Troubles on the Border
Marking Japanese Case-Particle Boundaries in Grammatical Annotation

Simone dalla Chiesa
Università degli Studi di Milano La Statale, Italia

Abstract   This paper aims to determine the extent and effects of the phenomenon of inconsistent case-particle boundary marking in the grammatical annotation of Japanese. It is focused on establishing what represents 'inconsistent boundary marking', how it is dealt with in informational terms, what effect it has on communication, and why it should be avoided. To this purpose, I will first build a typology of the tokenization strategies in the grammatical annotation of Japanese. I will then individuate several forms of inconsistent boundary marking and, more in general, of poor grammatical annotation, and discuss them according to the types of inconsistency and their different epistemic effects.

Keywords   Japanese. Grammatical annotation. Tokenization. Case-particles. Boundary symbols. Script conversion.

Summary   1 Introduction. – 1.1 The Sample. – 1.2 The Structure of This Paper. – 1.3 Unanswered Questions. – 2 Purpose and Structure of Linguistic Annotation. – 3 The Grammatical Annotation of Japanese. – 4 Word-Boundary Marking in the Grammatical Annotation of Japanese. – 5 The Morphological Categorization of Japanese Case Particles. – 6 A Discussion of the Sample. – 6.1 Inconsistent Case-Particle Boundary Marking in Collections. – 6.2 Inconsistent Case-Particle Boundary Marking in Single Works. – 6.3 The Case of Shibatani. – 7 Poor Segmentation and Sloppy Annotation. – 7.1 Sloppy Annotation. – 7.2 Weak Editing. – 8 Linguistic Annotation and Information. – 9 Sloppy Annotation and Conventionality. – 10 Coping With Sloppy Annotation. – 11 Conclusions.

> [A] sense of common interest; which sense each man
> feels in his own breast, which he remarks in his fel-
> lows, and which carries him, in concurrence with
> others into a general plan or system of actions, which
> tends to public utility.
>
> (Hume [1751] 1983, 95)

1 Introduction

Whether Japanese case markers are free word-forms, affixes, clit-ics, or morphemes of another kind has been the object of some de-bate. In a recent paper, Kageyama (2016) has provided what is proba-bly a definitive settling of the question by determining that Japanese case markers are intermediate between independent words and clit-ics, and form a morphological class of their own. Japanese tradition-al grammar calls this class *fuzokugo* 附属語. Obviously, the issue is mainly of interest to morphologists, but it is also relevant to any lin-guistic work containing annotated examples of Japanese, for in gram-matical annotation the marking of morph boundaries is achieved by way of conventional symbols that differ according to the nature of the morphemes being separated.

Given that linguistic annotation presupposes and reflects linguis-tic analysis (Iannàccaro 2000; Lehmann 2004a, 208), regardless of whether an author takes a realist or a nominalist stance on the na-ture of the objects represented in annotation (Hutton 1993, 166-7), one might assume that in the annotation of Japanese the boundary mark-ing of case particles is based on the morphological theory which has won the greatest consensus. Or, alternatively, that the use of bound-ary symbols is explicitly motivated, possibly as part of a general ex-planation given in some introductory remarks. This is not the case, though, for even a cursory glance at few random works will show how the treatment of the boundaries of Japanese case markers is erratic, to say the least. In grammatical annotation, Japanese case particles are variously represented as independent words, suffixes or clitics in works by different authors, chapters of collections, papers by indi-vidual authors and even within single articles. In extreme cases, the symbol used to mark the boundary of a given particle may even vary within one and the same annotated example. Moreover, the marking strategy adopted is hardly ever discussed. The question therefore nat-urally arises as to whether the actual use of boundary symbols reflects a theoretical choice, or is rather decided superficially, even at random.

This paper aims to determine the extent and the effects of the phe-nomenon of inconsistent case-particle boundary marking in the gram-matical annotation of Japanese. More specifically, it is focused on es-tablishing what represents 'inconsistent boundary marking', how it is dealt with in informational terms, what effect it has on communi-

502

Annali di Ca' Foscari. Serie orientale
56, 2020, 501-562

e-ISSN 2385-3042
ISSN 1125-3789

cation, and why it should be avoided. To this purpose, I will first build a typology of the word-boundary marking (or *tokenization*) strategies in the grammatical annotation of Japanese as they appear in English-language linguistics after the rule codification by Lehmann (2004b) and the *The Leipzig Glossing Rules* (hereafter LGR, Bickel, Comrie, Haspelmath 2004). I will base this typology on the stable equilibrium of the past fifteen years, not on a diachronic analysis of trends before or after 2004. Within this frame, I will individuate several forms of inconsistent boundary marking and, in a more general sense, of poor grammatical annotation, and discuss them individually according to the types of inconsistency and their different epistemic effects.

1.1 The Sample

For this research I parsed a sample of 184 English-language works (journal articles, chapters in collections, monographs) in linguistics containing annotated examples of Japanese. I chose this corpus from the material of my own library on the basis of the following criteria.

a. I only considered material published on or after 2006, allowing two years for the 2004 codification to settle and for the process of editing and publishing the papers to be complete before testing for compliance with the norms.

b. I gave preference to collections, both for practical reasons and because the different approaches to boundary marking can be better appreciated when appearing close to each other within a single volume.

c. I excluded from the sample works authored by those typologists with no specific knowledge of Japanese (such as Stassen 2009 or Creissels 2014, for example, writing on the typology of possession and functive phrases respectively). This is because I noticed that many such authors have no informed opinion about the morphology of Japanese case markers and are prone to adopt their sources' different styles of annotation, even when inconsistent with each other. As an exception, though, I included in the sample all relevant chapters in morphology collections and in general works on case (such as Malchukov, Spencer 2009) for I thought that the contributors to such volumes should have enough knowledge about case morphology to represent Japanese case particles accurately.

d. I made no specific choice of authors. However, since the contributors to the collections in my sample are among the leading scholars of the field, the sample ended up including works by the most authoritative linguists of Japanese (Jacobsen, Kageyama, Kuno, Miyagawa, Shibatani) and by the main experts on case (such as Blake and Malchukov).

503

Annali di Ca' Foscari. Serie orientale | e-ISSN 2385-3042
56, 2020, 501-562 | ISSN 1125-3789

1.2   The Structure of This Paper

This paper is structured as follows. After introducing linguistic annotation in general, with special regards to its nature as a coordinative convention, I will introduce, more specifically, the grammatical annotation of Japanese and illustrate the three attested methods of showing Japanese case-particle boundaries. With the aid of few concrete examples, I will point out the distinct theoretical stances on the morphological nature of Japanese case markers that each method presupposes. The works forming the sample will be listed at this moment, grouped by type of boundary marking strategy. Then I will briefly introduce the theory on the morphology of Japanese particles, noticing how the practicalities of grammatical annotation interact with it to originate the problematic boundary marking actually observable in the literature. In this light, I will then discuss the sample in more detail, focusing on inconsistent boundary marking in collections and single works. I distinguish two degrees of inconsistency: minor inconsistencies, when a differential treatment of particle boundaries might be ascribed to some rationale, and major inconsistencies, when it may not. Within each group, I will deal with individual works in chronological order of publication. This choice is based on practical reasons only, for the purpose of this paper is not to investigate the evolution of the use of boundary symbols, either in the whole sample or in the works of individual authors. Thus, I will not make any effort to establish diachronic trends. The authoritative linguist Masayoshi Shibatani, however, stands out in this and in many other respects, so that I deemed it necessary to briefly discuss his case in a specific section (§ 6.3). Since this paper does not deal with the validity of theories but only with the ways they are projected in annotation, I also considered irrelevant the quantitative distribution of the different types of boundary symbols, that is, the relative success of the distinct theories on the morphology of Japanese case markers as it may be inferred from the sample. I will therefore provide the distributional figures but I will not comment on them.

Inconsistent case-particle boundary marking is actually part of a larger phenomenon, which I call 'sloppy boundary marking', to which poor segmentation also belongs. Instances of poor segmentation are symbol mismatching, bad alignment, use of conflicting boundary markers in one and the same annotated example. Then again, sloppy boundary marking is only one subtype of 'sloppy annotation', an affliction endemic to all the grammatical annotation of Japanese which includes the use of non-standard symbols and of conventional symbols in unconventional ways, poor category labelling, copy-and-paste errors and so on. The vast typology of sloppy annotation is not the focus of this article, because to deal with it would complicate the picture and require a deeper, more extensive pars-

ing of the sample. However, the worst cases of sloppy annotation may be seen as representing a problem of cohesion, as they often destroy the epistemic value of annotation. I therefore decided to briefly address sloppy annotation and more specifically poor segmentation in a dedicated section.

In conventional terms, sloppy annotation represents a systematic violation of the norms of annotation and should be rejected on both practical and ethical grounds. However, it is persistent, unsanctioned so far, and evidently accepted by the community of linguists. To explain this I will discuss sloppy annotation in epistemic terms, as an instance of bad data encoding prone to convey 'misinformation'. This term refers to information with semantic contents which is *unintentionally* false but is wrongly believed to be true (Fetzer 2004; Floridi 2011; Lewandowsky et al. 2012, 124-5). I will show that the acceptance of sloppy annotation is due to the ability possessed by all linguists to filter out notational inconsistencies and thus to avoid being misinformed and acquiring false beliefs. Eventually, I will revert to the specific topic of inconsistent particle-boundary marking and, in the conclusive section, I will assess it practically and express my advice on the necessity of preventing it.

1.3 Unanswered Questions

I must admit that my motivation in writing this paper was a desire to investigate the reasons of the sense of sloppiness I often feel when faced with instances of grammatical annotation of Japanese. My choice of terms in naming certain aspects of bad annotation reflects this. Sloppiness implies carelessness; sloppy annotation, thus, originates in a lack of interest in accuracy, not in poor notational skills or bad knowledge of linguistics. How can it be then that so many linguists and editors, Japanese and non-Japanese, show so little motivation towards accurately representing Japanese in the metalanguage of annotation? The question is important, because to the reader who does not already know Japanese, annotation is essential to acquire new knowledge about the language. Moreover, good annotation gives readers the possibility of extracting from the examples more information than that to which the author has given salience.

The processes that lead to sloppy annotation take place in the minds of annotating authors, and are therefore impossible to ascertain without direct interviews, or at least without an in-depth analysis of the whole scientific production of each author. Certain attitudes, or deep motivations, may be inferred, but any judgment based on such inferences is pure speculation. To be honest, though, after so much research I have formed an opinion on the matter, and I will provide a concise account of it in the last subsection of the conclusions.

505
Annali di Ca' Foscari. Serie orientale   e-ISSN 2385-3042
56, 2020, 501-562   ISSN 1125-3789

2 Purpose and Structure of Linguistic Annotation

In the study of language, the term 'annotation' refers to any descriptive or analytic notation applied to raw language data, which may include any kind of transcription (phonemic, phonetic, relating to discourse structure), tagging, syntactic analysis, 'named entity' identification and so on (Bird, Liberman 2001, 23). Generally speaking, such a system of notation has the function of providing the community of linguists with a common, artificial metalanguage whereby they can describe and talk about linguistic phenomena. In particular, through shared symbols and category labels grammatical annotation allows experts in the field to understand schematically how general rules work in a given language even without knowing that language.[1]

In other words, linguistic annotation originated for the purpose of solving a coordination problem. In lack of any authority formally appointed to institutionally codify it, linguistic annotation has evolved as a convention in a true Lewisian sense (Lewis [1969] 2002, 85 ff.), i.e. an arbitrary system of norms practiced by a large population[2] for the strong reason that its usefulness for the community will improve as the number of members complying with it increases (according to the principle of "compliance dependence", Marmor 2009, 10-11). Over the years, a core of experts has taken stock of the trends in annotation practice as they appear in published annotated works, compared the usefulness of the notational devices with the needs of the community of linguists, recommended changes and proposed principles of standardization, in a process of "encyclopedic codification" (Marmor 2009, 50). For the devices relevant to this paper, which concerns grammatical analysis, the milestones have been Lehmann (1982, see his rationale for encyclopedic codification: 199-200), followed by Bakker et al. (1994), Croft (2003), Lehmann (2004b) and the two first editions of the LGR (Bickel, Comrie, Haspelmath 2004, 2008). Other relevant works include Schultze-Berndt (2006), Himmelmann (2006) and the recent Ide, Pustejowsky (2017). The lack of relevant publications on the topic after 2008 suggests that by that year, and probably even as early as 2004, sufficient codification had been reached to establish a coor-

---

1 With 'not knowing' a language I refer collectively to the conditions of not being a native speaker of that language, of not being able to use it in any real-word communication event, and of merely being unfamiliar with it (cf. Lehmann 1982, 200).

2 According to the philosophy dealing with conventions, the absolute size of the relevant population is important because in a large population it is hardly possible for members to reach formal agreements. Coordination conventions provide a solution to this problem. The community of linguists may be considered too small and too well networked for needing a coordination convention on a matter as technical as annotation. However, it is sparse; and it is a fact that no authority exists to regulate annotation or provide standard style-rules.

506

Annali di Ca' Foscari. Serie orientale
56, 2020, 501-562

e-ISSN 2385-3042
ISSN 1125-3789

dinating equilibrium, and no further major rule change was needed. For several years, then, the community of linguists has been able to rely on a powerful set of communicative tools which, by their very conventional nature, are expected to be employed consistently by every expert in the field.

