verbs

A verb with perfective morphology V.PERF can be identified through its inflected suffix e.g. *A, I, E, EN* concatenated to the root form of a verb. The annotation examples can be seen in 25 for a verb *kahA* 'said' in (a) and *giyA* 'went' in (b). The repetition of same verb can be annotated as V.PERF.REP. More examples can be seen in 3(d, e), 8(b), 9(d), 13(c) and 16.

(25) a.

rasUl	nE	kahA
prophet/N	CM	said/V.PERF

'The prophet said'

b.

voh	pOlIs	kE	pAs	giyA
he/P.PERS	police/N	CM	to/POSTP	went/V.PERF

'He went to the police'

2.20.8 Root Verbs

A verb with root form is a verb to which suffixes can be added (Schmidt, 2013). An annotated example can be seen in 26 for a verb *A* 'come', whose infinitive form is *AnA* 'to come'. More examples can be seen in 3(b) and 13(c). The repetition of same verb can be annotated as V.ROOT.REP.

(26)

voh	yahAN	dOrAnE	taftIS	A
he/P.PERS	here/ADV.SPT	during/PREP.TMP	investigation/N	come/V.ROOT

giA
go/V.LIGHTV.PERF

'He came here during the investigation'

2.20.9 Subjunctive Verbs

A subjunctive verb is a verb used to express hypothetical actions or conditions (Dic, 2014; Schmidt, 2013). Annotated examples can be seen in 2(e) and 15(b) for the subjunctive form of the verbs *AyIN* 'come' and *jAO* 'go' respectively.

2.20.10 Verb With Tense

There are sentences, the structures of which look like copular constructions but the argument requirement (subject and predicate link) for their predicates cannot be fulfilled. It means that either the subject or the predicate link is missing in these types of sentences. The structure of these types of sentences is closer to existential copula construction in English. For example, the sentence *There is the God* has an existential copular construction (Raza, 2011). The translation of this sentence in Urdu is *xUdA*/God

hE/is with one argument for an existential copula verb *is*. Due to incomplete arguments in these type of sentences only, the copula verb V.COP.PRES/PAST is reduced to V.PRES/PAST for present and past tense as follows.

(27) mErI xAhiS tHI/hE
my/P.POSS desire/N be/V.PAST/PRES
'It is my desire'

2.21 Special VALA

The VALA is a special word in Urdu, which normally appears in a noun or an adjective phrase. It can also express the action that is going to start in a special way as can be seen in 28(a). Another reading of the same sentence is also mentioned. A single tag VALA is used to represent all types of *vAlA* morphological forms. Example given in 28(b) has a nominal reading.

(28) a. mEN kAm karnE vAlA hUN
I/P.PERS work/N to-do/V.INF going/has/VALA be/V.COP.SUBTV
'I am going to do work or I am a working person' (two different readings)

b. mEN dUdH vAlA hUN
I/P.PERS milk/N has/VALA be/V.COP.PRES
'I am the milkman'

2.22 Verb Auxiliaries

Verb auxiliaries VAUX denote the tense, aspect, modality, voice, mood, emphasis, etc., of the sentence predicate (Aarts et al., 2014). In Urdu, a predicate or a complex predicate in the main verb phrase of the sentence precedes verb auxiliaries e.g., *hO*/V.COP.ROOT *gayI*/V.LIGHTV.PERF *hE*/VAUX.PRES 'has/have become' contains a tense auxiliary VAUX.PRES along with the complex predicate *hO gayI*. Verb auxiliaries are divided into 11 subcategories discussed as follows.

2.22.1 Imperfective Auxiliaries

The method of identification for the imperfective auxiliary VAUX.IMPERF is the same as was discussed in Section 2.20.2 of imperfective verbs V.IMPERF. It is a single sub- category with no any further divisions. An annotated example for this subcategory is given as follows.

(29) kEs invesTigESan pOlIs kE pAs calA
case/N investigation/N police/N CM to/POSTP walk/V.PERF
jAtA hE
go/VAUX.IMPERF be/VAUX.PRES
'The case (usually) goes to investigation police'

2.22.2 Infinitive Auxiliaries

The identification of infinitive auxiliaries VAUX.INF is the same as was discussed in Section 2.20.3 of infinitive verbs V.INF. It is also a single subcategory, whose annotated example is presented for *jAnE* 'to go' as follows.

(30) Ap a2rab qabAyl mEN pehcAnE jAnE
he/P.PERS Arab/N.SPT tribes/N in/CM recognize/V.INF go/VAUX.INF
lagE tHE
take/VAUX.PERF be/VAUX.PAST
'He had become known in Arab tribes'

2.22.3 Modal Auxiliaries

A modal auxiliary VAUX.MOD expresses a range from possibility to necessity (Abbas and Nabi Khan, 2009; Bhatt et al., 2011). It is subcategorized into three morphological forms, which include VAUX.MOD.IMPERF (modal auxiliary with imperfective morphology), VAUX.MOD.PERF (modal auxiliary with perfective morphology) and VAUX.MOD.SUBTV (modal auxiliary with subjunctive morphology). These modal auxiliaries are different from the modal verbs discussed in Section 2.20.6, in which the modal verbs were acting as the predicate of the sentence but here the modal auxiliaries are following the predicate of the sentence. The examples for modal auxiliaries are as follows for *saktE* 'can', *cAhIyE* 'should' and *paREN* 'has/have to'.

