Acta Linguistica Asiatica

Volume 6, Issue 1, 2016

ACTA LINGUISTICA ASIATICA
Volume 6, Issue 1, 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

Just before summer, when the air around university is filled with students' adrenalin due to numerous tests, we are pleased to announce our summer ALA issue. It was compiled bearing in mind that the outcome of such efforts is mainly students' of course, however, ours also; the outcome of teachers and researchers. In a very broad sense, this issue places importance on a successful second language pedagogical process, be it readability, pronunciation, generalization and application of grammatical rules, or their methodological issues. It supports the idea that reciprocal improvements on students' as well as teachers' and researchers' sides undoubtedly deliver best results in the language pedagogy as well as in linguistic research. Improvements that build upon expertise and considerable amount of real-life data. Improvements aspired to.

Kristina HMELJAK SANGAWA in her article analized a collection of Japanese texts which had been linguistically simplified for learners of Japanese as a foreign language, and compared them to their original versions. The main aim of such analysis was to uncover different strategies that are used to make texts more accessible to learners. The author, however, makes some further steps and discusses the application of such strategies to assessing, selecting, and devising texts in a language classroom.

Zuzana POSPĚCHOVÁ offers a detailed introduction to the method of prosodic transcription (PTR) for Standard Chinese established by phonetician Oldřich Švarný. The PTR method has taken several decades to form and it is nowadays a well established way of teaching Chinese prosody in the language courses around the Czech Republic. The article offers a short sample text, students' opinion on PTR, and an outline of the use of PTR in academic research. It concludes with the suggestion that PTR could be an international system of transcription capturing prosodic features worldwide.

The idea in **Mateja PETROVČIČ**'s article also emerged from her experience with students of Chinese as a second language and their problems in the learning process. She highlights the so called *liheci*, a special type of Chinese polymorphemic verbs. Such verbs are known to sometimes accept one or more elements to infuse in between their morphemes, however, the author points out that word sketches such as Sketch Engine hardly offer any information on the behaviour of such words. She gives suggestions on how to include them.

Liulin ZHANG offers a discussion on the two commonly recognized imperfective aspect markers in Mandarin Chinese *zai* and 着 *zhe*, and argues their qualifications as imperfective aspect markers based on the differences in their origins, historical evolutions, and corpus data.

Alexander AKULOV is critical towards the methods in comparative linguistics that base on the characteristics of lexems of the compared languages. He points out that

such methods do not suppose verification and therefore allow different, even opposing conclusions. In his article he suggests the comparison of grammars of the languages involved, and by using Prefixation Ability Index (PAI) and Verbal Grammar Correlation Index (VGCI) tackles the problem of the Buyeo language group. His findings prove that Japanese and Korean belong to the same language group, and not just to the same language family.

Finally, **Pankaj DWIVEDI** and **Somdev KAR** contributed a survey article on a Hindi dialect called Kanauji. The article exposes problems researchers have to deal with on the field when monitoring and documenting spoken language of a certain area, and fitting the findings into concepts such as a language and a dialect.

Nina Golob

RESEARCH ARTICLES

AN ANALYSIS OF SIMPLIFICATION STRATEGIES IN A READING TEXTBOOK OF JAPANESE AS A FOREIGN LANGUAGE

Kristina HMELJAK SANGAWA
University of Ljubljana, Slovenia
kristina.hmeljak@guest.arnes.si

Abstract

Reading is one of the bases of second language learning, and it can be most effective when the linguistic difficulty of the text matches the reader's level of language proficiency. The present paper reviews previous research on the readability and simplification of Japanese texts, and presents an analysis of a collection of simplified texts for learners of Japanese as a foreign language. The simplified texts are compared to their original versions to uncover different strategies used to make the texts more accessible to learners. The list of strategies thus obtained can serve as useful guidelines for assessing, selecting, and devising texts for learners of Japanese as a foreign language.

Keywords: readability; simplification; Japanese as a foreign language; textbook analysis; reading

Povzetek

Branje je eno od temeljev učenja drugega ali tujega jezika in je lahko posebej učinkovito, ko jezikovna težavnost besedila ustreza bralčevemu nivoju jezikovnega znanja. Članek nudi pregled dosedanjih raziskav na področju berljivosti in poenostavljanja japonskih besedil ter predstavlja analizo zbirke poenostavljenih besedil za učence japonščine kot tujega jezika. Iz primerjave poenostavljenih besedil z njihovo originalno različico izhaja seznam različnih strategij, ki so jih pisci uporabili, da izboljšajo dostopnost besedil za učence. Seznam teh strategij lahko služi kot iztočnica za ocenjevanje, izbiranje in sestavljanje besedil za učence japonščine kot tujega jezika.

Ključne besede: berljivost; poenostavljanje; japonščina kot tuj jezik; analiza učbenika; branje

1 Introduction

Reading is one of the bases of second language learning, and it can be most effective for the purpose of improving a reader's language skills when the text being read is not only appealing but also of the appropriate difficulty level for its reader. The development of reading skills through extensive reading can be supported on one hand by selecting appropriate material from existing texts and grading them according to objective or subjective readability criteria; and on the other hand by

Acta Linguistica Asiatica, 6(1), 2016.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.6.1.9-33

adapting existing texts to the level of the intended audience, i.e. simplifying and abridging existing material. Both approaches have been extensively researched and implemented for major languages, especially English (see for example DuBay 2004 for an overview), and some research has been conducted on the readability and simplification of Japanese texts. However, factors affecting readability for learners of Japanese as a second language have not been thoroughly researched yet.

In the present paper, after reviewing previous research on the readability and simplification of Japanese texts, a collection of texts that have been simplified for L2 Japanese learners is analysed and compared to the original Japanese texts from which the simplified versions were adapted, in order to investigate text characteristics that could be considered important factors in determining the readability of a text for readers of Japanese as a foreign language.

2 Readability and simplification of Japanese texts: previous research

Research in this area stems from different backgrounds and is targeted at different groups of weak readers. Some early work does not clearly specify in which context and by whom the measure is meant to be used, but most research is targeted either at young native speakers or at persons with disabilities.

Probably the earliest work on readability of Japanese texts (Morioka, 1952), inspired by Flesch's Reading Ease formula, reports on preliminary research at the National Institute for the Japanese Language to determine the criteria needed to develop a similar formula for Japanese. Regrettably, the project seems to have been discontinued.

Another early attempt at measuring the readability of Japanese texts was made by Sakamoto (1962), who manually analysed Japanese language textbooks for elementary school grades 1 to 6, using school grades as the scale of difficulty, and found that the ratio of frequent vocabulary, sentence length and the proportion of kanji characters in the text correlate with school grades.

A similar way of estimating the difficulty of written sentences is also proposed in a writing stylebook by Yasumoto (1983), who uses the average number of characters per sentence and the percentage of Chinese characters as indicators of text difficulty, but does not combine these two factors into a single formula.

Two decades after Sakamoto's research, when computers were already available for lengthier calculations, Tateishi et al. (1988a, 1989b) proposed the first readability formula for Japanese on the basis of four surface characteristics: the proportion of types of characters (Roman letters, hiragana, katakana and kanji); the length of continuous strings of the same type of character; the length of sentences; and the number of commas per sentence.

A more recent and very productive stream of research is work on readability formulae to predict the difficulty level of texts for Japanese school-children, to be

used in mother tongue education (Shibasaki & Tamaoka, 2010). Other formulae have also been developed by Sato et al. (2008, for young native speakers of Japanese), and Lee & Hasebe (2016, for learners of Japanese as a foreign language).

Another approach to the readability of Japanese, proposed by Sano and Maruyama (2008) is based on Halliday's concept of lexical density within the framework of Systemic Functional Grammar (Halliday, 1993). In this approach, lexical complexity is defined as the ratio of content words to ranking clauses in a text.

A second stream of research on Japanese readability is work on information accessibility and text simplification, aimed at facilitating communication with handicapped and elderly readers (Ichikawa, 2006), and paraphrase generation to assist handicapped readers with limited linguistic capabilities (Yamamoto et al., 2000, Inui and Yamamoto 2001, Inui and Fujita 2004, Nakano et al. 2005, Sato et al. 2004).

A third stream of research which bears on readability is work on computer-aided text revision, where readability criteria are used to highlight potentially incomprehensible passages and suggest more readable substitutions (Hayashi, 1992, Inui and Okada 2000, Ono et al. 2006, Oono and Inazumi 2007). These projects often use advice on clear writing from style manuals such as Kabashima 1979, Kinoshita 1981, Honda 1982, Mishima 1990 etc., which do not deal with numerical measurements of readability, but give hints on what factors can affect readability and should be considered in its measurement.

Linguistic factors which have been found to correlate with text readability in previous research can be divided according to the traditional levels of linguistic analysis: script (ratio of character type, punctuation, phonetic guides etc.), vocabulary, syntax (sentence length, clause length, ellpisis etc.), text and discourse (length, cohesion etc.). Statistical correlations between these factors and collections of graded texts have been described in previous research. However, the factors influencing the readability of a text for learners of Japanese have not been yet thoroughly researched. The following sections present a comparison between simplified texts and their originals and an analysis of the strategies used in this process.

3 Data

The texts are reading passages in a textbook for intermediate learners of Japanese, the second in the set of textbooks developed by the International Student Center of Sanno University: 日本語を楽しく読む本・中級 *Enjoyable task reading in Japanese: Intermediate*, published by Bonjinsha in 1991 and reprinted multiple times, a popular textbook for reading instruction.

These texts were chosen because they are one of the very few available collections of pairs of authentic and simplified Japanese texts targeted at foreign learners of Japanese.

The reading passages included in the textbook were selected and simplified by the textbook authors, experienced teachers of Japanese as a foreign language. Selection criteria, as stated in the foreword, were: content (that should be interesting to adult learners of Japanese: worth reading, intellectually challenging,), text type (as varied as possible, including narratives, expository and scientific writing, in order to offer learners the opportunity of practicing different reading strategies). Another criterion that is not stated in the foreword but was evidently applied, is length: not exceeding the length that can be read in a 90-minute lesson. The longest texts are approximately 1300 characters long, spanning one to two pages.

The foreword mentions that texts were rewritten for their target audience, learners of Japanese, while the afterword mentions that all textbook material was developed and used for two years in the Japanese course of Sanno University before being re-edited for publication in book form. Only vocabulary is mentioned in the foreword as a simplification criterion, but it is conceivable that strategies applied to text rewriting were based on the authors' experience as language teachers and empirically verified or found to be useful in their language classes.

The textbook has been used by the present author for Japanese language instruction in a class of 2nd year students of Japanese and received a positive response from the students, indicating that it is a good example of readable writing for students of Japanese. The pairs of texts were also given to read to a group of eight advanced learners of Japanese, who were asked to choose the easier in a pair of texts, the original and its simplified version. All participating students indicated the simplified versions as the easier to read.

In the forewords to each textbook in the series (この本を使う先生へ *To the teachers using this book*), the authors mention vocabulary as the main criterion used both when assessing the difficulty of texts included and also when rewriting texts for their intended readers. All texts in this textbook series are graded from one to three stars (one star indicating the easiest texts and three stars indicating the most difficult texts) according to the percentage of vocabulary included in the text but not present in the vocabulary lists used as a yardstick, and thus expected not to be known by readers at the given level. These percentages and vocabulary lists are shown in Table 1 on the next page.

Table 1: Vocabulary lists used as yardstick and percentage of new vocabulary

Textbook level	No. of expected known words	Vocabulary list used as yardstick	Words not covered by yardstick list in passages marked by stars:		
			☆	☆☆	☆☆☆
Pre-intermediate (1996)	2000	list of 2030 words in Nihongo kyōiku no tame no kihon goi chōsa (National Language Research Institute 1984)	5-6%	6-10%	10-20%
Intermediate (1991)	3500	unpublished vocabulary list compiled by the authors' research team	5%	5-10%	10-15%
Pre-advanced (1993)	6000	list of 2030 words in Nihongo kyōiku no tame no kihon goi chōsa (National Language Research Institute, 1984)	up to 30%	up to 30%	30% or more
		list of 6000 words in Nihongo kyōiku no tame no kihon goi chōsa (National Language Research Institute, 1984)	up to 15%	15% or more	15% or more

In the preface to the first volume the authors explicitly mention that vocabulary was overall the main criterion used in assessing text difficulty, while adding that the average length of sentences was also used as a secondary indicator of structural complexity, but no concrete data are given for these aspects of complexity. In the prefaces to the second and third volume in the series, only vocabulary is mentioned as the yardstick for assessing text difficulty.

Similarly, the level of proficiency which is expected from readers of each of the three volumes is defined in terms of hours (or months) of Japanese instruction received, which is supposed to reflect their vocabulary knowledge: readers who have studies Japanese for a certain period of time are expected to know a certain number of words, which should approximately correspond to the vocabulary prescribed for a certain level of the Japanese Language Proficiency Test (JF and AIEJ, 2004).

Table 2 shows the number of words which readers (learners of Japanese) are expected to know after different periods of study, as stated in the forewords.

Table 2: Expected proficiency of learners of Japanese using the textbook series Enjoyable Task Reading in Japanese

Volume	Expected Japanese instruction time	Expected No. of known words	Expected JLPT level of vocabulary knowledge
Pre-intermediate	7 months – 1 year or 400 - 600 hours	2000–3500	2nd–3rd level
Intermediate	9 months – 1 year or 450 - 600 hours	3500–5000	2nd–3rd level
Pre-advanced	1 year – 1 year and a half or 600 – 900 hours	5000–10000	1st level

As can be seen from the above descriptions, the authors of the textbooks have carefully controlled the vocabulary used in the reading passages, considering it the main factor of text difficulty.

Each reading passage is also preceded by lists of 10 to 20 keywords used in the text, with exercises to learn or reinforce vocabulary knowledge, including written form (Chinese characters), morphological, syntactic and collocational patterns, again emphasising the importance of depth and breadth of vocabulary knowledge for reading comprehension.

The textbook is divided into 9 chapters, each chapter containing one or two reading passages. All reading passages were analysed, except reading no. 4, where the original text was in English and only the simplified text used as reading material was in Japanese.

The following table presents the data used in this analysis: the titles of the simplified passages as they appear in the textbook, the title of their originals, the length of both (expressed in number of characters) and their sources.

Table 3: Analysed data: simplified texts in *Enjoyable Task Reading - Intermediate* and their originals (length expressed in number of characters)

Chap.	**Title of simplified text**	**Length**	**Original title**	**Length**	**Source**
1	禁酒	449	禁酒	352	『月刊アサヒ』1989 年 6 月号、朝日新聞社 p. 342
2	いつもとちがう	682	いつもとちがう	527	『ポケットジョーク 9 トラベル』(植松黎編訳)　角川書店 p. 18

Chap.	Title of simplified text	Length	Original title	Length	Source
3	賢い農夫	1327	賢い百姓	1285	矢崎源九郎『世界の民話』（矢崎源九郎訳編） 社会思想社 pp. 255~256
5	振り向き賃	526	振り向き賃	909	深田祐介『ちょっといい旅いい話』（深田祐介監修）旅行開発株式会社ジャルパック出版センターpp. 188-189
6	ママ、手が凍るんだよ行康君、手すり変わるよ	707 829	ママ、手が凍るんだよ行康君、手すり変わるよ	702 922	朝日新聞 1985 年 2 月 20 日朝刊 朝日新聞 1985 年 3 月 1 日朝刊
7	みんなって何人？	1277	みんなって何人？	2640	斉藤勇『人間関係の分解図』誠信書房 pp. 73~80.
8	握手１ 握手２	1147 851	握手１ 握手２	2408 1199	阿部謹也『逆光のなかの中世』日本エディタースクール出版部 pp. 97~100 樋口清之『日本の風俗の謎』大和書房 pp. 34~35
9	趣味	1219	趣味	1972	守屋毅『日本文明 77 の鍵』（梅悼忠夫編）創元社 pp. 136~138
	Total:	**9014**	**Total:**	**12916**	

4 Procedure and results

All pairs of original and simplified texts were scanned, OCR-processed and the resulting files were manually checked to correct OCR errors. Pairs of files were then automatically compared using the document comparison software JDiff X (Matsumoto, 2010), all differences found were transcribed into a spreadsheet file and marked according to type, linguistic level and content of modification. Modifications within the same sentence or clause which stem from different rewriting strategies, or are carried out at different linguistic levels, were counted separately. For example, the following rewriting of one original sentence into shorter sentences involved multiple strategies at distinct linguistic levels.

Original sentence:

「亡くなった飲み友だちと約束してね、僕が飲みに行くときは、必ずオレの分も注文して飲んでくれという遺言を実践しているんだ」

Literally: *Having made a promise to a drinking pal who's dead, I'm executing his will that says, "always order my part too and drink it" when I go drinking*. – Quotation marks are not used in Japanese.

Equivalent modified sentences:

「先週、僕の親友が亡くなったんだが、彼が亡くなる前に約束してね」

Literally: *A good friend of mine died last week, and having made a promise before he died* ...

「はあ」

Literally: *Oh.*

「僕が飲みに行く時は、必ず、彼の分も注文して飲むということになったんだ。それで、その約束を実行しているってわけさ」

Literally: ... *it was decided that when I go drinking, I would always order his part too and drink it.*

Firstly, one strategy was simplification at the syntactic level: both adnominal clauses in the first and last part of the original complex sentence (亡くなった飲み友だち *drinking pal who died*; ... オレの分も注文して飲んでくれという遺言... *the will that states always order my part too and drink it*...) were split into separate simple sentences, avoiding adnominal modification, a known hurdle for learners of Japanese.

Secondly, a simplification at the discourse level retained the same entity (僕, the first person narrator) as the subject of all clauses, avoiding the shift from the first person narrator (僕 *boku - I*) to the dead friend (using a more informal first person pronoun (オレ *ore - I*) within a short clause of reported speech (not marked by quotes: 必ずオレの分も注文して飲んでくれ *ore no bun mo chuumon shite nonde kure - always order my part too and drink it*) and then back to the first person narrator (遺言を実践している *yuigon o jissen shite iru - I am executing his will*), which could be confusing. This simplification also brought with it the omission of the very informal pronoun オレ (*ore - I*), thus resulting in a standardisation of register.

Thirdly, at the semantic level, two pieces of information were made more explicit: a concrete time setting (先週 *senshuu - last week*) was added, and the pronoun 僕の (*boku no - my*) was added to the noun 飲み友だち／親友 (*nomitomodachi / shin'yuu - drinking pal / good friend*).

Fourthly, at the level of vocabulary, three simplifications were made by substituting the less common word 飲み友だち (*drinking pal*) with a less specific but

more common one: 親友 (*good friend*), and the words 遺言を実践 (*yuigon o jissen - execute a will*) with 約束を実行 (*yakusoku o jikkô - keep a promise*).

Fifthly, two explicitations occurred at the script and punctuation level: the word とき (*toki - when*) written in hiragana was rewritten with its commonly used and unambiguous Chinese character 時, and a comma was added after the adverb 必ず (*kanarazu - always*).

One further boundary explicitation was carried out by inserting a back-channelling expression (はあ *Haa - Oh*) by the other participant in the conversation, in the middle of the longest remaining sentence.

In all such cases of multiple modifications, each modification was counted and transcribed separately, resulting in a list of 815 modification occurrences in the whole corpus. All transcribed modifications are reported in Appendix 1. Repeated modifications of the same item: e.g. the rewriting of わたし as 私 for three times in the same text was counted as 3 modifications; in such cases, the modification was transcribed once and the number of modifications was noted in the second column of the table in the appendix.

5 Analysis

Differences found between the simplified texts and their originals were grouped into three categories: simplification (including deletion), explicitation and standardisation (including visualisation). Strategies belonging to these categories were found to be used at different levels of linguistic analysis: from script, to vocabulary, morphology, syntax, semantics, to discourse and style. Let us consider each one in turn.

5.1 Simplification

Strategies of simplification were the most commonly used, amounting to 472 (of which 96 deletions) out of the 815 modifications found.

5.1.1 Script simplification

Script simplification occurred a few times, where non-standard or low-frequent Chinese characters were rewritten with hiragana:

> 尋ねた。 → たずねた。; 年を取った → 年をとった; 全然 → ぜんぜん; 土産 → おみやげ; 嫌がります → いやがります; 出来た → できた; 皆様 → みなさん; 頑張った → がんばった; 頑張りなさい → がんばりなさい; 暖かく → あたたかくて; 間もなく → まもなく; 誰と誰 → だれとだれ; 石鹸 → せっけん; 挨拶 → あいさつ.

5.1.2 Vocabulary simplification

Vocabulary simplification was the most frequent of all changes, obtained by:

- substituting less common with more common **content words**:

  誰しも → 多くの人が; 珍奇な → 珍しい; 花をさかす → 花を作る; 法則が確立される → 法則が発見された; 奇妙な感慨にとらわれている → 不思議に思った; 育種学 → 植物学; 事柄 → こと; 実用にうつして → 使って; 食糧 → 食べ物; 増産する → 生産を増やす; 奇怪な → 珍しい; 熱中した → 夢中になった; 高度な → 難しい; 数式 → 数学の問題; 受容する → 受け入れる; しかるべき先生 → 専門の先生; 特殊な階層 → 特殊な人々; 介して → 通して; 大衆化し → 広がり; 国民的な趣味に拡張した → 国民的な趣味になった; その道の達人 → 一流の人たち; 生活のたし → 生活の助け; 先生連中 → 先生たち; 一覧表 → リスト; 現在 → 今; 盛況をきわめている → 流行している; カルチャー・センターのカリキュラム → カルチャー・センターの内容; かたっぱしから → 何でも; 都市大衆 → 庶民; 愛好した → 愛した; 絆 → 関係; 手をさしのべる → 手を差し出す; 仕来り → 習慣; 述べる → 言う; 口上 → 言葉; 大変巧みに → じょうずに; 握手にこたえて → 握手して; 帰宅した → 家に帰った; あわてて → 急いで; ぬぐって → ふいて; 古来の → 昔からの; etc.

- substituting less frequent **functional vocabulary** (which is usually learned later in courses of Japanese as a foreign language), with more basic expressions:

  のみ → だけ; かならずしも…というわけではない → …ばかりではない; といえば → は; しかも → そして; といったような → というような; のもとで → では; において → では; にもかかわらず → けれども; のおかげであった → があったからだ; にあたっており → である; ようにおもわれる → と思う; …にせよ、…にせよ → 例えば、…でも、…でも; に対しても → にも; やたらと…したがる → すぐに…したがる; のだが → のに; 通っているかにみせねばならない→ 通っているようにみせなければならない; …すら → …も; 当然と言えば当然でしょう → 当然かもしれません; 帰りがけ → 帰る時; …握られっぱなしで → その間…握られていて; やたらと → すぐに; etc.

Difficult words are sometimes substituted with an explanation or definition: 古典 → 昔の文学作品; sometimes even substituted with words with a completely different meaning, if it does not change the overall gist of the text, such as 語学 instead of the less frequent word 手工芸 as an example of a hobby:

現代日本人も、あらたにみいだした文化を、テニスにせよ、手工芸にせよ、かたっぱしから「お稽古ごと」にしてしまう → 現代日本人も、現代

の文化を（例えば、テニスでも、語学でも）、何でも「お稽古ごと」にしてしまう。

Often, the substitution of a vocabulary item brings with it also syntactical modifications, such as:

- changes in part of speech, e.g. from noun to verb and vice-versa: の教授をして → を教えて; あらたにみいだした文化 → 現代の文化; ヨーロッパ伝来の → ヨーロッパから入って来た; 複数の相手のばあいには → 相手が何人かいる場合は; 婦人が優先される → 婦人たちから始める; 右手を失った人 → 右手のない人; or from verb to adverb to express a modal meaning: 相手はダメダメと首をふるに決まっています。 → 相手は、必ず、ダメダメと首を振ります; or from noun to adverb: 文化の実態が…であった → 文化が、実は、…であった;
- shortening from phrase to word, e.g.:

  まるでいいあわせたかのように → 必ず;
- argument distribution, resulting in the use of different particles:

  日本が西洋とあらためて接触した → 日本に西洋文化が入って来た; 日本人の間にもかなり浸透している → 日本人の間でもかなり一般的になっている; 間をもたせるのに苦労する → どうしたらいいか困る; 石鹸だらけであった → せっけんがたくさんついていた; 一体となる意識を目覚めさせ → 相手と一つだという意識を持つため; 四枚目ので女房を養います → ４枚目のは女房のために使います. The substitution of a difficult verb (ものがたる) with a basic verb (わかる) here results in a change in subject, from inanimate to animate: その文化の実態が趣味の「お稽占ごと」であったことを、この一覧表がものがたっている。 → このリストを見ると、江戸の庶民文化が、実は、「お稽古ごと」であったことがわかる。

Sometimes, vocabulary simplification is carried out by means of a paraphrase, where part of the original meaning is lost: 改めて見直し、考えさせられるチャンスに恵まれます。 → もう一度考えさせられます。; 私の中に、価値観の逆転が起きた → 私の考え方が変わった.

Vocabulary simplification at times also involves a change in cohesive devices, such as deictics instead of synonyms or paraphrases: 江戸時代に趣味として蓄積してきた高度な教養 → このような高度な教養; 店のおやじさん → 彼.

5.1.3 Morphological simplification

Morphological simplification could be seen in the use of more basic grammatical forms instead of markedly formal or markedly colloquial forms, such as:

- substituting the formal truncated connective form of affirmative predicates with the te-form: 会社を出　→　会社を出て; 冬暖かく、夏は気持ちいい　→　冬はあたたかくて、夏は気持ちいい; and substituting the formal negative connective form ずに with the general ないで: 言わずに　→　言わないで; 気づかずに　→　気がつかないで; せず、　→　しなかった。; and generally using basic instead of very formal forms: 同じく　→　同じように; みせねばならない　→　みせなければならない; になることでしょう。　→　になると思います。

5.1.4 Syntactic simplification

Syntactic simplification was also frequent, by means of:

- dividing sentences with coordinate clauses into separate, shorter sentences:

  きっと私も、この子がいなかったら、気づかずに、そのまま生きて行ったんだろうな、と思い、最近、行康に「ありがとう」としみじみ言ったものでした。　→　きっと私も、この子がいなかったら、気がつかないで、そのまま生きて行っただろう、と思います。行康には心から「ありがとう」と言いたい。

  彼女はあわてて手の甲をぬぐって私に右手の甲をさし出し、お互いの右手の甲を合わせて握手の代わりとしたことがあった。　→　彼女は急いで手の甲をふいて右手の甲を差し出した。お互いの右手の甲を合わせて握手の代わりにしたことがあった。

  握手は古来誓いや約定を固めるためになされた法行為であって、手の平の柔らかい部分を互いに合わせて一体となる意識を目覚めさせ、契約を公示する行為とされてきた。　→　握手は昔から契約の成立を確認するための行為であった。つまり、法律的な行為であった。手の平の柔らかい部分を合わせるのは、相手と一つだという意識を持つためだった。

  封建制のもとで領主に臣従する家臣は両手を合わせ、領主がその手を自分の手でつつんだとき、臣従契約が成立した。　→　封建時代では、家臣になりたい人間が両手を合わせる。そして、主人がその手を自分の手で包むと、主従の契約が成立した。

- separating subordinate clauses and turning them into separate sentences, especially in the case of adnominal modifiers:

  日本の伝統的な文化とされる茶道にせよ華道にせよ、いずれも、もともと特殊な階層の文化であったものが、こうした「お稽古ごと」を介して大衆化し、国民的な趣味に拡張したものなのである。　→　日本の伝統的な文化である茶道も、もともとは、特殊な人々の文化であった。が、こうした「お稽古ごと」を通して広がり、国民的な趣味になったものなのである。

The separation of complex sentences into shorter, simpler ones, at times resulted in (or was motivated by a desire of) bringing subject and predicate of the original complex sentence nearer to each other:

この実験は、人が集団の圧力を感じ、集団に従おう、あるいは従わざるをえないというような気持ちになるのは、三人以上であることを明らかにしています。 → この実験から、次のようなことがわかります。私たちが、集団の圧力を感じて他の人と同じ意見を言うようになるのは、３人以上の人が同じことを言った時であること。

5.1.5 Discourse simplification

Discourse simplification was obtained in three cases by maintaining topic continuity and avoiding topic-shift from one agent to another:

必ずオレの分も注文して飲んでくれという遺言を実践しているんだ → 必ず、彼の分も注文して飲むということになったんだ。それで、その約束を実行しているってわけさ;

四枚目のので女房を養いますが、女房はわたしに何の利益ももたらしませんから、これは捨てることになります。 → ４枚目のは女房のために使いますが、これは利益にはなりませんから、捨てることになります;

さて、大臣たちが集まったとき、王さまは → お城に帰ると、大臣たちを集めて、王さまは、こう言いました。

5.1.6 Deletion

Deletion was the most drastic form of simplification, used quite often: 95 instances of deletion were found, including deletion of:

- modal forms, e.g.:

  どうも土産が足りない → おみやげが足りない;

- intensifiers, e.g.:

  それよりかなり以前 → それ以前; 何の利益ももたらしません → 利益にはなりません; ただの一度でも → 一度でも

- aspectual forms, e.g.:

  捨ててしまいます → 捨てます; 引き返していって → 引き返して;

- extra-textual references, e.g.:

  「ママ、手が凍るんだよ」(一二、一四ページ)を読んで → 「ママ、手が凍るんだよ」を読んで

- rhetorical devices, e.g.:

  誤解のないようにいっておくが → しかし

- semantically redundant, non-essential information or details, e.g.:

  林野庁業務部販売推進室長羽賀正雄さん → 林野庁業務部の羽賀正雄さん; 見知らぬ十二歳の坊やの姿に、八十歳を超した両親の光景を重ね合わせた → 80 歳を越したいなかの両親のことを思い出した; 越後からやってくるたび、東京の街なかで → 東京に出てくると; 江戸時代の都市民の文化 → 江戸時代の文化; 法則の基礎知識をもっていた → 法則を知っていた;

- redundant paraphrases, or information that can be inferred from the context, e.g.:

  「みんなではなく、たった三人じゃないか」 → 「３人だけ？」; その趣味の代表的な事例 → その例; …に出版された…案内書 → …の…案内書; 都市にすむ学者 → 都市の学者.