Grammatical annotation is a complex system of smaller, specific conventions, regulating different levels or fields of linguistic analysis. When language examples are provided in the literature, for instance, utterances are reproduced in the original standard orthographies, or in other phonetic or phonemic transcriptions; they are freely translated into a background language which is not necessarily the writer's native language, but is often the *lingua franca* of English; they are segmented into components which are then re-described and re-codified by means of an artificial metalanguage of labels and symbols. These notational fields are classified in a number of levels or tiers (the latter notion is credited to Edwards 2001): Lehmann (2004b, 1836-7) proposes seven possible levels, Schultze-Berndt (2006) discusses twenty tiers/fields. The number of tiers used in actual instances of annotation is of course decided by authors according to the type of source text and to their analytical goal. Many tiers are often irrelevant and thus remain unexpressed.[3]

The number of possible tiers is high because one and only one distinct type of data is to be represented in each tier. Hence, each tier should occupy one dedicated line of text and, conversely, each line should contain one tier only. One would expect then that actual annotation should be structured as a stack of many lines, each encoding a different type of information (see 1 and 2 below). As a matter of fact, though, in the literature on syntax and semantics the norm is of formatting annotation in three lines only, corresponding to the three "main levels" of transcription, grammatical annotation and translation (Schultze-Berndt 2006, 215, 217). Since the very beginning of the encyclopedic codification of annotation the actual unit of notation has been the *line*, so that, in order for a trilinear format to be maintained, several tiers have to be compressed into one single line (Lehmann 1982, 218-9; Duranti 1997, 158-9; Edwards 2001, 327; Lehmann 2004b, 1837). In other words, grammatical annotation is afflicted by an unfortunate confusion between the notion of 'line', which is actually a unit of text, and that of 'tier', which corresponds more properly to a type of information or of analysis. The effect is that in the only two lines carrying grammatical information, it is often unclear which linguistic phenomena are represented and what value the encoding symbols have.

In the grammatical annotation of Japanese, this problem is particularly evident when decoding the information represented in Line

---

[3] On the overall necessity of frugality in annotation, see Schultze-Berndt 2006, 217.

1, where boundary symbols are first shown. In the next two sections I will address this problem by analyzing several concrete examples of annotated Japanese.

3 The Grammatical Annotation of Japanese

Japanese is a head-final SOV language in which nouns bear no grammatical gender or number and do not inflect. Case is marked with a set of postnominal particles. The right boundary of a NP's nominal head is also the site of the topic and other focus markers, which are added to or replace case markers. Case stacking may also occur but is very limited. For the purpose of this paper the difference between case markers, postpositions (see discussion on Tsujimura 2007, 2013 in § 6), focus markers and their combinations is irrelevant, and I will refer to them collectively as 'case markers', 'case particles' or occasionally 'postnominal particles'. The latter two expressions will be used when the need arises to avoid confusion with the word 'marker' used to refer to the notational symbol encoding the boundary between nominal and particle.

In this section I will provide a line-by-line analysis of two actual examples of trilinear annotation of Japanese.

1. Case particles as free word-forms (Takezawa 2016, 473)

Taroo	*ga*	*kabe*	*o*	***aka-ku / makka-ni***	*nut-ta*
Taro	NOM	wall	ACC	red-KU / crimson-NI	paint-PST

'Taro painted the wall *red/crimson*'.

1′
Line #	Tier type(s)	Actual contents and structure of example
Line 0	\ref	1. Case particles as free word-forms (Takezawa 2016, 473)
Line 1	\orth \phonet \mo	*Taroo ga kabe o aka-ku / makka-ni nut-ta*
Line 2	\it	Taro NOM wall ACC red-KU / crimson-NI paint-PST
Line 3	\ft	'Taro painted the wall *red/crimson*'.

2. Case particles as suffixes (Fukuda 2006, 90)

Takeshii-ga	*Kenji-no-kaoj-o*	***makkuro-ni***$_{*i/j}$	*nur*	*-ta*
T-NOM	K-Gen-face-Acc	black	paint	-Per

'Takeshi painted Kenji's face black'.

2′
Line #	Tier type(s)	Actual contents and structure of example
Line 0	\ref	2. Case particles as suffixes (Fukuda 2006, 90)
Line 1	\orth \phonet \mo \gr	Takeshi$_i$-ga Kenji-no-kao$_j$-o **makkuro-ni**$_{*i/j}$ nur -ta
Line 2	\it	T-Nom K-Gen-face-Acc black paint -Per
Line 3	\ft	'Takeshi painted Kenji's face black'.

508

Annali di Ca' Foscari. Serie orientale | e-ISSN 2385-3042
56, 2020, 501-562 | ISSN 1125-3789

In this paper I will quote the examples exactly as they appear in the sources. I will faithfully retain the original punctuation, transcription system, capitals, italics and bold characters. Examples will only differ from the originals in font type and by the addition of a header line (Line 0 above), hosting a progressive number and a reference to the source.

In this section I will explain the structure and contents of examples (1-2) with the use of (1'-2'), wherein two columns are added to the left of (1-2) to display line numbers and tier labels. The latter are the short names used to code tiers by Schultze-Berndt (2006), following the format of the Shoebox/Toolbox software developed by the Summer Institute of Linguistics.[4] This software and several more recent tools of the same kind are used for generating annotation automatically, mostly in treebanking. It follows that tier labels are ordinarily omitted in the simpler, hand-made grammatical annotation of the sample. The type of data, the processing and the output are very similar, though, so that I found it useful to quickly refer to tiers by means of Schultze-Berndt's labels. For clarity's sake, in the discussion that follows I will accompany each mention of tier labels with a reference to their functions and contents.

Line 0   This header line contains no linguistic data but only a reference Id and/or metadata used for indexing the example. In this paper, I will use Line 0 for showing the source of the example (as in 1-2) and for specifying the theoretical stance on the nature of Japanese case markers conveyed by the use of boundary symbols.

Line 1 (Level of Transcription)   In both (1-2) Line 1 hosts:

• A representation of the utterance in Latin script. This script is that of the standard orthography of the annotated language, or is a phonemic or phonetic transcription of it (tiers \phonem and \phonet respectively) (Lehmann 1982, 209-10; 2004b, 1836-7; Schultze-Berndt 2006, 220-1; Himmelmann 2006, 258). In the grammatical annotation of Japanese, this transcription follows the conventions of either of the two more common Latin notations of Japanese. These systems are the *Shūsei hebonshiki* 修正ヘボン式 (or Modified Hepburn system, henceforth Hepburn), and the *Naikaku kunreishiki* 内閣訓令式 (or Cabinet official directive system, hereafter Kunrei).[5] The former, dating from 1905, is widely used in public signs and when transcribing proper names. It is a type of broad phonetic transcription in which allophones are graphically distinguished (e.g. by writing <ta> and <chi> for [ta] and [t͡ɕi] respectively). The latter is the system officially pre-

---

scribed by the government since 1954, and is currently serving as the ISO standard 3602 for the Romanization of Japanese.[6] It is a phonemic transcription in which graphic confusion between allophones is maintained (<ta> and <ti> for [ta] and [t͡ɕi]). Neither system is a transliteration of the Japanese syllabaries *hiragana* 平仮名 and *katakana* 片仮名, the two overlapping sets of native graphemes which make it possible to render in writing the sounds of the language. In linguistics, a slightly modified Kunrei is most commonly used, but for the purpose of this paper the system adopted in the examples is of little interest. What is of great importance, however, is that both the history and the current status of the Hepburn and the Kunrei systems qualifies them both as *alternate established orthographies*, functionally substituting the standard logo-syllabic writing system currently used by the speech community.[7] Consequently, there is no need of representing utterances in the original non-Latin orthography, as is suggested for those languages whose writing systems do not use a Latin script (Schultze-Berndt 2006, 220-1; Bickel, Comrie, Haspelmath 2008, 2). Line 1 thus expresses both the orthographic transcription tier (\orth) and either one of the phonetic (\phonet) or phonemic (\phonem) transcription tiers.

• Morphemic segmentation. Line 1 also displays morphological analysis and expresses the morpheme break tier, labeled \mo. In both (1-2), morpheme breaks are shown by the hyphenation of the boundary between the verbal base and the suffix *ta*, and in (1) by the hyphen separating the adverbializers *ku* and *ni* from the adjectival bases. Affix boundaries are not marked in the Latin orthography of any language, Japanese included, so that inserting them in Line 1 may cause a "disfiguration" of the original orthography (Lehmann 2004b, 1852) or even prove impossible (Lehmann 2004b, 1837). To solve the problem, a quadrilinear representation has been proposed wherein a fourth line is added before Line 1 for displaying the utterance in unaltered orthography, irrespective of the original script (Schultze-Berndt 2006, 220-1, 237-8; Bickel, Comrie, Haspelmath 2008, 2).

• A theoretical statement on morphemic classification, implicitly expressed by boundary marking. The representation of morpheme breaks in Line 1 presupposes a background segmentation guided by morphological theory (Schultze-Berndt 2006, 240). This is what happens in (1), where case particles are distinguished from affixes because the boundaries with the nouns they mark are shown by blank

spaces. Blanks imply that particles are items of the same morphological status as the nouns they accompany. In contrast to this, in (2) the left boundaries of case particles (and even the right boundary of the genitive particle *no*) are hyphenated just like those of suffixes. The authors' choices are indeed expression of two distinct theoretical stances. I will address these matters in the next section of this paper.

In (2) only, Line 1 also hosts:

• An indexing of constituents. Here, the result phrase *makkuro-ni* is co-indexed with the two nominal arguments with the aid of subscripted letters. Such indexing is proper of a specific tier, that of grammatical tagging, coded \gr (Schultze-Berndt 2006, 241-3). Also belonging to \gr is the representation of reduced tree diagrams by bracketing (not shown in 2 but present in 7 and 8). In annotated Japanese it is very frequent to express grammatical tagging in the same line as the morpheme break tier \mo. This habit makes the informational burden of Line 1 even heavier, leading to illegibility in extreme cases (Lehmann 2004b, 1854).

• Underlying verbal forms. The sentence-final verb is represented in (2) as *nur-ta*. This is the underlying form of the surface form *nutta*, which is transcribed as such in (1). The function of representing underlying forms is also typical of morpheme break tier \mo (Schultze-Berndt 2006, 240) so that, in trilinear annotation, underlying forms end up furtherly cluttering Line 1.

Line 2 (Level of Grammatical Annotation)   In (1-2) Line 2 hosts:

• Morphemic segmentation and word-boundaries marking in biunique mapping to those in Line 1 (Lehmann 2004b, 1837). This means that exactly the same number of blanks and hyphens (and also equals signs, discussed below) must occur in the transcription (Line 1) and in the glosses (Line 2). Moreover, when one of those symbols marks a boundary in one line, it must be matched by a corresponding identical symbol in the other line, with left vertical alignment of the segmented words (Lehmann 1982, 219; Lehmann 2004b, 1851, 1856; Schultze-Berndt 2006, 239; Bickel, Comrie, Haspelmath 2008, 2). Ony (1) fully complies with this norm. In (2), Line 1 *makkuro-ni* is segmented with a hyphen which is not matched by another hyphen in the Line 2 gloss 'black'.

• The glossing of the object language's morphs by means of a reference to the morphemes of a metalanguage. In case of functional morphemes, abbreviated grammatical category labels in the metalanguage are used. Such morpheme-by-morpheme glossing constitutes the gloss tier \it (Schultze-Berndt 2006, 239; "Interlinear morphemic gloss" or IMG in Lehmann 2004b and Bickel, Comrie, Haspelmath 2008). Examples (1-2) present no major issue in this regard. Since this paper focuses on the boundary marking of Japanese case particles, in the first part of it I will not comment on how much the actual examples of annotation I quote comply with the norms of glossing.

511

Annali di Ca' Foscari. Serie orientale | e-ISSN 2385-3042
56, 2020, 501-562 | ISSN 1125-3789

However, after pointing out several issues of lack of accuracy and consistency in the use of boundary symbols, I will argue that the annotation of Japanese is afflicted by a more general problem, which I will call 'sloppy annotation'. Being a significant part of sloppy annotation, bad glossing and bad category labelling will then become relevant, and I will address them in § 7.

Line 3 (Level of Translation)

This line hosts a free or idiomatic translation in a background language (tier \ft) (Lehmann 2004b, 1838; Schultze-Berndt 2006, 234-7; Bickel, Comrie, Haspelmath 2008, 2). In this paper, translations of the examples are included to facilitate comprehension and to the effect of reproducing the quoted materials in complete form, but are irrelevant for the main discussion.

4 Word-Boundary Marking in the Grammatical Annotation of Japanese

As mentioned above when discussing (1-2), in Japanese the representation of the boundary between a nominal and a case marker can reveal two contrasting theoretical stances, i.e. that case markers are free word-forms or that they are suffixes. In a small minority of annotated works, a third stance is taken, that of treating case markers as clitics. In the following section I will show several examples of annotation and discuss in more detail which theory on Japanese case markers is implicitly supported by the boundary symbols adopted. I will assign any given work to one boundary marking type on the basis of the author's use of symbols even if the annotated examples are admittedly taken from secondary sources. This is because the convention requires that authors change their sources' glossing methods to fit their own purposes (Bickel, Comrie, Haspelmath 2008, 2), so that all annotation is to be taken by default to reflect an author's theoretical persuasion. In sum, authors are always directly responsible for the meaning expressed through their notational devices. I will consider the morphological categorization of Japanese case markers in a subsequent separate section.

3. Case markers as free word-forms, or of undetermined nature (Shimojo 2018, 405)
   Tarō   ga   Hanako   ni sono hon   o   yatta.
          NOM   to that book ACC   gave
   'Taro gave the book to Hanako'.

According to the norms of grammatical annotation, blank spaces carry information as symbols for the boundaries between free word-forms (Lehmann 2004b, 1852; Schultze-Berndt 2006, 221, 239). This seems to be the meaning assigned to blanks in (3), as one would ex-

512

Annali di Ca' Foscari. Serie orientale
56, 2020, 501-562

e-ISSN 2385-3042
ISSN 1125-3789

pect given the technical nature of Shimojo's work. As a matter of fact, however, Line 1 of (3) is not graphically different from one of the many occurrences of Romanized Japanese in the linguistic landscape, where only blanks are ever used, with no theoretical significance. There is no positive proof that a theory-based transcription is used in (3). In the technical terms explained above, instead of hosting the morpheme break tier \mo as expected, Line 1 might be a mere expression of the orthographic transcription \orth, and maintain its default boundary marking. This point is crucial and needs some clarification.