(31) a. voh baRE h2Ads2E kA sabab ban
they/P.PERS big/ADJ.DEG accident/N of/CM reason/N become/V.COP.ROOT
saktE hEN
can/VAUX.MOD.IMPERF be/VAUX.PRES
'They can become the reason of a big accident'

b. kAm xatam kar dEnA cAhIyE
work/N finish/N do/V.LIGHT.ROOT give/V.LIGHTV.INF should/VAUX.MOD.PERF
'The work should be finished'

c. vATar kOnsal kO qAnUnI mUSgAfIyAN
Water/N.PROP Council/N.PROP CM regulation/ADJ anomalies/N
dUr karnI paREN gIN
far/ADJ do/V.LIGHT.INF has/haveto/VAUX.MOD.SUBTV will/VAUX.FUTR
'The Water Council will have to remove the regulation anomalies'

2.22.4 Passive Auxiliaries

In sentences with passive auxiliaries VAUX.PASS, the theme/patient becomes the grammatical subject of the main verb. It is divided into five subcategories, which includes VAUX.PASS.IMPERF (passive auxiliary with imperfective morphology), VAUX.

PASS.INF (passive auxiliary with infinitive morphology), VAUX.PASS.PERF (passive auxiliary with perfective morphology), VAUX.PASS.ROOT (passive auxiliary with root form), and VAUX.PASS.SUBTV (passive auxiliary with subjunctive morphology). The given examples have a morphological annotation of passive auxiliaries *jAtA* 'use to go', *jAnA* 'to go', *giyA* 'went', *jA* 'go' and *jAyEN* 'may go' respectively. These different forms of *jA* 'go' auxiliary are considered passive only, when they are preceded by a predicate or a complex predicate with perfective morphology (Raza, 2010).

(32) a. jAnvarON kO pAnI pilAyA
animals/N CM water/N make-someone-drink/V.PERF
jAtA hE
go/VAUX.PASS.IMPERF be/VAUX.PRES
'The animals are watered'

b. kaSmIrION kA bHI sOcA
Kashmiri/N.SPT of/CM also/PT.INTF think/V.PERF
jAnA cAhIyE
go/VAUX.PASS.INF should/VAUX.MOD.PERF
'Kashmiri's should also be considered'

c. tehsIldAr ka tabAdlah kiyA giyA
Tehsil-officer/N of/CM transfer/N do/V.PERF go/VAUX.PASS.PERF
'The Tehsil officer has been transferred'

d. sUDAn mEN nasal-kUSI kI
Sudan/N.PROP.SPT in/CM genocide/N do/V.PERF
jA rahI hE
go/VAUX.PASS.ROOT continue/VAUX.PROG be/VAUX.PRES
'Genocide is being commuted in Sudan'

e. kaSmIr sE fOjEN nikAl
Kashmir/N.PROP.SPT from/CM armies/N takeout/V.ROOT
lI jAyEN
take/V.LIGHTV.PERF go/VAUX.PASS.SUBTV
'The armies may be taken out from Kashmir'

2.22.5 Perfective Auxiliaries

The identification of perfective auxiliary VAUX.PERF is the same as was discussed in Section 2.20.7 of perfective verbs. It is an independent single subcategory, whose annotation is given in the following example.

(33) lOgON kE ravaIyON mEN tabdeelI AtI gayI
people/N of/CM behavior/N in/CM change/N come/V.IMPERF go/VAUX.PERF
'The change used to come in people's behaviors'

2.22.6 Progressive Auxiliaries

The progressive auxiliary VAUX.PROG can be identified easily through its morphological form after a verb or an auxiliary. Its morphological forms include *rahA, rahE, rahI, rahIN*. An annotated example can be seen in 32(d) for a progressive auxiliary *rahI* 'continue'.

2.22.7 Root Auxiliaries

The identification of an auxiliary with a root form VAUX.ROOT is the same as discussed in Section 2.20.8 for verbs with root morphology. An annotated example is given as follows.

(34) faqIr baRHtE jA rahE hEN
beggars/N increase/V.IMPERF go/VAUX.ROOT continue/VAUX.PROG be/VAUX.PRES
'The beggars are increasing'

2.22.8 Subjunctive Auxiliaries

A subjunctive verb auxiliary VAUX.SUBTV describes an uncertain action or state contingent on something else like permission, wish, request, etc., (Schmidt, 2013). It has no further divisions. An annotated example of subjunctive auxiliary is given as follows.

(35) kiyA mEN andar AyUN ?
what/QW I/CM In/ADV.SPT come/VAUX.SUBTV ?/M.S
'May I come in?'

2.22.9 Tense Auxiliaries

The tenses of auxiliary VAUX are divided mainly into three tense divisions, which include VAUX.FUTR (future tense auxiliary e.g. *gA, gI, gE, gIN*, etc.), VAUX.PAST (past tense auxiliary e.g. *tHA, tHI, tHE, tHEN*, etc.) and VAUX.PRES (present tense auxiliary e.g. *hE, hEN*, etc.). The annotation of the future tense auxiliary can be seen in 36(a). The annotation of past tense auxiliary is presented in 36(b). Similarly, the annotation of last subcategory VAUX.PRES is annotated in 36(c).

(36) a. ham naqSE par naz4ar dORAEN gE
we/P.PERS map/N on/CM vision/N run/V.LIGHT.SUBTV will/VAUX.FUTR
'We will look on the map'

b. ham naqSE par naz4ar dORA
we/P.PERS map/N on/CM vision/N run/V.LIGHT.PERF
rahE tHE
continue/VAUX.PROG be/VAUX.PAST

'We were looking on the map'

c. ham naqSE par naz4ar dORA
we/P.PERS map/N on/CM vision/N run/V.LIGHT.PERF
rahE hEN
continue/VAUX.PROG be/VAUX.PRES
'We are looking on the map'

Tags presented in Figure 1 have been completed along with the examples. The discussion in this Section 2 concludes the SSP guidelines for the URDU.KON-TB treebank.