In some instances, whole paragraphs were omitted, such as the underlined part in the following example, which includes culturally-bound terms, where not only the words, but also the words' referents are probably not known to learners, and at the same time, being only exemplifications of the previous general statement, are not essential to convey the general meaning of the passage:

古典には、手を握り合って飛び上がったとか、手を取り合って喜んだといった文章がたくさん出てきます。<u>『古事記』にもあれば、さらに軍記文学にも、家臣たちが成功して帰ってくると武将が手を取って喜んでくれる、という場面がよく見られます。</u>手を取るとは、つまり握手です。

→ 昔の文学作品には、手を握り合って飛び上がったとか、手を取り合って喜んだといった表現がたくさん出てきます。手を取るとは、つまり握手です。

5.2 Explicitation

Strategies of explicitation were also very frequent: 203 instances of explicitation were found, encompassing all linguistic levels.

5.2.1 Semantic explicitation

Semantic explicitation was the most common, occurring in 110 cases, such as:

- adding concrete time or place settings, sometimes even with an extensive description:

ある男が、バーのカウンターにすわった。 → ある男が、バーのカウンターにすわった。はじめての男だった。; 思わず、言った。 → 大田区役所へ行った時、思わず、こう言った。; では二杯、ご用意しましょう → では、２杯いっしょに、ご用意いたしましょう; やっぱり後ろから → この時も、やっぱり後ろから; ニューヨークで小さな衣料品会社をやっているスタインバーグ氏は、もう二十年間も毎日同じレストランで昼食を食べている。 → 佐藤さんは、東京の銀座で、小さな出版会社を経営しているが、なんでもきちょうめんである。仕事もきちんと片づけるし、食事も同じである。佐藤さんは、昼に、もう 20 年も同じ店で同じものを食べている。

- using a more specific word instead of a more general hypernym, e.g.:

  とうとう、 → 次の日、; 連絡をとった → 電話をした; 店のおやじさん → 店の主人; こいつ → この客; やっている → 経営している

- using a hypernym and adding a definition:

  … ためになされた法行為であって → … ための行為であった。つまり、法律的な行為であった; or adding a hypernym to a word that readers might not know, instead of a definition: 水割りを → ウィスキー、水割りを;

- in one case even adding loan-word synonyms as furigana (here written in parentheses):

  利己的遺伝子の乗り物 → 利己的遺伝子［セルフッシュジーン］の乗り物［ヴィークル］

Almost half of the semantic explicitations occurred at the **script** level (42 cases): words that were written in hiragana in the original text, but can be and usually are written with Chinese characters, were rewritten using these characters, which made them less ambiguous, both visually, in terms of word delimitation, and semantically, distinguishing between homophones:

いっている → 言っている; いままで → 今まで; とき → 時; ばあい → 場合; ひとり → 一人; もっている → 持っている; らくに暮らせる → 楽に暮らせる; わたし → 私; よんで → 呼んで, etc.

5.2.2 Boundary explicitation

Boundary explicitation was the next most common type of explicitation (70 cases), mostly by means of added punctuation:

- adding commas to separate ambiguous or just long strings of hiragana:

もう二十年間もうちのスープを　→　もう 20 年も、うちのスープを; 男はそれからもときどき　→　男は、それからも、ときどき; ところがある日　→　ところが、ある日; あれお客さん　→　あれ、お客さん;

- adding commas to separate ambiguous strings of kanji:

  実際手は　→　実際、手は;

- adding commas to separate phrases, often topical phrases:

  必ずオレの分も　→　必ず、彼の分も; いったいどうしたというんですか？　→　いったい、どうなさったんですか。; わたしは飲まなくったって　→　私は、飲まなくても; 大人でも似たようなもの　→　大人でも、同じようなもの;

- adding commas between clauses:

  時間をやりくりして大田区役所へ…　→　時間を見つけて、その地下道へ…

- adding parentheses for emphasis or reported speech:

  三人というのはみんななのです　→　３人というのは「みんな」なのです; みんながゴルフをはじめたので……というとき　→　「みんながゴルフを始めたので…」という時

- dividing one paragraph into two: ...」

  と百姓は答えました。王さまは　→　…」と農夫は答えました。<p> 王さまは; という話になる。ところが　→　という話になる。<p> ところが; 「さっきあなたはー、ooo円とおっしゃったけれど、二、〇〇〇円でいかがでしょう」となる。振り向き賃が七、〇〇〇円についちまうわけです。→「さっきあなたは 1000 円とおっしゃったけれど、2000 円ではいかがでしょう」となる。<p> 振り向き賃が 7000 円になってしまうわけです。

- inserting back-channelling expressions in dialogues:

  「亡くなった飲み友だちと約束してね、僕が飲みに行くときは、必ずオレの分も注文して飲んでくれという遺言を実践しているんだ」　→　「先週、僕の親友が亡くなったんだが、彼が亡くなる前に約束してね」<p>「はあ」<p>「僕が飲みに行く時は、必ず、彼の分も注文して飲むということになったんだ。それで、その約束を実行しているってわけさ」; 「そうですか。わかりました。飲みますよ。あれ、スプーンはどこですか?」　→　「そうですか、わかりました。飲みますよ。飲めばいいんですね」<p> 「ああ」<p> 「じゃあ、飲みます。あれ、スプーンはどこですか」

Boundary explicitation was also obtained by splitting long complex sentences into shorter ones, which also implies syntactic simplification, as mentioned in the previous sub-section, and by using Chinese characters instead of hiragana where possible, as

mentioned in the previous subsection, to mark the delimitation between words, which is not marked by blank spaces in Japanese standard script.

5.2.3 Syntactic explicitation

Syntactic explicitation was obtained by:

- adding an omitted argument to a predicate:

  交配をかさねて → 朝顔を交配させて; 手袋は、滑ってしまうから、つけてはいかれない。→ 手袋をすると滑ってしまうから、手袋をすることはできない。

- substituting an intransitive expression, where all agents are not immediately obvious, with a transitive one, where agent and patient are more clear (here the semantic value of the verb is also more specific):

  一杯目が終わったら → 一杯目をお飲みになったら;

- using polite or humble forms which disambiguate the subject of the predicate:

  一杯目が終わったら → 1杯目をお飲みになったら; ご用意しましょう → ご用意いたしましょう; いったいどうしたというんですか? → いったい、どうなさったんですか。

- adding particle の to split a compound noun into a noun with a nominal modifier:

  臣従契約 → 主従の契約; 現代日本 → 現代の日本; 遺伝法則 → 遺伝の法則.

Other, less common cases of **explicitation**, were cohesion and phonetic explicitation.

5.2.4 Cohesion

Cohesion explicitation, by adding cohesive elements, eg.:

- deictics (その, この, 上の etc., such as in その結果が図です → その結果が、上の図です) or
- temporal expressions (その時, そして etc.)

5.2.5 Phonetic explicitation

Phonetic explicitation by using hiragana instead of Chinese characters, where the pronunciation of the characters could be ambiguous: その後の事 → そのあとの

事、　宝物　→　宝もの; or by adding furigana on difficult Chinese characters: 疑惑　→　疑惑［ぎわく］; 氷解　→　氷解［ひょうかい］; 爽やか　→　爽やか［さわやか］.

5.3 Standardisation

The third strategy used in the adaptation of texts for foreign language learners was the use of more standard or basic linguistic forms, i.e. forms which are usually learned at the beginning of Japanese language courses, instead of stylistically marked forms, which are usually learned later. 128 modifications were counted in this category, including the following means.

5.3.1 Script standardisation

Script standardisation, using:

- standard punctuation instead of non-standard brackets, question marks and other punctuation:

  技術や知識・芸能を　→　技術や知識や芸能を; どうかいたしましたか？　→　どうかいたしましたか。; ですか？　→　ですか。; 《皆様の励ましは、…宝物とさせていただきます》　→　『みなさんの励ましは、…宝ものとさせていただきます』; 冬の朝。金属の手すりは、氷のように凍っています。　→　冬の朝、金属の手すりは、氷のように冷たくなっています。; ので……　→　ので…; or using standard characters: 八十歳を超した　→　80歳を越した.

5.3.2 Tense levelling

Changing single predicates in non-past form to past form in texts which are otherwise written in the past form, to make the tense uniform throughout the text:

確立されるのは　→　発見されたのは;

5.3.3 Formality levelling

F**ormality** levelling from de-aru to plain style:

であった　→　だった; or from plain to formal style: あった　→　ありました; だから　→　ですから; in texts which are otherwise written in this style, to make the style uniform throughout the text

5.3.4 Formality standardisation

Substituting colloquial or otherwise register-marked forms with standard (more polite) forms, which are generally learned earlier in Japanese language courses:

> 飲まなくったって → 飲まなくても;話さん → 話さない; 飲んでみたまえ → 飲んでみてくれないか; 言え → 言わなければならない; 「…どうしてです」 → 「…どうしてですか」; ようがす。 → わかりました。

Exceptionally, in one case the opposite occurred: a colloquial, shortened form was used instead of a politer one: 振り向いてはいけない → 振り向いちゃいけない, probably for stylistic effect.

5.3.5 Visualisation:

Throughout the texts, numbers written in Chinese characters in the original texts, printed vertically, were replaced with arabic numerals in the textbook reading passages which are printed horizontally:

> 百枚 → 100 枚; 百度 → 100 回; 四枚 → ４枚目; 四十五分 → 45 分; 十二時 → 12 時; 十七世紀 → 17 世紀 六、〇〇〇円と → 6000 円と; 五年生 → 5 年生; 五十四段 → 54 段; 二枚 → ２枚; 二杯 → 2 杯; 二月二十日付の → ２月 20 日付けの; 二十年間 → ２０年間; 二、五〇〇軒 → 2500 軒; 三人 → 3 人; 三つ → ３つ; 一・三キロ → 1.3 キロ etc.

One modification was found that does not clearly belong to any of the four categories proposed, but could tentatively be categorised as standardisation, or could also be termed familiarisation or domestication. It was only used in one text: the setting of a story, originally happening in New York to a Mr. Steinberg, apparel vendor, was reset in Ginza, one of the most famous Tokyo districts, with Mr. Sato, publisher, as the main character:

> ニューヨーク → 東京の銀座; スタインバーグ → 佐藤; 小さな衣料品会社をやっている → 小さな出版会社を経営している.

5.4 Other modifications

Some modifications were found for which no clear motive could be guessed: they may have been made for stylistic purposes, according to the rewriter's tastes, or may be the results of multiple modifications, where the original motive became blurred in

subsequent modifications. One substitution was probably just a spelling mistake, resulting in a colloquial きてます instead of きます in the sentence:

おやじは必ず後を追っかけてきます。 → おやじは必ず後を追っかけてきてます。

Other cases where the motive for the substitution were not clear were:

- one separation of a clause indicating reported speech into a separate paragraph, probably for dramatic effect:

  「それ、みたまえ!」とスタインバーグ氏は言ったのである。 → 「わかったかね」<p> そう佐藤さんは言ったのである。

- the substitution of a more standard full stop with a less standard comma after a polite predicate:

  そうですか。わかりました。 → そうですか、わかりました。

- one substitution of a causal connective with a more polysemic and not easier connective:

  身体を接触させず、ただ目で意志を相手に伝えることの方が多いからである。→ 相手の体に触れない。目で意志を相手に伝える方が多いのである。

In six cases, commas separating phrases in short sentences were deleted, which is slightly surprising, given that in other 39 cases, commas were added in such positions:

階段の途中で、息子は立ち止まってしまいます。 → 階段の途中で息子は何回も止まってしまいます。; だが、それは行康君にはひどくつらいのだった。 → だがそれは行康君にはひどくつらいのだった。; 行康君は、記者に、そう言った。 → 行康君は記者にそう言った。; と、区の担当者はいっている。 → と区の担当者は言っている。; 社会部に、行康君の母親の栗田敦子さん(四五)から次のような手紙が届いた。 → 社会部に行康君の母親の栗田敦子さん(45)から次のような手紙が届いた。; たいがい、三、四人で終わりです。 → だいたい３、４人で終わりです。

6 Discussion

As could be seen in the previous section, multiple strategies were used when rewriting texts which were originally written for native speakers of Japanese, to be included in a reading textbook for intermediate learners of Japanese as a foreign language. A summary of all modifications, counting their number at different levels of linguistic analysis and by type of strategy, is given in the following tables.

Table 4: Number of modifications by level of linguistic analysis

Number of modifications by level of linguistic analysis	No. of occurrences
script	175
punctuation	56
vocabulary:	
content words	193
function words	44
vocabulary + syntax	167
morphology	20
modality	18
syntax	42
semantics	45
cohesion	10
discourse	32
formality	12
intertextuality	1
Total	**815**

Table 5: Number of modifications by strategy type

Number of modifications by strategy type	No. of occurrences	Percentage
simplification (of which 96 deletions (12% of total)	472	58%
explicitation	203	25%
standardisation (of which 80 visualisations)	128	16%
not categorised	12	1%
Total	815	100%

While the quantities of modifications on different linguistic levels cannot be objectively compared, since they refer to linguistic elements which occur in different scales of magnitude (the number of elements of vocabulary in a text is always larger than the number of phrases, clauses, sentences and paragraphs, thus making a comparison

impossible), it is still interesting to see how modifications were made on all levels, multiple times.

The modifications at different linguistic levels confirm the central role of vocabulary as declared in the foreword to the textbook and as could be inferred from the structure of the textbook containing many vocabulary exercises. It is not, however, the only level at which modifications were made: a considerable number of modifications was made at the script level, and all other linguistic levels were also touched by the modification process. Vocabulary modifications in many cases (167) brought with them also syntactic changes, and other aspects of the text were modified irrespective of the vocabulary used: many were structural modifications, touching syntax and discourse, to simplify syntactic and discourse structures or make them more explicit, cohesive devices were introduced, and stylistic changes (standardisations) were made at the level of formality.

As for type of strategy used, simplification was the most common strategy, accounting for more than half of the occurrences. It occurred at the level of vocabulary, where less frequent words were substituted with more common synonyms, explanation, definitions or paraphrases, at the level of morphology, where less common predicate forms were substituted with more basic ones, at the level of syntax, where long sentences with coordinate and subordinate clauses were split into smaller units, at the level of discourse, where roles were switched to maintain topic continuity, paragraphs of narrative were divided and shorter conversation turns introduced in dialogues.

However, alongside simplification, explicitation was another strategy that should not be overlooked, as it accounts for one quarter of the number of modifications, indicating that the authors of the rewritings considered it a useful device and found it useful when rewriting texts for their students.

It was used on the semantic level, adding information that could otherwise be inferred from the text, or cultural background that is not likely to be know to learners, or just inventing concrete settings to help readers reconstruct the narrative being told. Semantic disambiguation also occurred at the script level, where strings of hiragana were often rewritten in Chinese characters to disambiguate homophones.

Structural explicitation was also observed at different levels: boundaries between linguistic units were made clearer by the use of punctuation, script or layout; omitted predicate arguments were made explicit, polite forms were used to disambiguate the subject of the sentence, and some phonetic information was added by means of rewriting Chinese characters in hiragana or by adding furigana.

The third strategy used, standardisation, was used at the script level, to standardise punctuation and to make the text visually more familiar (using Arabic instead of Chinese numbers), and at the discourse level, both to uniform the use of the same tense or level of formality within one text (choosing one of two possibilities, such as past/non-past, or formal/informal, which are both known to learners), or to standardise the text as a whole by removing marked forms which are typical of less

standard registers (very colloquial, literary etc.) and not likely to be known by intermediate learners.

7 Conclusion and further work

Overall, it could be seen that the rewriters used some strategies which could be applied to the simplification of texts for most weak readers of Japanese, not only foreign learners of Japanese: short sentences and frequent vocabulary are two aspects of language that have been found to be easier to read in most research on readability in different languages.

However, it is also clear that the authors (rewriters) of these texts, teachers of Japanese as a foreign language, were conscious of the typical progression of formal Japanese language instruction, and tended to prefer vocabulary, morphology and syntactic structures that are learned earlier in language courses. These are very often also the most frequent forms in the Japanese language as a whole and learned earlier by Japanese children (especially in the case of content words), but some linguistic elements, such as standard polite language (as opposed to very colloquial or very formal speech) are typical of beginning language courses for foreigners, while colloquial language (including vocabulary and contracted or otherwise colloquial morphology), which is learned quite early by Japanese children, is learned later in formal language instruction and therefore relatively difficult for foreign learners of Japanese.

All the strategies which were highlighted in this analysis could be useful as guidelines when assessing the readability of texts for foreign learners of Japanese, both in an overall assessment of readability, and when devising methods and systems to pinpoint difficult aspects of particular texts as a first step to text simplification. Especially in the first case, when assessing overall readability, i.e. grading multiple text on one scale of readability, further and more extensive analysis of the weight of each of these aspects on overall readability is needed. In both cases, it would be useful to devise a system for automatic discovery and assessment of particular aspects of readability.

References

DuBay, W. H. (2004). *The principles of readability*. Costa Mesa, California: Impact Information.

Halliday, M. A. K. (1993). Some grammatical problems in scientific English. In *Writing Science: Literacy and Discursive Power*, Pittsburgh, 1993 (pp. 69-85). University of Pittsburgh Press.

Hayashi, Y. (1992). A Three-level Revision Model for Improving Japanese Bad-styled Expressions. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics* (pp. 665-671).

Honda, K. [本田勝一] (1982). *Nihongo no sakubun gijutu* 日本語の作文技術 [*Japanese writing techniques*]. Tokyo: Asahi shimbun 朝日新聞.

Ichikawa, A. [市川熹] (2006). Fukushi jouhou gaku towa nanika 福祉情報学とは何か. *Gekkan gengo* 月刊言語, 35(7), 26-34.

Inui, K. [乾健太郎] & Fujita, A. [藤田篤] (2004). Iikae gijutsu ni kansuru kenkyuu doukou 言い換え技術に関する研究動向 [Paraphrase research trends]. *Shizen gengo shori* 自然言語処理, 11(5), 151-198.

Inui, H. [乾裕子] & Okada, N. [岡田直之] (2000). Nagai bun wa tsune ni wakarinikui ka? Wakarinikusa no youin to sono izon kankei - Is a long sentence always incomprehensible? A structural analysis of readability factors 長い文は常にわかりにくいか? : わかりにくさの要因とその依存関係. *NL SIG Technical reports 情報処理学会研究報告自然言語処理*, 2000(11), 63-70. [http://id.nii.ac.jp/1001/00048682/]

Inui, K. & Satomi, Y. (2001). Corpus-based acquisition of sentence readability ranking models for deaf people. In *Natural Language Processing Pacific Rim Symposium Tokyo* (pp. 205-212).

Japan Foundation [国際交流基金] & Association of International Education Japan [日本国際教育協会] (2004). *Nihongo nouryoku shiken shutsudai kijun - Japanese language proficiency test : test content specifications* 日本語能力試験出題基準. Tokyo: Bonjinsha 凡人社.

Kabashima, S. [樺島忠雄] (1979). *Bunshou sahou jiten* 文章作法事典. Tokyo: 東京堂 Tokyodo.

Kinoshita, K. [木下是雄] (1981). *Rikakei no sakubun gijutsu* 理科系の作文技術. Tokyo: Chuokoron shinsha 中央公論新社.

Koide, K. [小出慶一] (1991). *Nihongo o manabu hitotachi no tame no Nihongo o tanoshiku yomu hon - Enjoyable Task Reading in Japanese* 日本語を学ぶ人たちのための日本語を楽しく読む本. Tokyo: Sanno University International Student Center 産能短期大学国際交流センター.

Lee, J.-H. & Hasebe, Y. (2016). Readability measurement for Japanese text based on leveled corpora. In *Papers on Japanese Language from an Empirical Perspective*, Ljubljana: Academic Publishing Division of the Faculty of Arts, Univ. of Ljubljana.

Matsumoto, S. (2010). *JDiff X Document Comparison Plug-in for Jedit X*. [http://www.artman21.com/en/jdiff_x/].

Mishima, H. [三島浩] (1990). *Gijutsusha, gakusei no tame no technical writing* 技術者・学生のためのテクニカル・ライティング. Tokyo: Kyouristu shuppan 共立出版.

Morioka, K. [森岡健二] (1952). Yomiyasusa no kiso kenkyuu「読みやすさ」の基礎的研究. In *Kokuritsu kokugo kenkyuujo nenpou - Annual report of National Language Research Institute* 国立国語研究所年報 (pp. 91-108).

Nakano, T. [中野智子], Endo, A. [遠藤淳], Sugawara Sh. [菅原昌平], Inui, K. [乾健太郎], & Fujita, A. [藤田篤] (2005). Lexical Paraphrasing for Improving Accessibility to the Web - Web サイトへのアクセシビリティ向上を目的とした難語の平易化. *IEICE technical report Welfare Information technology* 電子情報通信学会技術研究報告 WIT 福祉情報工学 25(25), 11-14.

Ono, T. [小野貴博], Suganuma, A. [菅沼明], & Taniguchi, R. [谷口倫一郎] (2006). Nihongo bunshou suikou shien ni okeru kakariuke o gokai sareru bun no chuushutsu - Extraction of the sentences whose modification relation is misunderstood for a writing tool 日本語文章推敲支援における係り受けを誤解される文の抽出. *IPSJ SIG technical reports* 情報処理学会研究報告 FI 情報学基礎, 2006(94), 99-104.

Oono, H. [大野博之] & Inazumi, H. [稲積宏誠] (2007). - Development of an education support tool for improvement of ability for sentence making 技術文章作成能力の育成を目指した教育支援ツールの開発. In 電子情報通信学会 第 18 回データ工学ワークショップ論文集 - *DEWS2007*.

Sakamoto, I. [阪本一郎] (1962). Bunshou no goi hijuu no sateihou - Readability no kenkyuu no kokoromi 文章の語彙比重の査定法---Readability の研究の試み---. *Dokusho kagaku* 読書科学, 6(1), 37-44.

Sano, M. & Maruyama, T. (2008). Lexical Density in Japanese Texts: classifying text samples in the Balanced Corpus of Contemporary Written Japanese (BCCWJ). In *Proceedings of ISFC 35: Voices Around the World*, Sydney, 2008 (pp. 359-364).

Sato, S., Utsuro, T., Tsuchiya, M., Asaoka, M., & Matsuhoshi, S. (2004). Natural Language Processing Technologies to Enhance Readability. In *Proc. of International Conference on Informatics Research for Development of Knowledge Society Infrastructure* (pp. 46-53).

Sato, S., Matsuyoshi, S., & Kondoh, Y. (2008). Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

Shibasaki, H. [柴崎秀子] & Tamaoka, K. [玉岡賀津雄] (2010). 国語教科書を基にした小・中学校の文章難易学年判定式の構築 - Constructing a formula to predict school grades 1-9 based on Japanese language school textbooks. *Nihongo kyouiku kougakukai rombunshi - Japan journal of educational technology* 日本教育工学会論文誌, 33(4), 449-458.

Tateisi, Y., Ono, Y., & Yamada, H. (1988). A computer readability formula of Japanese texts for machine scoring. In *Proceedings of the 12th conference on Computational linguistics* (pp. 649--654). Association for Computational Linguistics.

Yamamoto, S. [山本聡美], Inui, K. [乾健太郎], Nogami, M. [野上優], Fujita, A. [藤田篤], & Inui, H. [乾裕子] (2000). 聾者向け文章読解支援のための文可読性基準の調査 - Exploring the Readability Criteria for Congenitally Deaf People: A Step toward Computer-Aided Text Reading. *IPSJ SIG technical reports* 情報処理学会研究報告 NL 自然言語処理, 135(17), 127-134.

Yasumoto, B. [安本美典] (1983). *Settoku no bunshô gijutsu* [*Techniques of persuasive writing*] 説得の文章技術. Tokyo: Kôdansha 講談社.

PROSODIC TRANSCRIPTION OF STANDARD CHINESE AND ITS USE IN TEACHING

Zuzana POSPĚCHOVÁ
Department of Asian Studies, Faculty of Arts,
Palacký University Olomouc, Czech Republic
zuzana.pospechova@upol.cz

Abstract

The present paper's main aim is to introduce a method of prosodic transcription (PTR) for Standard Chinese established by Oldřich Švarný in the background of the Czech Republic. It is used to describe suprasegmental features of Standard Chinese, namely the stress prominence and linear segmentation of sentences. PTR is applied in teaching Chinese prosody in the language courses around the Czech Republic. This paper also contains a short sample text, students' opinion on PTR and an outline of the use of PTR in academic research.

Keywords: Standard Chinese; prosody; PTR; Oldřich Švarný; teaching Chinese

Povzetek

Glavni namen članka je predstaviti metodo prozodične transkripcije (PTR) za kitajski jezik, ki jo je osnoval češki sinolog Oldřich Švarný. S to metodo je moč opisati suprasegmentalne značilnosti standardne kitajščine, kamor sodita stavčni poudarek in linearna segmentacija povedi. Na Češkem se PTR uporablja pri poučevanju kitajske izgovorjave. Članek predstavi kratko, vzorčno besedilo, prouči odzive in mnenja študentov ter poda oris, kako bi lahko prozodično transkripcijo uporabljali v akademskih raziskavah.

Ključne besede: standardna kitajščina; prozodija; PTR; Oldřich Švarný; poučevanje kitajščine

1 Introduction

This paper deals with a brief introduction and a simple description of the principles and methodology of prosodic transcription (PTR), which is used to transcribe Standard Chinese. The notational system of PTR was invented by a Czech sinologist, Prof. Švarný, based on the analysis of huge amounts of audio recordings of spoken Chinese.

This transcription enables us to describe prosodic features of modern spoken Chinese. It determines various levels of stress (tone) prominence of syllables (not only stressed and unstressed) and a linear segmentation of sentences (organization of syllables into more or less closely connected binary groups and groups of more than

Acta Linguistica Asiatica, 6(1), 2016.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.6.1.35-45

two syllables). The two prosodic characteristics constitute rhythmic structure of spoken Chinese. Their transcription helps beginner students to pronounce Chinese sentences more or less correctly, and has a considerable effect on the progress of their further studies. From the experience obtained in our department[1], we can confirm the spoken realization of prosodically transcribed text is much more natural than the realization of the same text in pinyin or even in Chinese characters.

Finally, this paper also contains students' evaluation of the usefulness of PTR.

2 Basic facts about Professor Švarný and his work

In the introductory part of this paper, I would like to mention some basic facts about the outstanding Czech sinologist and phonetician Oldřich Švarný (1920–2011), whose lifelong research focused primarily on the sound structure of modern spoken Chinese and particularly on its prosodic (suprasegmental) features.

Švarný studied Chinese language and phonetics as a postgraduate at the Faculty of Arts of Charles University in Prague, where in 1952 he defended his thesis entitled *Prosodic Characteristics of Syllable and their Modifications in the Continuous Speech*. From 1951 he worked at the Oriental Institute of the Czechoslovak Academy of Sciences in Prague, where, among other projects, he established the Czech transcription of Standard Chinese, which continues to be used in press and fiction even today. In 1963 he defended his dissertation entitled *Discussing the Concept of the Morpheme and Word in Modern Spoken Chinese.* These two dissertations laid the basis for the entirety of his theoretical work.

All his life, he focused his research on phonological–syntactical analysis of frequent 'syllabosemems' (cf. morphemes), which became the basis for his lifelong work – *A Learner's Dictionary of Modern Chinese* with sixteen thousand example sentences. He published individually or with his co-workers (his former co-workers Tang Yunling Rusková and Guang Minzhe were also one of his Chinese informants, later he was working together with his student and successor David Uher).