In the standard Japanese writing system, graphic spaces do not exist, but a distinction between morpheme types can be visually understood thanks to the fact that lexical words are commonly written in Chinese-derived logograms (*kanji* 漢字), while functional morphemes are rendered phonetically in syllabic script (*kana* 仮名). (The resulting multi-script orthography is known as *kanjikanamajiri* 漢字仮名交じり, or '*kanji-kana* mix'). If such a lack of graphic spaces were maintained in the alternate Latin orthographies of Japanese, utterances would be transcribed as unbroken strings of characters. This is of course unacceptable, making it necessary to adopt some breaking device. Reflecting a native speaker's intuitive notion of wordhood, but perhaps also to keep graphic words within a constant, moderate length, in the common Latin script of Japanese nominal prefixes and verbal and adjectival suffixes are conventionally attached to bases, while all other morphemes are separated by blanks. This is consistent with the orthographic trend pointed out by Himmelmann (2006, 257) of using the least possible number of boundary marking devices. A frugal segmentation of this kind is observable in the Hepburn transcription of the title of Hattori's papers in the references ([1949] 1960, 1950). In this transcription, the segmentation has no technical function.[8]

In instances of annotation like (3), where no boundary symbol other than blank spaces is present, the Latin script of Line 1 could be an occurrence of the default pretheoretical Latin orthography. As

---

[8] In the current standard *ISO3602*, which explicitly adopts the Kunrei system (ISO 1989, 1), it is declared that

> In a language using a syllabic System of writing is usually written without rules governing the division between characters and/or words, the conversion System must include such rules taking account of the morphological and grammatical structure of the language. (ISO 1989, v)

The main text of the ISO document, though, contains no provision on how to recognize and separate words. This is all it has to say on the matter:

> In all Japanese documents, a sentence in *kanzi* and *kana* is spelt in a sequence without divisions by words, in romanized Japanese texts separation into words is necessary. (ISO 1989, 1)

Obviously, the ISO itself is unconcerned about morpheme boundaries, and leaves segmentation to be decided by following the unwritten conventions already attached to the Kunrei orthography.

513

Annali di Ca' Foscari. Serie orientale | e-ISSN 2385-3042
56, 2020, 501-562 | ISSN 1125-3789

Lehmann points out, the use of a language's original orthography in the line hosting the morpheme break tier \mo is always problematic, for original orthographies employ several types of delimiter (blanks, hyphens and punctuation marks) which do not necessarily represent grammatical boundaries and may therefore interfere with the glossing (Lehmann 2004b, 1837).

Indeed, when in Japanese annotation Line 1 is a mere transcription in the Latin orthography, blank spaces occurring there are not symbols. They are given no theoretical meaning by the author. But then, according to the very norms of annotation, blanks are copied into Line 2, in biunique correspondence to Line 1 blanks. Now, a Line 2 graphic device charged with meaning by the rules of annotation is always decoded consequently. There is no way of opting out of the convention: writing Line 2 without giving a technical function to morpheme breaks is impossible, even if the morphology of Line 2 elements is irrelevant for the topic under discussion. In this sense, grammatical annotation does not allow for observance of Grice's maxims of Relation and Quantity, which require to be relevant and not overinformative (Grice 1975, 45-6).

It so happens that when Line 1 blank spaces, originally meaningless, intrude into Line 2, they acquire a proper meaning there, becoming boundary symbols for free word-forms. At this point, since symbols are supposed to have the same meaning in both Lines, Line 2 blanks unavoidably project their meaning back into Line 1. Line 1 blanks are now taken to represent the boundaries between free word forms just as they do in Line 2. Such a 'semantic infection' of Line 1 occurs even if an author wanted to use blanks pretheoretically, or atheoretically, giving them no technical function or meaning in either the object-language or the metalanguage. It happens even without an author's knowing it.

That blank spaces are often used meaninglessly is confirmed by the fact that a proliferation of blanks may even affect word-internal morpheme breaks:

4. All morphs represented as free word-forms (Shirai 2015, 225)
    Ken   wa   sin de iru.
    Ken   TOP   die RES NONPST
    'Ken is dead'.

Sinde is actually a converbal form derived from *sinu* 'die' (see glossing in 8 and 9), but the use of blanks spaces in (4) implies that the morph *de* is considered a word by itself, bearing the grammatical meaning of 'resultative' (I will comment on Shirai's glossing when discussing again this example in 29 below). Here the problem is that the author, perhaps under the influence of Latin orthography, only knows blanks as morpheme separators. An effect of this is that neither line contains

514

Annali di Ca' Foscari. Serie orientale | e-ISSN 2385-3042
56, 2020, 501-562 | ISSN 1125-3789

real morphemic analysis, even if it seems to do so. There would be no confusion if morpheme breaks were systematically expressed in a distinct line and not forcibly and intrusively inserted into the line already hosting the orthographic or phonetic transcription.

In sum, interpreting blanks as word-boundary symbols when occurring between nominals and case particles, and so as being expression of morphological analysis, might be a side effect of a take-over of Line 2 IMG by meaningless blanks. The possibility of such takeover makes it ultimately undecidable whether in annotated examples like (3) case markers are intentionally represented as free word-forms or not. Such ambiguity does not occur when more than one type of boundary symbol is present in Lines 1-2. For example, in (1) hyphens allow the reader to infer that the default, uncoded morpheme-breaking system of the Latin orthography, where the use of hyphens is not contemplated, is not employed in Line 1, and that this Line expresses the morpheme break tier \mo as expected. The effect is that blank spaces are correctly taken to encode meaning in both Lines 1 and 2, and inform that case particle are independent words.

In my sample, the works for which no such disambiguation is possible even by way of comparing annotated examples with each other are eighteen, nearly one fifth of all the works in which blanks are used. Among them, five belong to *The Oxford Handbook of Case* edited by Malchukov and Spencer (2009) and only contain one or two Japanese examples each. Eight are part of the *Handbook of Japanese Psycholinguistics* edited by Nakayama (2015). This volume contains seven more chapters that are relevant for the present analysis. A simple blanks-only-boundary-marking strategy thus accounts for nearly one half of the chapters. This is consistent with the nature of the collection, in which no complex morphologic analysis is required. The use of blanks in this volume can therefore be traced back to the atheoretical word-boundary marking discussed above. Two more ambiguous works are chapters in the *Handbook of Japanese Syntax* edited by Shibatani, Miyagawa and Noda (2017); three are in Hasegawa (2018b).

5. Case markers as free word-forms (Sato 2014, 317)

3-nin	*no*	*gakusei*	*ga*	*Itoo-sensei*	*ni-yotte/*	*ni*	*maikai*	*sas*
3-people	GEN	student	NOM	Ito-prof.	by/	by	every-time	call

are-ta.
PASS-PAST
'Three students were called on by Prof. Ito every time'.

As opposed to blanks, hyphens are used in grammatical annotation to separate segmentable word-internal morphemes (Lehmann 2004b, 1852; Bickel, Comrie, Haspelmath 2008, 2). Just like in (1), in (5) both blank spaces and hyphens are used. The co-occurrence of both break-

515
Annali di Ca' Foscari. Serie orientale | e-ISSN 2385-3042
56, 2020, 501-562 | ISSN 1125-3789

ing devices implies that they are used as symbols with contrasting meaning, and therefore that case markers are here considered to be independent words, not affixes.

An unambiguous treatment of case markers as free word-forms is by far the most common in my sample, occurring in a group of 95 works. I list them in chronological order, omitting references to single chapters in collections:

List 1. Case Markers Unambiguously Considered as Free Word-Forms

Shibatani 2006;
Otoguro 2006;
1 chapter in Rehbein, Hohenstein, Pietsch 2007;
3 chapters in Frellesvig, Shibatani, Smith 2007;
Shibatani 2009c;
2 chapters in Malchukov, Spencer 2009;
1 chapter in Frellesvig, Sells 2013;
1 chapter in Giriko et al. 2014;
5 chapters in Nakayama 2015;
1 chapter in Kubozono 2015;
Shibatani, Kageyama 2015 (the "Introduction" to the whole series in the earliest HJLL volume);
7 chapters in Minami 2016;
18 chapters in Kageyama, Kishimoto 2016;
8 chapters in Kageyama, Jacobsen 2016;
18 chapters in Shibatani, Miyagawa, Noda 2017b;
13 chapters in Hasegawa 2018b;
all 23 chapters in Pardeshi, Kageyama 2018.

In (6) the second most used strategy for separating nominals and case particles is shown. Here, the boundaries of postnominal particles are marked with hyphens, like in the morphemic segmentation of verbal forms. Case markers are thus unambiguously treated as affixes or bound morphemes:

6. Case markers as bound morphemes (Tsuboi 2010, 420)
   Taro-wa   Hanako-ni   home-rare-ta.
   Taro-TOP Hanako-DAT praise-PASS-PST
   'Taro was praised by Hanako'.

In my sample, case-marker hyphenation consistently occurs in 50 works:

516

Annali di Ca' Foscari. Serie orientale   │   e-ISSN 2385-3042
56, 2020, 501-562   │   ISSN 1125-3789

List 2. Case Markers Unambiguously Considered to Be Affix-Like

Fukuda 2006;
Tsujimura 2007;
Tomioka 2007;
1 chapter in Rehbein, Hohenstein, Pietsch 2007;
4 chapters in Frellesvig, Shibatani, Smith 2007;
13 chapters in Miyagawa, Saito 2008;
Harada, Larson 2009;
6 chapters in Malchukov, Spencer 2009;
Tsuboi 2010;
Nakazawa 2012;
Motomura 2013;
Tsujimura 2013;
9 chapters in Saito 2014;
1 chapter in Giriko et al. 2014;
1 Kishimoto, Kageyama, Sasaki 2015;
1 chapter in Kubozono 2015;
2 chapters in Minami 2016;
1 chapter in Kageyama, Jacobsen 2016;
3 chapters in Hellan, Malchukov, Cennamo 2017;
1 chapter in Hasegawa 2018b.

The choice of either blanks or hyphens appears to be even more strongly theory-based whenever case markers are distinguished from clitics:

7. Case markers as free word-forms contrasting with both clitics and affixes (Shibatani 2017, 281)
   [*Ame ga   kyuuni   huridas-i*]=*soo   da.*
   rain NOM suddenly fall.start-NMLZR=EVID COP
   'It appears that rain will start falling suddenly'.

8. Case markers as bound morphemes contrasting with clitics (Shibatani 2007a, 47)
   *Gakkoo-ni sake-o **non-de** **ku-ru**=to-wa   nanigoto=ka?*
   school-to sake-ACC drink-CON come-PRES=COMP-TOP   whatever =Q
   'Whatever (it is the matter with you) to come to school (after) drinking sake?'.

In grammatical annotation, the boundaries between clitics and hosts are shown by equals signs (Lehmann 2004b, 1852; Bickel, Comrie, Haspelmath 2008, 2). In (7-8) the use of boundary symbols reveals that the author is conscious of the existence of clitics in Japanese and of the need to mark clitic boundaries in a specific fashion. He identifies as clitics the evidential marker *soo* in (7) and the two sentence-final particles *to* and *ka* in (8), and shows the relevant boundaries consequently. Shibatani's contrasting ways of treating case markers

as free word-forms in (7) and as bound morphemes (alike to verb-internal morphemes) in (8) are therefore to be reckoned as expressing two opposite morphologic analyses. In his career, that same author has also taken a third, different theoretical stance and represented Japanese case markers as clitics:

9. Case markers as clitics (Shibatani et al. 2014, 364)
    Teisyu=ga   *kaet-te*   *kuru=no=o*   *matte*   *iru*.
    husband=NOM   return-CON   come=NMLZM=ACC   wait   be
    '(I am) waiting for (my) husband to return (home)'.

In (9) the use of boundary symbols associates case markers to clitics and keeps them distinct from free word-forms and word-internal morphemes. Shibatani is the only author in my sample to have ever firmly taken the stance of considering Japanese case markers clitic-like, as can be seen also in Sibatani 2007b and Shibatani 2013. His case is of particular interest and I will discuss it later in a separate subsection (§ 6.3).

5 The Morphological Categorization of Japanese Case Particles

In this section I will briefly introduce the main hypotheses on the morphologic categorization of Japanese case markers, without discussing the validity of each theory.

According to the influential analysis of Hattori ([1949] 1960, 1950), Japanese case and focus markers are non-independent free forms: non-independent because they may never be used alone to form a whole utterance, and free because they can be uttered separately (by means of pauses) from other adjoining forms. They are therefore different from affixes, which are non-independent and bound forms, and from free-standing lexical words. Hattori named the class of case- and focus-markers *fuzokugo*, a term taken from Japanese school grammar which he himself translated as 'clitic' ([1950] 1960, 103). Hattori's association of Japanese case particles with clitics went basically unchallenged until Vance (1993), who applied Zwicky, Pullum's (1983) and Zwicky's (1985) diagnostics and concluded that Japanese postnominal particles are certainly not affixes, but neither are they clitics. Rather, Vance concluded, focus markers are independent words, while case markers "seem no less word-like than English prepositions" (Vance 1993, 29).

Hattori's and Vance's divergent analyses only became relevant for grammatical annotation much later, when detailed conventions for marking morpheme boundaries were thoroughly codified by Lehmann (2004b) and the LGR (Bickel, Comrie, Haspelmath 2004).

518

Annali di Ca' Foscari. Serie orientale   e-ISSN 2385-3042
56, 2020, 501-562   ISSN 1125-3789

One might think that from that moment on the segmentation of functional morphemes has been implemented by way of either spaces or equals signs, according to the theory espoused by each author. However, that has not happened, as my corpus reveals. Hyphens have not been abandoned and are extensively used even today, while the use of equals signs is virtually nonexistent. Why?