3 The URDU.KON-TB Treebank

The development of the URDU.KON-TB treebank was performed in three steps: the collection of sentences in the form of a corpus, manufacturing of an annotation scheme and the employment of this annotation scheme on the said corpus. These steps are overviewed as follows. In initial development of the URDU.KON-TB treebank (Abbas, 2012), a POS, a syntactic and a functional tag sets were proposed. It was an original work done after getting motivation from the Penn treebank[8] and the Urdu Lexical Functional Grammar (LFG) built during a project called PARGRAM[9]. This work has some issues, which were resolved and updated after the annotation evaluation (Abbas, 2014a & 2014b). The updated versions of the tag sets have been presented and discussed earlier in Section 2.

3.1 Corpus Construction

In initial development (Abbas, 2012), a 19 million words corpus (Ijaz and Hussain, 2007) was used that was available at CRULP/CLE.[10] This corpus was collected from the Jang[11] and the BBC[12] newspapers. This corpus had licensing constraints due to which it is not publicly available anymore (Urooj et al., 2012). One thousand (1000) sentences taken from this corpus are then extensively modified to become free from licensing constraints, because we want to share our corpus freely under a Creative Commons Attribution/Share-Alike License 3.0 or higher. The next four hundred (400) sentences are collected from the Urdu Wikipedia.[13] The data collected from the Urdu Wikipedia

[8] http://www.cis.upenn.edu/~treebank/home.html

[9] http://ling.uni-konstanz.de/pages/home/pargram_urdu/

[10] A center for language engineering in Pakistan at http://cle.org.pk/

[11] http://jang.com.pk/

[12] http://www.bbc.co.uk/urdu/

[13] https://ur.wikipedia.org

is already under that license. Thus, the size of the corpus is limited to fourteen hundred (1400) sentences. The size of corpus is kept limited within the context of doctoral work (Abbas, 2014b), however, an extension project[14] to increase the size of the treebank up to 2000 sentence is completed and will be published soon. Overall, the corpus contains the text of local & international news, social stories, sports, culture, finance, history, religion, traveling, etc.

3.2 Annotation Scheme

The annotation scheme of the URDU.KON-TB treebank consists of semi-semantic POS (SSP), semi-semantic syntactic (SSS) and functional (F) tag sets. The term semi-semantic (partly or partially semantic) is used with the POS because some tags are encoded with semantics but not all e.g. N.SPT (a spatial noun) tag for a word *house*, ADJ.TMP (a temporal adjective) tag for a word *previous* in *previous year*, etc. At the SSP level, a dot '.' is used to add morphological (Figure 3) and semantical (Figure 4) labelings of subcategories into the main categories (Figure 2) as discussed in Section 3.3. Overall, for the SSP, SSS and F annotation, a combination of phrase structure (PS) and hyper dependency structure (HDS) has been adopted. The DS is called HDS because it is not limited to make constituents on the basis of headwords, but also on the basis of the head-constituents, when you have to make a constituent from its nested constituents. The details are given in Abbas (2014b). The POS, morphological, syntactical, semantical, clausal and functional information (Abbas, 2014b) all together, makes a rich annotation scheme for the URDU.KON-TB treebank. The need for such type of schemes is highly advocated by some researchers, such as Clark et al. (2010), Skut et al.(1997), etc.

3.3 Employment of Annotation

A simple POS tag set was devised first, which contained twenty two (22) main POS-tag categories displayed in Figure 2. The description of the tags is given in the respective cells of the figure. The figure includes some non-familiar tags like HADEES and MARKER to represent the Arabic statements of prophets in Urdu text and a phrase or a sentence marker similar to punctuation marks but not all, respectively. The labels for morphological and semantic subcategories are presented in Figures 3 and Figure 4 respectively, which can be added to 22 main categories of POS tags by using a dot '.' symbol. The SSP tag set was refined during the manual annotation process of sentences and further refined after the evaluation process with the Krippendorf's α statistical model (Krippendorf, 2014) and also presented in Abbas (2014a). The final refined form of the SSP tag set is given in Figure 1. In case of morphology, if a main verb V has a perfective morphology, then the tag becomes V.PERF. Similarly, the case of spatial noun N.SPT is discussed in the beginning of Section 3.2. The semantic tags like SPT

[14] http://clsp.org/projects.html

(spatial), TMP (temporal), MNR (manner), etc. are not possible with verbs, auxiliaries, conjunctions, etc., as can be seen in Figure 1.

(37) حامد نے شیر کو جنگل میں بندوق سے مارا ۔

hAmed	nE	SEr	kO	jangal	mEN	bandUq	sE	mArA	.
N.PROP	CM	N	CM	N.SPT	CM	N	CM	V.PERF	M.S

'Hamid killed the lion in the jungle with a gun.'

An example of the SSP annotation is given in Example 37. The Urdu script is written from right to left. The row beneath the Urdu script is the transliteration of the sentence as proposed in Malik et al. (2010). The tokens of the sentence are tagged according to the SSP tag set. *hAmed* is a proper name (N.PROP). *SEr* and *bandUq* are common nouns (N), while *jangal* is a spatial common noun (N.SPT).[15] *nE, kO, mEN*, and *sE* are case markers (CM) for ergative, accusative, spatial/locative and instrumental cases, respectively. The syntactic differentiation of the case markers is done according to the studies in Butt and King (2004).