Švarný's suprasegmental theory based on PTR continues to be used at the Faculty of Arts at Palacký University Olomouc, and his works are still the basis for teaching Chinese prosody.

3 System of prosodic transcription

As mentioned above, the PTR of Chinese was one of Švarný's main research objectives. He was influenced by other related publications, such as *Beginning Chinese* (1963) by DeFrancis, *Dictionary of Spoken Chinese* (1966) and *Fonetika kitajskogo jazyka* (1972)

[1] Department of Asian Sudies, Faculty of Arts, Palacký University Olomouc, the Czech Republic

by Spěšněv (Třísková, 2011, p. 45). Švarný and his co-workers worked on PTR for four decades and due to this long time development there are four different versions of the transcription. In spite of the fact that PTR underwent several evolutionary steps, the essential idea of all its versions is kept throughout without any substantial changes. The differences appear only in the graphic notational system.

During the whole time of development the main thought was to keep the transcription as simple as possible and express all the necessary prosodic features without becoming too cumbersome (Švarný, 1991b, p. 234). The sporadic alternations that were made to PTR in the following years were to utilize advances in technology, such as the change from hand-written to computer-typed transcription.

PTR's main aim is to capture prosodic features of spoken Chinese. These are essentially two; stress (tone) prominence and linear segmentation of sentences. As is well known, stress is one of the most important factors in Chinese, due to the fact that stress influences the modification of tones – they become weakened or emphasized. From an acoustic perspective stress prominence in Chinese can be manifested in the changes of pitch or syllable duration, or in a combination of both of these characteristics (Duběda, 2005, p. 55–56).

All rules of PTR described below were determined by Švarný on the basis of his long-term research work. The main method was a precise auditive analysis of thousands of of utterances nearing a spontaneous speech spoken by Chinese native speakers. This enabled him to develop a complete system of registering all relevant prosodic features as described below. However, this method was not determined purely by subjective auditive impression, it was also supported by objective measurements. In the late 1960's and early 1970's Švarný stayed in California at the phonetic laboratory of the University of Berkeley where he could process his speech recordings with machines, compiled graphs of a F_0 pitch. Having done so, he could support the adequacy of the set prosodic rules and the essential method of PTR.

Generally known prosodic terminology was slightly adapted for the use in European languages. Švarný preferred the term ictus to the term accent, because the quality of accent in European languages differs from the accent in Chinese, which is more flexible, e.g. it can move within a single word, changing the word's meaning at the same time. He also established a term for prosodic phrase: colon (pl. cola) and a term for prosodic word: segment. Because our research work employs the whole theory, we also accept these terms. However, the adaptation of this terminology is mostly relevant for researchers interested in Chinese prosody and PTR as a theoretical method. Students who use prosodically transcribed texts in their studies are unaffected by this term inconsistency.

PTR employs Chinese pinyin with slight modifications. Basic words and phrases in pinyin are given special marks that capture the "suprasegmental level of the utterances". In the process of transcribing, we mark rhythmical groupings of speech as first and tone syllable prominences as second.

As for the rhythmical grouping, we distinguish three types of pauses: junctures, blank spaces and breaks (Švarný 2000, p. 150). Junctures are used between binary sequences and (or) odd syllables within the frame of rhythmical segments. This kind of pauses is marked by a hyphen between the sequences (e.g. chī-kǎoyā) (Uher & Švarný, 2014, p. 11). Blank spaces are used between rhythmical segments which do not belong together and there is a short pause between them (e.g. qǐng-women-qu^{4} ta^{1}-jiā). Breaks indicate the boundaries of cola, the pause lasting approximately ¾ of a second. They are marked by common punctuation: comma, exclamation mark, point, or semicolon (e.g. zuótiān, zhāng-lao^{3}shī). Cola endings are also characterized by relative rising or relative falling intonations (Švarný, 1991b, p. 235). For foreign students it is difficult to master the prosodically correct pronunciation of relatively long cola at once. Instead, it is easier for them to acquire first the prosodically correct pronunciation of two to about four or five syllable long rhythmical segments and then join them together into cola (Švarný, 1991b, p. 236).

As for the tone syllable prominences, we distinguish stressed tone prominence, full tone prominence, weak tone prominence and atonicity. Stress tone prominence is always arse (ictus-bearing). Both full tone prominence and weakened tone prominence can be differentiated into arses (ictus-bearing syllables) and theses (non-ictus-bearing syllables). Atonic syllables are differentiated into neutralized syllables (with a slight residuum of tone) and toneless syllables (always without a tone) (Švarný, 2000, p. 150). Altogether we call this system as seven degrees of syllable prominence (Švarný, 1998, p. 24–25). Below you can see how these syllables are marked:

- stressed tone syllables: bold red font
- full tone ictus-bearing syllables: traditional tone mark
- full tone non-ictus-bearing syllables: traditional tone mark
- weakened tone ictus-bearing syllables: numerical indices$^{1/2/3/4}$ (the number is determined by the tone)
- weakened tone non-ictus-bearing syllables: numerical indices$_{1/2/3/4}$ (the number is determined by the tone)
- atonic neutralized syllables: no marking
- atonic toneless syllables: no marking

Appearing in this form, the transcription is simple and expresses all the necessary prosodic features. However, there is still a danger of confusion among full tone ictus-bearing/non-ictus-bearing syllables, weakened tone ictus-bearing/non-ictus-bearing syllables and atonic neutralized/toneless syllables. For these purposes Švarný established implicit rules:

1. If a syllable is not marked in any way and occurs at the beginning of a rhythmical segment, it is automatically deemed as a weakened tone non-ictus-bearing syllable. Marking by indices $_{1/2/3/4}$ is waived in this case (Švarný, 1998, p. 26).

2. For sequences of two, three and four full tone syllables within a frame of a rhythmical segment Švarný sets the default acronymic rule. This rule allows us to avoid another graphic representation of these syllables. In disyllabic rhythmical segments the ictus rests on the second syllable; in the trisyllabic and tetrasyllabic rhythmical segments there are ictuses on the first and last syllables (Švarný, 1998, p. 26–27; Švarný, 2000, p. 153).
3. The distinction between an atonic neutralized syllable and an atonic toneless syllable has to be inferred by the student himself (Švarný, 2000, p. 153).

If students follow all these linear rhythmical grouping and syllable prominence rules correctly, they will be able to produce Chinese utterances similar to the speech of native speakers.

However, in this theory there is a time handicap: to obtain a transcription, it is first necessary to have an audio recording of speech produced by a Chinese native speaker, which is then transcribed accordingly. Such transcription process itself is relatively lengthy and time-consuming, and efforts [2] to come up with "universal prosodic transcription" which would allow linguists to transcribe texts according to the universal rules is understandable.

4 Sample text and its audio recordings

The sample text used here as a demonstration was taken from the *Textbook of Chinese Conversation* (Uher, 2012). This textbook is used for teaching beginner students in the first semester of Chinese studies at our department, thus the text is well known to them. The original text is in Chinese pinyin and the same text in PTR version is used from the second or third semester onwards. The text was transcribed according to an audio recording of native speakers. The sample text is as follows:

Chinese characters version:

> 昨天张老师请我们去他家吃饭。张老师的家很干净也很漂亮。张老师的饭菜做得非常好。他的拿手菜是红烧肉。他做的红烧肉真棒。张老师还告诉我们，在中国的外国留学生常常说:不到长城非好汉。不吃烤鸭真遗憾。等我以后到了北京， 我一定去看长城，吃烤鸭。

Pinyin version:

> Zuótian Zhāng lǎoshī qǐng wǒmen qù tā jiā chīfàn.
> Zhāng lǎoshī de jiā hěn gānjing, yě hěn piàoliang.
> Zhāng lǎoshī de fàncài zuò de fēicháng hǎo.
> Tā de náshǒucài shì hóngshāoròu.

[2] Represented by PhDr. Hana Třísková, Ph.D. from the Oriental Institute of the Academy of Sciences of the Czech Republic.

Tā zuò de hóngshāoròu zhēn bàng.
Zhāng lǎoshī hái gàosu wǒmen, zài Zhōngguó de wàiguó liúxueshēng chángchang shuō:
Bú dào chángchéng fēi hǎohàn.
Bù chī kǎoyā zhēn yíhàn.
Děng wǒ yǐhòu dàole Běijīng, wǒ yídìng qù kàn chángchéng, chī kǎoyā.

Prosodic transcription version:

Zuótiān, zhāng-lao^{3}shī, qǐng-women-qu^{4} ta^{1}-jiā chī-fàn.
Zhāng-lao^{3}shī-d-jiā, hen^{3}-gānjing yě-hen-piàoliang.
Zhāng-lao^{3}shī-d fàncài, zuò-d **fēi**chang-hǎo.
Ta^{1}-d-náshoucài-shi hóng-shāo-ròu.
Tā-zuo^{4}-d hóng-shāo-ròu, zhēn-bàng.
Zhāng-lao^{3}shī, hái-gàosu-wo^{3}men, zai^{4}-zhōngguo-d wàiguo-liúxueshēng chángchang-shuō:
Bu^{2}-dao^{4}-chángchéng, fēi-hǎohàn.
Bu^{4}-chī kǎoyā, zhēn-yíhàn.
Deng^{3}-wo-yihòu dao^{4}le-běijīng, wo^{3}-**yí**ding qu^{4}-kan chángchéng, chī-kǎoyā.

A trial test was made to find out which production of these two text versions — pinyin and PTR — sounds more natural to a native speaker. Four anonymous beginner students were asked to read a sample dialogue in both pinyin and PTR versions and their reading was recorded. Two of them had read the pinyin version first and the PTR version second and two of them reversely. The recordings were carefully listened to by two native speakers who were neither informed about the existence of two different text versions nor about the overall aim of the trial test. Both of them proved the PTR version recordings to sound more natural.

One student audio recording and our findings were presented at the international conference on Chinese linguistics[3] where the majority of participants were overseas Chinese native speakers and most of them considered our findings as valid.

The above mentioned facts lead us to believe that PTR could be helpful in teaching and studying Chinese and has a considerable effect on the quality and naturalness of speech of non-native speakers.

[3] The 27th North American Conference on Chinese Linguistics, University of California, Los Angeles, April 3-5, 2015.

5 Students' opinion

Even though influence on the quality of speech has been proven, there is yet another important aspect of using PTR in teaching that is worth mentioning. Namely the students' opinion on PTR and its use in teaching: Is PTR user friendly or not?

To find out students' opinion about the use of PTR and its effects in the acquisition of Chinese pronunciation we have carried out a short survey. We gathered a total of 30 questionnaires, in which students answered 7 basic questions and expressed their own opinions about PTR. The following are the questions used in the survey, and their answers.

1. Was it hard for you to learn how to read prosodically transcribed text? (possible answers: hard, moderately difficult, easy)

None of the students chose the possibility 'hard', and 23% chose the possibility 'easy'. The largest group, 77% of students, chose the possibility 'moderately difficult'.

This is an answer to the most pressing question about PTR. The obtained result satisfactorily shows that PTR is manageable and not too complicated.

2. How do you appraise the PTR graphic notational system? (possibille answers: complicated, rather complicated, simple, rather simple)

None of the students chose the possibility 'complicated', while 'rather complicated' was chosen by 16% of students. 32% chose the possibility 'simple' and 52% the possibility 'rather simple'. If we count together the options simple/rather simple and complicated/rather complicated it comes to 84% and 16% respectively. In other words, PTR notational system is relatively simple for 84% of students.

It could be said that for the overwhelming majority of students (77%) the difficulty of PTR is moderate and they are able to read prosodically transcribed text quite easily and without noticable troubles. The notational system is simple or rather simple for 84% of them. This shows that if students gain the insight into PTR, they will be able to use it easily and will find its notational system simple as well.

3. Do you think that reading from prosodically transcribed text sounds more natural than reading from pinyin? (possibile answers: less natural, more or less the same, more natural)

Results show that 13% of students chose the possibility 'less natural', 30% chose the possibility 'more or less the same', and 57% chose the possibility 'more natural'.

This shows that for more than half of the students reading from prosodically transcribed text sounds more natural than reading from pinyin. Nevertheless, an essential question on whether a beginner student is able to evaluate the naturaleness of speech and its slight nuances remains. It is well known that such judgements are rather difficult even for non-native teachers of Chinese. The best solution seems to be to rely on the opinion of native speakers.

4. Does PTR have a positive effect on you at the beginning of your studies? (possibile anwers: no, rather no, yes, rather yes)

Only 3% of students chose the possibility 'no', 37% chose the possibility 'rather no', 40% chose the possibility 'yes', and 20% chose the possibility 'rather yes'. We can see the unequivocal answer 'yes' reaches 40% while the unequivocal answer 'no' has only 3%. If yes/rather yes are counted together and no/rather no as well, the result is 60% and 40% in favour of the former.

That clearly shows that PTR has had a positive effect at the beginning of studies on 60% of students (in their opinion) and to us this is a very promising result.

5. What are the positives and negatives of PTR? (Answers were produced by the students.)

Negatives:

- confusing
- superfluous

We perceive that the existence of two text versions and two different notational systems could be somewhat confusing and PTR could be viewed as something superfluous, a "superstructure" of pinyin. The opinion will be one of the topics for future discussion between teachers and students.

Positives:

- easy learning
- correct realization of pronunciation and linear segmentation
- satisfactory understanding of the rhythm of speech and the placement of syllable prominences
- gradual improvement of speech
- step by step integration of PTR in teaching

As there are more positive than negative point stressed in the survey, this leads our team to think that PTR has a positive influence on beginner students, and that students are satisfied with its use in teaching.

6. Is PTR a benefit for beginner students? (possible answers: no, yes)

12% of students chose the possibility 'no' while 88% of them chose the possibility 'yes'.

This question differs from the previous question No. 4 (Does PTR have a positive effect on you at the beginning of your studies?) in that it refers to all students and is therefore supposed to be more objective. For the question No. 4, which was based on subjective experiences, 60% of students answered that PTR had a positive effect on their early Chinese language acquisition. The percentage of a positive answer in this latter, more objective question is almost 30% higher. 88% of students think that PTR is a benefit for beginners, and this makes our team believe that the positive influence and

the effect of PTR is even wider than could be concluded from the asnwers on question No. 4.

7. What particular benefits or drawbacks are there to PTR? (Answers were produced by the students.)

Drawbacks:

- two versions of one text are confusing

This is the same problem as in the previous question No. 5 (What are the positives and negatives of PTR?), and will be also one of the topics for future discussion between teachers and students.

Benefits:

- learning to place stress on correct syllables, correct linear segmentation, correct syllable prominences
- easier orientation in a system of Chinese language
- understanding of natural pronunciation and its rules
- fixing of speech melody
- creation of correct habits in speech
- help in communication with native Chinese speakers
- "The speech sounds more chinese."

The last benefit that the speech sounds more Chinese could be a kind of principle of our work in the field of prosody. In general, our effort is focused on the achievement of correct production of suprasegmental units and to approximate it to a native speech. Table 1 shows results of the closed-ended questions:

Table 1: An overview of the closed-ended questions

Question	**Available answers**			
1. Was it hard for you to learn how to read prosodically transcribed text?				
	hard	moderately difficult	easy	
	0%	77%	23%	
2. How do you appraise the PTR graphic notational system?				
	complicated	rather complic.	simple	rather simple
	0%	16%	32%	52%
3. Do you think that reading from prosodically transcribed text sounds more natural than reading from pinyin?				
	less natural	more or less the same	more natural	
	13%	30%	57%	
4. Does PTR have a positive effect on you at the beginning of your studies?				
	no	rather no	yes	rather yes
	3%	37%	40%	20%

Question	Available answers	
6. Is PTR a benefit for beginner students?		
	no	yes
	12%	88%

6 Discussion

The aim of this paper was to present PTR, a traditional method used for teaching Chinese language at the Department of Asian Studies in the Czech Republic, where it has been deep-rooted for several decades. Our experience show that there is still much effort to adapt it to the modern contemporary situation and continue the heritage of Oldřich Švarný. Our effort does not consist only in practical teaching, but there are also plenty of theoretical works using Švarný's data and principles in academic spheres. By the application of PTR we are able to precisely describe all relevant prosodic features of modern spoken Chinese. From this point of view PTR is the first step for subsequent research. PTR enables us to count syllables of different stress prominence and average length of segments or cola, and analyze different rhythms contained in speech, their positions in a sentence and the number of ictuses.

We use the above mentioned methods for various types of research, e.g. comparing data gained formerly by Švarný with data gained currently in order to find out a potential language change, or in connection with sociolinguistics in order to find out suprasegmental differences between male and female speech.

PTR is used only with the Czech (and Slovak) language students of Chinese so far, however, its potential is thought to be wider. It is not a national transcription based and adapted on phonetic system of a particular language and is thus not restricted for usage in one language only. PTR could be an international system of transcription capturing prosodic features worldwide. As described above, PTR is warmly adopted by students, it helps them to produce Chinese sentences correctly. The first step of the process is reproduction and acquirement, the next step is imitation and the highest step is unsupported production.

References

Duběda, T. (2005). *Jazyky a jejich zvuky: univerzálie a typologie ve fonetice a fonologii [Languages and their sounds. Universals and typology in phonetics and phonology]*. Praha: Karolinum.

Švarný, O. (1952). *Prosodic Characteristics of Syllable and their Modifications in the Continuous Speech*. Unpublished manuscript. Prague: Charles University.

Švarný, O. (1963). *Discussing the Concept of the Morpheme and Word in Modern Spoken Chinese*. Unpublished manuscript. Prague: Oriental Institute.

Švarný, O. (1991a). Prosodic Features in Chinese (Pekinese). *Archív Orientální, 3*, 234–254.

Švarný, O. (1991b). Prosodic Features in Chinese (Pekinese): Prosodic Transcription and Statistical Tables. *Archív Orientální, 3*, 234–254.

Švarný, O. et al. (1998). *Hovorová čínština v příkladech III.* [*Colloquial Chinese in Sentence Examples III.*]. Olomouc: Palacký University.

Švarný, O. (1998–2000). Učební slovník jazyka čínského [*A Learner's Dictionary of Modern Chinese*]. 4 vols. Olomouc: Palacký University.

Švarný, O. (2000). Prosodical Transcription of Modern Chinese. Experimental Research and Teaching Practice. In Palková Z. (Ed.), *Papers in Phonetics and Speech Processing* (149–159). Frankfurt am Main: Hector Verlag.

Třísková, H. (2011). Prozodická transkripce čínštiny O. Švarného: čtyři historické verze [Oldřich Švarný´s prosodic transcription of Mandarin: four successive versions]. *Nový Orient, 4* (66), 45–50.

Uher, D. (2012). *Učebnice čínské konverzace 2* [*Textbook of Chinese Conversation*]. Dosud nepublikovaný prototyp [Manuscript submitted for publication].

Uher, D., & Švarný, O. (2014). *Prozodická gramatika čínštiny*. [*Prosodical Grammar of Chínese*]. Olomouc: Palacký University.

WORD SKETCHES OF SEPARABLE WORDS *LIHECI* IN CHINESE

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

Abstract

Separable words (*liheci*) are a special type of Chinese verbs with unique syntactical features in a sense that some elements come in between the two morphemes of a verb for a sentence to be grammatically acceptable. Not all separable words are extendable to the same degree. To understand the behaviour of words, it is generally advised to check word sketches, because they are based on large text corpora. This article examines how Chinese separable words are treated in Sketch Engine and discusses on the appropriateness of the available Chinese corpora for word sketches. It further stresses the importance of including information on inserted elements in word sketches and gives suggestions on how to include them.

Keywords: Chinese; separable words; *liheci*; word sketches; Sketch Engine

Povzetek

Ločljive besede (*liheci*) so poseben tip glagolov s svojevrstnimi sintaktičnimi lastnostmi. Zanje je značilno, da moramo določene stavčne člene vstaviti med oba morfema glagola. Pri ločljivih besedah je posebej problematično to, da ne moremo nikoli natančno vedeti, do kakšne mere je beseda ločljiva in s katerimi vzroci jo lahko razširimo. Pri orisu rabe besed so nam lahko v veliko pomoč besedne skice, ki izhajajo iz besedilnih korpusov. V članku proučimo, kako ločljive besede obravnava Sketch Engine, kateri kitajski korpus je najbolj primeren za besedne skice in predlagamo, kako bi besednim skicam dodali tudi informacije o ločljivosti glagolov.

Ključne besede: kitajščina; ločljive besede; *liheci*; besedne skice; Sketch Engine

1 Introduction

Separable words in Chinese have been investigated by numerous researchers for several decades. According to their research orientation and focus, Huang (2006) divides them into two periods. The first period lasted from 1950s to 1970s, and the second period stretches from the 1980s to the present.

Huang (2006) notes that due to the problems related to transliteration this group of verbs caught linguists' attention even before the term *separable word* (*liheci* 离合)

Acta Linguistica Asiatica, 6(1), 2016.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.6.1.47-57

was defined. Chinese writing system does not have explicit word boundary markers, such as the spaces between words. When the official romanization system for Standard Chinese *Hanyu pinyin* was to set orthographic rules, it encountered the problem of word boundaries. Even after several decades of discussions, scholars had yet to agree whether these "items" are words, word phrases, words as well word phrases, or something in-between.

After the China's opening up policy in 1980s, learning and teaching Chinese as a foreign language became the major point of interest. Separable words became the subject of research in relation to foreign language acquisition. They seemed to be a common problem for foreign students regardless of their native language. It was obvious that a non-native speaker had difficulties understanding whether a verb should be used as a unit or separately, and if the latter, which elements could be inserted in between the two morphemes (Huang, 2006).

This second period has also been related to the development of information processing and machine translation, which brought new insights into the existing research topics. Analyzing language by means of corpus linguistics is also one of the novelties in this period. Several recent papers refer to the data from Peking University CCL Online Corpus.

2 Separable words *liheci*

Separable words are disyllabic verbs that are separable in certain circumstances. Even more, in these circumstances, some separable words should undertake at least one element in between its syllables (morphemes), or else the sentence would be grammatically incorrect. There are several types of separable verbs, but the majority of them has the morphological structure "verb–object", for example *tiao//wu* (跳舞) "to dance" (lit. to jump dance), *jian//mian* (见面) "to meet" (lit. to see a face), *bang//mang* (帮忙) "to help" (lit. to help//busy).

4. 他 **跳** 了 一个小时 的 **舞**。
   ta **tiao** le yi ge xiaoshi de **wu**
   he **dance** LE one hour DE **dance**
   He was dancing for an hour.

5. 我们 只 **见** 过 一次 **面**。
   women zhi **jian** guo yi ci **mian**
   we just **see** GUO once **face**
   We've met only once.

6. 他 **帮** 了 我 一个 大 **忙**。
   ta **bang** le wo yi ge da **mang**
   he **help** LE I one big **help**
   He helped me a lot.

Scholars advocate separable words in two ways; either as *words* or as *word phrases*. Those who interpret them as *words* support their ideas with the following three facts. Firstly, separable words may be uttered in isolation with semantic or pragmatic content; secondly, several morphemes of separable words are bound morphemes; and finally, although separable words can be extended, their extension patterns are very limited. On the other hand, scholars who claim that separable words are *word phrases* say that separable words carry syntactic features of word phrases such as splitting, flexible word order, and often carry a special idiomatic meaning (Zhou, 2010, p. 123).

Different authors propose various categorizations of separable words, classifying them into up to ten different groups. However, it is generally agreed that there are at least the following three types of separable words (Huang, 2006, p. 85):

- V–O type (*dongbin shi* 动宾式)
- V–Complement type (*dongbu shi* 动补式)
- S–V type (*zhuwei shi* 主谓式)

As mentioned above and demonstrated in examples 1–3, the two morphemes of a separable word may demand one or more additional elements in between them. Such additional elements may be an aspectual particle and a durational phrase, as in example 1, an aspectual particle and a phrase expressing number of occurrences, as in example 2, or others. Example 3 is extended with an aspectual marker, followed by the recipient of an action and an attribute.

Scholars have come to several conclusions on which elements can be inserted in separable words. In a very simplified manner, we present Zhou's (2010) conclusions because they are well organized and systematic.

- aspectual particles (*le* 了, *guo* 过 and *zhe* 着)
- complements (quantitative, resultative, directional, potential)
- attributes
- some question forms and patterns
- a combination of these elements

The most intriguing part concerning the inserted elements is their degree of separability. Some separable words can be extended with all the above patterns whereas others are limited to some of them (He, 2009, p. 65). Wang (2008; 2010) provides us with corpus–driven findings, where he concludes that the majority of separable words are related to our everyday's life and activities. Such separable words are also very flexible and allow various combinations of extension. Wang draws insightful conclusions about the semantics of separable words but due to length limitations of this paper, we will not go into details.

In this paper, we will rather focus on word sketches, which may provide collocational and grammatical features of words.

3 Word sketches

Following the explanation on the Sketch Engine's website, a word sketch is "a corpus–based summary of a word's grammatical and collocational behavior" (Getting Started with Sketch Engine, 2016). This is a very useful tool not only for researchers, but also for language teachers, language learners and other users, because it shows "the word's collocates categorized by grammatical relations such as words that serve as an object of the verb, words that serve as a subject of the verb, words that that modify the word etc." (Word Sketch, 2016).

Sketch Engine may include several corpora for the same language. For standard Chinese, there are nine text corpora available for subscribed users.[1] However, their word sketches vary remarkably. The main reason for such divergence is not the size of the underlying corpora but the availability of syntactical descriptions. Namely, word sketches depends on the available grammatical definitions supplied to Sketch Engine (Getting Started with Sketch Engine, 2016).

Figure 1 shows three sets of grammatical relations that are available for Chinese corpora. From the user's perspective, these can be selected from the *advanced options* of word sketch tool.

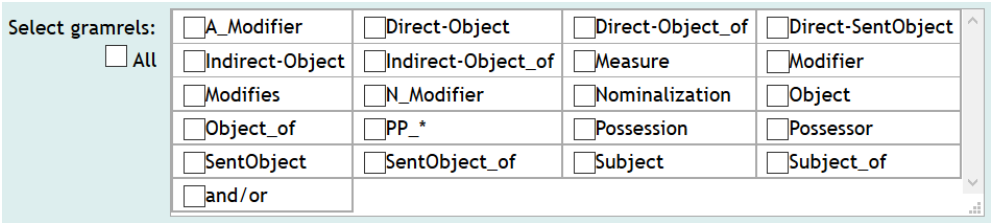

Figure 1a: Chinese grammatical relations (Set 1)

Figure 1b: Chinese grammatical relations (Set 2)

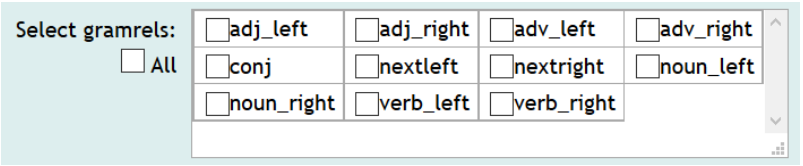

Figure 1c: Chinese grammatical relations (Set 3)

[1] The list of Sketch Engine's text corpora is available at https://www.sketchengine.co.uk/corpora/.

Grammatical relations are defined as regular expressions over Part-of-speech-tags (POS-tags), and are saved in the so-called *gramrel files*.[2] They are typically created for nouns, verbs, and adjectives, but can be enriched with other definitions, as well.

Presently, the best Chinese gramrel file is related to the Chinese GigaWord 2 corpus, both the Mainland (simplified) version and the Taiwan (traditional) version. Corpus zhTenTen [2011] is compared to other Chinese corpora in Sketch Engine much larger, and would therefore generate better word sketches, but has less sophisticated definitions for grammatical relations. Its wordsketches are therefore not as informative as in Chinese GigaWord 2 corpus (see Table 1).

Table 1: Chinese corpora and their corresponding grammatical relations

Text corpus	**Number of tokens**	**Grammatical relations**
Chinese GigaWord 2 Corpus: Mainland, simplified	299,338,099	Set 1, Figure 1a above
Chinese GigaWord 2 Corpus: Taiwan, traditional	455,526,209	Set 1, Figure 1a above
zhTenTen [2011]	2,106,661,021	Set 2, Figure 1b above
OPUS2 Chinese Simplified	299,338,099	Set 2, Figure 1b above
OPUS2 Chinese Traditional	622,382	Set 3, Figure 1c above
Internet-ZH	277,931,664	N/A
ChineseTaiwanWaC	349,198,060	Set 2, Figure 1b above
ChineseTaiwanWaC (Universal Sketch Grammar)	349,198,060	Set 3, Figure 1c above

[2] Gramrel file related to Figure 1a:
https://the.sketchengine.co.uk/bonito/corpus/wsdef?corpname=preloaded/cgw2_sc

Gramrel file related to Figure 1b:
https://the.sketchengine.co.uk/bonito/corpus/wsdef?corpname=preloaded/opus2_zh_TW

Gramrel file related to Figure 1c:
https://the.sketchengine.co.uk/bonito/corpus/wsdef?corpname=preloaded/opus2_zh

4 Word sketches of separable words

Although word sketches are primarily created for nouns, adjectives and verbs, information on grammatical and collocational behavior of separable words is still very limited. Recall that separable words are a special type of verbs.