One reason for not discontinuing the use of hyphens is, as trivial as it may sound, the strength of the convention itself. Up to the first pioneering attempt at codifying grammatical annotation by Lehmann in 1982, the hyphen had been used as a convenient general-purpose morpheme-boundary symbol, to the extent that Lehmann himself remarked:

> By now, the only morpheme boundary symbol used in L1 [source] texts is the hyphen. One might ask whether it would not be desirable to distinguish various kinds of morpheme boundaries. (Lehmann 1982, 215)

At the time of Lehmann's writing, then, the hyphen was polysemic and carried little information. It was so versatile that it survived unscathed the thorough codification of 2004. An important role was also played by the status of orthography held by the Latin script of Line 1. As Lehmann points out,

> No orthography distinguishes clitic boundaries from word and morpheme boundaries. If L1 is represented in conventional orthography, then the simplest solution for an IMG is not to distinguish them either. (Lehmann 2004b, 1852)

Once the atheoretical word-boundary marking of the Hepburn or Kunrei orthography has intruded into the second line of annotation, we are left with very little choice of symbols for marking case-particle boundaries: the only alternative to meaningless blanks are all-purpose hyphens.

In more general terms, the same factors pointed out by Himmelmann (2006) as part of a general discussion on the principles for representing wordhood in transcription seem to be at work here. First, there may be inferred to exist what I call a 'cognitive constraint' which imposes limits on how many morpheme-boundary symbols can be used consistently by writers and can be understood by readers without constantly referring to the conventions (Himmelmann 2006, 273 fn. 2). When strongest, such a constraint originates a general tendency to reduce the options for distinguishing morpheme boundaries to the smallest possible number of two: marking no boundary, using a space. It will be recalled that the two alternate Latin orthographies of Japanese use this very strategy. When increased complexi-

519

Annali di Ca' Foscari. Serie orientale | e-ISSN 2385-3042
56, 2020, 501-562 | ISSN 1125-3789

ty is needed, as it happens in theory-informed segmentation, hyphens are first added. Indeed, even in the Latin orthographies of Japanese a limited use of non-standard hyphens is occasionally attested, although only for separating roots in compounds. With the introduction of hyphens an equilibrium is reached. Now the constraint hinders further complexity, practically inhibiting the use of additional symbols like equals signs even when an explicit marking of clitic boundaries should be scientifically appropriate. This block leads to forcibly reducing clitics (or morphemes with some clitic-like property) to either affixes or independent words (Himmelmann 2006, 258), to the effect that clitics are wiped out from the metalinguistic representation of grammar.

To better illustrate this point, one can go one step further in the morphological analysis of Japanese case markers. Drawing from both Hattori (1950) and Vance (1993), Kageyama has recently claimed that *fuzokugo* are neither independent words or clitics. Rather, they constitute a distinct class of syntactic words, which he dubs "non-independent words". These "non-cliticized morphologically bound morphemes" are characterized by the fact that, though "glued to their bases", they occupy a position of their own in syntactic structure without phonologically adjoining to any surrounding word (Kageyama 2016, 509).

According to Kageyama, then, morphological objects of four different kinds exist in Japanese. Their properties are plotted as follows.

Table 1 Types of morphological objects in Japanese

	Independent words	*Fuzokugo*	Enclitics	Affixes
Independent status in syntactic structure	Yes	Yes	Yes	No
Phonological independence	Yes	Yes	No	No
Unboundness	Yes	No	No	No

(Kageyama 2016, 519)

Kageyama's notion of *fuzokugo* is absent from European linguistics (Kageyama 2016, 519). Actually, it might be that *fuzokugo*-like objects exist in other languages as well but have been unrecognized so far. At least for annotating Japanese a new symbol should be adopted in order to distinctively mark *fuzokugo* boundaries and hence distinguish them from morphemes of a different nature. I would suggest the sign <~>, available on all English keyboards and already considered for the role, as discussed below in relation to Ogawa (2009), Sakoda (2016) and example (22). But if three symbols are already too

520
Annali di Ca' Foscari. Serie orientale
56, 2020, 501-562

e-ISSN 2385-3042
ISSN 1125-3789

many, so that little effort is made to distinguish clitics from non-clitics, one can image how complicated it would be to add a fourth symbol to the list. Indeed, as far as grammatical annotation is concerned, *fuzokugo* seem to suffer the same fate as clitics: they are not recognized as an individual class of morphemes, are forcibly assimilated to clitics and treated just like them, i.e. dropped from the above table and represented as either independent words or affixes. See how Kageyama deals with their annotation:

10. Case markers (*fuzokugo*) represented as free word-forms, clitics as affixes (Kageyama 2016, 511)

Kanozyo	*wa*	[*otoko*	*ni*	*damas-are*]-*soo*		*da.*
she	TOP	[man	DAT	deceive-PASS]-appearance		COP

'She looks like getting deceived by the man'.

Under Kageyama's analysis, the case marker *ni*, the focus marker *wa* and the dummy item *da* are all *fuzokugo*, but throughout his article he treats their boundaries as those of free word-forms. Likewise, Kageyama considers the element *soo* to be a clitic (2016, 513), but marks its boundary as that of morpheme *are*, which is an affix. Such a reductionist treatment of morpheme boundaries occurs throughout Kageyama's paper – the very paper in which the author ascertains the nature of Japanese case and focus particles as entities distinct from clitics and remarks the significant difference between clitics and affixes. Obviously, Kageyama himself considers it unnecessary that such a complex set of distinctions be reflected in annotation.

The morphological nature of Japanese postnominal particles may be unclear and difficult to assess and, as Kageyama's case shows, properly conveying it in annotation can be troublesome. In particular, the current conventions do not provide an accurate choice of symbols under the *fuzokugo* and clitic theories. How to mark case-particle boundaries, all considered, is up to individual authors, in accordance with their theoretical persuasion and with the relevance of particle morphology relative to the topic they discuss. Of the three symbols made available by the current convention, the hyphen is surely the least accurate, but none is wrong in itself. What linguistic theory clearly says, though, is that there are no prosodic or morphological differences between types of postnominal particles (like that between case/focus markers and postpositions, which may only be distinguished syntactically) such as to justify the use of distinct boundary symbols in annotation. In any given work, therefore, coherence and theory-consistency demand that one and only one type of symbol be used for marking the boundaries of Japanese case particles.

521

Annali di Ca' Foscari. Serie orientale   e-ISSN 2385-3042
56, 2020, 501-562   ISSN 1125-3789

6 A Discussion of the Sample

In this section I will discuss my sample of 184 works with regards to clarity and consistency in boundary marking. As noticed above, the choice of boundary symbols is hardly ever motivated. The few authors who dedicate even the briefest comment to the morphology of Japanese case particles are the following.

Otoguro (2006) is part of a PhD dissertation, of particular interest in view of the author's theorization on the nature of Japanese case. Otoguro so concludes his analysis:

> Thus, there are plenty of reasons to believe that the nominal particles in Japanese are suffixes.
>
> However, the particles in question display some degree of separability from the host. (Otoguro 2006, 242)

> The conclusion […] is that the Japanese nominal particles seem to have mixed properties of morphological suffixes and independent syntactic units, i.e. phrasal suffixes or clitics […]. Therefore, I analyse the particles as bound elements in the morphological component, but non-projecting postpositions […] in c-structure. (Otoguro 2006, 243)

According to Otoguro, then, Japanese "nominal inflection" is realized by bound particles that are morphologically intermediate between clitics and affixes. Interestingly, his conclusion is different from that of Kageyama (2016), who uses the same diagnostics (like Zwicky, Pullum 1983 on clisis). Consequently, Otoguro's most logical choice of symbols for showing case-marker boundaries in annotation should be equals signs (as in Otoguro 2003), or possibly hyphens. Instead, just like Kageyama, Otoguro opts for blank spaces, the least accurate symbol according to his own analysis. Otoguro never uses equals signs for marking clitic boundaries.

In the entry "Japanese" of the fourteen-volume *Encyclopedia of Language and Linguistics* edited by Keith Brown, Shibatani explicitly associates Japanese case markers to prepositions: "postpositional particles are used instead of prepositions" (Shibatani 2006, 104) and consistently annotates them as free word-forms.

In the two latest editions of the *Introduction to Japanese Linguistics*, Tsujimura painstakingly distinguishes "case particles" from "postpositions" (2007, 121-30 and fnn. 209-11; 2013, 133-43, 233-4). Case particles, she explains, are the surface markers of grammatical relations, with no semantic content. Morphologically, they are bound morphemes like the verbal affixes expressing tense, form a single NP with the host noun, are outside the word class but are not part of nouns either. Postpositions are the markers of semantic re-

522

Annali di Ca' Foscari. Serie orientale   e-ISSN 2385-3042
56, 2020, 501-562   ISSN 1125-3789

lations, and bear a specific or inherent meaning. Even if not constituting a word class (Tsujimura 2007, 122; 2013, 133), they nevertheless represent a lexical category (2007, 210; 2013, 233), the Japanese counterpart of English prepositions. As such, they head the PP formed with the preceding noun. However, postpositions too are morphologically bound, for they cannot stand independently but "always occur with accompanying nouns in order to form a meaningful unit" (Tsujimura 2007, 122). On such a basis, Tsujimura consistently hyphenates the morpheme boundary of both case particles and postpositions.

A remarkable case is that of Iwasaki (2013), a general introduction to Japanese language. Along with Shibatani, Kageyama (2015), discussed below, Iwasaki's work contains both a statement on the nature of Japanese postnominal particles and a clarification about the symbols used for marking their boundaries:

> In presenting clauses/sentences, words are separated by spaces. Particles (case particles and others) can be regarded as enclitics, but for the sake of simplicity and readability (and in accordance with the tradition), a space is put to separate a word from the following particle (except for Chapter 1), as if particles were regular, independent words, e.g. *boku wa koko de hon o katta*. (I TOP here INS book ACC bought) "I bought a book here". (Iwasaki 2013, xix)

Iwasaki's reasons for preferring blanks to hyphens are the same as Kageyama's, whose case was discussed above. Both authors thus express a refusal to introduce specific clitic-boundary markers as a third kind of symbol. It is puzzling, then, why Iwasaki makes an explicit exception for Chapter 1, where he hyphenates particle-boundaries in the twenty-two annotated examples that contain postnominal particles. He provides no justification for it, and neither can I find one.

Two contributors to Hasegawa 2018b take an explicit position on the matter. One, Horie, explicitly states that Japanese case markers are bound morphemes (2018, 66). Horie's boundary marking is quite inconsistent, though, and I will discuss it in § 6.2.2. The second one is Nakamura (2018), who distinguishes case particles from both declensional suffixes and independent words, and identifies them as clitics:

> These case particles are phonologically bound to the preceding words, but the fact that other elements may intervene between the case particles and the nouns they mark and that their scope may extend over more than one NP when they are coordinated [...] indicates that the case particles are phrasal clitics rather than nominal declensions. (Nakamura 2018, 249)

That said, Nakamura is facing the decision as to whether to project such a distinction on boundary marking. Following the general

trend of ignoring clisis in annotation, Nakamura disregards equals signs – which no other contributor to the volume ever uses anyway, with the unsurprising exception of Shibatani (2018a) – and chooses to highlight the bounded nature of case markers by using hyphens.

In my sample, other authors just touch upon the topic and seem to be intentionally noncommittal about the nature of Japanese case markers. In "Case marking" Kishimoto (2017) writes:

> Even though the grammatical functions of arguments are coded morphologically by means of postnominal case markers in Japanese, the relationship between the two is not always straightforward. (Kishimoto 2017, 447)

The volume with Kishimoto's article is one of the *Handbooks of Japanese Language and Linguistics* (HJLL) published by De Gruyter Mouton (Shibatani, Miyagawa, Noda 2017b). If, in search for clarification, the readers refer to the general introduction to the series, identical in each volume, they will find a straightforward but equally noncommittal explanation of the boundary marking strategy adopted in HJLL:

> In line with the general rules of Romanization adopted in books and articles dealing with Japanese [...] HJLL transliterates [*sic*] example sentences by separating word units by spaces. [...] [W]ord-internal morphemes are separated by a hyphen whenever necessary, although this practice is not adopted consistently in all of the HJLL volumes. (Shibatani, Kageyama 2015, xviii-xix)

> Special attention should be paid to particles like *wa* (topic), *to* 'with' and *e* 'to, toward', which, in the HJLL representation, are separated from the preceding noun or noun phrase by a space [...]. Remember that case and other kinds of particles, though spaced, form phrasal units with their preceding nouns. (Shibatani, Kageyama 2015, xix)

> Because case markers can be set off by a pause, a filler, or even longer parenthetic material, it is clear that they are unlike declensional affixes in inflectional languages like German or Russian. Their exact status, however, is controversial; some researchers regard them as clitics and others as (non-independent) words. (Shibatani, Kageyama 2015, xx)

In the annotated examples of the HJLL series, one gathers, hyphens are explicitly given a meaning in accordance with the general conventions, but blank spaces are not. While the use of hyphens for showing case-particle boundaries is more or less explicitly rebutted on theoretical grounds, blanks are instead introduced into Line 1 and into

morphological analysis as part of the pretheoretical boundary marking of the object language's Latin orthography. Blanks also occur in the metalanguage, but bear no meaning there as well. This is what I have noticed and discussed in relation to (3).

As remarked, the few authors quoted above represent exceptions to the norm, which is of not taking any explicit stance on the nature of Japanese case markers and of not expounding the rationale behind the choice of boundary symbols. In most cases, the actual boundary marking used in annotation simply works as an implicit statement of what an author believes Japanese case markers morphemically are. Experts do not agree as to whether this is acceptable. The LGR deny that an IMG may be a way of stating an analysis (Bickel, Comrie, Haspelmath 2008, 2), while Lehmann has an open attitude (2004b, 1836). For this author, especially in the morpheme break tier and in the IMG, it is possible to conceive representations from which readers can abduce an author's entire grammatical theory (Lehmann 2004a, 208). Of course stating a morphological analysis exclusively by ways of morpheme break symbols may only work if the distribution of symbols is coherent.