The tagset in Figure 1 represents the complete SSP tagset. The discussion on each tag is presented in Section 2. As a repeated example, consider the ADJ (Adjective) in Figure 1, which is divided into five subcategories of tags DEG (Degree), ECO (Echo), MNR (Manner), SPT (Spatial) and TMP (Temporal). Relevant examples are provided in 38.

(38) a. اچھا لڑکا

acHA	laRkA
good/ADJ	boy/N

'Good Boy'

b. اہم ترین شخصیت

aham	tarIn	Saxs2iat
important/ADJ	most/ADJ.DEG	personality/N

'Most important personality'

c. برا ورا کام

burA	vurA	kAm
ugly/ADJ	ADJ.ECO	work/N

'Ugly work'

[15] In the presence of a sense of place/location or direction to/from place/location in a word, SPT tag is used e.g. Pakistan and the country, are the two words. Pakistan is the proper name of a place (country) and is tagged as N.PROP.SPT. However, country is a common noun but having a sense of place. So, it is tagged as N.SPT. This distinction is not different from spatial adverbs e.g. there, here, etc.

d. جابرانہ حکومت

jaberaanah hakUmat
cruel/ADJ.MNR government/N
'Cruel Government'

e. گزشتہ سال

guzaStah sAl
previous/ADJ.TMP year/N
'Preveious Year'

f. ملتانی کھسہ

mUltAnI kHUsah
multani/ADJ.SPT shoe/N
'Multani shoe'

The example 38(a) is a simple case of ADJ, while 38(b) is a case of a degree adjective [16] annotated with ADJ.DEG. The comparative and superlative forms of adjectives can be made by introducing Persian suffixes *tar* 'more' and *tarIn* 'most' after the absolute form of adjectives e.g. *xUbs3Urat-tar* 'prettier' and *xUbs3Urat-tarIn* 'prettiest'. There are some words, which can play the role of a degree adverb or a degree adjective e.g. *zEyAdah* 'more/most/much', *bohat* 'more/enough', *kAfI* 'quite/too', etc. (Schmidt, 2013). If these words qualify adjectives, then this is the usage as degree adverbs, otherwise as a degree adjective. Example 38(c) is a case of reduplication (Abbi, 1992; Boegel et al., 2007). As reduplication has two versions, first in Urdu like other South Asian languages, the reduplication of a content word is frequent. Its effect is only to strengthen the proceeding word or to expand the specific idea of a proceeding word into a general form e.g. *kAm THIk-THAk karnA* 'Do the work right' or *kOI kapRE-vapRE dE dO* 'Give me the clothes or something like those'. Second version is the repetition of the original word e.g. *sAtH sAtH* 'with/along-with'. These two versions are named as full word reduplication and echo reduplication by Boegel et al. (2007), which are represented in our annotation as ECO (echo) and REP (repetition) respectively. The echo words normally begin with the letters *S* or *v* or *m*.

Example 38(d) is the case of adjective having a sense of manner annotated as ADJ.MNR. If an adjective qualifies an action noun, then a sense of action or something is produced, whose behavior or the way to do that action is confirmed through ADJ.MNR e.g. *z4AlemAnah t2abdIlIyAN* 'brutal changes'. If an adjective comes individually, then its mannerism can be resolved independently through its sense or by building a sense with the predicate. If there is a sense of manner then an adjective of manner can exist like in copular construction e.g. *voh GEr-h2Az3ir hE* 'He is absent'. An exercise of adjectives and adverbs of manner for the English language can be seen at Cambridge

[16] This division is used to represent absolute, comparative and superlative degree in adjectives and adverbs.

University from which this idea is taken.[17] Example 38(e) is case of an adjective having a temporal sense. Finally, example 38(f) is the case of an adjective having a spatial sense. The adjective used here is the derivational form of a city/place name Multan, which is a spatial proper noun. But it appears here as an adjective and annotated as ADJ.SPT[18] like in this sentence e.g. *voh Ek pAkistAnI laRkA hE* 'He is a pakistani boy'.

The example 38 for adjectives exploited its POS tags along with semantic tagging like TMP, SPT, MNR, etc. However, to give an introduction about morphology and verb functions, another POS category V from Figure 1 is discussed as follows. A few high quality studies were conducted on verbs for morphologically rich language (MRL) Urdu by Butt and Rizvi (2010), Butt and Ramchand (2001) and Butt (2010). The rules for identifying different forms of verbs were adopted from these studies. The V annotates the predicate/main-verb of the sentence and is divided mainly into 11 subcategories, which include COP (copula verb), IMPERF (imperfective morphological form of the verb), INF (infinitive form of verb), LIGHT (1st light verb with nouns and adjectives), LIGHTV (2nd light verb with verbs), MOD (modal verb), PERF (perfective morphology), ROOT (root form), SUBTV (subjunctive form), PAST (past tense of a verb) and PRES (present tense of a verb). Their description is also given in Figure 1. These tags are further divided into subcategories depicted in Figure 1. All these tags represent different morphological forms and the function of a verb that it governs. Some annotated sentences containing different verb forms and functions are given in example 39.