In this research, we focus on Chinese GigaWord 2 Corpus, because it provides the most comprehensive word sketches among Chinese corpora. Queries have shown that separable words are treated as words in their disyllabic form, but have been assigned different POS-tags. Table 2 presents categorization of 21 basic separable words, listed on HSK3 vocabulary list.[3]

Table 2: Separable words and their POS-tags (HSK3)

Verb	Pinyin	Literal meaning[4] → Meaning	POS-tag[5]
睡觉	shui//jiao	to sleep//a sleep → to sleep	VA12
刮风	gua//feng	to blow//wind → to blow	VA3
见面	jian//mian	to see//a face → to meet	VA4
结婚	jie//hun	to tie//a marriage → to marry	VA4
跑步	pao//bu	to run//a step → to run	VA4
起床	qi//chuang	to get up//a bed → to get up	VA4
上网	shang//wang	to go up//a net → to go online	VA4
说话	shuo//hua	to speak//words → to speak	VA4
跳舞	tiao//wu	to jump//a dance → to dance	VA4
洗澡	xi//zao	to bathe//a bath → to bathe	VA4
游泳	you//yong	to swim//swimming → to swim	VA4, VA
帮忙	bang//mang	to help//busy → to help	VC2
离开	li//kai	to depart//to start → to leave	VC2
完成	wan//cheng	to finish//to complete → to complete	VC3
生病	sheng//bing	to arise //illness → to fall ill	VH11

[3] HSK stands for "Hanyu Shuiping Kaoshi" or Chinese Proficiency Test. HSK3 is the third of six levels.

[4] Literal meaning of each morpheme is provided for better understanding of the corresponding bisyllabic word.

[5] List of POS-tags for Chinese: https://www.sketchengine.co.uk/symbols-of-parts-of-speech/

Verb	Pinyin	Literal meaning[4] → Meaning	POS-tag[5]
发烧	fa//shao	to dispatch//heat → to have a fever	VH11
生气	sheng//qi	to arise//steam → to be angry	VH21
着急	zhao//ji	to take action//urgent→ to worry	VH21
担心	dan//xin	to carry//a heart → to be anxious	VK1
放心	fang//xin	to put down//a heart → to be at ease	VK1
注意	zhu//yi	to focus // an idea → to pay attention	VK1

Although the list is very short and as such not the best representative sample of separable words in standard Chinese, we can see at a glance that the major part of separable words is tagged as VA4. Based on this idea, I have further analyzed Wang's (2008) list of 207 separable words and got roughly similar results, as shown in Table 3.

Table 3: POS-tagging for 207 separable words

POS-tag	Number of separable words	Percentage
/VA4/	80	39%
/VH11/	47	23%
/VB11/	10	5%
/VH21/	9	4%
/VB12/	7	3%
/VA13/	6	3%
/VK1/	5	2%
/VC2/	4	2%
/VC31/	4	2%
Sub Total	**172**	**83%**
Others (less than 1 % each)	35	17%
Total	**207**	**100%**

In overall, the results show that almost 62% of all separable words are classified as VA4 or VH11. We assume that this percentage is even higher if we eliminate some "suspicious" items. For example, it is highly disputable whether 注意 "to pay attention"

is a separable word or not. The dictionary of 5000 graded words for New HSK provides an example of separate usage, but this is not very common.

你身体不好，健康状态要多**注**点儿**意**。
Ni shenti bu hao, jiankang zhuangkuang yao duo **zhu** dianr **yi**.
You are in poor health. Take care of your health condition. (Li, 2013, p. 381)

However, most of other authors do not consider this verb as separable. No cases of separable use were found in Chinese GigaWord 2 Corpus, nor in CCL corpus (Wang, 2008). Furthermore, even if this verb can be used separately, such examples are probably not frequent enough to be relevant for word sketches.

We have already noted (see Chapter 2 above) that in certain patterns, some elements must be inserted between the first and the second morpheme. Although this is a very important syntactical feature of Chinese separable words, word sketches offer no such information.

Based on corpus query language (CQL), we further analyzed the selected 21 separable words in their separate forms. We formulated the following CQL expression:

"*A*"[word!="\，|\;|\：|\。|\？|\！|\)" & tag!="PARENTHESISCATEGORY"]{1,}"*B*" within <p/>[6]

Query results were very fruitful and relatively accurate. The concordance list included all desired extensions, mostly without noise. Figure 1 shows one segment of the results for verb *bangmang* 帮忙 "to help".

	word	Frekvence
p | N	帮 了 大 忙	67
p | N	帮 个 忙	16
p | N	帮 大 忙	12
p | N	帮 了 我们 大 忙	12
p | N	帮 了 忙	12
p | N	帮 了 我 的 大 忙	8
p | N	帮 的 忙	7
p | N	帮 这 个 忙	5
p | N	帮 了 我们 的 大 忙	5
p | N	帮 了 他 的 忙	5
p | N	帮 了 他 一 个 大 忙	4
p | N	帮 点 忙	3
p | N	帮 什么 忙	3
p | N	帮 了 我 的 忙	3
p | N	帮 了 我 一 个 大 忙	3
p | N	帮 了 很 大 的 忙	3
p | N	帮 了 他 的 大 忙	3
p | N	帮 了 他 一点 忙	3
p | N	帮 过 忙	2
p | N	帮 俺 忙	2

Figure 1: Collocations of verb *bangmang*'s extended form

[6] Letters *A* and *B* represent the *first* and *second* morpheme of a separable word in question.

Among other results it is worth mentioning an example with 11 tokens inserted between both parts of the separable word, as shown and explained in Figure 2. Despite a far distance between the two morphemes, the structure shows syntactically correct relation.

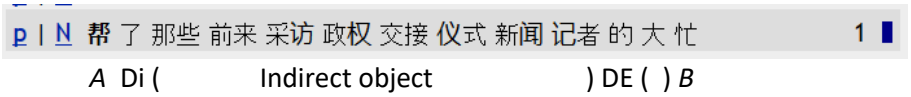

Figure 2: Collocations of verb *bangmang*'s extended form

However, mMorphemes of separable words do not tend to be so far apart as in the example from Figure 2. Analysis has shown that there are usually one to five tokens in between (Figure 3).

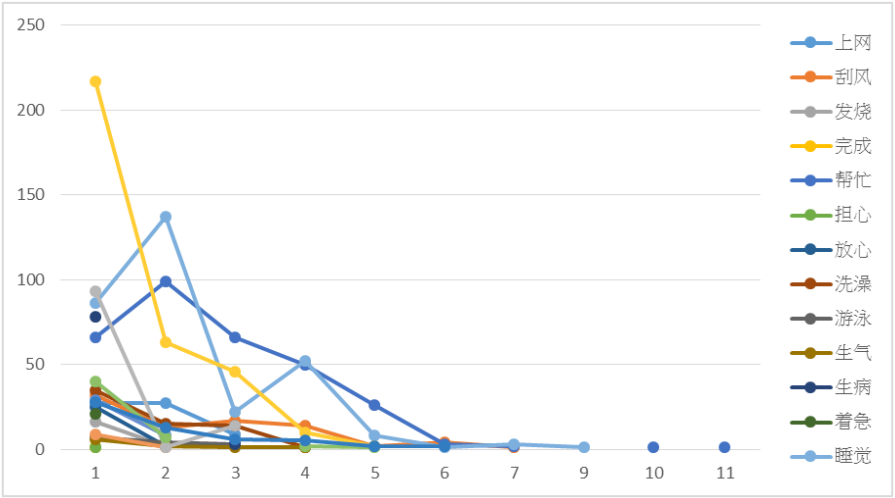

Figure 3: Number of inserted tokens in separable words

Separable words in their separate forms are generally treated individually and are not considered as the same verb anymore. Therefore We further investigated which part-of-speech tags are assigned to such separate forms and the results are shown in Table 4. Because our focus is not on the meaning of every single separable word but rather on their associated POS-tags, we intentionally left out English expressions.

Table 4: POS-tags of separable forms for HSK3 separable words

SW as a unit	SW in its separate form
睡觉/VA12	睡/VA12 [...] 觉/Nad; /VK1
刮风/VA3	刮/VC2 [...] 风/Na

SW as a unit	SW in its separate form
见面/VA4	见/VE2 […] 面/Na; /Ncda; /Nfa
结婚/VA4	结/VC31 […] 婚/Nad
跑步/VA4	跑/VA11 […] 步/Nf
起床/VA4	起/VC31; /Di; / Ng […] 床/Nab; /Nfa
上网/VA4	上/VC1; /Ng; /Nes […] 网/Nab
说话/VA4	说/VE2 […] 话/Nac
跳舞/VA4	跳/VA11; /Na […] 舞/Nac; /VC2
洗澡/VA4	洗/VC2 […] 澡/Na
游泳/VA4; /VA	游/VA11; /Nbc […] 泳/b
帮忙/VC2	帮/VC2; /P37 […] 忙/VH11
离开/VC2	离/VC2 […] 开/VC
完成/VC3	完/VH11 […] 成/VH11
生病/VH11	生/VC31 […] 病/VH11
发烧/VH11	发/VH11; /VD1; /VJ […] 烧/VC2
生气/VH21	生/VC31 […] 气/Naa; /VK
着急/VH21	着/VC2 […] 急/VH
担心/VK1	担/VC2 […] 心/Na
放心/VK1	放/VC33 […] 心/Na
注意/VK1	N/A

In 6 cases, the first morpheme is interpreted as a VC2 verb, in 4 cases as a VC31 verb, and in 3 cases as a VA11 verb. There are further 2 instances of a VE2 and VH11 verb, and the remaining 4 verbs merged to some minor groups.

The second morpheme is a noun in most cases, as expected. Recall that the majority of separable words have the internal morphological structure verb–object (V–O). The HSK3 list of separable words is too short to draw reliable conclusions, but we do believe that the above findings show correlalations and patterns that could, with some further support, get generalized.

5 Final thoughts

The preliminary research on separable words in Chinese GigaWord 2 corpus has shown that separable words are not treated as a special subtype of verbs. Instead, the disyllabic verb and its monosyllabic counterpart are tagged as two different verbs,

usually belonging to different POS-tags. It is impossible to expect POS-tags to be changed, however, it would be possible to track the relations between disyllabic and monosyllabic couterparts. To do so, a large number of separable words should be analyzed.

Collocations are therefore generated separately for disyllabic and monosyllabic forms of these verbs. Consequently, word sketches do not provide information about which elements should be inserted between the two morphems of separable words. This is undoubtly a very important syntactical feature of Chinese verbs.

Since CQL queries provide quite accurate results, and it is already known which patterns may be formed with separable words, it might be possible to create additional definitions of grammatical relations for the relevant gremrel file.

References

Getting Started with Sketch Engine. (2016, 3 12). Retrieved from Sketch Engine: https://www.sketchengine.co.uk/getting-started/

He, Q. [何清强]. (2009). Influence of Separation Extent on Acquisition of Verb-object Separable Words [分离度对动宾式离合词习得的影响]. *Journal of Ningbo University (Liberal Arts Edition), 22(6)*, pp. 65–70.

Huang, X. [黄晓琴]. (2006). A Summery of the Research on Separable Word [离合词研究综述]. *Journal of ILi Normal University*, pp. 84–87.

Li, L. [李禄兴]. (2013). *A dictionary of 5000 graded words for new HSK (Levels 1, 2 & 3)* [新 HSK 5000 词分级词典（一~三级）]. Beijing: Beijing Language and Culture University Press.

Wang, H. [王海峰]. (2008). *The Study on the Separable Words Separated Form Function of Mandarin Chinese* [现代汉语离合词离折形式功能研究]. *PhD thesis.* Beijing: Beijing Yuyan Daxue.

Wang, H. [王海峰]. (2010). A corpus-based semantic study of the Chinese separable words [基于语料库的现代汉语离合词语义特征考察]. *Journal of Hebei Normal University (Philosophy and Social Sciences Edition), 33(1)*, pp. 96–100.

Word Sketch. (2016). Retrieved from Sketch Engine: https://www.sketchengine.co.uk/word-sketch/

Zhou, W. [周卫华]. (2010). Expanding Forms and Features of Separable Words in Modern Chinese [现代汉语离合词的扩展形式及特点]. *Sanxia luntan, 6*, pp. 123–127.

A Character-based Constructional Approach to Chinese Imperfective Aspect Markers *Zai* 在 and *Zhe* 着

Liulin ZHANG
University of Hawaii at Manoa, United States
lz2@hawaii.edu

Abstract

在 *zai* and 着 *zhe* are commonly recognized imperfective aspect markers in Mandarin Chinese, though there are noticeable differences between their distributions and functions. By resorting to origins, historical evolutions, and corpus data for the meanings and functions of these two characters, it is observed that they are both polysemies displaying semantic networks organized around a central sense respectively, and thus the characters 在 and 着 are distinct in form and meaning pairings. 在 is a construction indicating presence within a certain range while 着 generally denotes 'reach to'. Related to their basic meanings, 在 and 着 exhibit some constraints respectively when marking imperfective aspect. From this character-based constructional account, 在's and 着's qualifications as Chinese imperfective aspect markers are theoretically arguable.

Keywords: Chinese; imperfective aspect marker; grammaticalization; character; construction

Povzetek

在 *zai* in 着 *zhe* sta splošno poznana označevalca nedovršnega glagolskega vida v mandarinščini, ki pa se po pojavnosti in vlogi tudi med seboj precej razlikujeta. Ob pregledu izvora, zgodovinskega razvoja in korpusnih informacij o pomenu in funkciji the dveh označevalcev lahko zaključimo, da oba nosita večpomenskost, ki je za vsakega označevalca drugače organizirana v pomensko mrežo s središčem v njegovem osnovnem pomenu. 在 *zai* in 着 *zhe* se tako med seboj razlikujeta tako v obliki kot tudi pomenskih navezavah. 在 *zai* je struktura, ki nakazuje prisotnost v določenem območju, medtem ko 着 *zhe* v splošnem pomeni 'dospeti'. Na osnovi njunih primarnih pomenov pa označevalca izkazujeta vsak svoje omejitve pri vzpostavljanju nedovršnega glagolskega vida. S teoretičnega vidika je tako njuna vloga kot označevalca nedovršnega glagolskega vida vprašljiva.

Ključne besede: kitajščina; označevalec nedovršnega glagolskega vida; gramatikalizacija; pismenke; struktura pismenk

Acta Linguistica Asiatica, 6(1), 2016.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.6.1.59-79

1 Introduction

Aspects are different ways of viewing the internal temporal constituency of a situation (Comrie, 1976, p. 3; Bybee, 2003, p. 157). The contrast of perfective and imperfective is the most basic distinction of aspect. The **perfective** indicates that the situation is to be viewed as a bounded whole, looking at the situation from outside, without necessarily distinguishing any of its internal structure. The **imperfective** looks at the situation from inside, or looks inside its temporal boundaries, and it is crucially concerned with its internal temporal structure (Kibort, 2008).

According to Li & Thompson (1981, p. 185), Chinese has the following system of verbal aspect:

(1) i. Perfectivity: 了 *le* and perfectivizing expressions

ii. Imperfectivity (durative): 在 *zai,* 着 *zhe*

iii. Experiential aspect: 过 *guo*

iv. Delimitative: reduplication of verb

This perspective is generally agreed on. 在 *zai* and 着 *zhe* are therefore commonly recognized as two imperfective aspect markers in Chinese (Huang, Li, & Li, 2009, p. 101). This does not mean they are treated the same by linguists. The most prominent difference is their distributions. 在 *zai* is preverbal while 着 *zhe* occurs post-verbally, as shown in (2).

(2) a. 在　下雨。
zai　xiayu
ZAI　fall rain
It is raining.

b. 下着 雨。
xia-zhe　yu
fall-ZHE　rain
It is raining.

Beside their distributions, meanings have not escaped from the attention of researchers either. 在 *zai* is argued to feature a dynamic meaning while 着 *zhe* is claimed to be relatively static (Kwan-Terry, 1978; Smith, 1991, p. 271).

Confronted with distinct distributions and meanings of the two imperfective aspect markers, a problem arises naturally that where do these differences come from. The present study suggests a character-based constructional approach to solve this problem. Section 2 describes and summarizes the forms and meanings (functions) when 在 *zai* and 着 *zhe* co-occur with various event types, thus provides an overall picture for the constructions in question. The character-based constructional approach is introduced in section 3, and the applications for 在 *zai* and 着 *zhe* are laid out in

section 4, combined with historical data to illustrate their processes of grammaticalization and to explain the forms and meanings of them as distinct constructions. Closely related with their meanings, some constraints of the imperfective aspect marking 在 *zai* and 着 *zhe* are discussed in section 5, and Chinese imperfective aspect marking system is revisited. Section 6 is a summary and provides some implications for the character-based constructional approach in Chinese linguistics study.

2 Forms and functions of 在 *zai* and 着 *zhe*

2.1 Event types of Chinese verbs based on time notions

Since 在 *zai* and 着 *zhe* behave differently when co-occurring with various types of events, we find it necessary to begin our description with a summary of event types denoted by Chinese verbs based on time notions.

According to Vendler's (1967, p. 106) distinction of four categories of verbs, with the refinements by Dowty (1979) and Foley & Van Valin (1984, p. 33), states hold for an unbounded period of time. Achievements occur at a single moment, with an immediate end point. Activities go for a period of time, with no defined end point. Accomplishments go on for a period of time, but with a defined end point. Travis (2010, p. 120) introduced the following matrix to represent Vendler's four categories.

(3)

	– process	+ process
– definite	state	activity
+ definite	achievement	accomplishment

Tai (1984) argued that Chinese verb does not really have the subcategory of accomplishments. Most of the results in Chinese are expressed by word compounds. This view is accepted by most Chinese linguists (Chen, 1998; Jiang, & Pan, 1998, p. 333; Xuan, 2013; among others).

It can also be observed in Chinese, some verbs present a combination of an achievement and the resultant state, like 坐 *zuo* "sit down, sit", and 站 *zhan* "stand up, stand". We will refer to them as achievement–state for the convenience of description. Similarly, some verbs denote a combination of an activity and the resultant state, like 穿 *chuan* "put on, wear", and 堆 *dui* "pile up, lie in pile". We will refer to them as activity–state[1].

Therefore we can draw an outline of the event types denoted by Chinese verbs according to time notions.

[1] Vendler (1967, p.109) has already pointed out many activities have a "derived" state sense. He also noticed a group of verbs with conceptual divergences of their own, like *to know*, *to understand*. The Chinese activity-state verbs we mentioned here present similar properties.

Table 1: Event Types of Chinese Verbs Based on Time Notions

Event Types	Examples
States	是 *shi* "is/am/are"; 喜欢 *xihuan* "like"; 讨厌 *taoyan* "dislike"
Achievements	死 *si* "die"; 忘记 *wangji* "forget"; 到 *dao* "arrive"
Activities	说 *shuo* "speak"; 跑 *pao* "run"; 走 *zou* "walk"; 写 *xie* "write"
Achievement–states	坐 *zuo* "sit down, sit"; 站 *zhan* "stand up, stand"; 开 *kai* "open"
Activity–states	穿 *chuan* "put on, wear"; 包 *bao* "pack, hold inside"; 堆 *dui* "pile up; lie in pile"

Because the feature [+definite] is incompatible with imperfective aspect, achievement and accomplishment will not be discussed in the present study focusing on Chinese imperfective aspect marker.

2.2 Forms and functions of 在 *zai*

Zhang (2000) pointed out as a time adverb, 在 *zai* denotes progression of activity or continuation of state. Yang & Bateman (2002) referred to 在 *zai* as an activity–durative aspect marker, which is in line with Kwan-Terry's (1978) and Smith's (1991, p. 271) opinion we mentioned in section 1 that 在 *zai* has a dynamic meaning.

Corpus data shows 在 *zai* actually never co-occur with state verbs or achievement–state verbs. The verbs following it have to be activities. As for activity–state verbs like 穿 *chuan* "put on, wear", when co-occurring with these verbs, 在 z*ai* only indicates the progression of the activity.

Table 2: Co-occurrence of 在 *Zai* with Different Types of Verbs

Event Types	在 zai + V	Examples
States	—	是 *shi* "is/am/are"; 喜欢 *xihuan* "like"
Activities	Progression of Activity	说 *shuo* "speak"; 跑 *pao* "run"
Achievement –states	—	坐 *zuo* "sit down, sit"; 开 *kai* "open"
Activity–states	Progression of Activity	穿 *chuan* "put on, wear"; 包 *bao* "pack, hold inside"; 堆 *dui* "pile up; lie in pile"

As is shown in Table 2, basically the form "在 *zai* + activity" conveys the meaning that the activity is in progress.

2.3 The grammatical functions of 着 *zhe*

There are numerous studies devoted to the particle 着 *zhe* in Mandarin. The widely accepted opinion concerning its function suggests 着 *zhe* signals progression of activity and continuation of state (Liu, 1985; Gao, 1986, p. 172; Lü, 1999, pp. 665–666). Nevertheless, Dai (1991), Yuan (1992) and Fang (2000) argue that 着 *zhe* has only one function, and Guo (1997) proposes 着 *zhe* has more than two functions.

This disagreement is understandable if we take into consideration all the possibilities when 着 *zhe* co-occurs with verbs, as presented in the following discussion.

1. "V *zhe* (+ object)" denotes an action in progression or a state in continuation, but not all state verbs are allowed in this form. Permanent states like 是 *shi* "is/am/are", 姓 *xing* "be surnamed" are among the few exceptions.

(4) 人们 跳着， 唱着。
Ren-men tiao-zhe, chang-zhe
people-PL dance-ZHE, sing-ZHE
People are dancing and singing. (Lü, 1999, p. 666)

Chen (1980) noticed when 着 *zhe* co-occurs with activity verbs that are volitional, the clause sounds unfinished. He suggested 着 *zhe* has a subordinating function and usually serves as background information for the main event in discourse. The self-sufficiency of "V *zhe* (+ object)" increases as the verb becomes less volitional, and the whole structure is more state-like rather than activity-like at the same time. So when the verb is an achievement–state or an activity–state, and the subject is not the agent of the verb, the clause purely displays a state, without any activity meaning. Existential sentence (locative inversion) is among this situation.

(5) 墙 上 挂着 一幅 画。
Qiang shang gua-zhe yi-fu hua
wall top hang-ZHE one-CL painting
There is a painting hanging on the wall.

2. "V_1 *zhe* (+object$_1$) +V_2 (+ object$_2$)" denotes two events happening at the same time. V_1 can be the means of V_2 and V_2 can be the purpose of V_1. The V_1 in this form needs to be an activity and "V_1 *zhe* (+object$_1$)" serves as background information.

(6) 说着 看了 我 一眼
shuo-zhe kan-le wo yi yan
speak-ZHE look-PERF I a glance
gave me a glance while speaking (Lü, 1999, p. 666)

(7) 藏着 不肯 拿出来
cang-zhe bu ken na chulai
hide-ZHE not willing take out
hide (something) and not willing to take it out (Lü, 1999, p. 666)

3. "V_1 *zhe* +V_1 *zhe* +VP" denotes the VP happens unexpectedly when V_1 is in progress. The V_1 in this form also needs to be an activity and "V_1 *zhe* +V_1 *zhe*" is a subordinating form as background information.

(8) 说着 说着 不觉 到了 门口
shuo-zhe shuo-zhe bujue dao-le menkou
talk-ZHE talk-ZHE unconsciously arrive-PERF doorway
arrive at the doorway unconsciously while talking (Lü, 1999, p. 666)

4. "S + V *zhe* + AP" denotes the subject displays some kind of property through the experience of the verb. The verb here needs to be a perception/cognition/emotion verb, corresponding to different event types based on the time notions.

(9) 这个 主意 听着 不错。
Zhe-ge zhuyi ting-zhe bu cuo
this-CL idea sound-ZHE not bad
This idea sounds not bad.

(10) 这把 椅子 坐着 很舒服。
Zhe-ba yizi zuo-zhe hen shufu.
this-CL chair sit-ZHE very comfortable.
This chair is rather comfortable to sit on.

In general, 着 *zhe* has the following functions when co-occurring with different event types.

Table 3: Co-occurrence of 着 *Zhe* with Different Types of Verbs

Subcategories of Verbs	**V+ 着 *zhe***	**Examples**
States	Continuation of the state	爱 *ai* "love"; 喜欢 *xihuan* "like"
Activities	Progression of Activity (subordinating)	说 *shuo* "speak"; 跑 *pao* "run"
Achievement –states	Continuation of the resultant State	坐 *zuo* "sit down, sit"; 开 *kai* "open"

Subcategories of Verbs	V+ 着 *zhe*	Examples
Activity–states	Progression of Activity (subordinating) / Continuation of the resultant State	穿 *chuan* "put on, wear"; 包 *bao* "pack, hold inside"; 堆 *dui* "pile up; lie in pile"

The form "(S+) V+ 着 *zhe*" therefore entails two constructions:

Table 4: Constructions involved by "(S+) V+ 着 zhe"

Frame	Form	Meaning and Function
Activity	agent+ activity + 着*zhe*	a subordinate activity in progress, as background information
State	non-agent + state + 着 *zhe*	the non-agent subject is in some kind of state

3 A character-based constructional approach

Ever since Langacker (1987) argued syntactic patterns, are form and meaning pairings, but at a more abstract (schematic) level than words, lexicon–syntax continuum has become a fundamental notion among constructionalists and cognitive linguists. Croft (2003) and Baredal (2011) question the dichotomy between lexical rules and syntactic constructions. This idea was further demonstrated by Bybee (2006) by positing there is no unitary "grammar" of language but rather a continuum of categories and constructions ranging from low frequency, highly specific, and lexical to high frequency, highly abstract, and general. Boas (2008) also points out the importance of the lexicon–syntax continuum. Langacker (2008) stated clearly that there is no clear border between lexicon and grammar.

Based on the concept of lexicon–syntax continuum, the imperfective aspect markers in Chinese, 在 *zai* and 着 *zhe*, are also likely to be derived from some specific lexical items, exhibiting complicated polysemous networks ranging from lexical meanings to grammatical functions.

As for Chinese, characters can provide crucial hints for us to plot polysemous networks of words, considering the special properties of Chinese characters as a writing system. Saussure referred to Chinese writing system as an "ideographic system" (1983, p. 26). More specifically, Chinese characters are also declared to be "logographic writing system" (Diringer, 1962; Fabar, 1992), "morpho–syllabic" (DeFrancis, 1989) "morphemic writing system" (Hill, 1967; Su, 2001). Although these classifications are

based on different perspectives, they all acknowledge the semantic functions of Chinese characters. In this sense, each character is a construction, as constructions are pairings of form and meaning/function[2]. In order to capture the semantic network of words, we can turn to corpora of classical Chinese and see how the meanings and functions of a certain character changed over time. This is what we call *Character-based Constructional Approach*.

Actually it has been noted there are usually some kinds of systematically related and therefore explainable connections between different meanings and functions of the same lexical items (Tyler, 2012, p. 6) and that linguistic units, i.e. lexical items, morphemes and syntactic constructions, can subsume a range of distinct but related meanings organized with respect to a central meaning (Tyler, 2012, p. 22), which means by taking the character-based constructional approach, it will be able to reveal the central meaning/function of the construction represented by the character.

In our study of Chinese imperfective aspect markers, we search characters 在 and 着 in the classical Chinese corpus, Yuliaoku Zaixian (http://www.cncorpus.org/) to extract their meanings and functions, including imperfective aspect marking, at different times. Data is analyzed for the central meanings of these two characters, as constructions.

It is not really a novel idea to approach Chinese function words from characters. As early as 1825, the Germany linguist and philosopher Wilhelm von Humboldt spoke of a threefold isolation in Chinese, "The Chinese writing expresses, by a single sign, each simple word and each integral part of composed words; it suits the grammatical system of the language perfectly. The latter offers . . . a threefold isolation, of ideas (concepts), words, and characters". Wenzel (2010) further shed light on the relationship of Chinese grammar, phonological system and writing system, "The Chinese language is basically monosyllabic, has a non-alphabetic script, and offers almost no morphology (no inflections)". In the same vein, Xu (2004) and Pan (2006) proposed a "character-based method" in Chinese linguistic study and Chinese teaching. However, what is innovative about the character-based constructional approach is systematically incorporating constructional grammar with character-based method, including the fundamental tenet of lexicon-syntax continuum as well as the emphasis on the bottom-up corpus-based research method. Boas (2008) discusses how construction grammar is supposed to deal with the interactions between lexical entries and grammatical constructions, and points out that further research should be done with a bottom-up corpus-based approach. The present study is carried out along this line.