One interesting instance of using boundary symbols as theory statements is Ogawa (2009), the only chapter devoted to Japanese in the *Oxford Handbook of Case* (Malchukov, Spencer 2009). In the annotated examples, Ogawa hyphenates Japanese case-particle boundaries, but uses the sign <~> whenever he mentions them, as in the list "sensei~*ga* (teacher-nominative), ~*wo* (accusative), ~*ni* (dative)" etc. (Ogawa 2009, 780). He gives no clarification about the meaning of the tilde. A similar case is Sakoda, who uses blanks in the examples and tildes when mentioning (and glossing) constructions, as in "~*ga aru* '~NOM exist'" (Sakoda 2016, 145). Equally unexplained tildes also slip into Narrog (2017), as shown below in (22). If these authors are giving the tilde a value as a boundary symbol, they would indirectly suggest that Japanese case markers are neither ordinary affixes or free word-forms. In this case, they would also implicitly invite their readers to ignore the boundary marking in the annotated examples, where, alas, they only use conventional symbols.

6.1 Inconsistent Case-Particle Boundary Marking in Collections

In my sample, the norm for collections is to host both authors that hyphenate particles boundaries and authors that use blanks. This variation is expected, of course. I consider it inconsistent because, when unexplained, a use of conflicting boundary symbols disconcerts readers with no specific knowledge of Japanese morphology. In the following list, for each parsed volume I will schematically show the ratio of articles with hyphens to those with blanks by means of figures

525

Annali di Ca' Foscari. Serie orientale
56, 2020, 501-562

e-ISSN 2385-3042
ISSN 1125-3789

separated by a slash. I will list the volumes of the HJLL series separately, further below. In § 6.2 I will specifically deal with those cases in which several conflicting boundary symbols are used in one and the same single work (monograph, article or chapter).

List 3. Ratio of Hyphens to Blanks in Collections

• *Current Issues in the History and Structure of Japanese* (Frellesvig, Shibatani, Smith 2007): 4/3.
• *The Oxford Handbook of Japanese Linguistics* (Miyagawa, Saito 2008): 13/0; in two other chapters (Takahashi 2008; Heycock 2008) both hyphens and blanks are inconsistently used.
• *The Oxford Handbook of Case* (Malchukov, Spencer 2009): 6/7.
• *Japanese/Korean Linguistics*. Vol. 20 (Frellesvig, Sells 2013): 1/1.
• *Japanese/Korean Linguistics*. Vol. 22 (Giriko et al. 2014): 1/1; moreover, in a third chapter equals signs are used.
• *Japanese Syntax in Comparative Perspective* (Saito 2014): 10/0.
• *Transitivity and Valency Alternations* (Kageyama, Jacobsen 2016): 0/10; in two other chapters (Shibatani 2016; Malchukov 2016) more than one symbol is used.
• *The Cambridge Handbook of Japanese Linguistics* (Hasegawa 2018b): 1/13; in one other chapter (Horie 2018) both symbols are used.

 In the works in my sample, then, the tendency is for editors to leave authors free to decide which boundary-marking system to use. Editors also allow authors to express their stance on the morphology of Japanese case markers either explicitly (almost never done) or implicitly, through the very boundary symbols they use (the norm). The listed volumes are among the most important English-language collections on Japanese linguistics in the years considered, so that intuitively such an editorial attitude may be considered representative of the editorial strategy of the period.

 As for the recent HJLL series, given the statement in the general introduction to the series, quoted above, one would expect all authors to mark postnominal particle boundaries with blank spaces only. This is not the case, though, as several freak instances of hyphenation do pop up.

List 4. Ratio of Hyphens to Blanks in the HJLL Series

• HJLL *Psycholinguistics* (Nakayama 2015): 0/17; in two other chapters both hyphens and blanks are used (Murasugi 2015; Sakamoto 2015).
• HJLL *Phonetics and Phonology* (Kubozono 2015): 1/1.
• HJLL *Applied Linguistics* (Minami 2016): 2/8.
• HJLL *Lexicon and Word Formation* (Kageyama, Kishimoto 2016): 0/18.

• HJLL *Syntax* (Shibatani, Miyagawa, Noda 2017b): 0/19; in four other chapters the authors use more than one symbol (Iwata 2017; Kitagawa 2017; Miyagawa 2017; Narrog 2017; Shibatani, Miyagawa, Noda 2017a).
• HJLL *Contrastive Linguistics* (Pardeshi, Kageyama 2018): 0/23.

In the HJLL series, one gathers, compliance to the general, explicit rule of using blanks is ultimately entrusted to each author's sense of responsibility, with no editorial attempt to enforce it. The success of such a policy is limited, though, as the six inconsistent cases show.

6.2 Inconsistent Case-Particle Boundary Marking in Single Works

By 'single works' I refer to monographs and dissertations, articles in journals and chapters in collections, either written by a single author or by multiple authors. In this sense, all the chapters contained in the collections discussed above are single works.

As expected, a consistent use of one case-particle boundary symbol only is so well attested among the single works in my sample as to render it unnecessary to provide the exact figures. However, in eighteen cases, slightly less than 10% of the sample, the authors mark the boundaries of postnominal particles with different symbols (usually with a mixture of hyphens and blanks). Such a differential treatment is not backed by any theory (that is, is invalid, or theory-inconsistent), and, although rare, is confusing enough as to justify a specific discussion. I classified the instances of such poor boundary marking in two levels, according to the degree of inconsistency. It should be noticed that the epistemic effect of bad boundary marking may actually be worse when inconsistency is low (see § 9.1 below).

6.2.1 Minor Cases of Inconsistency

I considered of low inconsistency works in which, (a) only one type of boundary symbol is used but with certain recurring particles or combination of words, whose boundaries are instead shown by way of other symbols; and (b) such a treatment is regular enough as to suggest that those particles or phrases may have some special status. In simpler words, I classified in this group those uses of conflicting boundary symbols for which some rationale might be inferred, even if weak. I also assigned to this group the works in which (a) the inconsistency is limited to so few instances as to be obviously due to copyediting errors; and (b) it does not affect the robustness of the general discussion. The most interesting cases are the following.

In Fukuda (2006) case-particle boundaries are always hyphenated, but so are also the boundaries between nominal heads and their gen-

527

Annali di Ca' Foscari. Serie orientale | e-ISSN 2385-3042
56, 2020, 501-562 | ISSN 1125-3789

itive-marked modifiers, as shown in (2): *Kenji-no-kao-o*. The segmentation hints to the fact that the bound between those elements is so tight as to resemble that between base and affix. Such a hyphenation is maintained throughout the article and is consistent with Fukuda's general treatment of case particles. However, it is incorrect, for other modifiers may occur right before the head noun (see discussion of 17).

11. Case markers in local functions as free word-forms (Murao 2009, 102)
     *Niwa   ni   neko/*ki-ga   iru*
     yard   in   cat/tree-NOM   be.
     'There is a cat/tree in the yard'.

In Takahashi (2008) and Murao (2009), with some effort a rationale might be detected according to which certain particles in local functions (such as Locative *de*, Dative *ni*), the focus marker *mo* and few sentence-final particles (such as Q *ka* and quotative *to*) are independent words, whereas all other postnominal particles are affix-like. This interpretation is confirmed by the fact that both authors gloss local particles and focus markers with English free word-forms (prepositions, adverbs, pronouns). One such example is (11), where Dative *ni* is glossed 'in'.

12. Case markers in local functions as affix-like (Ono 2013, 420)
     [*John no Oosaka-de no kooen*]
      John GEN Osaka-in GEN speech
     [Untranslated in the original source, but meaning 'John's speech in Osaka']

A similar but opposite case is that of Ono (2013). As shown in (12), the author only uses blanks but for Locative *de* (variously glossed 'in', 'at' and 'on'), whose boundary he makes a (not always successful) effort at hyphenating. With such hyphenation, Ono suggests that *de* belongs to a morphological class distinct from that of all other case markers.

Iwasaki also belongs here. As seen above, Iwasaki explicitly states (2013, xix) that in Chapter 1 he will not use blanks. And, indeed, in that chapter he marks postnominal particle boundaries with hyphens. Nevertheless, in four of the twenty-two relevant examples there, blanks do pop up for tokenizing random instances of particles *o*, *wa*, *ga* and *ni*. Thanks to the author's preliminary explanation about his boundary-marking strategy, though, the reader has no problem in interpreting such freak hyphenation as a mere copyediting error.

13. Genitive marker *no* as free word-form (Saito, Lin, Murasugi 2014, 13)
     Taroo-wa   san   -satu   no   hon   -o   katta
     Taroo-TOP   three-CL   no   book-ACC   bought
     'Taroo bought three books'.

Of special interest is Saito, Lin, Murasugi (2014), represented by (13). The authors use hyphens, like all other contributors to the same volume. As an exception, though, they show with blanks the boundary of Genitive particle *no*, which they call a "modifying marker", and specifically a "contextual Case marker", and do not gloss. By so doing the authors probably only want to stress a functional (not morphological) peculiar relation occurring between a nominal and this marker, whichever it may be. Given that the topic of the chapter are NPs, a reader will interpret such a choice of symbols as based on a morphological theory according to which the marker *no* does not belong to the same morphological class as all other postnominal particles.

14. Allative case markers *e* as clitic (Narrog 2017, 345)

    Taroo   *wa*   **_gakkoo=e_**   *arui.te*   *it.ta/ki.ta.*
    Taroto   TOP   school ALL   walk-GER   go-PST/come-PST
    'Taro walked (walk went/came) to school'.

Narrog marks with an equals sign the Line 1 boundary of Allative *e* (shown in example 13), and in one occurrence that of Limitative *made* as well (in a second occurrence, *made* is hyphenated). However, he shows with blanks the boundaries of other local case markers like *ni* (also glossed ALL) and Ablative *kara* (Narrog 2017, 333, 345-6, 348). Again, no theory supports a morphemic status as clitics of *e* and *made* only, as distinct from all other postnominal particles. But all annotation in Narrog (2017) is problematic, with repeated violations of lexical conventions. This is an issue of cohesion, and I will discuss (14) again in § 7 as an instance of 'poor segmentation'.

15. Focus marker *mo* as affix-like (Miyagawa 2017, 604)

    Dareka   *ga*   *dono-hon-mo*   *yonde-iru.*
    someone   NOM   every book   read-ing
    'Someone has read/is reading every book'.

Over-hyphenation also occurs in Miyagawa (2017), same volume as Narrog (2017). Departing from the official policy for the HJLL series, Miyagawa hyphenates case-particle boundaries in all the notes but one, the boundary of genitive *no* in some instances, and the boundaries between all elements in phrases having the structure '*dono* N *mo*', like *dono-hon-mo*; *dono-hashi-mo*; *dono-sensei-mo*; *dono-mado-mo*; *dono-omocha-mo*. In these constructions, the first element is a *wh* adjective meaning 'which', the middle element is the head noun ('book', 'bridge', 'professor', 'window' respectively), and the rightmost element is a focus particle whose addition transforms the whole NP in an indefinite expression with the meaning 'every N'. By so doing, the author probably wants to suggest that such phrases form a fixed pattern with a specific, non-compositional meaning. There is no mor-

529
Annali di Ca' Foscari. Serie orientale   e-ISSN 2385-3042
56, 2020, 501-562   ISSN 1125-3789

phological property of either *dono* or *mo* which may justify hyphenation, however, because other items can occur between *dono* and the head noun (*dono akai bōshi mo* 'which red hat FOC' 'all red hats') and between a host noun and *mo*.

Other cases of low inconsistency are Murasugi (2015), Sakamoto (2015), Shibatani (2016) and Kitagawa (2017), for a total of eleven works in this group.

6.2.2   Major Instances of Inconsistency

I assigned to this group those works in which the boundaries of a given particle are marked with blank spaces in certain examples, and with hyphens in others. Often this problem consists in the fact that, in a work where blanks are used, hyphenated examples pop up from nowhere in two or three places. In such cases there is no logic behind the use of different boundary symbols, and the reader cannot determine which of them reflects the true morphology of the particle in question or the author's actual theoretical stance.

In some cases, I could ascertain that the inconsistency originates from boundary marking having carried over from the original sources of the annotated examples. This might seem a trivial fault, as it could be interpreted as an author's candid admission of incompetence in Japanese grammar or as a sign of respect for the theoretical persuasion of the cited source(s). Boundary symbols and glosses, however, are not part of the data (as they are in this paper, which makes of the former its main epistemic object), to be reproduced in unaltered form. Rather, they are part of the analysis. When citing examples from a published source, authors may change them if they favor a different analysis or glossing system (as explicitly done in Iwasaki 2013, xix),[9] but if they cite them in original form, they implicitly adhere to the source's theoretical persuasion. In other words, since Line 2 boundary markers are theory-reflecting meaningful symbols, an author who uses conflicting boundary markers is expressing adherence to conflicting theories. The seven works belonging to this group are discussed in chronological order in the next paragraphs.

16. Occasional hyphenation (Heycock 2008, 69)

Motto	*anzen-ni*	*ki-o*	*tuketa-hoo-ga*	*ii-desu-yo*
more	safety-DAT	attention-ACC	pay-rather-GA	good-COP-PART

'You'd better pay more attention to your safety'.

---

[9] Remarkably, Iwasaki 2013 is the only work in my whole sample whose author informs the reader that the annotated examples taken from secondary sources have been re-glossed.

In Heycock (2008), the author marks particle boundaries with blank spaces. However, for no apparent reason, she occasionally hyphens the boundary of Nominative case particle *ga*. Moreover, in one example, here reproduced as example (16), she hyphenates all particle boundaries. In this example and in one instance of hyphenated *ga* (62) Heycock uses examples taken from Tomioka (2007, 899 and 882 respectively). Interestingly though, Heycock does not seem to appreciate the original format of her source's annotation, because she re-glosses Tomioka's examples by writing category labels in small capitals (they were not, in the original) and by changing the gloss of sentence final particles from 'particle' to 'PART' (see *yo* in 16). Despite so actively revising her source's annotation, Heycock does not extend her effort to particle boundary marking, which she leaves unaltered.