(39) a. مہنگائی نے لوگوں کا جینا دوبھر کیا تھا

mehangAI	nE	lOgON	kA	jInA	dUbHar	kiyA	tHA
N	CM	N	CM	N	N	V.LIGHT.PERF	VAUX.PAST

'The inflation had made the life of people hard.'

b. گرانفروشوں کے خلاف قانون حرکت میں لایا جائے

giraN-faroSoN	kE	xilAf	qAnUn	harkat	mEN	lAyA	jAyE
N	CM	POSTP.MNR	N	N	CM	V.PERF	VAUX.PASS.SUBTV

'The law should be practiced against inflators'

c. محمد صلی اللہ علیہ والہ وسلم نے فرمایا کہ الحسین منی و انا من الحسین
یعنی حسین مجھ سے ہے اور میں حسین سے ہوں .

mUhammad	sal-lal-la-ho-a2lEhE-va-AlEhI-salam	nE	farmAyA	keh		
N.PROP	PRAY	CM	V.PERF	C.SBORD		
“	al-hUsynON-mInnI-vA-anA-mInal-hUsyn	”	ya2nI	‘	hUsyn	
M.P	HADEES	M.P	ADV	M.P	N.PROP	

[17] http://www.cambridge.org/grammarandbeyond/wp-content/uploads/2012/09/Communicative_Activity_Hi-BegIntermediate-Adjectives_and_Adverbs.pdf

[18] Spatial adjectives are used to describe a place/location, direction or distance e.g. multAnI 'Multani', aglI 'next', and dUr 'far' respectively.

mUjH sE hE aOr mEN hUsyn sE hUN ’ .
P.PERS CM V.COP.PRES C.CORD P.PERS N.PROP CM V.SUBTV M.P M.S
'Muhammad (May Allah grant peace and honor on him and his family) said that “al-hUsynON-mInnI-vA-anA-mInal-hUsyn” means ‘Hussain is from me and I am from Hussain’.'

d. تم نے حج تو کر لیا ہو گا ؟

tUm nE haj tO kar liyA hO gA ?
P.PERS CM N PT.EMP V.ROOT V.LIGHTV.PERF VAUX.SUBTV VAUX.FUTR M.S
'You will have made the pilgrimage?'

e. جب یہاں کھیت ہوتے تھے

jab yahAN kHEt hotE tHE
ADV.TMP.REL ADV.SPT N.SPT V.IMPERF VAUX.PAST
'When, here would have been crop fields'

f. ان کا مطالبہ تھا

Un kA mUtAlibah tHA
P.PERS CM N V.PAST
'It is their demand.'

The sentence in example 39(a) is a case of noun-verb complex verb predicate, which was first proposed by Mohanan (1994). The words *dUbHar kiyA* 'made hard' is a noun-verb complex predicate. The noun *dubHar* and the verb *kiyA* with a perfective morphological form *yA* at the end are annotated as a N and a V.LIGHT.PERF respectively. Similarly, a perfective verb *liyA* 'took' after a root form of verb *kar* 'do' is an example of the verb-verb complex predicate depicted in 39(d). This construction is adopted from the studies given in (Butt, 2010). The light verb after a N or an ADJ lies in the 1st category of light verbs and annotated as V.LIGHT in our annotation, while the light verb after a verb lies in the 2nd category of light verbs and annotated as V.LIGHTV. The next sentence in 39(b) is a passive sentence. A passive construction can be concluded with the inflected form of a verb *jAnA* 'to go' proceeded by another verb with perfective morphology as can be seen in 39(b). The subjunctive form of auxiliary verb tagged as VAUX.PASS.SUBTV is preceded by a perfective verb *lAyA* 'brought', which is then annotated as V.PERF. The subjunctive form of verb is acting as an aspectual auxiliary and not as a V.LIGHTV, which was discussed in (Butt and Ramchand, 2001) and adopted as it is. The rules for identification of verb function and other morphological forms can be found in Section 2.

To explore some other unusual tags, a long sentence is presented in 39(c). After the name of prophets or righteous religious-personalities, some specific and limited prayers called *s3alAvAt* 'prayers' e.g. *sal-lal-la-ho-a2lEhE-va-AlEhI-salam* 'May Allah grant peace and honor on him and his family', *a2lEh salAm* 'peace be upon him', etc. in Arabic is most likely in Urdu text and annotated as PRAY. Similarly, the statements of

prophet Muhammad (PBUH) called *h2adIs2* 'narration' e.g. *In-namal-aa2mAlo-bin-niyAt* 'The deeds are considered by the intensions' in Arabic is also a tradition in Urdu text and annotated as HADEES. In religious text of Urdu, this kind of phenomenon is most likely in Arabic script rather than the Urdu script. This annotation with PRAY and HADEES is performed only, when prayers or narrations appear in Arabic language in Urdu text as can be seen in 39(c). The phrase markers like comma, double quotes, single quotes, etc. are annotated with M.P and sentence marker like full stop, question mark, etc. are annotated with M.S as presented in the same example. The tense is divided into present, past and future. A predicate of the sentence with present and past tense is possible as annotated in 39(f) but not with future tense, because future tense always behaves as verb auxiliary in Urdu. The tense of verb auxiliaries like present, past and future is annotated in 39(a, d, e). A verb with imperfective morphology e.g. *tA, tI, tE, tEN* at the end of a verb is annotated with V.IMPERF as given in 39(e).

This section concludes the concept of SSP tags used in the annotation of the URDU.KON-TB treebank. There are twenty-two tags, which are divided into further subcategories as presented in Figure 1. The POS annotation evaluation via Krippendorf's Alpha α is detailed in (Abbas, 2014a & 2014b), however an overview is presented next in Section 4. This evaluation came up with POS tags issues related to readability. After evaluation, the problematic POS tags are either removed or revised and a final SSP tagset is obtained and presented.

4 Evaluation and Results

This Section describes the evaluation of the annotation guidelines of the URDU.KON-TB treebank presented in Section 2 and 3. The evaluation is the process of calculating inter-annotator agreement (IAA), which provides a quantitative answer as to the overall consistency plus feasibility of the annotation scheme. For the evaluation of the URDU.KON-TB treebank annotation, the most advanced measure known as the Krippendorf's α coefficient (Krippendorf, 2004) is used. The output of the annotators is recorded and processed. The reliability of the SSP annotations is evaluated. The issues faced in annotation evaluation are removed via respective revisions and are reported shortly in forthcoming sections.