[2] Since the notion meaning is sort of problematic for some general syntactical constructions, Goldberg (2006, p. 3) began to use "function" instead: "conventionalized pairings of form and function".

4 Constructions 在 *zai* and 着 *zhe*

4.1 The construction 在 *zai*

In oracle bone inscriptions at least 3000 years ago, 在 *zai* first appeared as a verb meaning "be living; exist; be in/at … (some place)".

(11) 王 在 兹， 大 示 左。
Wang zai zi, da shi zuo.
emperor ZAI here senior master support
The emperor is here, with the support of a senior master.[3] (Chen, 2001, p.176)

(12) 朕 在 位 七十 载。 (Yi·Qian, about 1000 BC)
Zhen zai wei qishi zai.
I ZAI position seventy year
I was in the position for seventy years.

According to the etymological dictionary, 说文解字 *Shuowenjiezi* "The Explanation of Simple Graphs and Analysis of Compound Graphs" compiled by 许慎 Xu Shen (30 BC – 124 BC), the seal style and the explanation of 在 *zai* is as follows.

(13)  ，存也，从土才声。
在, to exist. (Meaning) from 土 *tu* "earth" and phonetic 才 cai.

No later than the Han Dynasty (206 BC–220 AD), 在 *zai* developed the function of a preposition which indicates "in/on/at certain time, location or range".

(14) 王夫人 在 壁后 听之。
Wang Furen zai bihou ting zhi.
Mrs. Wang ZAI next room behind listen to it
Mrs. Wang listened to it behind the next door.
(Liu, Yiqing.404–444. *Shishuoxinyu*)

(15) 在 药 则 未 为 良 时。
Zai yao ze wei wei liang shi.
ZAI medicine whereas not yet is good time
Whereas in terms of making medicine, (it is) not good time yet.
(Shen, Kuo. 1031–1095. *Mengxibitan*)

The progression marking function of 在 was not developed until Ming Dynasty (1368–1644). The earliest appearance of "在+VP" structure detected in corpus is from *Pingyaozhuan* (1620).

[3] This translation needs to be further studied, but it is apparent that *zai* is a verb in this sentence.

(16) 众人 都 在 笑。
Zhongren dou zai xiao.
everybody all ZAI laugh
Everybody is laughing.
(Luo, Guanzhong.1620. *Pingyaozhuan*)

The polysemy network (mainly functions, in this case) of 在 and the timeline of its development can therefore be represented in Figure 1.

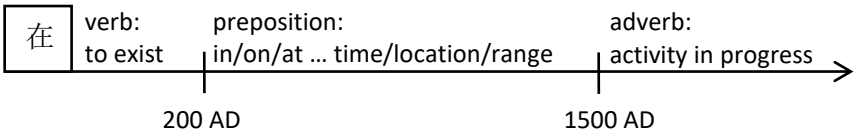

Figure 1: Development of 在

It can be seen that the meaning/function of 在 originates from the temporal domain and gradually extended to the temporal domain, which reflects the cognitive feature that spatial sense is "more central" than the temporal sense (Lakoff, 1987, pp. 416–417), and the general conceptual metaphor which maps spatial notions onto non-spatial domains (Taylor, 2008; Langacker, 1987, 1990; Talmy, 2000; and Boroditsky 2000). At the same time, even though the domains are changed, the sense "presence within a certain range" is well preserved. The verbal meaning of "to exist" can be interpreted as to occupy some space, a range in the spatial domain, and when this range happens to be in the temporal domain, according to the event types we presented in section 2, it denotes an activity.

Therefore, the form "A 在 B" can easily be understood as a presence construction, which means the presence of A in the range B. The biggest constraint is B has to cover a range in certain domain.

4.2 The construction 着 *zhe*

According to Wang (2004, pp. 357–361), 着 was originally a pure verb which means "adhere to; come into contact with; reach to". This use can at least be traced back to Warring States Period (475 BC to 221 BC). It is normally read as *zhuo* for this meaning in contemporary Chinese.

(17) 风 行 而 著於 土。
feng xing er zhuo-yu tu
wind walk and then adhere to earth
The wind walks and then adheres to the earth. (*Zuozhuan,* about 450 BC)

(18) 胡人 衣食之业 不 着於 地。
Huren yishizhiye bu zhuo-yu di.
Hu people life no adhere to place
The life of Hu people is not fixed in one pleace.
(Sun, Wu. *Bingfa* "The art of War", about 450 BC)

From the late Han Dynasty (206 BC–220 AD), 着 began to exhibit other functions like following another verb and serving as resultative verb complement (RVC).

(19) 蓝田 爱念 文度， 虽 长大，
Lantian ainian Wendu, sui zhangda,
Lantian care for Wendu although grou up

犹 抱着 膝上。
you bao-zhuo xishang
still hold-ZHUO on the knee

Lantian cares for Wendu. Although (Wendu) has already grown up,
(Lantian) still holds her on the knee. (Liu, Yiqing.404–444. *Shishuoxinyu*)

After the Southern and Northern Dynasties (220–589), the verb function of 着 disappeared in Chinese, but it is preserved in Japanese. 着く *tsuku* still means "reach; arrive at" in Japanese now. Ever since the Tang Dynasty (618–907), there could be an object after 着.

(20) 想得 家中 夜深 坐，
Xiang-de jia-zhong shenye zuo,
think about home in midnight sit

还 应 说着 远行人。
hai ying shuo-zhao yuanxingren.
still should talk about travelling person

(I) think families should be talking about the travelling person while sitting at home at midnight. (Bai Juyi, 804. *Handan Dongzhiye Sijia*)

Seeming to be rather similar to an aspect morpheme, 着 is normally pronounced as "zhao" or "zhuo" for this function in modern Mandarin and apparently bears some kind of lexical meaning of "come into contact with; adhere to; reach to". The typical progressive or durative aspect marker usage of 着 was first seen in the Song Dynasty (960–1279) and did not become common until the Yuan Dynasty (1271–1368). According to the search result from corpus, the state–continuation meaning, as shown in (21), was developed slightly earlier than the subordinating activity–progression sense, as shown in (22).

(21) 惠州 近日 科折 秣米 一事，
Huizhou jinri kezhe shumi yi-shi,
Huizhou now trade-off rice one-issue

正 违着 此 赦文。
zheng wei-zhe ci shewen.
exactly violate-ZHE this remit

The recent issue of rice trade-off in Huizhou right goes against this remit.
(Su, Shi. 1037–1101)

(22) 如 战阵 厮杀，擂着 鼓， 只是 向前 去。
Ru zhanzhen sisha, lei-zhe gu, zhishi xiang-qian qu
like battlefield fight beat-ZHE drum just toward forward go
Just like the fight on the battlefield, beating the dram, just go forward.
(Zhu, Xi. 1263. *Zhuziyulei*)

So the development of functions of 着 along the timeline can be summarized in Figure 2.

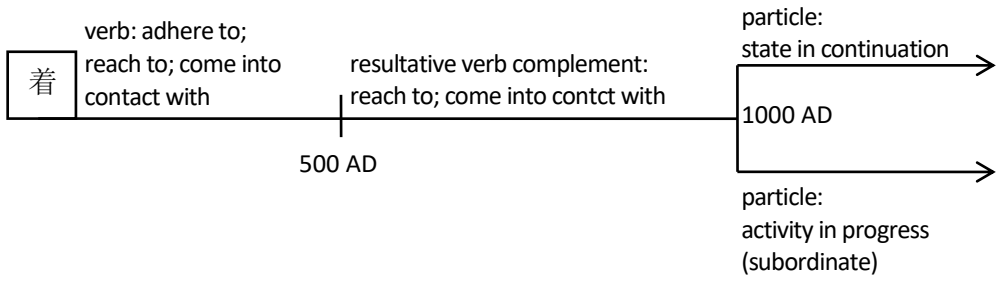

Figure 2: Development of 着

Similar to 在, the meaning of 着 also extends from the spatial domain to the temporal domain, reconfirming the "central" role of spatial perspective (Lakoff, 1987, pp. 416–417) in human cognition. The motivation underlying this extension seems to lie in its original meaning "adhere to; reach to". The earliest form "A 着 B" means "A reaches to/ comes into contact with B", from which the form "A V 着 B" was derived and 着 specifies the result, "reach to/ in contact with B". The progression of activity meaning occurs when "A V 着 B" is mapped onto the temporal domain, thus B is realized by another activity. So the "agent+ activity + 着" form we listed in section 2 can better be represented as "agent+ activity A + 着 + activity B" denoting activity A reaches to activity B in the temporal domain, just like in example (6) and (8), repeated here as (23) and (24).

(23) 说着 看了 我 一眼
shuo-zhe kan-le wo yi yan
speak-ZHE look-PERF I a glance
gave me a glance while speaking (Lü, 1999, p. 666)

(24) 说着 说着 不觉 到了 门口
shuo-zhe shuo-zhe bujue dao-le menkou
talk-ZHE talk-ZHE unconsciously arrive-PERF doorway
arrive at the doorway unconsciously while talking (Lü, 1999, p. 666)

Moreover, if the verb in "A V 着 B" represents a static state but not an activity, the "reach to" meaning can be realized without the presence of B. Here the form "A V 着 (B)" means A reached to the state V 着 (B), and this is where the state–continuation meaning comes from.[4]

(25) 门 开着。
Men kai-zhe.
door open-ZHE
The door is open.

To summarize, the central meaning of the character 着 is "reach to; in contact with". This meaning is retained in different constructions involving 着, including "A 着 B" and "A V 着 (B)".

5 Revisit Chinese imperfective aspect marking system

Assuming 在 *zai* and 着 *zhe* are the two imperfective aspect makers in Chinese, just like Li & Thompson stated in 1981, we should be able to claim under any circumstances, Chinese imperfective aspect is marked by either 在 or 着. However, there are actually some constraints involved with 在 and 着 respectively. Besides, there are some other plausible imperfective aspect markers in Chinese.

5.1 Constraints of aspect marking 在 *zai*

We have already shown 在 can co-occur with activity verb to denote activity in progress, but exception arises when the verb assigns locative as an argument. Generally all locatives need to be put between 在 and the verbs, appearing as adjunct phrases,

[4] The locative inversion sentence we mentioned in section 2 involves another ground-anchored construction (Liu, 2015) and the positions of constituents A and B are different from the basic form of 着 in 'A V 着 B'.

probably because 在 is also the commonly used pronoun to introduce locative in Chinese.

(26) a. ? 他们 在 搬 往 纽约。
Ta-men zai ban wang Niuyue.
3rd -Pl ZAI move to New York

b. 他们 在 往 纽约 搬。
Ta-men zai wang Niuyue ban.
3rd- Pl ZAI to New York move

They are moving to New York.

(27) a. *他们 今天 在 住 在 纽约。
Ta-men jintian zai zhu zai Niuyue.
3rd-Pl today ZAI stay at New York

b. 他们 今天 在 纽约 住。
Ta-men jintian zai Niuyue zhu.
3rd-Pl today ZAI New York stay

They are staying at New York today.

5.2 Constraints of aspect marking 着 *zhe*

In the first place, it has already been mentioned "agent+ activity + 着 *zhe*" is not self-sufficient. The function of 着 here is actually linking one activity to another, essentially having nothing to do with the aspect.

Another important constraint concerning 着 *zhe*'s aspect marking function is it cannot be negated. More precisely, it basically does not appear in negative form. As we talked about in section 4, the basic meaning of 着 is "reach to" and this meaning is mapped from the spatial domain to the temporal domain. So if entities, activities or states do not come into contact (either in the spatial domain or in the temporal domain), we simply do not need 着 *zhe*. The negative forms of (25) is displayed in (28).

(28) a. ? 门 没 开着。
Men mei kai-zhe.
door not open-ZHE

b. 门 没 开。
Men mei kai.
door not open

The door is not open.

5.3 Other plausible imperfective aspect markers in Chinese

Some other morphemes (characters, according to the character-based constructional approach) beside 在 *zai* and 着 *zhe* can also express imperfective aspect independently under certain circumstances, like 正 and 呢.

(29) 老师 进来 的 时候， 他们 正 吵 得 厉害。
Laoshi jin-lai de shihou, ta-men zheng chao de lihai.
teacher came-in Link time, 3rd-Pl ZHENG fight Link heatedly
They were fighting heatedly when the teacher came in.

(30) 别 说话， 奶奶 睡觉 呢。
Bie shuohua, nainai shuijiao ne.
do not talk grandma sleep NE
Grandma is sleeping. Stop taking!

If we look at the 正 from the character-based constructional perspective, its central meaning is "no deviation, right", consistent with its definition in 说文解字 *Shuowenjiezi* "The Explanation of Simple Graphs and Analysis of Compound Graphs".

(31) 正，是也，从止，一以止。
正, right, no deviation. (Meaning) from 止 zhi "foot", to walk toward one direction.

In example (29), the "no deviation" meaning is mapped onto the temporal domain, thus indicates two or more events happen exactly at the same time. Imperfective meaning is conveyed without the presence of 在 *zai* or 着 *zhe*.

As for the particle 呢 *ne*, there are various opinions regarding its functions. Considering the fact that 呢 *ne* normally occurs in the middle of discourse, this study follows Alleton (1981) and Shao's (1989) opinion that the basic function of 呢 *ne* is "to remind, appealing to the communicators' active participation." So in spoken Chinese, as long as there is proper context, it can denote imperfective aspect independently.

5.4 Section summary

From the above analysis, the relationship between 在 *zai*, 着 *zhe* and imperfective aspect marking can generally be shown as in Figure 3.

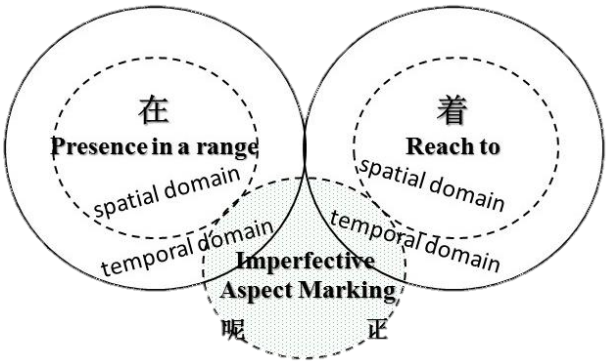

Figure 3: Relationship between Functions of 在 zai, 着 zhe and Imperfective Aspect Marking

Both 在 *zai* and 着 *zhe* have their distinct central senses, which underwent extension from spatial domain to temporal domain. In modern Chinese, they both can express imperfective aspect conditionally, but many constraints are observed at the same time. Additionally, imperfective aspect can also be expressed by other morphemes/characters in Chinese. Therefore, under the character-based constructional account, the roles of 在 *zai* and 着 *zhe* as Chinese imperfective aspect markers are questionable. We can only say, they can indicate imperfective aspect under certain circumstances, just like some other characters such as 正 *zheng* or 呢 *ne*.

6 Conclusion and implication

The character-based constructional approach believes that in Chinese, each character is a form-meaning pairing. By studying characters through historical development and with the assumption of the lexicon–syntax continuum, there can be a new perspective to look at Chinese lexicon and syntax.

Through this approach, it is discovered that the basic meaning of the character 在 is to indicate presence in a certain range and 着 is "to reach to". Their meanings and functions were originally developed in the spatial domain and were mapped onto the temporal domain later on, which reflects general human cognition principle. The process of grammaticalization is clearly exhibited here, consistent with Humboldt's hypothesis (1925) about evolutionary stage of language.

(32) Content word > grammar word > clitic > inflectional affix
(Hopper and Traugott's, 2003, p. 7)

Hopper and Traugott noted it is no coincidence that Humboldt's four stages correspond quite closely to a typology of languages that was in the air during the first decades of the nineteenth century (2003, p. 20). Chinese is basically known as an isolating

language, corresponding to the stage of "grammar word" stage in the cline according to them. It is therefore self-explanatory that Chinese grammar words are polysemous. The semantic network of a single Chinese grammar word is organized around a central sense, which, according to the character-based constructional approach, can be accessed through the corresponding character(s). For this reason, the imperfective aspect marking use of 在 z*ai* and 着 z*he* are also constrained by their basic meanings respectively, and so do some other plausible imperfective aspect markers in Chinese. In other words, as an isolating language, the imperfective aspect marking system is not maturely developed.

Hopefully, the character-based constructional approach will be able to provide some novel insights for Chinese linguistics study, and to help explain some mysterious constructions under other frameworks, like the famous 把 *ba* structure. On the other hand, blurring the traditional boundary between spoken and written language, this account may also be able to facilitate classical Chinese study and Chinese dialect study.

Last but not least, the character-based constructional approach may also bring some pedagogical implications as the biggest challenge for Chinese learners with an alphabetic native language is claimed to be Chinese characters (Allen, 2008; Bell, 1995; DeFrancis, 1984; Everson, 1988; Guder, 2005; McGinnis, 1999). Since the character-based constructional approach base vocabulary and grammar on characters, characters will be an indispensable medium instead of extra burden for teachers and learners. Introduction of individual characters can be more coherent and explanatory at the same time.

References

Allen, J. R. (2008). Why learning to write Chinese is a waste of time: A modest proposal. *Foreign Language Annals, 41,* 237–251.

Alleton, Viviane (1981). Final particles and expression of modality in Modern Chinese. [J].*Journal of Chinese Linguistics 9,* 90–115.

Baredal, J. (2011). Lexical vs. structural case: A false dichotomy. *Morphology 21*(1), 619–654.

Bell, J. S. (1995). Relationship between L1 and L2 literacy: Some complicating factors. *TESOL Quarterly, 29,* 687–704.

Boas, H. C. (2008). Determining the structure of lexical entries and grammatical constructions in Construction Grammar. *Annual Review of Cognitive Linguistics.* http://sites.la.utexas.edu/hcb/files/2011/02/06boa.pdf

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, *75,* 1–28.

Bybee, J. L. (2003). Aspect. In: W. J. Frawley (Eds.), *International Encyclopedia of Linguistics* (pp. 157–158). Oxford, UK: Oxford University Press.

Bybee, J. L. (2006). *Frequency of use and the organization of language*. Oxford, UK: Oxford University Press.

Chen, G. (1980). On comparison between the usage of -zhe and English progressive. In Z. J. Yang & R. H. Li (Eds.), *Collections of Articles of Comparison between English and Chinese* (pp. 379–390). Shanghai: Shanghai Foreign Language Education Press.

Chen, N. (2001). Jiaguwen dongci cihui yanjiu [Study of the verbs in oracle bone inscriptions]. Chengdu: Bashu Publishing House.

Chen, P. (1998). Lun xiandai hanyu shijian xitong de sanyuan jiegou [On tripartite organization of the temporal systmen in modern Chinese]. *Zhongguo yuwen, 6,* 401–422.

Comrie, B. (1976). *Aspect.* Cambridge, UK: Cambridge University Press.

Croft, W. (2003). Lexical rules vs. constructions: a false dichotomy. In H. Cuyckens, T. Berg, R. Dirven & K.-U. Panther (Eds.), *Motivation in language: Studies in honour of Günter Radden* (pp.49–68). Amsterdam: John Benjamins.

Dai, Y. (1991). Xiandai hanyu biaoshi chixuti de *zhe* de yuyifenxi [Semantic analysis of *zhe* denoting continuative aspect in modern Chinese]. *Language Teraching and Linguistic Studies, 1991(2),* 92–106.

DeFrancis, J. (1984). *The Chinese language: Fact and fantasy*. Honolulu: University of Hawaii Press.

DeFrancis, J. (1989).*Visible Speech: The diverse oneness of writing systems*. Honolulu: University of Hawaii Press.

Diringer, D.(1962). *Writing*. Westport, Connecticut: Praeger.

Dowty, D. R. (1979). *Word meaning and montague grammar*. Dordrecht: D. Reidel Publishing Company.

Everson, M. E. (1988). Speed and comprehension in reading Chinese: Romanization vs. characters revisited. *Journal of the Chinese Language Teachers Association, 23,* 1–15.

Fabar, A.(1992).Phonemic segmentation as an epiphenomenon. Evidence from the history of alphabetic writing. In P. Downing, S. D. Lima & M. Noonan (Eds), *The linguistics of literacy* (pp. 111–134). Amsterdam: John Benjamins.

Fang, M. (2000). Cong *Vzhe* kan hanyu buwanquan ti de gongneng. [On the functions of incomplete aspect in Chinese through *Vzhe*]. In Chinese Language Magazine House (Eds.), *Grammar Study and Exploration (Vol. 9)* (pp. 212–221). Beijing: The Commercial Press.

Foley, W., & Van Valin, R. (1984). *Functional syntax and universay grammar*. Cambridge: Cambridge University Press.

Gao, M. (1986). *Hanyu yufa lun. [Studies on Chinese grammar]*. Beijing: The Commercial Press.

Goldberg, A. (2006). Constructions at work. The nature of generalization in language. Oxford, UK: Oxford University Press.

Guder, A. (2005, August). Struggling with Chinese: New dimensions in foreign language teaching. Paper presented at the International and Interdisciplinary Conference, University of Mainz, Germersheim, Germany.

Guo, R. (1997). Guocheng he feiguocheng- hanyu weicixing chengfen de liangzhong waizai shijian leixing [Process and non-process: two extrinsic temporal types of Chinese predicative constituents]. *Chinese Language*, *1997*(3), 162–175.

Hill, A. (1967). The typology of Writing systems. In W. A. Austin (Eds), *Papers in linguistics in honor of Leon Dostert* (pp. 92–99). The Hague: Mouton.

Hopper, P. J., & Traugott, E. (2003). *Grammaticalization*. Cambridge: Cambridge University Press.

Huang, C.-T. J., Li, Y.-H. A., & Li, Y. (2009). *The syntax of Chinese*. New York: Cambridge University Press.

Humboldt, W. (1825). Über die Entstehung der grammatischen Formen und ihren Einfluss auf die Ideenentwicklung. Berlin: Abhandlungen der Akademic der Wissenschaften zu Berlin.

Jiang, Y., & Pan, H. (1998). *Xingshi yuyixue yinlun [Introduction to formal semantics]*. Beijing: China Social Science Press.

Kibort, A. (2008). "Aspect." Grammatical Features. 7 January 2008. http://www.grammaticalfeatures.net/features/aspect.html.

Kwan-Terry, A. (1978). Two progressive aspect markers in Chinese. In N. G. Liem (Eds.), *South-east asian linguistic studies: vol. 4* (pp. 213–232). Canberra: Pacific Linguistics, the Australian National University.

Lakoff. G. (1987). Women, fire, and dangerous things. Chicago: University of Chicago Press.

Lakoff. G. (1990). The invariance hypothesis: Is abstract reason based on image schemas? *Cognitive Linguistics*, *1,* 39–74.

Langacker. R. W. (1987). *Foundations of cognitive grammar: theoretical prerequisite*. Stanford, California: Stanford University Press.

Langacker. R. W. (1990). Concept, image, and symbol: The Cognitive Basis of Grammar. Berlin: Mouton de Gruyter.

Langacker. R. W. (2008). Cognitive grammar as a basis for language instruction. In P. Robinson, & N.C.Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 66–88). London: Routledge.

Li., C. & Thompson, S. (1981). *Mandarin Chinese: A functional reference grammar*. Berkeley: University of California Press.

Liu, M. & Chang, J.-C. (2015, April). Redefining Locative Inversion in Mandarin: A Lexical–constructional Approach. Paper presented at the 27th North American Conference on Chinese Linguistics, Los Angeles, California. Abstract retrieved from http://chineselinguistics.org/Events/NACCL-27/NACCL27_Abstracts.pdf.

Liu, N. (1985). Lun *zhe* jiqi xiangguande liangge dongtai fanchou [On *zhe* and two related dynamic categories]. *Chinese Research,1985*(2), 117–128.

Lü, S. (1999). Xiandai hanyu babai ci [Eight hundred words in modern Chinese]. Beijing: The Commercial Press.

McGinnis, S. (1999). Students' goals and approaches. In M. Chu (Eds.), *Chinese Language Teachers Association monograph series: Vol. III. Mapping the course of the Chinese language field.* Kalamazoo, MI: Chinese Language Teachers Association.

Pan, W. (2006). Zibenwei lilun de zhexue sikao (Philosophical Thinking of character-based theory). *Yuyan jiaoxue yu yanjiu [Language teaching and research], 3,* 36–45.

Saussure, F. (1983). *Course in general linguistics*. Peru, Illinois: Open Court Publishing.

Shao, J. (1989). Yuqici *ne* zai yiwenju zhong de zuoyong [The Function of the Expressive Particle *ne* in Interrogatives]. *Zhongguo Yuwen [Chinese Language]. 3,* 170–175.

Smith, C. (1991). *The parameter of aspect. Dordrecht*. The Netherlands: Kluwer Academic Press.

Su, P. (2001). Xiandai hanzixue Gangyao [An outline of modern Chinese characters]. Beijing: Peking University Press.

Tai, J. H.-Y. (1984). Verbs and times in Chinese: Vendler's four categories. In D. Testen (Eds.), *Papers from the Parasession on Lexical Semantics* (pp.23–35). Chicago: Chicago Linguistics Society.

Talmy, L. (2000). Towards a cognitive semantics, Vol. 1: Concept structuring systems. Cambridge, MA: MIT Press.

Taylor. J. R. (2008). Protypypes in Cognitive Linguistics.In P. Robinson, & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp.39–66). London: Routledge.

Travis, L. deMena. (2010). *Inner Aspect*. New York: Springer Science & Business Media.

Tyler, A. (2012). Cognitive linguistics and second language learning–theoretical basics and experimental evidence. New York and London: Taylor & Francis Group, 2012.

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

Wang, L. (2004). *Hanyu shigao. [Chinese history]*. Beijing: Zhonghua Book Company

Wenzel, C. H. (2010). Isolation and involvement: Wilhelm von Humboldt, François Jullien, and more. *Philosophy East & West, 2010* (4), 458–475.

Xu, T. (2004). Hanyu yanjiu fangfalun chutan [In the methodology of Chinese study]. Beijing: The Commercial Press.

Xuan, Y. (2013). Mingci duanyu de youjiexing yu shuliang duanyu de guanxi [The relationship between boundedness of NP and Quantitative phrase]. In Y. Shen (Eds.), *Zouxiang dangdai qianyan kexue de xiandaihanyu yufa yanjiu [The Chinese grammar study that is approaching forward science]* (pp. 372–391). Beijing: The Commercial Press.

Yang, G., & Bateman, J. A. (2002). The Chinese aspect system and its semantic interpretation. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02) Vol. 1* (pp. 1–7). Taipei: Taiwan.

Yuan, Y. (1992). Qishi jushi *V+zhe* fenxi [Analysis of *V+zhe* imperative sentence]. *Chinese Teaching in the World, 1992* (4), 259–275.

Zhang,Y. (2000). Xiandai hanyu xuci [Functional words in modern Chinese]. Shanghai: Academia Press.

PREFIXATION ABILITY INDEX AND VERBAL GRAMMAR CORRELATION INDEX PROVE THE REALITY OF THE BUYEO GROUP

Alexander AKULOV
Independent scholar; St. Petersburg, Russia
aynu@inbox.ru

Abstract

All suggestions about reality of the Buyeo group were based on the representation of a language as a heap of lexemes: such method allows different scholars to make different conclusions and does not suppose verification. Language is first of all structure/grammar, but not a heap of lexemes, so methods of comparative linguistics should be based on comparison of grammars. Prefixation Ability Index (PAI) and Verbal Grammar Correlation Index (VGCI) are typology based tools of comparative linguistics. PAI allows us to see whether languages are potentially related: if values of PAI differ more than fourfold, it's a sign of unrelatedness, if PAI values differ less than fourfold, there is a possibility for some further search to find proves of relatedness. VGCI completely answers questions about relatedness/unrelatedness: if VGCI value is 0.4 and more then languages are related, if VGCI is 0.3 and less then languages are unrelated. PAI of Japanese is 0.13, PAI of Korean is 0.13; it means they can be related. VGCI of Japanese and Korean is 0.57, it's almost the same as VGCI of English and Afrikaans that is 0.56, so it means that Japanese and Korean belong to the same group, but not just to the same family.