In Malchukov, Haspelmath, Comrie (2010), the authors only provide three annotated examples of Japanese. In two, they use blanks (2010, 12, 33), with no source quoted; in one (2010, 29), identical to the quoted source (Miyagawa, Tsujioka 2004, 16, 19), hyphens.

In Malchukov (2016), the author provides four Japanese language examples, all taken from the works of other scholars. I could ascertain that in all examples Malchukov adopts his sources' original annotation. In one example case boundaries are shown by blanks (Malchukov 2016, 395), in three by hyphens (396, 401).

17. Genitive marker *no* as affix-like (Iwata 2017, 246)

Hanako-no-te	*ni*	*hari*	*o*	*sasu*
Hanako-GEN hand	in	needle	ACC	stick

'Stick a needle in Hanako's hand'.

Iwata (2017) is a chapter in the HJLL volume on syntax. The author follows the general policy of the series and tokenizes case particles by way of blanks. However, he also sprinkles his article with a number of hyphens, and uses hyphens in all figures. Of special interest is the peculiar treatment of Genitive *no*, represented in example (17). Iwata hyphenates not only the (left) boundary of the Genitive marker, so distinguishing it from all other case particles, but also the boundary to the right, that with the modified head noun. In addition to *Hanako-no-te* in (17) other instances are *koyubi-no-tume*, glossed 'little finger-GEN nail' (Iwata 2017, 262) and *kabe-no-daibubun* 'wall-GEN most' (236, 242). This arbitrary secondary hyphenation is invalid because an adjective may occur between the two elements, as in *koyubi no nagai tume* 'little_finger GEN long nail'. Interestingly, but confusingly, the additional hyphens are not maintained in Line 2, where they are replaced by blanks. As I remarked on example (2), such a use of hyphens is not supported by morphologic theory. In the specific case of Iwata it also lacks in cohesion (because of the mismatching of symbols in Line 1 and 2), and is accompanied by an inconsist-

531

Annali di Ca' Foscari. Serie orientale   e-ISSN 2385-3042
56, 2020, 501-562   ISSN 1125-3789

ent use of hyphens throughout the article, so that I could infer no background logic to it.

Shibatani, Miyagawa, Noda 2017a is the specific introduction to the HJLL volume on syntax. As expected, the authors show case-particle boundaries with blank spaces. Surprisingly, however, hyphens also occur in five examples (2017a, 15, 18, 20-1), showing no consistent pattern. Of them, two (2017a, 20-1) are associated to a reference to chapters by other authors in the same volume (Kishimoto 2017; Saito 2017). Of those authors, though, only Saito (2017, 710) mistakenly uses a hyphen. The deviation from the norm is probably to be ascribed to Shibatani, Miyagawa and Noda only. I consider this instance as major not only because of the erratic nature of such a deviation, but also because of the importance of the chapter within the volume.

A puzzling case is Horie (2018), a chapter in *The Cambridge Handbook of Japanese Linguistics* edited by Hasegawa, bearing the important title "Linguistic Typology and the Japanese Language". The author states his purpose by saying that:

> Inspired by the recent findings in Linguistic Typology and its related disciplines, this chapter is intended to provide a structural and functional typological profile of the Japanese language. It first provides a general typological sketch of the Japanese language in terms of structural properties such as word order, agglutinating morphology, case marking, and the degree of differentiation between noun and verb. (Horie 2018, 65)

He then adds some remarks on morpheme boundaries:

> Morphologically, Japanese is categorized into an agglutinating language in which free and bound morphemes are tightly glued, as it were, in a strict co-occurrence order, while keeping the morpheme boundaries distinct, with each bound morpheme coding a different grammatical meaning, as shown in (1) [=18]. (Horie 2018, 66)

Next it comes the very first annotated example:

18. Case markers represented as bound morphemes (Horie 2018, 66)
    Chichi-ga   haiku-o   hajime-mashi-ta.
    father-NOM   poem-ACC   begin-POL-PST
    'My father started composing *haiku* poems'.

accompanied by the commentary:

> As shown in [18], the morpheme boundaries are clear and each bound morpheme encodes a distinct grammatical meaning. (Horie 2018, 67)

532

Annali di Ca' Foscari. Serie orientale | e-ISSN 2385-3042
56, 2020, 501-562 | ISSN 1125-3789

Despite his distinction between "tightly glued" free and bound morphemes, Horie leaves no doubt that Japanese case particles are of the very same nature as the POL auxiliary and the PST suffix. The hyphenation implies that all these forms are bound morphemes. But after briefly discussing the sentence above and two more under the same heading, Horie abruptly renounces hyphens and marks case-particle boundaries with blanks throughout the chapter. What might have happened here is that the author, after making his point, just backed off into atheoretical boundary making, to the confusion of the unwarned reader.

6.3 The Case of Shibatani

Among the authors in my sample, Masayoshi Shibatani stands out because during the time span under analysis he has marked case-particle boundaries by using all three attested symbols. Such a variation can hardly be missed, for a contribution of his is included in almost all the collections published in the period.

His case is relevant for several reasons. First of all, Shibatani is probably the most authoritative Japanese linguist on the scene. As a typologist, he has to deal with languages in which case marking is expressed in any possible grammatical way, and with languages in which clisis is an important phenomenon. Consequently, Shibatani is well aware of the necessity of distinguishing affixes from clitics from independent words in annotation, and is accustomed to doing so in his papers. Needless to say, he is also frequently addressing syntactic issues related to Japanese language (even if, for as much as I know, he has never treated the morphological nature of Japanese case marker in a specific paper). As a consequence, the expert reader expects that Shibatani will deal with the marking of Japanese case-particle boundaries with the same wisdom he applies to other languages. The changes in his use of case marker boundary symbols in annotating Japanese, then, cannot be simply ascribed to carelessness or to a shaky knowledge of glossing conventions. Rather, they reflect a true evolution in Shibatani's theoretical opinions. It is under such an assumption that in this subsection I will comment on his change in Japanese particle boundary marking as revealed in my sample of works.

I parsed the fourteen works of Shibatani which are relevant for the present paper (they include multi-author articles) plus two slightly older works, to see whether a diachronic pattern emerges. I could detect the following phases.

Phase 1   From 2003 to 2009, Shibatani marked Japanese case-particle boundaries alternately with blanks (2004, 2006) or hyphens (2003, 2007a, 2007b, 2009a). Especially in the most recent years, a slight preference for the latter symbols may be observed. One of such

cases is (8). I compared Shibatani's use of symbols with the instances of annotated Japanese by other contributors to the same volumes (not included in the sample of 184 works under analysis in this paper), and I could not notice any pattern of compliance with (unstated) general norms: all contributors showed Japanese case-particle boundaries as they liked. In two works, Shibatani's use of hyphens can be interestingly contrasted with that of equals signs for marking clitic boundaries, in both Japanese (2007a, the clitic is nominalizer *koto*) and other languages (Chukchi) (2009a).

Phase 2 From 2009 to 2014,[10] Shibatani mostly used equals signs (2009b, 2013; Shibatani et al. 2014). He also continued to so mark clitic boundaries in Japanese (the nominalization marker *no* in example 9) and other languages (2009b).

Phase 3 Since 2015, Shibatani has definitely abandoned equals signs in favor of blank spaces (Shibatani, Kageyama 2015; Shibatani, Miyagawa, Noda 2017a; Shibatani 2018a, which are all chapters in HJLL volumes, where blanks are indeed required as a general policy). Remarkably Shibatani has retained the use of equals signs for distinguishing the boundaries of Japanese true clitics (2016, 2017, 2018a), as shown in (7), and of clitics in other languages (2018b).

I interpret the data as follows. In a first phase of low general interest (despite the seminal works of Hattori [1949] 1960, 1950; Vance 1993, discussed above), Shibatani complied with the conventional reductionist use of two marking symbols only, alternating between blanks and hyphens. This strategy was probably grounded in the perception that, morphologically, Japanese case particles are somehow intermediate between free word-forms and affixes (the two poles of Table 1). In general, Shibatani preferred hyphens, but when the type of publication required some judgment on the morphology of Japanese case markers, he expressly identified them as postpositions, associated them to prepositions (2006, 104), and consistently marked their boundaries with blanks. At a certain point, however, he saw the need to distinguish case markers from both affixes and independent words. He so begun to annotate them as clitics, the third commonly acknowledged morphemic class, intermediate between words and affixes (Table 1). Such a strategy may reveal either a persuasion that Japanese case markers are clitic (strong stance), or a mere will to distinguish them as neither free word-forms of affixes (weak stance). After that, Shibatani adhered to Kageyama's claim (2016) that Japanese case markers form the unique class of *fuzokugo*, and felt the need to distinguish between *fuzokugo* and true clitics in annotation. To do so, he maintained the conventional equals signs for marking

---

10 In his CV, Shibatani lists no publication for the years 2010-12 and no publication on Japanese for 2015, https://linguistics.rice.edu/matt-shibatani.

clitic boundaries, whereas reverted to blanks for *fuzokugo*. Even if he occasionally supported this choice with statements about the adpositional nature of Japanese particles (Shibatani 2009c, 754), I do not believe that Shibatani's return to blanks reflects a reversal in his theoretical persuasion, given the previous complex evolution of his notational strategies. I rather believe that Shibatani, again following Kageyama (2016), judged the introduction of a new symbol to be unnecessary, and just adopted the boundary marking symbol already in use for the class next to *fuzokugo* in Table 1, that of independent words.

In sum, the change in Shibatani's particle-boundary marking reflects an active, well pondered evolution in his beliefs about the status of Japanese case particles, and a consequent search for the boundary marker whereby best to represent their unique nature. One might not agree with Shibatani's theoretical stance or choice of symbols, but his case, as reconstructed above, is an example of sound scientific practice.

7 Poor Segmentation and Sloppy Annotation

Up to this point, I have dealt with contradictions in the use of boundary symbols at text level. At this level, inconsistent case-particle boundary marking consists in the fact that certain cohesively annotated examples convey information conflicting with that of other well-formed examples. In collections, this type of inconsistency confounds the reader, who becomes uncertain about the morphology of Japanese case markers. In single works the problem may be considered to be one of coherence, as bad tokenization hinders the comprehension of the overall message an author wants to transmit.

As a matter of fact, however, the main problem related to boundary marking in the grammatical annotation of Japanese does not concern the semantic choice of symbols, but consists in bad matching between the symbols used in Lines 1 and 2. I call this phenomenon 'bad segmentation', as it involves all boundary symbols, not only those showing word breaks. In the metalanguage of annotation, the matching of symbols between the two Lines can be considered a form of anaphora. Bad segmentation is therefore a problem of cohesion, for it comes from bad grammar, and has the power to disrupt the epistemic value or function of individual instances of annotation.

As seen above, the conventional norms prescribe that words are to be vertically aligned in Lines 1 and 2, and that blanks, hyphens or equals signs must match in both. Moreover, words and attached affixes must be left-flushed. In my sample, the deviations from this set of prescriptions include the following:

535

Annali di Ca' Foscari. Serie orientale | e-ISSN 2385-3042
56, 2020, 501-562 | ISSN 1125-3789

• Using a Boundary Symbol in Line 2 with no Corresponding Break
in Line 1

19. Blanks in Line 2 vs. no breaks in Line 1 (Tamaoka 2015, 615)
    Siryoo ga   keikakutekini   zyunbisareta.
    reference ACC   deliberately   prepare PASS PST
    'The reference was deliberately prepared'.

20. Blanks in Line 2 vs. no breaks in Line 1 (Shirai 2015, 222)
    Hanasiteimasu.
    speak PROG POL NONPST
    'He is speaking'.

In (19-20), Line 2 blanks have no counterpart in Line 1. When no seg-
mentation is necessary in Line 1 but the meanings of Line 1 morphemes
are to be shown as part of the glosses, the relevant Line 2 bounda-
ries (only) should be marked with colons or periods, not with blanks
(Lehmann 2004b, 1852; Schultze-Berndt 2006, 239; Bickel, Comrie,
Haspelmath 2008, 3-5). Shirai (2015) and Tamaoka (2015) are chap-
ters in the same HJLL volume. The latter has only three groups of an-
notated Japanese examples; here Lines 1 blanks are used in two cas-
es, hyphens in one (see discussion of example 4 in § 4). With reference
to example (20), Shirai mentions "the aspect marker –*te i-*(*ru*)" (2015,
222), using a segmentation which he deems relevant but not enough
to be represented in the annotated examples.

Occasionally, it also happens that hyphens, rather than blanks, are
present in Line 2 but not in Line 1. One case is example (5), where
the Line 1 word *maikai* is glossed 'every-time'. In this case, either the
Line 1 word had to be segmented *mai-kai*, or the Line 2 gloss had to
be rendered as 'every_time'. This would comply with the convention,
which states that when a Line 2 translational equivalent consists in
more than one graphic word, the elements must be separated with
underscores (Bickel, Comrie, Haspelmath 2008, 4).

A variation is the use of a hyphen in Line 1 in correspondence with
a Line 2 blank space. This can be observed, among other examples,
in (17), where Line 1 *Hanako-no-te* is mismatched by Line 2 'Hana-
ko-GEN hand'; and in (15), where Line 1 *dono-hon-mo* corresponds
to 'every book' in Line 2. A similar variation is the use of an equals
sign in Line 1 and of a blank space in Line 2, as in (14) (= 23 here be-
low), where *gakkoo=e* is mismatched by 'school ALL'.