4.1 Setup

For the reliability evaluation of annotation guidelines presented in Section 2 of the URDU.KON-TB treebank for Urdu, it was essential that annotators should be the native speakers of Urdu possessing linguistics skills. To fulfill this purpose, an undergraduate class of 25 linguistics students has been adopted in the training course of annotation at Department of English, University of Sargodha, Pakistan.[19] This training was given to

[19] http://uos.edu.pk/

students as a partial part of their major course of linguistics. During this training course, thirty-two (32) lectures on annotation guidelines with practical sessions were delivered. The duration of each lecture was of 3 hours. The class was further divided into five groups and during their initial practical sessions, one student with high caliber of understanding was selected (but not informed) secretly from each group for the final annotation. The annotation task of 100 random sentences was divided into 10 home assignments. Each assignment contained 10 sentences. After twenty days of this course, the annotation assignments were given to all students along with the selected students with an instruction not to discuss it with each other. These assignments were collected, marked and the students were awarded with grades. The annotation performed in their home assignments by the selected students was then recorded in Microsoft Excel and evaluated for the reliability of annotation or IAA by applying the Krippendorf's α coefficient. The details of the SSP annotation evaluation can be seen in Section 4.2.

4.2 SSP Tagset Evaluation & Results

The detail and definition of the SSP tagset was already described in Sections 2 and 3. The complete guidelines of the SSP tagging were given to students for annotation of sentences according to a procedure described in Section 4.1. The tagged sentences by the annotators were recorded in the form of a reliability data matrix. Details are given in the doctoral thesis by Abbas (2014b). From the reliability data matrix, values by tokens matrix was obtained which is also presented in Abbas (2014b). The α coefficient was computed according to the formula given in equation below and also described in Abbas (2014b). In this work, different variables of numerator and denominator of equation were computed and described. The value of the α coefficient obtained was 0.964 for the SSP tagging of annotators. The value of α obtained lied in the category of perfect agreement according to the Krippendorf. A perfect agreement of 0.964 has been found in case of the SSP annotation only, which means that the SSP annotation guidelines are reliable. The error analysis and discussion of issues related to SSP tag set evaluation is given in Abbas (2014b) but not discussed here due to the scope of this article.

$$metric\alpha = 1 - \frac{O_d}{E_d} = 1 - (n..-1)\frac{\sum_t \frac{1}{n_t - 1}\sum_p \sum_{q>p} n_{tp} n_{tq} metric_{pq}^{\delta^2}}{\sum_p \sum_{q>p} n_{.p} n_{.q} metric_{pq}^{\delta^2}}$$

ADV	30	30	0		100.00
ADV.DEG	10	10	0		100.00
ADV.MNR	20	20	0		100.00
ADV.NEG	10	10	0		100.00
ADV.SPT	20	20	0		100.00
ADV.TMP	70	70	0		100.00
C.CORD	30	30	0		100.00
CM	630	630	0		100.00
DATE.Y.CAL	20	8	12	DATE.Y	40.00
DIA.IZF	90	39	51	DIA	43.33
DIA.PESH	10	3	7	DIA	30.00
KER	150	71	79	CM	47.33
N	1010	1005	5	BLANK, N.SPT	99.50
N.PROP	80	80	0		100.00
N.PROP.SPT	10	10	0		100.00
N.SPT	60	56	4	N	93.33
N.TMP	60	60	0		100.00
P.DEM	50	50	0		100.00
P.INDF	10	10	0		100.00
P.PERS	200	197	3	P.PRO	98.50
P.POSS	10	8	2	P.PERS	80.00
P.POSS.REF	10	10	0		100.00
P.REL	10	10	0		100.00
POSTP	60	60	0		100.00
POSTP.CMP	30	25	5	POSTP	83.33
POSTP.SPT	20	20	0		100.00
POSTP.TMP	20	20	0		100.00
PREP	20	20	0		100.00
PT	20	9	11	PT.INTF	45.00
PT.INTF	30	13	17	PT	43.33
Q	30	30	0		100.00
Q.CARD	210	210	0		100.00
Q.FRAC	20	20	0		100.00
Q.ORD	40	36	4	Q.CORD, Q.CARD	90.00
QW	50	50	0		100.00
U	30	30	0		100.00
V.COP.IMPERF	20	20	0		100.00
V.COP.PERF	10	10	0		100.00
V.COP.ROOT	20	20	0		100.00
V.COP.TENS.PRES	110	57	53	V.COP.PRES, V.LIGHT.TENS.PRES, VAUX.TENS.PRES	51.82
V.INF	10	10	0		100.00
V.LIGHT.IMPERF	40	37	3	V.LIGHTV.IMPERF	92.50
V.LIGHT.KER	50	21	29	V.LIGHT.ROOT	42.00
V.LIGHT.PERF	40	36	4	V.LIGHTV.PERF, V.PERF	90.00
V.LIGHT.ROOT	30	30	0		100.00
V.LIGHT.TB.ROOT	30	13	17	V.LIGHT.ROOT	43.33
V.LIGHTV.PERF	60	54	6	V.LIGHT.PERF	90.00
V.LIGHTV.PROG.IMPERF	10	5	5	V.LIGHTV.IMPERF	50.00
V.PERF	200	189	11	VAUX.PASS.PERF	94.50
V.ROOT	20	20	0		100.00
V.TENS.PRES	50	21	29	V.PRES	42.00
VAUX.IMPERF	20	17	3	V.PASS.IMPERF	85.00
VAUX.LIGHTV.TENS.PRES	10	4	6	VAUX.PRES	40.00
VAUX.MOD.PERF	30	30	0		100.00
VAUX.PASS.PERF	40	40	0		100.00
VAUX.PROG.PERF	20	9	11	VAUX.PROG, V.COP.PERF	45.00
VAUX.TENS.FUTR	20	10	10	VAUX.FUTR	50.00
VAUX.TENS.PAST	10	3	7	VAUX.PAST	30.00
VAUX.TENS.PRES	220	101	119	VAUX.PRES	45.91