Keywords: the Buyeo group; Japanese; Korean; comparative linguistics; linguistic typology

Povzetek

Dosedanji predlogi o jezikovni skupini Buyeo so povečini osnovani na osnovi leksemov raziskovanih jezikov. Takšna metoda omogoča raziskovalcem različne zaključke in ne zahteva njihovih preveritev. Članek poudari, da jezik ni le kopica leksemov, temveč je primarno definiran s s vojo strukturo/slovnico, zato bi morale metode v primerjalnem jezikoslovju temeljiti predvsem na primerjavi slovnic primerjanih jezikov. Index možne uporabe predpon (Prefixation Ability Index ali PAI) in Korelacijski index glagolske slovnice (Verbal Grammar Correlation Index ali VGCI) sta dve orodji primerjalnega jezikoslovja, ki temeljita na jezikovni tipologiji. PAI nam omogoča vpogled v možnost jezikovne povezanosti raziskovanih jezikov; ko njegova vrednost preseže 0.4, le-ta nakazuje njihovo nepovezanost, njegova vrednost pod četrtino pa obuja možnost, da z nadaljnjimi raziskavami dokažemo povezanost med raziskovanimi jeziki. Nadalje pa orodje VGCI dokončno odgovori na vprašanje o jezikovni povezanosti med jeziki; če je njegova vrednost višja od 0.4, lahko rečemo, da so jeziki med seboj povezani, v nasprotnem primeru pa ne. PAI tako japonskega kot korejskega jezika je 0.13. To nakazuje, da obstaja možnost o njuni medsebojni povezanosti. Njun VGCI pa je 0.57, kar nakazuje, da jezika ne pripadata samo isti jezikovni družini, ampak jih lahko uvstimo v isto jezikovno skupino, podobno kot isti skupini pripadata angleščina in jezik afrikaans, katerih VGCI je 0.56.

Ključne besede: skupina Buyeo; japonščina; korejščina; primerjalno jezikoslovje; jezikovna tipologija

Acta Linguistica Asiatica, 6(1), 2016.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.6.1.81-97

1 Problem introduction

Buyeo is a conventional name of hypothetical stock that includes Japanese language, Ryukyuan languages and Korean language.

Ogura Shimpei, Sergei Starostin, Cristopher Beckwith have suggested the reality of the Buyeo stock (Beckwith, 2007; Ogura, 1934; Starostin, 1991). Alexander Vovin has suggested that Japanese and Korean are not related (Vovin, 2010).

Main problem of all these suggestions (as well as of most of hypotheses about certain languages relationship) is that they are not based on any verifiable methods. All such suggestions are based mainly on the idea that language is just a heap of lexemes but not grammar. Such approach does not suppose any verification and so different scholars are allowed to make contradictory conclusions about the same material: probably some conclusions are right, but the absence of appropriate methods of verification makes it impossible to understand what is right and what is wrong. Actually it pushes comparative linguistics outside of the field of science: science always supposes verification and also supposes rejection of unproven hypotheses, while methodology based on "artist sees so" principle does not suppose any verification and so contradictory conclusions can coexist.

As long as language is first of all grammar, conclusion about genetic affiliation of certain language should be made on the base of analysis of grammar (Akulov, 2015d).

Current paper are represents the proofs based on typology/grammar which show that Japanese and Korean are closely related.

2 Methods

2.1 Prefixation Ability Index (PAI)

Prefixation Ability Index (PAI) allows us to see whether two languages can potentially be genetically related.

PAI is a method to estimate the percentage of prefixes in a language. It presupposes that any language has its own prefixation ability, which is then measured as percentage of prefixes among affixes. In order to estimate percentage of prefixes (PAI), the following steps should be undertaken:

1) Count total number of prefixes;
2) Count total number of affixes;
3) Calculate the ratio of total number of prefixes to the total number of affixes.

It is generally believed that PAI of genetically related languages is close in its values; and tests of PAI on the material of firmly assembled stocks (Indo-European, Austronesian, Afroasiatic) show that PAI values of distant relatives can differ maximum fourfold. (A detailed description of PAI method can be seen in Akulov, 2015a.)

Thus PAI can be used as a tool that allows us to see whether certain languages can potentially be related: no conclusions can be made when PAI values differ fourfold or less (for instance in the case of Indo-European and Austronesian), but if PAI values of certain languages differ, for instance, tenfold (the case of Ainu and Nivkh; see Akulov, 2015a, p. 13) it is an evidence that considered languages are not related.

PAI could be calles a safety valve of comparative linguistics: if its values do not differ more than fourfold then there are no obstacles for further search for genetic relationship; if values differ about fourfold then should be found ferroconcrete proves of genetic relationship (like for instance those that were shown in the case of Semitic group and Coptic language); if values differ sevenfold – tenfold or even more then considered languages belong to completely different stocks.

It is possible to say that PAI shows direction in which looking for potential relatives of certain language can be perspectives.

2.2 Unrelatedness

An important point of current consideration is possibility of proving the unrelatedness of languages.

This is a necessary tool of any classification as well as possibility of proving of relatedness: if there would be no possibility to prove unrelatedness then even a single stock hardly could be assembled.

Possibility of proving of unreltedness is discussed and proved in the following papers: Akulov (2015c) and Brown (2015).

2.3 Verbal Grammar Correlation Index (VGCI)

VGCI is thought to be the main tool in a search for language relatedness, so a more detailed description of VGCI method is given below.

2.3.1 VGCI method background

As seen in a previous section, PAI allows us to see whether languages are potentially related. However, in order to be able to say whether two languages are related, we need the that would pay attention to grammar and consequently give precise results.

As long as language is structure, i.e. grammar, language relatedness should be understood on the comparison of their grammars.

Grammar is first of all positional distributions of grammatical means, i.e. ordered pair of the following view: <A; Ω> where A is a set of grammatical meanings and Ω is a set of operations defined on A or positional distributions.

In order to understand whether two languages are genetically related we should analyze the degree of correlation among sets of grammatical meanings and estimate the proximity of positional distributions of common grammatical meanings.

2.3.2 Why does the method entail verbs?

Why is itpossible to give conclusions on the relatedness or unrelatedness of languages considering only verbal grammar? The answer lies in the fact that there are many languages with a poor or almost no grammar of nouns while there is no language without verbal grammar. In other words, there are languages with no grammatical case or gender (even very closely related language can differ in that case, for instance, English and German, or Russian and Bulgarian), but there are no languages without modalities, moods, tenses, and aspects. Therefore a verb is thought to be the backbone of any grammar, and the backbone of comparative method.

2.3.3 General scheme of VGCI calculation

As written in section 2.3.1, the following steps should be taken to estimate grammar correlation:

1. Correlation of grammar meanings sets is estimated in the following way. First, the intersection of two sets of grammatical meanings should be found. After calculating the intersection ratio to each set, arithmetical mean of both ratios should be taken. The value represents the index of sets of grammar meanings correlation.
2. Sets of meanings alone do not yet fully describe grammar systems. The second step is to estimate the correlation of positional distributions of common grammatical sets of meanings. Intersection of two sets of grammatical meanings would give us information on the degree of positional correlation.
3. In order to calculate values of VGCI we should take a logical conjunction of the correlation of the degree of grammatical meanings and the degree of correlation for positional distributions of common grammatical meanings. In other words, VGCI as a multiplication of the two indexes.
4. It is obvious that languages which are genetically closely related demonstrate higher values of VGCI – the more sets of grammatical meanings are alike, the higher the intersection ratio, and consequenlty, the more alike positions of common grammatical meanings –, while languages with low or no relatedness will demonstrate low values of VGCI.
5. According to the previous step, there should be a threshold value of VGCI which determines the border of stocks, i.e.: if certain languages execute values lower than the threshold, such language evidently do not belong to the

same stock. In order to determine the threshold I will compare distant languages of well assembled stocks.

The above method enables a direct comparison of natural languages that exist or have existed, but not of their reconstructions or constructed languages. Descriptions of the latter are under the influence of personal views of the authors and can not be verified anyhow.

I am also to note that the method supposes comparison of meanings and their positional distributions only an does not pay any attention to material exponents at all. It is not a response to radical adepts of *megalocomparison* (the term introduced by James Matisoff: see Matisoff 1990), which harshly ignores typological issues. It is rather a matter of reality and practice since material correlation (regular phonetic correspondence) between languages that are only related weakly can be very complicated. The method is intended to prove genetic relatedness or unrelatedness by pure typology.

Therefore the attention is not paid to technical meanings such as markers of transitivity for example, but rather to the so-called contensive grammatical meanings such as markers of tenses, aspects, modalities etc. In other words, the attention is paid to those grammatical categories that have certain contents expressed by lexical means. Only if necessary, items that express technical meanings (meanings of agreement) can also be taken into account.

It can be a rather complicated task to distinguish obligatory features of verbs from the facultative ones, so first of all attention should be paid to the following categories:

a) tense and aspect;
b) mood and modality;
c) voice;
d) agent, patient, object, subject, numbers

There can be certain categories used as evidences for a kind of modality or spatial orientation/versions, and are considered as a development of triggers system. Therefore it is very important to make precise descriptions of the languages compared though sometimes same items can be described in a slightly different way.

2.3.4 Results of VGCI testing: values of thresholds

Tests of VGCI on the material of firmly assembled stocks have given us the following values:

VGCI of English and Russian ≈ 0.52;
VGCI English and Lithuanian ≈ 0.43;
VGCI English and Latin ≈ 0.41;
VGCI English and Persian ≈ 0.38;
VGCI of Khmer and Vietnamese ≈ 0.53;

VGCI Hawaiian and Lha'alua ≈ 0.39;
VGCI (Chinese and Tibetan) ≈ 0.39.

On the other hand, tests of VGCI on the material of unrelated languages have shown us the following:

VGCI (Chinese and English) ≈ 0.32;
VGCI (Chinese and Latin) ≈ 0.30;
VGCI (Khmer and Latin) ≈ 0.29;
VGCI (English and Tibetan) ≈ 0.13.

If value of VGCI is around or above 0.4 then languages are related, i.e. they belong to the same stock). If on the other hand value of VGCI is about 0.3 or less than 0.3 then languages are not related. Values such as 0.39 and 0.38 are closer to 0.4, while 0.31 and 0.32 are closer to 0.3. The closer languages are related the higher is their corresponding VGCI.

2.3.5 Measurement error

Details on measurement error are described in a separate paper (Akulov 2015b). It was calculated to about 2%.

3 Applying PAI and VGCI methods to the Buyeo problem

The main problem of the Buyeo stock is the question of relatedness of Japanese and Korean. In this paper I try to show the relatedness of Japanese and Korean with the use of PAI and VGCI methods.

3.1 PAI suggests that Japanese and Korean can potentially be related

Lavrent'yev (2002) calculated the PAI for Japanese to be 0.13, and Mazur (2004) reported that the PAI for Korean is 0.13.

It is rather interesting that the PAI values demonstrate such similarities. However, as it has been noted in section 3.2.1, no conclusions can be made from similar PAI values. Nevertheless, such similarity is promising and further research might bring us to the proofs of their relatedness.

3.2 VGCI proves close relationship of Japanese and Korean

3.2.1 VGCI of Japanese and Korean

3.2.1.1 List of Japanese forms

The following list of Japanese forms has been compiled by Lavren'tyev 2002.

1. Active: zero marker
2. Agent: [prp-]
3. Attemptive: -sfx + -pp
4. Causative: -sfx
5. Conditional (real): -sfx
6. Conditional (unreal): -sfx
7. Centrifugal version: -sfx + -pp
8. Centripetal version: -sfx + -pp
9. Deontic1 -sfx + -pp + -pp
10. Deontic2 -sfx + -pp + -pp + -pp
11. Deontic3 -sfx + -pp
12. Desiderative: -sfx/-sfx + -pp
13. Directive (benefactive): -sfx + $-pp_1$/-sfx + $-pp_2$
14. Directive (from subject): -sfx + -pp
15. Directive (to subject): -sfx + $-pp_1$/-sfx + pp_2
16. Hortative: -sfx
17. Imperative: -sfx/inner fusion + -sfx
18. Indicative: zero marker
19. Iterative/Frequentative: -sfx + -pp
20. Interrogative: -pp
21. Negation: -sxf + $-sfx_1$/-sfx + $-sfx_2$
22. Passive: -sfx
23. Passive causative: -sfx
24. Past continuous: -sfx + -pp
25. Past perfect: -sfx + pp/-sfx + pp/-sfx + -pp/-sfx + -pp
26. Past simple: -sfx/-inner fusion + -sfx
27. Patient: [prp-]
28. Permissive: -sfx + -pp
29. Politeness (formal) -sfx
30. Politeness (plain) -sfx
31. Potential: -sfx + -sfx/-sfx + -pp + -pp + -pp
32. Present continuous: -sfx + -pp
33. Present-Future: -sfx
34. Present perfect: -sfx + -pp/-sfx + -pp/-sfx + -pp/-sfx + -pp
35. Prohibitive: -sfx + $-pp_1$/-sfx + $-pp_2$
36. Subject: [prp-]

It is obvious that each grammatical meaning is followed by certain schemes of letters and signs. These are notations representing general schemes of positional implementation for the grammatical meaning concerned.

Notations of positional implementations are the following: prp- – preposition; prfx- – prefix; -infx- – infix; crfx-crfx – circumfix; crp-crp – circumposition; -RR- – reduplication; inner fusion – irregular changes inside the root; suppletivism; R – root; -sfx – suffix; -pp – post position. In case of a different form of one position (i.e. forms

used in different contexts), they are numbered as prp_1-/prp_2-/prp_3- and distinguished by slash. Positional elements that are components of the same implementation are expressed as prp- + -sfx., where - means that certain positional element can optionally be omitted and if written in square brackets, it is not obligatory.

Such a notation shows grammatical meanings and their positions in relation to a nuclear position rather than their absolute positions in a linear model of word or phrase. To state an example, it is of no importance which prefix is placed closer to the nuclear position; for the current tasks it is sufficient to know that al prefixes are placed left from the nuclear position.

It is important to note that this way of notation carries information on places and technical means of expressions concerning grammatical meanings. The so-called "school grammar" offering the number of verbal stems in a certain language, for example, is not of my interest. I consider language as something like a dark box with many holes, and implementation of certain grammatical meanings is the light coming out of those holes. My task is to record in what holes the light appears, and then to compare recordings of different boxes (i.e. different languages).

3.2.1.2 List of Korean forms

List of Korean forms has been compiled by Mazur 2004.

1. Active: zero marker
2. Agent: [prp-]
3. Attemptive: -sfx/-sfx + -sfx
4. Causative: -sfx
5. Conditional (real): -sfx + -pp
6. Conditional (unreal): -sfx + -sfx + -pp
7. Deontic: -sfx + -sfx + -pp/-sfx + -pp
8. Desiderative1: -sfx + -pp + -pp
9. Desiderative 2: -sfx + -pp
10. Directive: (from subject) -sfx + -pp + -pp /-sfx + -pp
11. Future simple 1: -sfx
12. Future 2: -sfx + -pp +
13. Hortative: -sfx_1/-sfx_2/-sfx+ -sfx
14. Imperative: -sfx_1/-sfx_2/-sfx_3/R
15. Indicative: zero marker
16. Interrogative: -sfx
17. Negation: -sxf + -pp
18. Passive: -sfx
19. Past simple: -sfx/inner fusion + -sfx
20. Patient [prp-]
21. Permissive: -sfx + -sfx + -pp/-sfx + -pp
22. Plain style: -sfx/R
23. Polite style (middle) -sfx/-sfx + -sfx
24. Polite style (very formal): -sfx + -sfx

25. Potential: -sfx + -pp + -pp
26. Present continuous: -sfx + -pp
27. Present simple: -sfx
28. Prohibitive: -sfx + -pp + -pp
29. Subject [prp-]

3.2.1.3 Japanese^Korean

"^" is a sign for VGCI operation.

1. Active: J: zero maker ~ K: zero marker 1
2. Agent: J: [prp-] ~ K: [prp-] 1
3. Attemptive: J: -sfx + -sfx ~ L: -sfx/-sfx + -sfx 0.75
4. Causative: J: -sfx ~ K: -sfx 1
5. Conditional (real): J: -sfx ≠ K: -sfx + -pp 0
6. Conditional (unreal): J: -sfx ≠ K: -sfx + -sfx + -pp 0
7. Deontic: J: -sfx + -pp + -pp/-sfx + -pp + -pp + -pp/-sfx + -pp ~ K: -sfx + -sfx + -pp/-sfx + -pp: 0.66
8. Desiderative: J: -sfx/-sfx + -pp ~ K: -sfx + -pp + -pp/-sfx + -pp 0.5
9. Directive (from subject) J: -sfx + -pp ~ K: -sfx + -pp + -pp /-sfx + -pp 0.75
10. Hortative: J: -sfx ~ $-sfx_1$/$-sfx_2$/-sfx+ -sfx 0.66
11. Imperative: J: -sfx/inner fusion + -sfx ~ $-sfx_1$/$-sfx_2$/$-sfx_3$/R (1/2 + 1/4)/2 = 0.375
12. Indicative: J: zero marker ~ K: zero marker 1
13. Interrogative: J: -pp ~ K: -sfx 1
14. Negation: J: -sxf + $-sfx_1$/-sfx + $-sfx_2$ ~ K: -sxf + -pp 0.75
15. Passive: J: -sfx ~ K: -sfx 1
16. Past simple: J: -sfx/inner fusion + -sfx ~ K: -sfx/inner fusion + -sfx 1
17. Patient: J: [prp-] ~ K: [prp-] 1
18. Permissive: J: -sfx + -pp ~ K: -sfx + -sfx + -pp/-sfx + -pp 0.75
19. Politeness (formal): J -sfx ~ K: -sfx + -sfx 0
20. Politeness plain: -sfx ~ K: -sfx/R 0.75
21. Potential: J: -sfx + -sfx/-sfx + -pp + -pp + -pp ~ K: -sfx + -pp + -pp 0.75
22. Present continuous: J: -sfx + -pp ~ K: -sfx + -pp 1
23. Present simple: J: -sfx ~ K -sfx 1
24. Prohibitive: J: -sfx + $-pp_1$/-sfx + $-pp_2$ ~ K: -sfx + -pp + -pp 0.75
25. Subject: J: [prp-] ~ K: [prp-] 1

(25/29 + 25/36)/2*(11 + 7*0.75 + 2*0.66 + 0.5 + 0.37)/25 ≈ 0.57

The following is the brief explanation of notation scheme: first comes the name of a grammatical meaning that is common for both the compared languages (or meanings that are correlated), which is then followed an abbreviation of the name of the first of the compared languages and first language schemes of expressions of the grammatical meaning. A sign of correlation "~" or anti-correlation "≠" comes in between the two languages, then abbreviation of the name of the second language and

its ways of expressions of the grammatical meaning. Finally, the number that expresses degree of correlation is written down. If a certain meaning can be expressed in several ways, options are separated by a slash; in case there are some similar items expressing the same meaning, they are marked by lower index numbers. Also, if there is no difference in positional expressions schemes, this point is counted as 1, and if there is no correlation, corresponding point is counted it as 0, while in other cases particular degree of correlation is estimated. It is supposed that, for instance, the case of -sfx and -pp execute the same full correlation as -sfx and -sfx; while -sfx and -sfx + -sfx show zero correlation.

VGCI of Japanese and Korean is higher than VGCI of Khmer and Vietnamese (VGCI=0.53) or VGCI of English and Russian (VGCI = 0.52), and this brings us to the conclusion that Japanese and Korean belong to the same group rather than just to the same stock. In order to verify if this is so, VGCI of Japanese and Korean is compared with VGCI of languages that evidently belong to the same group.

3.2.2 VGCI values of closely related languages: English and Afrikaans

3.2.2.1 List of English forms

The list of English forms has beend compiled by Barhkhudarov et al., 2000.

1. Active voice: zero marker
2. Agent: prp-/ [prp-] +6 -sfx
3. Causative: prp-
4. Conditional mood: prp-
5. Deontic: prp-$_1$/prp-$_2$/prp-$_3$/prp-$_4$
6. Desiderative: prp-$_1$/prp-$_2$
7. Future continiuous: prp + prp + -sfx
8. Future perfect: prp- + prp- + inner fusion/ prp- + prp+ -sfx
9. Future perfect continuous: prp + prp + prp- + -sfx
10. Future simple: prp-
11. Horative: prp-
12. Imperative: R
13. Impossibility: prp-
14. Indicative: zero marker
15. Interrogative: prp-
16. Negation: prp-
17. Optative: prp-
18. Passive voice: prp- + -sfx/prp- + inner fusion
19. Past continuous: prp + -sfx
20. Past perfect: prp + -sfx / prp + inner fusion
21. Past perfect continuous: prp- + prp + -sfx
22. Past simple: inner fusion/suppletivism/-sfx
23. Patient: -pp

24. Plural number: prp- / [prp-] +3 -sfx
25. Possibility: prp-
26. Present continuous: prp + sfx
27. Present perfect: prp + sfx/ prp + inner fusion
28. Present perfect continuous: prp- + prp + -sfx
29. Present simple: 6 -sfx
30. Prohibitive: prp-$_1$/prp-$_2$
31. Singular number: prp- / [prp-] +3 -sfx
32. Subject: prp- / [prp-] +6 -sfx
33. Subjunctive mood: prp-

3.2.2.2 List of Afrikaans forms

List of Afrikaans forms compiled by Mironov, 2000.

1. Active: zero marker
2. Agent: prp-
3. Causative: prp-
4. Conditional: prp-
5. Deontic: prp-
6. Desiderative: prp-$_1$/prp-$_2$
7. Future perfect: prp + prfx-/prp-/prp- + prfx + inner fusion
8. Future simple: prp-
9. Imperative: R
10. Interrogative: prp-
11. Negation: -pp
12. Passive prp- + prfx/prp-/prp- + prfx- + inner fusion
13. Past simple: inner fusion/-sfx
14. Patient: -pp
15. Plural number: prp-
16. Potential: prp-$_1$/prp-$_2$
17. Present perfect ~ Past simple: prp- + prfx-/prp-/prp- + prfx- + inner fusion
18. Present simple: [prp-] + R
19. Prohibitive: crp-crp
20. Singular number: prp-
21. Subject: prp-
22. Subjunctive: prp-

3.2.2.3 English^Afrikaans

1. Active: E: zero marker ~ Af: zero marker 1
2. Agent: E: prp-/ [prp] +6 -sfx ~ Af: prp- 0.75
3. Causative: E: prp- ~ Af: prp- 1
4. Conditional: E: prp- ~ Af: prp- 1
5. Deontic: E: prp-$_1$/prp-$_2$/prp-$_3$/prp-$_4$ ~ Af: prp- (1+1/4) /2 ≈ 0.62
6. Desiderative: E: prp-$_1$/prp-$_2$ ~ prp-$_1$/prp-$_2$ 1

7. Future perfect: E: prp- + prp- + inner fusion/ prp- + prp+ -sfx ~ Af: prp + prfx-/prp-/prp- + prfx + inner fusion (1/2 +1/3) ≈ 0.42
8. Future simple: E: prp ~ Af: prp 1
9. Imperative: E: R ~ Af: R 1
10. Interrogative: E: prp- ~ Af: prp- 1
11. Negation: E: prp ≠ Af: -pp 0
12. Passive: E: prp- + -sfx/prp- + inner fusion ~ Af: prp- + prfx/prp-/prp- + prfx- + inner fusion 0
13. Past simple: E: inner fusion/suppletivism/-sfx ~ Af: inner fusion 0.66
14. Patient: E: -pp ~ Af: -pp 1
15. Plural number: E: prp-/ [prp-] +3 -sfx ~ Af: prp- 0.75
16. Potential: E: prp- ~ Af: $prp-_1$/$prp-_2$ 0.75
17. Present perfect: E: prp + sfx/ prp + inner fusion ~ Af: prp- + prfx-/prp-/prp- + prfx- + inner fusion (1/2 +1/3) /2 ≈ 0.41
18. Present simple: E: 6 -sfx ≠ Af: [prp-] + R 0
19. Singular number: E: prp-/ [prp-] +3 -sfx ~ Af: prp- 0.75
20. Subject: E: prp- / [prp-] +6 -sfx ~ Af: prp- 0.75
21. Subjunctive: E: prp- Af: prp- 1

(21/22 +21/33) /2* (9 +5*0.75 +0.62 +0.66 +0.42 +0.41) /21 ≈ 0.56

3.3 Buyeo languages form a group

These are the calculated VGCI values:

Japanese^Korean ≈ 0.57
English^Afrikaans ≈ 0.56

VGCI values of Japanese and Korean show a very similar relation as is attested by languages of firmly established language groups, which brings us to the conclusion that Buyeo languages belong to the same language group.

Schemes and diagrams represented below graphically show proximity of Japanese and Korean grammars. Schemes of English and Afrikaans are shown to illustrate similarities.

	A	B	C	D	E	F	G
1		prp	prfx	inner fus.	R	sfx	pp
2	Agent	J1; K1					
3	Attempt.					K3; J0,5	J0,5; K0
4	Causative					J1; K1	
5	Centrifug					J0,5; K0	J0,5; K0
6	Centrip.					J0,5; K0	J0,5; K0
7	Cond. R.					J1; K0,5	K0,5; J0
8	Cond. Un.					J1;K0,66	K0,33; J0
9	Deontic1					J1/3; K2/3	J2/3; K1/3
10	Deontic2					J0,25; K0	J0.75; K0
11	Deontic3					J0,5; K0,5	J0,5; K0,5
12	Desid.1					J1; K0,33	K0.66; J0
13	Desid.2					J0,5; K0,5	J0,5; K0,5
14	Direct 1					J2; K0	J2; K0
15	Direct 2					J2; K0	J2; K0
16	Direct 3					J0,5;K0,83	J0,5; K1,1
17	Fut. S.					K1; J0	
18	Fut. 2					K 0,5; J0	K0,5; J0
19	Hortative					J1; K4	
20	Imp.			J0,5; K0	K1; J0	J1,5; K3	
21	Interrog.					K1; J0	J1; K0
22	Iterative					J0,5; K0	J0,5; K0
23	Negation					J4; K0,5	K0,5; J0
24	Passive					J1; K1	
25	Pass.Caus					J1; K0	
26	Past. Con.					J0,5; K0	J0,5; K0
27	Past. Perf					J2;K0	J2; K0
28	Past. S.			J0,5; K0,5		J1,5;K1,5	
29	Patient	J1; K1					
30	Permiss.					J0,5; K1,1	J0,5;K0,83
31	Pol. Form					J1; K2	
32	Pol. Mid.					K4; J0	
33	Pol. Plain				K1; J0	J1; K1	
34	Potential					J1,5;K0,33	J1,5;K0,66
35	Pres. Con.					J0,5; K0,5	J0,5; K0,5
36	Pres. S.					J1; K1	
37	Pres. Perf					J2; K0	J2; K0
38	Prohibit.					J1; K0,33	J1; K0,33

Figure 1: Positional distributions of Japanese and Korean grammars.

Lines 1–38 represent grammar meanings whereas columns B–G are positional realizations. Japanese (J) is marked green and Korean (K) is marked red. Common positions are marked yellow. Numbers inside cells show the degree of use of corresponding positions by the two languages respectfully.

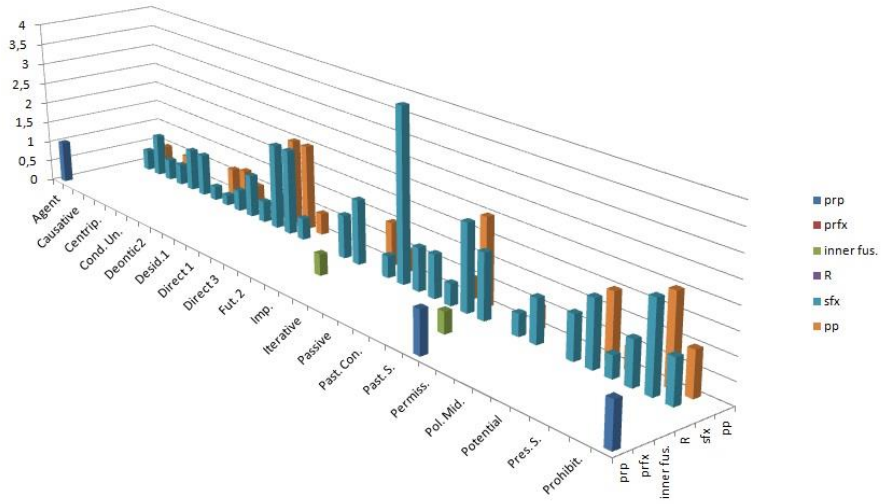

Figure 2: Positional distribution of Japanese grammar and its comparison with Korean.

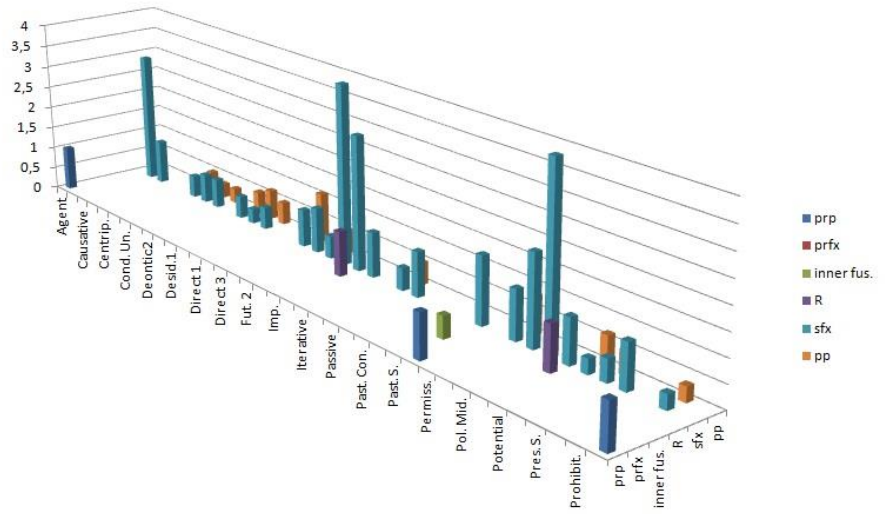

Figure 3: Distribution of Korean grammar and its comparison with Japanese.