21. Mismatching of boundary symbols in Lines 1-2 (Ono 2013, 423)
    [*Mary no kinoo-no kami*]
    Mary GEN yesterday GEN hair
    [Untranslated in the original source, but meaning Mary's hair, yesterday, …']

Another type of interlinear mismatching is the misuse of a hyphen in Line 1 and of a blank space in Line 2 is shown in (21). Other problems in (21) are the lack of alignment (discussed below) and the loss of cohesion caused by the conflicting treatments of *no* in Lines 1-2.

• Using Morpheme- and Word-Boundary Symbols in Line 1 only

This is a reversal of the mismatching of the previous paragraph. In (2), Line 1 *makkuro-ni* in is rendered as 'black' in Line 2; in (5) the hyphenated *ni-yotte* in Line 1 corresponds to the gloss 'by' in Line 2.

• Using Line 2 Unique Morpheme Boundary Symbols in Both Lines or in Line 1 only

This is perhaps the most confusing deviation from the norm. It mostly involves a misuse of period signs:

22. Line 1 tildes and periods in partial substitution of Line 2 hyphens (Narrog 2017, 345)
    Yamada~seNsei   *wa*   *ippai*   *noN.de*   *ki.ta/*mie.ta.*
    Yamada-Prof. to   TOP   a.drink   drink-GER   come-PST/come.HON-PST
    'Professor Yamada had a drink (and came)'.

23. Line 1 periods in substitution of Line 2 hyphens (Narrog 2017, 345) (= 14)
    Taroo   *wa*   **gakkoo=e**   *arui.te*   *it.ta/ki.ta.*
    Taroto   TOP   school ALL   walk-GER   go-PST/come-PST
    'Taro walked (walk went/came) to school'.

As remarked above, the norms of grammatical annotation prescribe the use of colons or periods for segmenting the Line 2 morphs whose boundaries are not shown in Line 1. This use of periods might be at the back of 'a.drink' in (22), since the form *ippai* is actually analyzable into *ip*, a morph standing for the numeral '1', and *pai*, an allomorph of the classifier for cups of drinks *hai*. When hyphens, blanks or equals signs are present, however, they must match in both lines. This does not happen with the clitic boundary symbol in (23). More important, the convention does not contemplate replacing Line 2 hyphens with periods in Line 1, as it happens in (22-23) and throughout Narrog 2017 and 2018. It should be added that Narrog actually takes (22-23) from Shibatani (2007b, 116, 119 respectively), where hyphens are used to segment all boundaries in both lines.

The peculiar use of period signs in (22-23) might be explained as follows.

In annotating texts for phonetical and phonologic purposes, periods are used to mark syllable boundaries (see for instance the annotation in Ito, Kubozono, Mester 2017). Now, it is a property of Japanese that morpheme breaks always fall at syllable boundaries. In

other words, syllable and morpheme boundaries always match. What might be inferred from the use of symbols in (22-23), then, is that the author intends to mark suffix boundaries with hyphens in the grammatical annotation tier of Line 2, while marking the corresponding syllable breaks by means of periods in the phonetic/phonemic tier of Line 1. The two Lines thus represent not merely two different kinds of data, but two types of annotation altogether.

This double strategy fails for three reasons. First, as I remarked, a non-mirroring use of symbols between Lines 1 and 2 is not contemplated in grammatical annotation. Second, the strategy is not justified by the general purpose of the two articles, which bear the title "The Morphosyntax of Grammaticalization in Japanese" (Narrog 2017) and "Modality" (Narrog 2018) and deal with grammar, not prosody or phonology. Now, syllable boundaries and syllable quality are not relevant *per se* in Japanese grammar. In the grammatical annotation of Japanese there is simply no need to mark syllable boundaries and to use the symbol <ɴ> for representing a non-contrasting allophone of /n/. Third, such a double strategy is applied inconsistently. In (23) a conventional morpheme-break symbol (the equals sign) appears in Line 1, where, according to this very strategy, it does not belong. In (22), prosodic syllable-breaks (period signs) occur in Line 2, where only *grammatical* annotation symbols should be used. But one also wonders why be so specific on the sound properties of /n/ while not considering other phonetic matters (like the devoicing of /i/, for instance).

Line 1 periods are not the only notational device used aberrantly in Narrog (2017). The confused segmentation of *mieta* in (22) implies that this form is composed of a root *mie* meaning 'come' and a portmanteau morph *ta* bearing both a connotative meaning ʜᴏɴ and a grammatical meaning ᴘsᴛ, or perhaps being further analyzable in two smaller components each with one of such meanings. This is not the case either: the ᴘsᴛ suffix is *ta*, as (23) helps clarify, while *mie* is the root of *mieru* 'being visible', a verb which replaces verb *kuru* (occurring as *ki.ta* in 22-23) 'come' in honorific speech. The source (Shibatani 2007b, 116) correctly segments *mie-ta* and glosses 'come.HON-PAST'.

As for the tilde <~>, it occurs seven times in Narrog (2017), always in Line 1, where it corresponds to a Line 2 hyphen. A tilde is used for showing the breaks between personal names and the title *sensei* 'Professor' (as in 22); between two roots in some (but not in all) compounds, between a verbal noun and the verbalizer dummy *suru* 'do', and between a converbal form and the auxiliary *kuru* 'come'. The tilde is not a conventional symbol in grammatical annotation. I know of no morphological theory supporting the distribution of the tilde observable in Narrog (2017). Concerning (21-22), I will abstain from commenting on the non-idiomatic translation in the free-translation tier of Line 3 and on proof-reading errors, causing the repetitions of meaningless syllables (a recurrent problem throughout Narrog 2017).

Another example of a confusing misuse of period signs is the following (extracted from a very long annotated text):

24. Misuse of period signs in Lines 1 and 2 (extracted from Hasegawa 2018a, 12)
    dete.ko-nai
    come.out-not
    [Unrendered in the free translation, but meaning 'do(es) not come out']

Hasegawa (2018a) is the introduction to the edited volume Hasegawa (2018b). A reader with knowledge of the norms of grammatical annotation who has accepted Hasegawa's deviant use of periods would probably be misled to believe that *dete* is a verbal root meaning 'come'; *ko* is an element meaning 'out', in some indistinct morphemic relation with the previous root; and *nai* is an affix meaning 'not'. Instead, *deteko* is a form derived from the grammaticalized verb *detekuru*, meaning 'come out', and segmentable into *de-te* (exit-cvʙ) and *ko* (come). Remarkably, the author's glossing has the effect of inverting the actual meaning of the two components, for it is *dete* that contributes the meaning 'out' to the construction. As for *nai*, it is the present, unmarked form of an adjective meaning 'inexistent', which in Japanese is agglutinated to verbs to express negation. At the level of the morphological analysis required in Hasegawa's chapter, such detailed segmentation is indeed superfluous. This is a typical case in which encoding information in excess opens the door to errors and misinformation. This problem could be avoided by simply segmenting *deteko-nai*, and glossing it <come_out-ɴᴇɢ> or <come_out-ɴᴇɢ:ᴘʀs>. An occasional misuse of periods also occurs in Hasegawa (2016).

• Bad Word Alignment

Lastly, there are cases in which words are not aligned in Line 1 and 2, or in which only partial alignment is implemented.

25. No vertical alignment of words (Ono 2013, 422)
    a. *eki de no kyaku no otosimono*
       station at GEN passenger GEN lost-thing
       'A lost article of a passenger at the station'.
    b. *kyaku no eki de no otosimono*
       station at GEN passenger GEN lost-thing
       'A lost article of a passenger at the station'.

When no alignment is present, as in (21) and (25a-b), the correspondence of Line 1 graphic words with Line 2 glosses can only be retrieved by *counting* words and finding the matching positions in the string. This does not facilitate the comprehension of the examples. When I

539
Annali di Ca' Foscari. Serie orientale
56, 2020, 501-562

e-ISSN 2385-3042
ISSN 1125-3789

first selected (25) for discussion, I only intended to use (25a) for I had not noticed that the two gloss lines in (25a) and (25b) are identical. The correct Line 2 sequence should be <passenger GEN station at GEN lost_thing>. This is obviously a copy-and-paste error; it went unnoticed by author, copy editors and at first by myself due to the very crowding of words in Lines 1 and 2 that the lack of alignment originates.

26. Partial vertical alignment with flush left (Shirai 2015, 228)

butyoo wa	*Ruusii-san ni*	*kopii o*	*sasemasita*
dept. chief TOP	Lucy Miss OBL	photocopy ACC	do CAUS POL PST

'The department chief made Lucy make photocopies'.

In (26), constituents are aligned vertically, but the strings of elements are left-flushed even if word-internal morpheme boundaries are marked with blank spaces (see also 4). This fact, along with the normal use of capital letters for category labels, suggests that the flushed elements are not suffixes but neither are they ordinary free word-forms. Even if making no use of Line 2 conventional symbols like colons, periods or underscores, this system does accurately convey an analysis and is therefore informative and useful to the reader.

7.1 Sloppy Annotation

The inconsistent marking of case-particle boundaries and poor segmentation are part of a vast array of deviations from the conventional norms of grammatical annotation which I call 'sloppy annotation'. Sloppy annotation involves a misuse of all notational devices, not only of boundary symbols, and it is not limited to the few works I dealt with in the previous sections. Rather, it is a widespread ailment affecting to various degrees many works whose boundary marking I judged to be otherwise consistent. I cannot go into a detailed analysis of it, because to produce a true typology of sloppy annotation would require a whole treatise of its own, especially for as much as category labels are concerned. So I will only produce a sketchy profile of it. Listed below are instances of sloppy annotation:

- Occasional errors due to inadequate proofreading, like the use of ACC instead of NOM in (19); the copy-and-paste error of (25); and others which I did not specifically pointed out.
- The insertion of underlying forms in the morpheme break tier \mo of Line 2 rather than in a dedicated line. As it happens for *nur-ta* in (2), the underlying forms of a language do not necessarily bear any resemblance with the sounds of actual speech or any correspondence with the sequence of characters of the orthography. Being objects of a different order altogether, under-

lying forms should never be represented in the same line(s) as the phonetic or orthographic transcription. If they are required, they ought to be shown in a distinct line; one more reason for expressing \mo separately. This is a general problem in annotation, stemming from the trilinear format itself. In the grammatical annotation of Japanese, the problem is particularly annoying because it is unclear from the beginning whether the transcription in Line 1 is orthographic or phonetic or phonemic.

- The abuse of representing Japanese case particles in Line 2 glosses by way of ambiguous translational equivalents (English prepositions) instead of category labels, as in the following:

27. No category labelling (Saito, Lin, Murasugi 2014, 1)
    isi   -de   no   koogeki
    stone-with *no* attack
    'An attack with stones'.

In (27) *de* is glossed 'with'; similarly, in (5) *ni* and *niyotte* are glossed 'by'. See also (8), where *ni* is glossed 'to'. This habit is strongly criticized by Lehmann (1982, 205; 2004b, 1840-1).

• The use of a Japanese morph as a gloss. Example (27) also shows a lack of true glossing, in that Genitive particle *no* is represented in the IMG of Line 2 by an italicized version of itself. Similarly, in Heycock (2008), particles *wa* and *ga* are not glossed but simply reproduced in small capitals as 'ᴡᴀ' and 'ɢᴀ' (the latter is shown in 16). The rendering of the adverbializer *ku* and *ni* in (1), respectively glossed 'KU' and 'NI', follows the same principle. Such a renouncement of glossing suggests that the properties of those morphemes are so unique as to defy grammatical description. Lehmann considers it an inadmissible "bankruptcy declaration of grammatical analysis" (Lehmann 1982, 205).

• Poor Labelling of Grammatical Categories, as in the Following:

28. Unclear labelling (Tomioka 2007, 891)
    Hayashi-san(-wa)   atama ii yo.
    Hayashi-honor-(top)   smart-part
    'Mr./Ms. Hayahi is smart'.

In (28) category labels are bad not because wrong, but because graphically indistinguishable from lexical words. Such cases are by no means rare. Conventionally, category labels and translational equivalents should be distinguished in Line 2 by writing the former in upper case letters (usually small capitals) (Lehmann 2004b, 1840, 1856; Schultze-Berndt 2006, 239; Bickel, Comrie, Haspelmath 2008, 3) "to facilitate the reader's understanding" (Lehmann 1982, 205). In (28), the lack of this helpful graphic distinction generates the most

unfortunate confusion as to the meaning of elements like *san*, *wa*, *atama* and *yo*. Are they lexical words, meaning, respectively, 'honor', 'top', 'smart' and 'part', as shown in the gloss tier? They are not: *san* is a honorific suffix, *wa* is the topic marker, and *atama* means 'head'; while it is the sentence *atama* (*ga*) *ii* 'head (NOM) good' as a whole that bears the sense of 'smart'. As for *yo*, it is a sentence-final assertive particle. Elsewhere Tomioka uses the full word "particle" for labelling not only *yo* but also sentence-final confirmative particle *nee* and the morph *syō*, which is actually part of the dummy portmanteaux *desyō* (2007, 894, 898-9). In his encyclopedic codification of grammatical annotation, Lehmann remarks that different morphemes are never to be rendered by means of one and the same label (2004b, 1839). Moreover, particles, of any kind, must not be glossed as such, since the gloss 'particle' denotes, at best, a word class, whereas a label should give information on the meaning or grammatical function of the relevant item (Lehmann 2004b, 1838-44). Heycock (2008) (briefly discussed in § 6.2.2, see 16) also takes some examples from Tomioka (2007) and glosses sentence-final particles in the same wrong fashion. Additionally, Heycock shows inconsistent glossing, as elsewhere in her chapter *yo* is glossed 'EMPH', and *desyoo* is left unsegmented and glossed 'be-would' (e.g. Heycock 2008, 73).

29. Wrong category labelling (Shirai 2015, 225) (= 4)

   Ken   wa   sin de iru.
   Ken   TOP   die RES NONPST
   'Ken is dead'.