Figure 5: Annotators SSP tags distribution and confusion

The format of data evaluation in the Krippendorf's α was different from the data displayed in Figure 5, however, a sample of annotators' SSP tags distribution and confusion is displayed in Figure 5. The annotated data of 100 sentences that were given to the annotators contained 1281 tokens, from which the data of 904 tokens is presented. The rest of the tokens have accuracy almost more than 90% due to which they are not depicted. It is attempted to show the tags of those tokens on which the annotators were remained confused or disagreed. The tags used in the initial version of the URDU.KON-TB treebank are displayed in the first column of the figure. Adjective (ADJ) appeared 54 times in the sentences, which were then annotated by 5 annotators. The frequency of adjective annotation is depicted in the second column of the figure after multiplying 54 with 5 numbers of annotators. It concludes 270 times of annotation for adjective. Among the 270 annotations of ADJ, annotators were remained 265 times in agreement or the annotators assigned 265 times the same/identical tag ADJ. The number of times the annotators remained disagreed or confused is mentioned in the different column. Similarly, the different or confused or the disagreed tags used by the annotators are depicted in the next column. Finally, by dividing the values in the identical and the frequency columns, the percentage accuracy of each tag in the first column of the figure is calculated.

The SSP tags in the initial version of the URDU.KON-TB treebank, which are correctly annotated and have 100% accuracy include ADV and ADV with its semantic labels for adverbs, coordination conjunctions with C.CORD, case markers with CM, N.PROP, N.PROP.SPT and N.TMP for proper nouns, spatial proper nouns and temporal nouns, P.DEM and P.INDF for demonstrative and indefinite pronouns, etc. The tags contained less than or equal to 50% accuracy include tense auxiliaries e.g. VAUX.TENS.PRES mostly annotated differently with VAUX.PRES, progressive auxiliary e.g. VAUX.PROG.PERF annotated differently with VAUX.PROG and its copular behavior with V.COP.PERF, KER as a light verb e.g. V.LIGHT.KER annotated with V.LIGHT.ROOT, diacritics e.g. DIA.IZF with DIA only, etc. Annotation of some tokens with tags was left by the annotators represented with BLANK in the column 'different/confused/ disagreed tags' for each tag in the first column of the figure.

The error analysis and evaluation of tags was performed on the basis of this data depicted in Figure 5. First, the tags with less or equal to 50% of accuracy are revised with annotators decisions e.g. DATE.Y.CAL, PT.INTF, V.LIGHT.TB.ROOT, etc., and second are the tags with accuracy a little more than 50% but they have common confused pairs like V.COP.TENS.PRES and VAUX.TENS.PRES modified to V.COP.PRES and VAUX.PRES, respectively as can be seen in Figure 1. The detailed discussion on error analysis and evaluation of the SSP annotation is presented in (Abbas, 2014b).

5 Conclusion

This concludes the complete SSP guidelines of the URDU.KON-TB treebank with preliminary and essential additional information needed to explain the SSP annotation procedure in full. After introducing the URDU.KON-TB treebank (Abbas, 2012; Abbas, 2014a; Abbas 2014b) and the parser based on the URDU.KON-TB treebank (Abbas, 2014c/2015), the demand of the complete guidelines in the community was raising, due to which it is attempted to present the SSP complete guidelines as a first step. The rest of the guidelines for the semi-semantic syntactic and functional annotations will be presented soon. This effort does not only strengthen the practice of producing the guidelines for the annotation schemes but also addresses the modern issue of how to prepare and evaluate guidelines (Mikulova and Stepanek, 2010) effectively with the state of the art evaluation techniques (Krippendorf, 2004; Hayes and Krippendorf, 2007). Corpus annotated with these SSP annotation guidline can be useful to applications in this domain like natural language processing, machine learning and many language specific analysis as discussed in (Zia, et. al, 2015a/2015b; Abbas, et. al, 2009/2010/2014; Abbas 2014d).

References

Aarts, B., Chalker, S., & Weiner, E. (2014). *The Oxford Dictionary Of English Grammar*. Oxford University Press.

Abbas, Q. (2012, March). Building a hierarchical annotated corpus of urdu: the URDU. KON-TB treebank. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 66-79). Springer Berlin Heidelberg.

Abbas, Q. (2014a). Semi-semantic part of speech annotation and evaluation. *LAW VIII*, *75*.

Abbas, Q. (2014b). *Building Computational Resources: The URDU. KON-TB Treebank and the Urdu Parser* (Doctoral dissertation).

Abbas, Q. (2014c). Exploiting language variants via grammar parsing having morphologically rich information. *LT4CloseLang 2014*, 36.

Abbas, Q. (2014d). A Stochastic Prediction Interface for Urdu. *International Journal of Intelligent Systems and Applications*, *7*(1), 94.

Abbas, Q. (2015). Morphologically rich Urdu grammar parsing using Earley algorithm, Natural Language Engineering (NLE), Vol.21(2), PP.1-36, Cambridge University Press, UK

Abbas, Q., & Khan, A. N. (2009). Lexical functional grammar for Urdu modal verbs. In *Emerging Technologies, 2009. ICET 2009. International Conference on* (pp. 7-12). IEEE.