3D diagrams are just unfolding of "conspectus" way of recording represented in 2D tables. Axis X shows the list of grammar meanings, axis Y signs of positional distributions, and axis Z shows to what degree a certain position is used by the two languages respectfully. Such 3D representations are useful in a sense that they show grammars of compared languages in the most illustrative way.

	A	B	C	D	E	F	G	H	I	J
1		prp	prfx	crp(h)	inner. F.	suppl.	R(n)	crp(e)	sfx	pp
2	Agent	Af1; E1							E6; Af0	
3	Causative	Af1; E1								
4	Cond.Md.	Af1; E1								
5	Deontic.	Af1; E4								
6	Desid.	Af2; E2								
7	Fut.Cont.	E0.66;Af0							E0.33;Af0	
8	Fut.Perf.	Af1.8E1.3	Af0.83;E0		Af0.33E0.33				E0.33;Af0	
9	Fut.Prf-C.	E0.75;Af0							E0.25;Af0	
10	Fut.S.	Af1; E1								
11	Hortative	E1; Af0								
12	Imp. Md.						Af1; E1			
13	Imposs.	E1; Af0								
14	Interrog.	Af1; E1								
15	Negation	E1; Af0								Af1; E0
16	Optative	E1; Af0								
17	Passive	Af1.83E1	Af0.83		E0.5;Af0.33				E0.5; Af0	
18	Past. Cont	E0.5; Af0							E0.5; Af0	
19	Past.Perf.	E1; Af0			E0.5; Af0				E0.5; Af0	
20	Past Prf-C	E0.66;Af0							E0.33; Af0	
21	Past S.				Af1; E1	E1; Af0			Af1; E1	
22	Patient									Af1; E1
23	Plural	Af1; E1							E3; Af0	
24	Possib.	Af2; E1								
25	Prs.Cont.	E0.5; Af0							E0.5; Af0	
26	Pres.Perf.	Af1.83;E1	Af0.83		E0.5;Af0.33				E0.5; Af0	
27	Prs.Prf-C.	E0.66;Af0							E0.33; Af0	
28	Pres.S.						Af1; E0		E6; Af0	
29	Prohib.	E2; Af0		Af0.5				Af0.5		
30	Singular	Af1; E1							E3; Af0	
31	Subject	Af1; E1							E6; Af0	
32	Subjunct.	Af1; E1								

Figure 4: Positional distributions of English (E) and Afrikaans (Af) grammars. English is marked red, common positions are marked yellow.

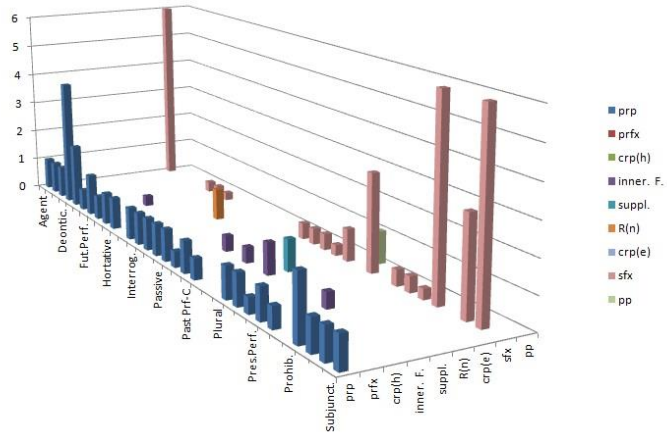

Figure 5: Positional distribution of English grammar compared to Afrikaans.

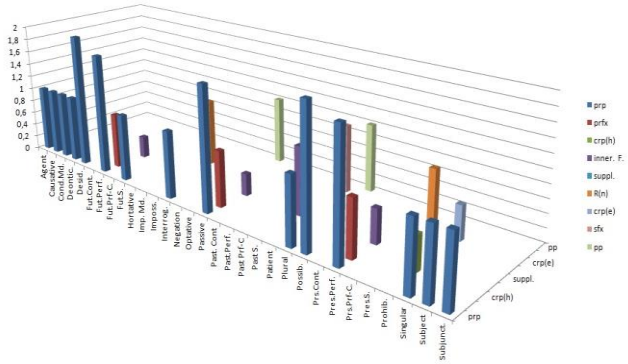

Figure 6: Positional distribution of Afrikaans grammar compared to English.

4 Conclusion and further perspectives of the Buyeo group

First, I suppose that it has been shown rather evidently that Japanese and Korean are not just languages of the same stock, but rather languages of the same language group.

Second, Ryukyuan languages show great proximity with Japanese so there is no problem at all to show their closeness with Korean.

Third, in the context of Altaic hypothesis it is traditionally supposed that the Buyeo group is related with Tungusic languages, Mongolian languages and Turkic languages. However, I suppose that the reality of the so-called Altaic stock/family is a highly doubtful issue since the PAI value of Buyeo languages is about 0.13, while Turkic languages show PAI value of around 0.012 (Tenishev, 1996), cf. a tenfold difference. According to section 3.2.1, such a difference of PAI values is a serious reason to doubt the relatedness of thelanguages considered. Anyway, whether the Buyeo group is related to other the above mentioned groups/stocks is matter of further research.

And finally, I suppose that the close relatedness of Japanese and Korean is a good evidence for the fact that Buyeo languages are not as ancient as argued by several scholars. The whole history of the Buyeo group probably counts about 1500 years only.

References

Akulov, A. (2015a). Prefixation Ability Index as a mean that allows us to see whether certain languages can potentially be genetically related. *Cultural Anthropology and Ethnosemiotics*, 1(2), 2–16.

Akulov, A. (2015b). Verbal Grammar Correlation Index (VGCI) method: a detailed description. *Cultural Anthropology and Ethnosemiotics*, 1(4), 19–42.

Akulov, A. (2015c). Whether is it possible to prove genetic unrelatedness of certain languages? *Cultural Anthropology and Ethnosemiotics*, 1(3), 2–4.

Akulov, A. (2015d). Why conclusions about genetic affiliation of certain language should be based on comparison of grammar but not on comparison of lexis? *Cultural Anthropology and Ethnosemiotics*, 1(3), 5–9.

Barhkhudarov, L.S.; Belyaevskaya, E. G.; Zagorul'ko, B. A.; Shveitser, A.D. (2000). Angliiskii yazyk [English language] In: N.N. Semenyuk, V.P. Kalygin, O.I. Romanova (eds.) *Yazyki mira. Germanskie i kel'tskie yazyki [Languages of the world. Germanic languages. Celtic langauges]* Moskow: Academia: 43–87.

Beckwith, C. (2007). *Koguryo, the language of Japan's continental relatives*. Leiden – Boston: Brill.

Brown, S. (2015). A bioimformatic perspective on linguistic relatedness. *Cultural Anthropology and Ethnosemiotics*, 1(4), 43–52.

Lavrent'yev, B. P. (2002). *Prakticheskaya grammatika yaponskogo yazyka [Practical grammar of Japanese]*. Moscow: Zhivoi Yazyk.

Mazur, Yu. N.(2004). *Grammatika koreiskogo yazyka (Morfologiya. Slovoobrazovanie) [Korean grammar. Morphology. Derivation]*. Moscow: Muravey

Matisoff, J. A. (1990). On Megalocomparison. *Language*, 66(1), 106–120.

Mironov S.A. 2000. Afrikaans yazyk [Afrikaans language] In: N.N. Semenyuk, V.P. Kalygin, O.I. Romanova (eds.) Yazyki mira. *Germanskie i kel'tskie yazyki [Languages of the world. Germanic languages. Celtic langauges]* Moskow: Academia: 87–101.

Ogura S. (1934). *朝鮮語と日本語 [Chōsengo to nihongo] [Korean language and Japanese language]*. Tokyo: Meiji Shoin.

Starostin, S. (1991). *Altaisakaya problema I proiskhozhdenie yaponskogo yazyka*. Moscow. Nauka.

Tenishev, E. R. (ed.) (1996). *Yazyki mira. Tyurksiye yazyki [Languages of the world. Turkic langauges]*. Moscow: Izadel'skii dom "Kyrgyzstan".

Vovin, A. (2010). *Koreo-Japonica A Re-Evaluation of a Common Genetic Origin*. Honolulu: University of Hawai‘i Press.

SURVEY ARTICLES

KANAUJI OF KANPUR: A BRIEF OVERVIEW

Pankaj DWIVEDI
Indian Institute of Technology Ropar
pankajd@iitrpr.ac.in

Somdev KAR
Indian Institute of Technology Ropar
somdev.kar@iitrpr.ac.in

Abstract

Hindi, in its totality, refers to a dialect continuum spoken mainly across northern India. This continuum is usually divided into two forms: Eastern and Western Hindi. Eastern Hindi is mainly made up of Awadhi, Chhattisgarhi and Bagheli dialects, while Western Hindi consists of Hindostani, Banagru, Braj Bhaka, Bundeli and Kanauji dialects.

After Linguistic survey of India (1894–1928) by George A. Grierson – there has been little or no work which specifically focuses on Kanauji. Trivedi (1993, 2005) and Mishra and Bali (2010, 2011) report some secondary data from Kanauji in their works, their focus of inquiry is not Kanauji though. Lewis, Simons & Fennig (2013) refers Kanauji as a language with very low identity.

This paper attempts to study the current sociolinguistic situation of Kanauji spoken in the Kanpur district of Uttar Pradesh (India). Some other goals of the paper are following: 1) to feel the pulse of language attitude, with reference to standard Hindi, of the people in Kanpur 2) to present basic linguistic information and 3) to direct attention of the other linguists to Kanauji, which unfortunately has not been the case so far despite of it being mother tongue of millions.

This study is result of eighteen days of a fieldtrip to Kanpur district and subsequent preparation of a small speech database of Kanauji. Importance of the work lies in the fact that no previous work, which specifically focuses on Kanauji, has been published so far. This is true at least in the open literature.

Keywords: Kanauji; language endangerment; sociolinguistics; Tirhari; varieties of Hindi

Povzetek

Ime hindski jezik se nanaša na kontinuum dialektov v severni Indiji. Zanj je značilna delitev na dve obliki: vzhodni in zahodni hindski jezik. Vzhodni hindski jezik sestavljajo dialekti Awadhi, Chhattisgarhi in Bagheli, zahodnega pa Hindostani, Banagru, Braj Bhaka, Bundeli in dialekti Kanauji.

Po jezikovnem pregledu v Indiji (*Linguistic survey of India*), ki ga je med leti 1894 in 1928 spisal George A. Grierson, se nihče več ni lotil raziskovanja dialekta Kanauji. Trivedi (1993, 2005) ter Mishra in Bali (2010, 2011) o njemu sicer poročajo, vendar je dialekt omenjen le posredno. Lewis in drugi (2013) ga omenjajo kot jezik z zelo slabo identiteto.

Članek je študija o trenutni sociolingvistični situaciji dialekta Kanauji, govorjenega na območju Kanpur v Uttar Pradesh v Indiji. Obenem si avtorja postavljata tudi naslednje cilje: 1. zaznati odnos na-

Acta Linguistica Asiatica, 6(1), 2016.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.6.1.101-119

ravnih govorcev do lastnega dialekta kot tudi do standardnega jezika, 2. predstaviti osnovne jezikovne informacije dialekta in 3. povečati zanimanje jezikoslovcev za dialekt Kanauji, ki je bil do sedaj zelo zapostavljen kljub temu, da ima več milijonov naravnih govorcev.

Ta študija je rezultat 18-dnevnega dela na terenu na območju Kanpur in kasnejše priprave podatkovne zbirke dialekta Kanauji. Pomembnost tega dela je predvsem v tem, da je ta raziskava edinstvena kar zadeva dialekt Kanauji, vsaj v javno dostopni znantveni literaturi.

Ključne besede: dialekt Kanauji; ogroženost jezika; sociolingvistika; Tirhari; raznolikost hindskega jezika

1 Introduction

Kanauji[1] (ISO 639–3 "bjj", written in Devnagri script as **कन्नौजी** and pronounced as /kən.nɔ.dʒi/) is a little known dialect[2] of Hindi and is hardly documented. It takes its name from town of Kanauj – the historic and one of the oldest cities dating back to ancient India. However, the language is not merely confined to district/town of Kanauj. Kanauj is a home to rich archeological and cultural heritage sites. History of Kanauj is believed to run as long as to the time of *Mahabharta*[3]. The ancient name of this place is Kanyakubja or Mahodya (Balmiki Ramayana, Mahabharata and Puran) and only later on the name Kanyakubja changed to Kannauj, the present name of the district.

In the legendary work Linguistic Survey of India written by George A. Grierson, Kanauji is classified under Western Hindi dialects together with four other dialects, namely, Hindostani, Banagru, Braj Bhaka and Bundeli (Grierson, 1916). Grieson considers Kanauji merely as a form of Braj[4] Bhakha and defines its area as east-central Doab[5] and country to its north. Apart from referring to it as a dialect of Western Hindi, in history or literary genre of Hindi language (Manuel, 1989), there are a few census related and other survey reports which have used the term *Kanauji* to refer to the Kanauji speaking population and not the dialect itself. According to Ethnologue (2013), Kanauji is presently, in its various forms, spoken in Kanpur, Farrukhabad, Etawah, Hardoi, Shahjahanpur, Pilibhit, Mainpuri, and Auraiya districts of Uttar Pradesh. However, the variety spoken in the district of Kanauj and Farrukhabad is referred as a

[1] Kanauji has been also spelled as Kannauji, Kanaoji or Kannoji by some authors and some scholars. Pankaj Dwivedi, one of the authors, is a native speaker of Kanauji spoken in Kanpur.

[2] Use of the term dialect here is only for the linguistic purpose, i.e., a linguistic variety. It does not allow for any sociolinguistic connotations such as standard language vs. dialect.

[3] Mahabharata/məhabʰarʈə/ is one of most important classical epics of Hindus. Its exact historical period is unclear. However, many historians date it as early as 10 century BCE, the Iron Age. Ramayana (balmɪkɪ ramajənə) and Purans (pʊrans) are also very significant religious scriptures in Hinduism.

[4] Alternative spelling for Braj is Brij. Both spellings are duly accepted by several scholars.

[5] Doab, in Hindi, refers to plains of Ganga River. The term has been used by several authors including by George A. Grierson in Linguistic Survey of India.

standard one. Kanauji is usually divided into three forms: Kanauji Proper (standard Kanauaji), Tirhari, and Transitional Kanauji (Lewis et al., 2013). It is spoken by more than six million people in total.

Figure 1: Distribution of the Kanauji dialect.

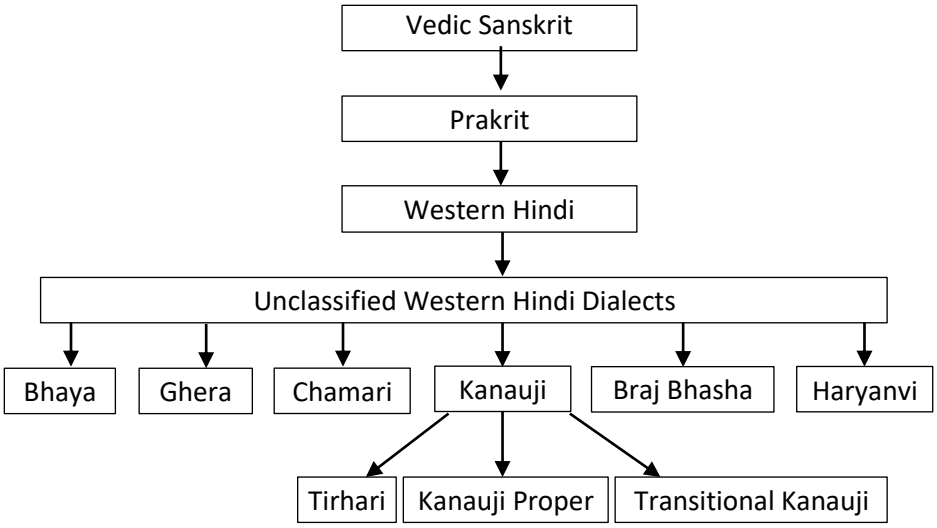

Figure 2: Evolution of the Kanauji dialect.

2 Kanpur and its linguistic demography

The term Kanpur is used to refer to Kanpur Nagar [6] (coordinates 26° 27′ 36″ N, 80° 19′ 48″ E) and Kanpur Dehat[7] (26° 20′ 39.48″ N, 79° 58′ 1.85″ E). It is the biggest city of the state of Uttar Pradesh and makes the main centre of commercial and industrial services. Kanpur Nagar (urban Kanpur) comprises of three subdistricts: Kanpur, Bilhaur, and Ghatampur, whereas Kanpur Dehat is made of five subdistricts: Akbarpur, Bhogani, Derapur, Rasulabad, and Sikandara. According to the 2011 census report, the total population of the Kanpur (both Nagar and Dehat) is estimated to be 6,368,043. A summary of the demographic information is presented in Table 1.

Table 1: Demographic summary of district Kanpur (taken from www.censusindia.gov.in).

District Name	Total Population	Sex Ratio	Density	Child Population	Literacy Rate: M/ F
Kanpur Nagar	1,795,092	852	1449	484,529	85.27/76.89
Kanpur Dehat	4,572,951	862	594	243,919	85.07/68.48

Kanauji of Kanpur is surrounded by at least four different dialects of Hindi. To the east of Kanpur, there is a district of Unao, where a variety of Awadhi is spoken; whereas, in order from north to west lie Hardoi, Kannauj and Auraiya respectively. In each of these places Kanauji is spoken; however, the variety spoken in Kannauj is considered to be the standard. To the south-east of Kanpur, between the plains of the River Ganga and Yamuna the district of Fatehpur is situated. Its northern region and language is influenced by "Awadhi" while the southern part shows effect of the "Bundeli". District of Hamirpur and Jalaun lie to the south and south-west of the Kanpur. The linguistic variety spoken in Hamirpur and Jalaun is Tirhari – the language of the river bank (Grierson, 1916). Both of these places are heavily affected by Bundeli and Bagheli. Tirhari of Hamirpur also carries a considerable touch of the Eastern Hindi.

3 Earlier works

Kanauji has little amount of literature to its credit and that too is available only in broken form. Some amount of folk literature and folk songs are found and preserved by the people in form of local oral literary traditions (Trivedi, 1997) . The main reason for this ignorance could be the supremacy of languages like Braj and Awadhi, which

[6] "Kanpur Nagar" refers to an area of Kanpur district. The term literary translates into "Kanpur City". However, Kanpur Nagar, unlike the name suggests, does have villages and rural areas.

[7] Kanpur Dehat refers an area of Kanpur District. The term literary translates into "Kanpur village". However, Kanpur Dehat, unlike the name suggests, does have city like towns.

were established with the purpose of literary creativeness. Supremacy of Braj and Awadhi is mainly due to religious and historical beliefs of people[8]. Along the same line, George A. Grierson (1890) reports that, as a literary language, Kanauji had been overshadowed by its more powerful neighbor Braj Bhakha. Most of the authors (of Kanauji language area) were Muslims and they wrote in Arabic and Persian. Among Hindu and Muslims authors writing in a vernacular, that is Kanauji, was not favorable. Grierson further mentions the authors from Tikampur/Tikawanpur town of Kanpur district, who lived in mid 17th century, such as *Chintamani Tripathi, Matiram Tripathi, Bhushan Tripathi and Nilkanth Tripathi* (Keay, 1933; Upadhyaya, 1934). They were siblings and together they published numerous works. All of them were patronized as poets in courts of many Mughal and Hindu Kings such as Shah Jahan, Aurangzeb, Shiv Raj of Sitara, Chhatrasal of Panna, etc. In His book titled *A history of Hindi literature*, Keay (1933) writes that Chintamani Tripathi (ʧɪnʈaməni ʈrɪpaʈʰi) was regarded as one of the great authorities on the subject of composition. Among his works are Chhand Bichar (ʧʰənɖ bɪʧar), a treatise on prosody, Kavya Vivek (kawjə vɪvek), Kavikul Kalptaru (kavɪkʊl kəlpʈəru), and Kavya Prakash (kawjə prəkaʃ). He was also author on Ramayana (ramajənə) in Kavitta (kəvɪʈʈ) and other metres. Other brothers, especially Bhushan, also excelled in their writings and invited laurels from all across the region for their work on different aspects of literacy creativeness (Nayyar, 2012). These works, however, were mainly composed either in Hindi or Urdu, not in Kanauji.

From the perspective of linguistics, the credit of being the first modern work on Kanauji can be given to *A Grammar of Modern Hindi* written by S. H. Kellog, originally published in the year 1876 by *Mission Press*, Allahabad, India. However, his work mainly focuses on Hindi; it discusses Kanauji as a dialect of Hindi and lists only a few examples explaining its morphology and phonology. Thereafter, some other works (Tiwari, 1960; Jaiswal, 1962; Saksena, 1971; Beams, 1974; Hopper, 1977; Shapiro, 1989; Hook, 1991; Masica, 1993, Kachru, 2006), which mainly focused on Standard Hindi, such as Khari Boli, and some other eastern or western variety of Hindi also paid some attention to Kanauji. Most of these works very briefly discuss areal distribution and position of Kanauji with reference either to Standard Hindi or Eastern and Western Hindi varieties. None of these works paid any closer linguistic attention to Kanauji. In the contrary, many Eastern and Western varieties of Hindi other than Kanauji received considerable attention of the linguists despite of the fact that compared to Kanauji they can be treated as minor in terms of their speaking population and the area covered.

The first major work on Indian languages including Kanauji is Linguistic Survey of India written by George A. Grierson during 1894–1928. There are a few other works

[8] Different religions and Godly incarnations have left deep impact on the Indian society and its literature. Most of the literary works in the middle age were meant for praise of the Gods, king, incarnations or religions. And the two most important incarnations of Lord Vishnu, Krishna and Rama, belong to Mathura (believed to be birth place of lord Krishna and language spoken there is Braj) and Ayodhya (believed to be the birth place of lord Rama and language spoken there is Awadhi), respectively.

besides, mainly representing supportive material while discussing literary works of the languages like Hindi, Awadhi and Bundeli (Trivedi, 1997, 2005), or works on NLP applications (Mishra & Bali, 2010, 2011; Kulshreshtha, Singh & Sharma, 2012; Kulshreshtha & Mathur, 2012). Works on NLP applications focused mainly on Hindi spoken in Kanauji speaking region rather than Kanauji. Published locally a long time ago, several of these works are no more available. In many cases, the local publication houses that published the books have been shut down due to the lack of commercial interests. *OpenLibrary.org* and *The Library of Congress Online Catalogue*[9] together have only found *few works*[10] on Kanauji given "Kanauji, Kanauji literature or Folk literature in Kanauji" as keywords.

4 Verbal repertoire of Kanpur

Kanpur has been a well-known industrial and educational hub of the state of Uttar Pradesh as well as of India for a long time. It also maintains a significant stake in small and large scale business houses and agriculture. For these two reasons it attracts people from all over India visit the region, and a large number of people moved in and settled down, which enriched the region's verbal repertoire. As a result, dialects spoken are very diverse; the majority speaks either Hindi or Kanauji, or both, however people speaking Punjabi, Bengali, Marathi, Urdu, Tamil, Oriya, and other dialects of Hindi such as Braj, Awadhi, Bihari, Bhojpuri, Bagheli Bundeli, etc. are also found in large numbers (Chaturvedi, 2015). Apart from these Indian languages and dialects, urban population of Kanpur possesses good knowledge of English for the very fact that most schools offer English language as a compulsory subject on all levels of their curriculum. Apart from the state governed Uttar Pradesh Madhyamik Shiksha Parishad (a state board of secondary and higher secondary school education), two of the most popular board of school education are Indian Council of School Education (ICSE) and Central Board of Secondary Education (CBSE), where English language is used for communication primarily. And it is the same with colleges, universities, technical institutes, and research organizations. There are some institutes, colleges and universities that also offer courses in foreign languages such as French, German, Chinese, Spanish, etc. and hence, people having good knowledge of these languages can easily be found. The use of English and other foreign languages is most common indication of upper class, higher educational and professional status, stronger socio-economic status, etc.

[9] The Library of Congress Online Catalog is the largest library in the world. It has millions of books, recordings, photographs, maps and manuscripts in its collections.

[10] Found works are listed as follows: **a)** Santarāma, Anila .(1975).Kanaujī loka sāhitya. Dillī : Abhinava Prakāśana; **b)** Tripāṭhī, Sureśa Candra (1977) Kanaujī loka sāhitya meṃ samāja kā pratibimba. Dillī : Rūpāyana Prakāśana; **c)** Gupta, Maheśa (1999) Loka-sāhitya kā śāstrīya anuśīlana : Bhāratīya loka-sāhitya, mukhyataḥ Pīlībhīta Janapada kā viśada adhyayana. Dillī : Śilpāyana.

Chaturvedi (2015) reports that such a variety of languages and dialects used in the area of Kanpur most likely indicates one of the following three future language situations: 1) a situation completely different from the present one, 2) a situation where more than one language is knowingly used, and 3) code-mixed language situation. With such a rich confluence of languages in one place, instance code mixing and code-switching are common or even inevitable in every day life. However, long and wide-spread presence of these languages has given rise to what is popularly referred as **Kanpuria Hindi.** Kanpuria Hindi is based on standard Hindi but the vocabulary consists of words borrowed from Hindi, English, and other Indian languages and dialects, though a significant proportion is nevertheless from **Kanauji** and other Hindi dialects such as Awadhi, Bhojpuri, Bihari, Bagheli, and Bundeli. Kanpuria Hindi is a very popular medium of networking/peer talk among youth. Its usage is wide spread in everyday communication, and people mostly use it in informal conversations. Standard Hindi or English (or Hinglish) is preferred in written and formal mode of communication.

In Kanpuria Hindi, the scale of code-mixing and code-switching varies depending on the urban vs. rural class, educated vs. less-educated/illiterate class, class belonging to higher socio-economic background vs. class belonging to lower socio-economic background, etc. While code-mixing observed with the former classes is connected to well-established languages such as Punjabi, Bengali, and Marathi, the later classes generally involve other dialects of Hindi. The reason can be found in the fact that people who have migrated from other states are financially sound due to their businesses or good-profile jobs. On the other hand, people speaking Kanauji or some variety of the neighboring Hindi dialects form a major percentage of low-profile workers and people from other lower classes.

The difference is so large that we can easily devide these two forms into urban Kanpuria Hindi and rural Kanpuria Hindi respectively. Rural Kanpuria Hindi borrows heavily from Kanauji, especially its lexical, morphological, and syntactic characteristics. It is interesting that people who generally speak urban Kanpuria Hindi may shift to rural Kanpuria Hindi for a better bargain during street shopping or to show solidarity with the people. Similarly, people speaking rural Kanpuria Hindi may shift to urban Kanpuria Hindi when in a restaurant, or speaking to public servants like post-officers, policemen, doctors, etc. Such a shift may also be used in order to make fun of either their own or the other linguistic group.

5 Present situation of Kanauji of Kanpur

As discussed in previous sections, the language of Kanauji in its original form (as a dialect of the indigenous people in the area) has been challenged significantly due to a heavy confluence of about a dozen languages and dialects. The present situation of

Kanauji is grave and complicated both in the urban and rural areas of the Kanpur district. While in the urban areas Kanauji has almost been replaced with Kanpuria Hindi and to a certain extent with Standard Hindi, Kanauji still functions as a mode of communication in rural areas. However, with a fast-growing educational system and urbanization, youths and children prefer speaking in Standard Hindi.

Lewis, Simons and Fennig (Ethnologue, 2013) report that despite of being a tongue of millions, Kanuaji has very low identity with no official status or proposed preservation plans. Ethnologue (2013) places Kanauji within the cloud of all living languages but in the category 6b-7 of the *EGIDS*[11] scale, which reads that language in question is in trouble and intergenerational transmission of the language is in the process of being broken.

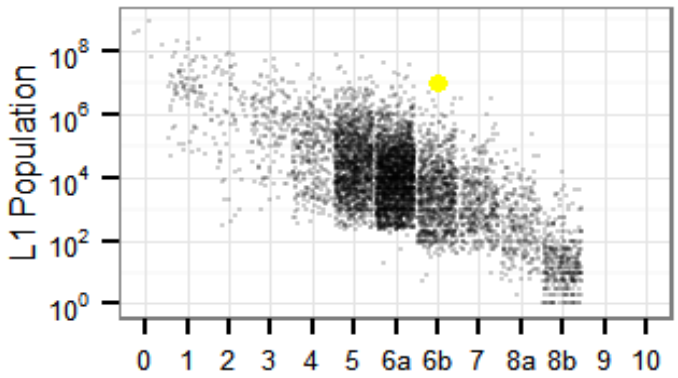

Figure 3: Yellow dot represents position of Kanauji on EDGIS Scale (taken from Ethnologue, 2013).