Abbas, Q., & Raza, G. (2014). A Computational Classification of Dynamic Urdu Copula Verb. *International Journal of Computer Applications*, *85*(10).

Abbas, Q., Ahmed, M. S., & Niazi, S. (2010). Language Identifier For Languages Of Pakistan Including Arabic And Persian. *International Journal of Computational Linguistics (IJCL)*, *1*(03), 27-35.

Abbas, Q., Karamat, N., & Niazi, S. (2009). Development of Tree-bank based probabilistic grammar for Urdu Language. *International Journal of Electrical & Computer Science*, *9*(09), 231-235.

Abbas, Q., Zia, T., & Khan, A. N. (2014). Syntactic and semantic analysis of Urdu modal verbs using XLE parser. *International Journal of Computer Applications*, *107*(10).

Abbi, A. (1992). Reduplication in South Asian Languages: An Areal, Typological, And Historical Study. Allied Publishers, New Delhi.

Ahmed, T., & Butt, M. (2011, January). Discovering semantic classes for Urdu NV complex predicates. In *Proceedings of the Ninth International Conference on Computational Semantics* (pp. 305-309). Association for Computational Linguistics.

Bhatt, R., Bögel, T., Butt, M., Hautli, A., Sulger, S., & King, T. H. (2011). *Urdu/Hindi modals*. Bibliothek der Universität Konstanz.

Bögel, T., Butt, M., Hautli, A., & Sulger, S. (2007). Developing a finite-state morphological analyzer for Urdu and Hindi. *Finite State Methods and Natural Language Processing*, 86.

Butt, M. (1995). *The structure of complex predicates in Urdu*. Center for the Study of Language (CSLI).

Butt, M. (2003). The light verb jungle [OL].

Butt, M. (2010). The light verb jungle: Still hacking away. *Complex predicates in cross-linguistic perspective*, 48-78.

Butt, M., & King, T. H. (2004). The status of case. In *Clause structure in South Asian languages* (pp. 153-198). Springer Netherlands.

Butt, M., & Ramchand, G. (2001). Complex aspectual structure in Hindi/Urdu. *M. Liakata, B. Jensen, & D. Maillat, Eds*, 1-30.

Butt, M., & Rizvi, J. (2010). Tense and aspect in Urdu. *Layers of aspect*, 43-66. Stanford: CSLI Publications.

Butt, M., & Sadler, L. (2003). Verbal morphology and agreement in Urdu. *Syntactic structures and morphological information. Mouton*, 57-100.

Clark, A., Fox, C., & Lappin, S. (2010). *The Handbook Of Computational Linguistics And Natural Language Processing*, 57. Wiley.com.

Facchinetti, R., Palmer, F., & Krug, M. (Eds.). (2003). *Modality in contemporary English* (Vol. 44). Walter de Gruyter.

Hayes, A. F., & Krippendorf, K. (2007). Answering The Call For A Standard Reliability Measure For Coding Data. *Communication Methods and Measures*, *1*(1), 77–89.

Hirsch, E. D., Kett, J. F., & Trefil, J. S. (2014). *The new dictionary of cultural literacy*. Houghton Mifflin Harcourt.

Ijaz, M., & Hussain, S. (2007, August). Corpus based Urdu lexicon development. In *the Proceedings of Conference on Language Technology (CLT07), University of Peshawar, Pakistan* (Vol. 73).

Kamran Malik, M., Ahmed, T., Sulger, S., Bögel, T., Gulzar, A., Raza, G., ... & Butt, M. (2010). Transliterating Urdu for a Broad-Coverage Urdu/Hindi LFG Grammar. In *LREC 2010, Seventh International Conference on Language Resources and Evaluation* (pp. 2921-2927).

Krippendorff, K. (2004). Reliability in content analysis. *Human communication research*, *30*(3), 411-433.

Leech, G. (2005). Adding linguistic annotation. , 17-29, Oxbow Books, Oxford.

Matthews, P. H. (2007). *The concise Oxford dictionary of linguistics*. Oxford University Press.

Mikulova, M., & Stepanek, J. (2010). Ways Of Evaluation Of The Annotators In Building The Prague Czech-English Dependency Treebank. In *LREC*.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, *3*(4), 235-244.

Mohanan, T. (1994). *Argument structure in Hindi*. Center for the Study of Language (CSLI).

Raza, G. (2010). Inferring Subcat Frames of Verbs in Urdu. In *LREC*.

Raza, G. (2011). *Subcategorization acquisition and classes of predication in Urdu* (Doctoral dissertation).

Schmidt, R. L. (2013). *Urdu, an Essential Grammar*. Psychology Press.

Skut, W., Krenn, B., Brants, T., & Uszkoreit, H. (1997, March). An annotation scheme for free word order languages. In *Proceedings of the fifth conference on Applied natural language processing* (pp. 88-95). Association for Computational Linguistics.

Stevenson, A. (Ed.). (2010). *Oxford dictionary of English*. Oxford University Press, USA.

Urooj, S., Hussain, S., Adeeba, F., Jabeen, F., & Parveen, R. (2012). CLE Urdu digest corpus. *LANGUAGE & TECHNOLOGY*, 47.

Zia, T, Akhtar, M. P., Abbas, Q. (2015a). Comparative Study of Feature Selection Approaches for Urdu Text Categorization. *Malaysian Journal of Computer Science*, *28*(2).

Zia, T., Abbas, Q., & Akhtar, M. P. (2015b). Evaluation of Feature Selection Approaches for Urdu Text Categorization. *International Journal of Intelligent Systems and Applications*, *7*(6), 33.