Despite the "troubled" situation, the child-bearing generation still uses the language. Hence, it is possible that revitalization efforts could restore transmission of the language. The scale of a given stage of endangerment in Figure 3 may be true for Kanauji in general, but not necessarily so for varieties in every part of the area. It is well-known that values on the endangerment scale may vary from EDGIS 6a to EDGIS 7) depending on sociolinguistic factors such as connectivity, education, economics, language attitude, migration, industrial development, and such, I believe, is the case with Kanauji of Kanpur; it is presently estimated at EDGIS-7. A further change of a grade would imply that the language is on the verge of extinction and could no more be restored as a form of communication. Such a fact urges us to get the language documented before it dies out.

[11] Ethnologue reports vitality of the languages of the world by using the Expanded Graded Intergenerational Disruption Scale or EGIDS and Graded Intergenerational Disruption Scale (GIDS) (Lewis and Simons, 2010 and Fishman's, 1991 respectively).

6 Sociolinguistic fieldwork and data collection

During our fieldtrip to 15 villages[12] of the Kanauji speaking Kanpur Nagar region with the purpose to collect speech samples[13], we surveyed 80 informants to get an insight into native people's towards the use of Kanauji. The relevant details are listed below:

Table 2: Basic data of the informants in the surveyed area.

Male	Female	Age	Education	Bilingualism (Tirhari–Hindi)
35	15	14–25 years	10th-undergraduate	Hindi-Kanauji unavoidably mixed. Hindi is highly dominant.
17	5	30–45 years	Illiterate to undergraduate	Kanauji is little dominant
5	3	50–65 years	Mostly Illiterate	Kanauji is more dominant

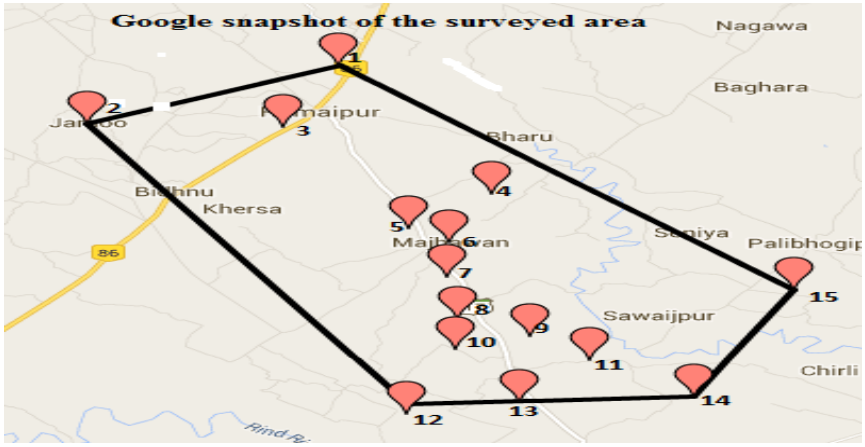

Figure 4: A snapshot of the surveyed area from the Google map.

Based on the responses of the informants, the following observations regarding language attitude of the people have been made.

[12] These fifteen villages in order of visit are: 1) Gadhewa Mohasinpur, 2) Jamu, 3) Ramaipur, 4) Shahpur Majhawan, 5) Jagdishpur, 6) Majhwan, 7) ArajiMajhwan, 8) Kulhauli, 9) Hardauli, 10) Kale ka Purwa, 11) Kundauli, 12) Kaji Khera, 13) Behta Gambhirpur, 14) Rajepur, 15) Bhukhnahi

[13] We collected speech samples using two channels simultaneously: first, by Olympus LS-100 96kHz/24 PCM linear recorder which is outfitted with two built-in 90° stereo condenser microphones, second using Sony Digital Flash Voice Recorder -ICD-PX312.

1. Most of high-school and university students do not even know the name of the Kanauji language – they call it either Hindi or a type of Hindi, sometimes Hindi of the uneducated people.
2. University students use standard Hindi (i.e., Khari Boli) in intra- or intergroup communication. However, they speak in Kanauji with their parents, grandparents and other older family members. They sometimes use Kanauji to look down upon or to make fun of someone, or to look funny.
3. Some youth – generally educated – admit that they feel ashamed using Kanauji in public though they are well familiar with the language. This confirms the general negative attitude of people towards Kanauji.
4. No school allows Kanauji to be used in classes, even in the areas where Kanauji speaking population is relatively large.
5. Informants aged between 10 and 35 said they do not use Kanauji in formal situations, for example if they go to see doctors, teachers, the village head, policemen, etc. On the other hand, old people said they use Kanauji and Hindi in such situations, added that they think the use of standard Hindi is more suitable in such situations.
6. None of the informants knew of any work such as books, dictionaries, newspaper or some classical literature on Kanauji.
7. Old consider such songs either as funny or as "songs of uneducated people". people know and sing folk songs in Kanauji but not the new generations. Youths and children
8. Since parents see no economic, educational or other benefits, they do not encourage the use of Kanauji among the children. They, however, emphasize the importance of using the standard Hindi and English.

The above observations were confirmed by the results obtained from a pilot study that included a smaller group of the informants.

Below is a piece of a conversation between a teacher in a primary school (language consultant A) and a data collector (a linguist). The teacher starts with a request to the data collector on whether he could help him (the teacher) to get a particular literary work from the city. After that the teacher talks about the training he (the teacher) had undergone in his school. An army man sitting nearby (language consultant B) also speaks to agree with the teacher. A look into the conversation shows how he (the teacher) switches from Kanauji to Hindi and then back to Kanauji. It is obvious that the teacher does not realize the shift and goes on talking in the same manner for around twenty minutes.

In this part of a conversation, every sentence in a dialogue is first written in Devnagri Script (marked as DS), followed by a broad IPA transcription (marked s BT) of the spoken sentence. The transcription is annotated using slanted brackets <> with ISO

639–3 language codes, that is, "bjj" for Kanauji, "hin" for Hindi and "eng" for English. The meaning is given in the last line.

Example:

Language consultant 1:

a) एत्ता काम करे जैव। एत्ती चिन्हारी अपन देहे जैव । (DS)
<bjj>eʈʈa kam kəre dʒɛw. eʈʈi tʃɪnhari əpən ɖehe dʒɛw **<bjj>** (BT)
Please do this favour to me. It will remind me of you. (M)

Data Collector:

b) मैं जिस दिन घर जाँउगा, उस दिन ढूढ़ के भेज दूँगा (DS)
<hin> mɛ̃ ʤɪs ɖɪn gʰər ʤaʊ̃ga, ʊs ɖɪn ɖuɽʰ ke bʰeʤ ɖũga**.<hin>** (BT)
I will send it the day I reach home. (M)

Language consultant 1:

c) सात साल होइगे पढ़ावत । (DS)
<bjj>saʈ sal hoɪge pəɽʰawəʈ<bjj> (BT)
I have been teaching for seven years. (M)

d) अभी जो **ट्रेनिग** चल रही है उसमें जब ये चीज आई **तो** दिमाग **हमारा** चकरा गया । (DS)
<hin> əbʰi dʒo <eng>**ʈreniŋ** <eng>tʃəl rəhi hɛ ʊsme dʒəb je tʃidʒ aɪ **ʈo** ɖɪmag <bjj>**həmara** <bjj> tʃəkra gəja.<hin> (BT)
When I came across this thing in the current training, my mind went blank. (M)

e) हेईं सासा का पढ़ै वाला लरिका आये।. नीतू का जानत है। (DS)
<bjj> hɛ̃ɪ sasa ka pəɽɛ wala lərɪka aje. niʈu ka dʒanəʈ hɛ. <bjj> (BT)
He (the trainer) comes from Sasa and knows the Neetu. (M)

f) जब उसने चालू किया तो **हमनें** कहीं पढ़ा था ये। दिमाग में था। लेकिन **होई कुछ…** (DS)
<hin>dʒəb ʊsne tʃalu kija ʈo <bjj>**həmne** <bjj> kahĩ pəɽʰa ʈʰa je. ɖɪmag me ʈʰa lekɪn <**hoɪ kutʃʰ…> <hin>”** (BT)
As soon as he started, I knew it as I read about it somewhere. It was on my mind. But then, may be something…” (M)

Language consultant 2:

g) हाँ हाँ, ध्यान न दीन होइहौ (DS)
<bjj>hã hã, ɖʰjan nə ɖɪn hoɪhaʊ. <bjj> (BT)
Yes, yes, you may not have paid attention. (M)

Language Consultant 1:

h) लेकिन वो **बेसिक** से ही आधारित है तो उसका ज्ञान आवश्यक है। (DS)
<hin>lekɪn wo <eng>**besɪk** <eng> se hi aɖʰariʈ hɛ. ʈo ʊska gjan awəsyək hɛ. <hin> (BT)
But, it starts from basic concepts. So, it's necessary to know it. (M)

i) **ट्रेनिंग** एक दिन के लिये थोड़ी न होती है। (DS)
<hin><eng>ʈreniŋ <eng>ek d̪ɪn ke lɪje t̪ʰoɽi nə hot̪ɪ hɛ. <hin> (BT)
Training does not help only for a day (meaning).

j) **आल al laɪpʰ** <eng>kɛ lɪje hot̪ɪ hɛ. <hin> (BT)
It's for a whole life. (M)**लाइफ** के लिये होती है. (DS)
<hin><eng>

A closer look into the above conversation gives us a fair idea of how people interchangely use (code-switch and code-mix) Hindi and Kanauji in communication. In the above piece of narration, the language consultant A starts his communication (a request) using a typical Kanauji sentence (statement a) spoken in Kanpur. In response to his request, the data collector answers in Standard Hindi (statement b). However, the language consultant A goes on explaining background for his request in Kanauji (statement c) and then switching to Hindi (statement d). This pattern is repeated in sentences e and f, respectively. While speaking in Hindi (statement d and statement f), he uses pronouns /həmara/ (meaning "I") and /t̪o/ (meaning "then") in Kanauji. In Hindi, the correct pronouns are /m̪æ̃/ and /t̪ab/ respectively. The language consultant A is about to say something more at end of statement f using Kanauji but he is cut short by the language consultant B, who agrees with the explanation given by language consultant A (statement g) using Kanauji. Elaborating further on the situation, language consultant A again speaks three more sentences in Hindi (statement h to j).

Language consultant A also uses words from English such as "training", "basic" and "all" in different statements. However, these words are adapted to the phonological inventory of Kanauji and therefore their pronunciation gets changed accordingly. For example, in the statement j the phrase "all life" is pronounced as /al # laɪpʰ/ in contrast to /ɔːl # laɪf/, as Kanauji lacks phonemes /ɔ/ and /f/. The word "training" which occurs in statements i and d is pronounced as /ʈreniŋ/ instead of /treɪnɪŋ/. Due to the absence of phoneme /t/ in Kanauji, phonetically its nearest phoneme / ʈ/ is adapted. A different placement of stress reduces the /eɪ/ into /e/.

Due to such constant code-switching and code-mixing, this piece of conversation may sound odd to native Kanauji speakers as well as Hindi speakers. During this narration, both languages consistently maintain Kanauji intonation, which can easily be recognized by a native speaker of Hindi.

7 Comparative linguistic sketch of Kanauji and Hindi of Kanpur

Kanauji has a few written records in Devnagari (Trivedi, 1997). Based on our 18 days of fieldwork we collected words, sentences, and free discourse samples from a range of domains such as *Basic wordlist, Demographic description, Cuisine, Family communication, Games, Culture and Traditions, Flora and Fauna, Mythological stories, Daily life activities, Children stories, Number systems, Free discourse, Minimal pairs,*

Representative sentences, Group conversation, and others. A text and a speech database[14] were prepared based on the collected data. The analyses of the database brought us to the following conclusions. Kanauji of Kanpur has 31 consonants, 10 native vowels and 1 foreign vowel "æ"[15] in its phonemic inventory. All vowels have their respective nasalized counterpart vowels. In writing, Kanauji exhibits the same number and types of consonants as standard Hindi. However, many consonants have lost their individual pronunciation in Kanauji, which is also observed in other dialects of Hindi. Table 3a shows a list of vowels, diphthongs and consonants in Kanauji of Kanpur.

Table 3a: Inventory of vowels in Kanauji of Kanpur.

Vowels	i	ɪ	e	ɛ	a	
Nasalized vowels	ĩ	ɪ̃	ẽ	ɛ̃	ã	
Vowels	ə	u	ʊ	o	ɔ	æ
Nasalized vowels	ə̃	ũ	ʊ̃	õ	ɔ̃	æ̃

Mishra and Bali (2010) list a slightly different number of diphthongs found in the Kanauji (Table 3b, right). We suppose that they have taken a slightly different variety of Kanauji, the variety spoken in the district of Kanauji, though it is not clearly mentioned which variety of Kanauji they are referring to. Kanpur, on the other hand, being repertoire of different languages, majorly includes varieties of Eastern and Western Hindi, therefore the number and type of diphthongs found may show some difference. From the database we prepared, diphthongs found in the Kanauji of Kanpur are listed on the left.

Table 3b: Diphthongs in Kanauji of Kanpur.

Kanauji of Kanpur	**Kanauji**
/aɪ/, /aʊ/, /əɪ/, /eɪ/, /eʊ/, /əʊ/, /oɪ/, /ʊa/, /ʊɪ/, /ɪʊ/ (ɪw), /eo/ (ew), /ae/	/əɪ/ (əe), /əʊ/ (əo), /ɑɪ/ (ɑe), /aʊ/ (ɑo), /ɪʊ/,/ʊɪ/ /ɛɪ/, /ɛʊ/ (eʊ), /ɔɪ/(oɪ), /ɔʊ/(oʊ), /ɛ/(jɑ), /ɔ/ (wɑ)

[14] Text database mainly consists of .doc and excel files, while speech database consists of wav files. Speech database is broadly annotated using PRAAT software. A detailed discussion on the preparation or annotation of these databases is beyond the scope of this a paper.

[15] English words like "cat", "bat", "man", "hat", etc. are commonly understood and frequently used by Kanauji speakers (even by uneducated ones). Somehow, most of them pronounce it with correct "æ" vowel sound. This inclusion of vowel sound could be due to far reaching impact of English media over Kanauji speakers.

Table 3C: Inventory of consonants in Kanauji of Kanpur.

Plosives	/p/,/pʰ/; /b/, /bʰ/; /ʈ/,/ʈʰ/; /ɖ/,/ɖʰ/; /ʈ/,/ʈʰ/; /ɖ/, /ɖʰ/; /k/, kʰ/; /g/, /gʰ/	16
Nasal	/m/, /n/, /ŋ/	3
Fricative	/s/, /h/ **/f/*, /ʃ/* /v/* /z/*[16]**	2 (+4)
Affricates	/ʧ/, /ʧʰ/; /ʤ/, /ʤʰ/	4
Laterals	/l/	1
Semi-Vowels	/w/, /j/	2
Trill	/r/	1
Tap/Flap	/ɽ/, /ɽʰ/	2
Total		**31 (+4)**

Kanauji is a CVC type language. Kanauji of Kanpur permits all four types of syllables, the "CV", "CVC", "VC" and "V" type. Consonant clusters are not prominent, and in cases of a borrowed consonant cluster, the cluster is broken by epenthesis. A cluster of a consonant followed by a semivowel (cf. C+w/j) is allowed in initial position to some extent. Clusters are more restricted in the coda position. Medial clusters in words usually belong to two different adjacent syllables, a coda of the preceding syllable and an onset of the following syllable respectively. An analysis of 772 words collected during the fieldwork provides us with the following results:

Table 4: An analysis of basic word list of Kanauji of Kanpur

Words	**Number**
Words with final clusters	9
Words with initial clusters	27
Words with medial clusters	192
Words with unbroken medial cluster	1
Words with no clusters	543
Total	**772**

[16] Consonants written in bold and marked with asterisk are spoken by speakers who know Standard Hindi or Urdu other than Kanauji, which is very common in Kanpur. Otherwise, these consonants are absent in native phonemic inventory of Kanauji.

Initial Consonant Clusters: rj, kw, kʰw, gw, gʰw, gj, ʧw, ʧw, ʤw, ʤʰw, ʈw, ʈj, ʈʰw, ʈj, ɖw, ɖj, ɖʰw, ɖʰj, ɖ̲w, ɖ̲j, ɖ̲ʰw, nj, nw, pw, pj, pʰw, pʰj,bw, bj, bʰw, bʰj, lw, lj, sw, sj, hw, hj, kj, kʰj, kr, ɖ̲ʰr, pr, sr, br and (**sl, bl, kl, ʈr, wy**).[17]

Medial Consonant Clusters: b-b, d-dʰ, ɖ̲-m,ɖ̲-r, ɖ̲-r, ɖ̲-w, ʤ-ʤ, ʤʰ-r, ʤ-ɾ, ʤ-r, ɖ-dʰ,ɖ-ɖ, pʰ-ʈ̲, g-dʒ, g-g, g-gʰ, g-r, h-k, h-l, h-r, h-ɾ, h-s, k-k, k-l, k-m, k-n, k-ɾ, k-s, k-ʈ, k-w, l-d, l-l, l-m, l-n, l-r, l-s, l-s, l-w, m-b, m-b, m-k, m-l, m-m m-n, m-pʰ, m-s m-ʈ,m-ʈ̲, m-tʃ, n-d, n-ɖ̲, n-ɖ̲, n-dʒ, n-ɖ, n-g, n-k, n-n, n-p, ɳ-r, n-ʈ, n-tʃ, n-w, ŋ-r, pʰ-ɾ, p-p, p-r, p-s, r-b, r-d, r-d, r-ɖ̲, r-dʒ, r-g, ɾʰ-w, [rj], r-k, r-r, r-s, r-ʈ̲, r-tʃ, r-w, ɾ-w, s-dʒ, s-k, s-m, s-n, s-r, s-s, s-t, ʈ̲-b, ʈʰ-k, ʈ-k, ʈ̲-k, ʈ̲-n, ʧ-k, ʧ-l, ʧ-r, ʧ-ʧ, ʧ-ʧ, ʧ-ʧʰ, ʧ-w, ʈ-ʈ, ʈ̲-ʈ̲, ʈ-w

Final Consonant Clusters: ɖ̲ɖ̲ʰ, lh, nd, ndʒ, rr, ʧʧʰ, pp, ll, kk, mm, ʈʈ,

Some other examples of the differences and similarities of Kanauji compared to Hindi are:

A. Kanauji of Kanpur does not distinguish between a voiceless palatal sibilant consonant /ʃ/ and a voiceless dental sibilant /s/. Usually, /s/ is preferred in all positions of the words (initial, medial or final). Examples:

Hindi	/saɖ̲i/ "plain"	/ʃaɖ̲i/ "marriage"
Kanauji	/saɖ̲i/ "plain"	/saɖ̲i/ "marriage"

B. A voiced dental sibilant /z/ and a voiceless fricative /f/ are absent from the consonant inventory of Kanauji of Kanpur and therefore they are usually pronounced as a voiced palatal affricate/dʒ/ and aspirated voiceless bilabial stop/pʰ/ respectively. Examples:

Hindi	/saf/ "clean"	/zəhər/ "poison"
Kanauji	/sapʰ/ "clean"	/ʤəhər/ "poison"

C. The h-elision can be observed in many cases. Though h-elision primarily looks sporadic in nature, it needs to be determined if there is some specific phonological environment for this phenomenon.

D. In some places semivowels /j/ and /w/ are replaced with a voiced palatal affricate/dʒ/ and a voiced bilabial stop /b/ respectively. Examples:

Hindi	/somwar/ "Monday"	/jəmʊna/ "river Yamuna"
Kanauji	/sombar/ "Monday"	/ʤəmʊna / "river Yamuna"

[17] Clusters in bold are found in the words which are borrowed from other languages such as Hindi and English and are in frequent use in Kanauji of Kanpur.

E. Rhotacism: /l/ is replaced by /r/ in intervocalic and in final positions if preceded by a vowel. Like in many other languages, such rhotacism is sporadic and frequent exceptions are found. Examples:

Hindi	/hərɪjali/ "greenery"	/baɖ̲əl/ "cloud"
Kanauji	/harɪjari/ "greenery"	/ baɖ̲ər / "cloud"

F. Lexical Similarity: According to Lewis, Simons and Fennig (2013), lexical similarity between Kanauji and Hindi measures around 83% – 94%. However, it is not clear which variety of the Kanauji is mentioned there. It is further claimed that 84% – 97% similarity is found between all varieties of Kanauji, which may be true if influence of fast-spreading Hindi is ignored and only native vocabulary is compared. A comparative list of ten words of Kanauji and Hindi are given below.

Table 5: Comparison between Hindi and Kanauji words

Words	Hindi	Kanauji
Finger	ʊŋgli	ə̃gʊ.ri
Eye	ə̃kʰ	ə̃kʰi
White	ʊʤla	ʊʤər
Twenty one	ikkis	ekəɪs
How	kɛse	kəɪse
Farms	kʰeʈ̲	kʰeʈ̲/kʰeʈ̲wa
Kitchen	ʧulha	ʧulh
Peel	ʧʰɪlka	ʧʰɪkla/ bokla
Gods	ɖ̲evʈ̲a	ɖ̲eʊʈ̲a
Marriage	bjah	bɪaw/bɪjaw

8 Factors causing the decline of Kanauji in Kanpur

Factors causing the decline of Kanauji are given as follows:

1. The district of Kanpur has been hub to industries and business for a long time and in the last few decades it has emerged as one of the most important centers of education in the state of Uttar Pradesh. People have therefore shifted to either Kanpuria Hindi or to standard Hindi for intergroup communication.

2. Medium of instructions in schools, colleges and universities is standard Hindi, or English in few cases, but not Kanauji. Hence, Kanauji does not cater to educational or professional needs of children and youths.
3. Standard Hindi is considered to be more prestigious.
4. Hardly any literature is written in or about Kanauji. The one that exists is restricted to show the language only as a part of the history of the Hindi language. Therefore, people may not find themselves attached to the language. Besdies, the lack of written literature deprives the language from being included as a teaching material in schools.
5. Religion plays a very important role in preserving either language or culture (Fellman, 1973). Kanauji has little religious grounding, especially compared to Awadhi, a neighboring Eastern Hindi dialect, which connects itself to Hindu religion and enjoys all the prestige.

9 Conclusion

The state of Kanauji (and therefore its language, Kanauji, too) has been a witness to one of the most prospective periods in the Indian history (Tripathi, 1989; Majumdar, 1951, 1955; Smith, 1908). Today, however, its native language lacks attention due to socio-political reasons.

Lack of attention to a dialect in comparison to other major and official languages is a common habit among the bureaucratic policies across the world, though not among linguists and other academics, writers, or language activists. Sadly enough, a close look at Kanauji reveals that despite it being a mother tongue of six million people, it has been ignored by both groups alike. Ironically, tens of languages in India and hundreds of languages across the world, which are far smaller in terms of their area and speaking population, have been documented, promoted and worked upon. As a result, some of these languages got not only saved from extinction but also promoted among community members and scholars.

Kanauji needs good documentation, reinvigoration and promotion efforts done in cooperation by linguists, officials and community members. There is also a dire need to collect data on Kanauji such as multipurpose speech and text corpora, language teaching materials, language technology application, etc. An active and positive role of print and electronic media can also help Kanauji regain its prestige and consequently reclaim its speakers.

In the present state of sheer indifference, Kanauji is being rapidly devoured by its neighboring languages and Standard Hindi, heading to what may be called a death of a major language (Khokhlova, 2014). For that reason it is of great importance to document and start promoting it immediately.

Acknowledgement

Authors acknowledge the kind support of their language consultants during data collection. Pankaj Dwivedi acknowledges Indian Institute of Technology Ropar for providing fellowship and funds for fieldwork and purchase of required equipment such as Olympus LS-100 96kHz/24 PCM linear recorder and tripods.

References

Beams, J. (1974). *Outlines of Indian philology with a map shewing the distribution of Indian languages.* New Delhi: Today and Tomorrow Printers and Publisher

Bhatia, T. K. (1987). *A history of the Hindi grammatical tradition: Hindi-Hindustani grammar, grammarrians, history and problems* (Handbuch Der Orientalistik, Vol 4). Leiden: The Netherlands.

Chaturvedi, S. (2015). A Sociolinguistic study of linguistic variation and code matrix In Kanpur. *Procedia-Social and Behavioral Sciences*, 192, 107–115.

Fellman, J. (1973). Concerning the 'revival' of the Hebrew language. *Anthropological Linguistics*, 15. 250–257.

Grierson, G. A. (eds.). (1916). *Linguistic survey of India*. Vol. 9. Kolkata: Superintendent Government Printing.

Hook, P. E. (1991). The emergence of perfective aspect in Indo-Aryan languages. In Elizabeth C. Traugott & Bernd Heine (eds.) *Approaches to grammaticalization*, 2, 59–89. Amsterdam: John Benjamins Publishing

Hopper, P. J. (eds.). (1977). *Studies in descriptive and historical linguistics: Festschrift for Winfred P. Lehmann.* Vol. 4. Amsterdam: John Benjamins Publishing Company.

Jaiswal, M. P. (1962). *A linguistic study of Bundeli* (Vol. 8). Brill Archive.

Kachru, Y. (eds.). (2006). *Hindi*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Keay, F. E. (1920). *A history of Hindi literature*. Mysore City: Wesleyan Press

Khokhlova, L. V. (2014). Majority language death. In: Hugo C. Cardoso (eds.), *Language endangerment and preservation in South Asia*. 19–45. Honolulu: University of Hawai'i Press.

Kulshreshtha, M. and Mathur, R. (2012). Hindi language and its dialects. In: *Dialect accent features for establishing speaker identity*, 15–20. New York: Springer-Verlag:

Kulshreshtha, M.; Singh, C. P. and Sharma, R. M. (2012). Speaker profiling: The study of acoustic characteristics based on phonetic features of Hindi dialects for forensic speaker identification. In: Amy Neustein & Hemant A. Patil (eds.), *Forensic speaker recognition: Law enforcement and counter-terrorism*, 71–100. New York. Springer

Lewis, P. M., Simons, G. F. and Fennig, C. D. (eds.). (2013). *Ethnologue: languages of the world, seventeenth edition*. Dallas-Texas: SIL International. Online version: http://www.ethnologue.com.

Majumdar, R. C. (1955). *The Age of imperial Kanauj* .vol. 4. Mumbai: Bharatiya Vidya Bhavan.

Manuel, P. L. (1989). *Ṭhumrī in historical and stylistic perspectives*. New Delhi: Motilal Banarsidass Publications.

Masica, C. P. (1993). *The Indo-Aryan languages*. Cambridge: Cambridge University Press.

Mishra, D. and Bali, K. (2010). *Hindi dialects phonological transfer rules for verb root Cələ*. Paper presented at 13th oriental COCOSDA-2010 conference in coordination with International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques held at Kathmandu. Nepal, 24–25 November

Mishra, D. and Bali, K. (2011). *A Comparative phonological study of the dialects of Hindi.* Paper presented at The 17th International Congress of Phonetic Sciences (ICPhS XVII), Hong- Kong, China, 17–21 August

Narula, S. S. (1955). *Scientific history of the Hindi language*. New Delhi: Hindi Academy.

Nayyar, S. (2002). *History of Hindi*. Retrieved from http://www.esamskriti.com/essay-chapters/History-of-Hindi-1.aspx on June 20, 2014.

Saksena, B. (1971). *Evolution of Awadhi: A branch of Hindi*. Vol. 12. New Delhi: Motilal Banarsidass Publisher.

Shapiro, M. C. (1989). *A primer of modern standard Hindi*. New Delhi: Motilal Banarsidass Publishers

Smith, V. A. (1908). The history of the city of Kanauj and of king Yasovarman.*The Journal of the Royal Asiatic Society of Great Britain and Ireland*, 765–793.

Tiwari, U. (1960). *The origin and development of Bhojpuri*. Kolkata. The Asiatic Society

Tripathi, R. S. (1989). *History of Kanauj: To the moslem conquest* (Vol. 11). New Delhi: Motilal Banarsidass Publisher.

Trivedi, G. M. (1997). *Ganga ke nichle doab ka bhasha sarvekshan*. Kolkata: Anthropological Survey of India Press

Trivedi, G. M. (1990). Eco formation. In Krishnan, Shree (eds.), *Linguistic traits across language boundaries*, 51–82. Calcutta: Anthropological Survey of India

Upadhyaya, A. S. (1934). *The origin and growth of Hindi language and its literature*. Patna: Patna University.