In (29) the labelling of grammatical functions is bad because of both the misuse of blanks and the choice of wrong category labels. Faced with (29), the reader who wants to extract the meaning or functions of items *de* and *iru* from the glosses cannot avoid assigning *de* the grammatical meaning of 'resultative', and *iru* that of 'non-past'. Instead, *de* is an allomorph of *te*, a converbal affix, while *iru* is the (non-past) dictionary form of an existential verb 'be'. As a matter of fact, the predicate does have the resultative meaning shown. But this meaning occurs when the whole *-te iru* ending is assigned to punctual verbs (or, in constructional terms, when a punctual verb is inserted into the *-te iru* construction). The label 'RES' has nothing to do with an inherent function or meaning of the individual morpheme *te*. Likewise, when elsewhere in the same article Shirai names the same two elements PROG and NONPST respectively (see 20), the progressive meaning does not reside in the morpheme *te* but rather originates from the association of *-te iru* with durative verbs. There is no doubt that the author knows all this. The problem is that not only did he not display such knowledge but also ended up encoding data which, albeit well-formed and meaningful, are false, and thus misinformed the reader.

Miyagawa (2017) also analyzes the *-te iru* construction in a misleading way. In (15) he glosses *yonde-iru* as 'read-ing', so conveying that *yonde* is a verbal root meaning 'read' and *iru* a suffix for deriving the gerundive of converbal form. At Miyagawa's level of morphologic detail, the correct segmentation and glossing should have been *yonde-iru* 'read:cᴠʙ-be' or 'reading-be'. Miyagawa's analysis is particularly confusing because he uses (15) for showing that *yonde iru* can express both progressive and resultative meanings, a fact that the gloss 'read-ing' does not convey at all.

7.2 Weak Editing

All the above errors and misrepresentations are not an exclusive responsibility of authors. They float into the final, printed version of many works because of a lack of power of editors. An exemplary case of weak editing is the HJLL series. As I remarked in § 6, in the general introduction to the series, identical in each volume (Shibatani, Kageyama 2015), a clarification is added saying that Japanese case markers are not free word-forms but in the HJLL volumes their boundary will be marked as if they were. HJLL authors do make an effort to comply with such a general prescription. Nevertheless, in eight of the ninety-two relevant chapters hyphens also pop up. Such inconsistency had to be detected and prevented from reaching the final volume by the editors, but it was not. As for the scientific editors, to sieve out inconsistencies is not properly their job; and besides one can hardly image a Shibatani or a Kageyama reading his colleagues' papers and pointing out to them their notational errors. As for the copyeditor (credited on page vi of each volume) and his staff, I ascribe their failure in enforcing compliance with the conventions to their little authority over the scientific editors and the illustrious contributors. An 8% deviance rate has little weight and might be considered unavoidable. But the HJLL volumes are expensive, each costing between €299 and €320 (US$344-70) in pdf or hardcover. For that price I believe a purchaser is legitimated to expect no proofreading error of such great misleading effect.[11]

---

[11]   Since it started, the HJLL series has acquired more and more consistency in the distribution of boundaries symbols. In the latest volume, on the contrastive linguistics of Japanese (Pardeshi, Kageyama 2018), all contributors use blanks, with no exception or incoherence. Probably, editors' control was stricter in this handbook because the volume is intended for linguists unfamiliar with Japanese, who rely heavily on annotated examples and expect full compliance with the rules.

543

Annali di Ca' Foscari. Serie orientale
56, 2020, 501-562

e-ISSN   2385-3042
ISSN   1125-3789

8 Linguistic Annotation and Information

To better understand the danger of misinformation originating in sloppy annotation, it is useful to stress the role of linguistic annotation as an encoding system. The author translates a source language text into a metalanguage following the specific lexicon and syntactic rules provided by a complex system of notational conventions. Boundary symbols properly belong to the lexicon, as each encodes a different meaning. The message, thus formatted, is made public to a community of expert linguists, all in possess of the decoding keys (lexicon and grammar). Each receiver applies those keys, decodes the message and so acquires information. In order to be epistemically effective and epistemologically valuable, this act of communication presupposes trust in the informer's observance of the Gricean maxim of Quality (Grice 1975, 46). This trust is of three kinds. The first is trust in the informer's honesty, thereby ruling out the possibility that the informer is conveying false information on purpose (i.e, is *disinforming* the reader, Fetzer 2004; Floridi 2011, 260). The second is trust in the author's expertise as an encoder, i.e. in the sender's knowledge of the metalanguage and competence to use it.[12] In short, the receiver needs to believe that the informer is not unintentionally misinforming her due to weak or wrong encoding. The third form of trust concerns the actual meaning of the message, and consists in the assumption that it is based on valid linguistic analysis and encodes information which can become knowledge.

Within this framework, every individual token of the several types of rule violation listed above as instances of sloppy annotation may only have one of two outputs. In theory, an occurrence of sloppy annotation might be uninformative. This happens when the use of the notational devices has so little cohesion as to make no sense. This is a very rare output, though, as it presupposes the use of devices not contemplated by the rules, or a totally messed up, jumbled use of conventional symbols. Sloppy annotation, instead, is technically still annotation. Each token of it consists in the misuse of one symbol only, which yet carries a conventional meaning and is decoded consequently. This is why sloppy annotation almost invariably conveys misinformation. There is no point in remarking the deleterious pow-

---

12 I do not consider rumour to be a factor in determining a sloppy-annotation output. For example, it might seem reasonable to ascribe to rumour the insertion of the syllable <to> in Line 2 in (22-23), and the copying of Line 2 from (25a) to (25b). However, I see the role of copyeditors as a filter which activates after the first, main encoding of information, with the function of cleansing the code of wrong or misplaced symbols. In such a role, copy editors are members of the input staff, and bear part of the responsibility for the quality of the code. In this sense, the first type of trust under discussion extends as deference in editors.

544

Annali di Ca' Foscari. Serie orientale | e-ISSN 2385-3042
56, 2020, 501-562 | ISSN 1125-3789

er of the incorrect data encoded in all the examples discussed above. Misinformation is indeed carried by the repetition of (25a) Line 2 in (25b), the mislabelling of elements in the gloss of the predicates in (22), (24) and (29); the typographic error ACC in (19), and so on. In most cases, the misinformation consists in the suggestion that Japanese case particle belongs to more than one morphological class, while in fact they do not. Allowing the atheoretical boundary marking of the Latin orthography to take over the theory-informed use of boundary symbols in Line 2 is also cause of misinformation. In this specific case, the product of noncompliance with a rule (the showing of case-particle boundaries as meaningless blank spaces) at source corresponds by chance to a meaningful signal according to that very same rule ("blanks mark the boundaries of independent words"). As an effect, the 'annotated' text only resembles a genuine instance of annotation, but is not. The receiver cannot know that, though, and will take the output as truly annotated text. Of course, if each singular instance of annotation is taken out of context, one cannot really determine in which role blanks are used. It may well be, but is equally undecidable, that the encoder uses blanks non-theoretically but that s/he collaterally believes that Japanese case markers are independent words. There would so occur a casual match of the encoded meaning with the encoder's beliefs. I may call this an instance of 'lucky information'. The epistemic problems caused by sloppy annotation are nearly endless.

9 Sloppy Annotation and Conventionality

Sloppiness thus endangers the utility of the whole system of linguistic annotation. As any other conventional system, linguistic annotation is both inspired by purely utilitarian ("I use this shared metalanguage otherwise no one would understand me") and ethical ("I talk the same language as others so that they'll make less effort to decipher me") reasons, and is dependent on the compliance of all actors (Marmor 2009, 52-4). In general, compliance does not necessarily require of an actor to be conscious of the fact of being following a norm, of the conventional nature of it, or of its reasons (Marmor 2009, 5-7). Linguistics and linguistic annotation, however, are very specialized fields, populated by a community of learned scientists well used to investigate the relationship between particular and universal, and to express themselves by means of metalanguages. All members of this community are legitimately expected to know better, and to be conscious of the nature, importance and ethical value of the technical conventions that regulate their world. As I pointed out in § 2.1, linguistic annotation is a coordination convention, which solves a coordination problem related to data sharing by providing a universal

545

Annali di Ca' Foscari. Serie orientale | e-ISSN 2385-3042
56, 2020, 501-562 | ISSN 1125-3789

packaging format based on the standardized analytical tools of descriptive grammar. The rule-breaking represented by sloppy annotation, therefore, diminishes the communicative value of the annotated text and puts the unaware reader at a disadvantage. So, disregarding the role of the convention is a denial of the value of the metalanguage itself and a betrayal of the very ethos that sustains the conventional system as such.

Moreover, linguistic annotation can also be considered a constitutive social convention, in the sense that it is a practice which does not exist unless an author actually engages in it, and that its rules can be defined only with reference to the activity they regulate (Marmor 2009, 33). As Marmor puts it, every set of constitutive conventions has, as a 'prologue', the voluntary engagement in the relevant activity. This is an 'if clause' which may be expressed as "if you want to play, these are the rules you ought to follow" (Marmor 2007, 134-5, 161). The same point was made by Grice in relation to the Cooperative Principle (1975, 49). Differently put, this means that those linguists who want to annotate can only do so by engaging in annotation, but if they do that, they must do it by the rules, otherwise they are not annotating at all. From a formalist point of view, then, the authors that slip into sloppy annotation are breaking some constitutive rule, no matter how marginal it may seem, and so effectively put themselves out of the game – both the game of linguistics and that of annotation. Sloppiness invalidates linguistic annotation, and changes an annotated text into something else.

According to the arbitrariness condition, however, conventional norms have as a definitional property the possibility of being substituted by alternate rules ultimately achieving roughly the same purpose (Marmor 2009, 43). Does sloppy annotation represent a collection of alternate conventions, then? At the present level of analysis, my answer for sloppy annotation in general is negative, for it is not uniform or widespread enough. My answer is negative for bad morpheme-boundary marking too, even though a significant number of linguists seem to be following it. For one thing, neutralizing the morphological distinction between clitics, *fuzokugo* and free word-forms, as the confusion of hyphens and blanks does, can hardly be considered epistemically equivalent to maintaining that distinction. Rather, it causes a significant loss in the encoding power of the marking symbols (Marmor 2009, 9). It does not achieve the same purpose of the current rule; and therefore it does not represent a viable substitute for it – certainly not at the present state of equilibrium of the convention as a whole. Moreover, alternate norms are only effective if they are recognized as such and universally adopted by the relevant community. Bad boundary marking could never be so recognized and thereby changed into a simpler marking convention, because it would introduce a language-specific bullet into a larger rule

system expressly designed to be universal. One can hardly imagine a rule like "For Japanese case particles only, boundaries can be indifferently shown by means of blanks, hyphens or equals signs" in an encyclopedic repository like the Leipzig Glossing Rules. No matter how conventional bad boundary marking may be, then, it represents a breaking of the larger set of norms that regulate linguistic annotation – a more complex system which antedates it and already includes perfectly viable prescriptions on boundary marking.

10 Coping With Sloppy Annotation

The question now arises as to how poor boundary marking and sloppy annotation in general may survive in the community of linguists. A first important point is that, like the rules of certain games, the norms of linguistic annotation work (up to a point) even if they are followed partially, or imprecisely. The home-rules of certain games work so well (i.e., satisfice), that the players may never even realize something is wrong. They believe they are still playing the same game, without feeling a need to change their ways of playing it or to learn the listed rules better.

The homologies with natural languages are evident. Natural languages work effectively in coding communication even when used incompetently, with opaque expressions or expressions that are unacceptable to a large part of their speaking communities. Beyond a certain threshold, of course, the source code is too corrupted and communication becomes impossible. When this happens, one has to consider if there is a community of speakers which still understand the source. If there is, then the source has become another language. The notational problem of imperfect encoding is the same in natural languages as in the artificial metalanguage of linguistic annotation. Until an abstract point is reached when cohesion is lost and sloppy annotation becomes unintelligible, it still displays somehow regular associations of meanings and signs. Therefore it still works, albeit imperfectly, so that, even if noticed and informally disapproved, is not subject to criticism and sanction. In other words, although under rigid formalism sloppy annotation is not annotation at all, in practice its home-rules do not change it enough to invalidate its function or constitute a different language altogether.

The disruption provoked by sloppy annotation is kept within acceptable limits because of the important role played by another actor, the informee or decoder, that is, the reader. The reader is the site of two distinct damage-control mechanisms. The first one is related to relevance.

As I noticed in § 4, there is no way of *not* representing Line 2 word boundaries theoretically, even if doing so is irrelevant to the

547
Annali di Ca' Foscari. Serie orientale   e-ISSN 2385-3042
56, 2020, 501-562   ISSN 1125-3789

discourse. Therefore, in boundary marking the Gricean maxim of relation ("be relevant") and, consequently, the maxims of quantity ("provide enough information, but not too much") (Grice 1975, 45-6), cannot be observed. This is actually true for grammatical annotation in general. With few exceptions (works on phonetics, phonology, discourse analysis), no matter how frugal it is, grammatical annotation is structured in such a way as to require a representation of grammatical contents more finely grained than strictly needed. Moreover, as discussed by Lehmann, a high degree of detail is advisable because an author cannot foresee all the uses his readers may make of the annotated examples (Lehmann 1982, 202; 2004a, 194; 2004b, 1839). Now, as Grice himself put it, the excess of information so generated may cause confusion:

> The second maxim ["Do not make your contribution more informative than is required"] is disputable; it might be said that to be overinformative is not a transgression of the C[ooperative] P[rinciple] but merely a waste of time. However, it might be answered that such overinformativeness may be confusing in that it is liable to raise side issues; and there may also be an indirect effect, in that the hearers may be misled as a result of thinking that there is some particular point in the provision of the excess of information. However this may be, there is perhaps a different reason for doubt about the admission of this second maxim, namely, that its effect will be secured by a later maxim, which concern relevance. (Grice 1975, 46)

The readers of the literature of linguistics know than the threshold (the "guarantee of relevance": Sperber, Wilson 1995, 49-50) is set low and that the annotated examples are prone to contain information in excess. In extreme cases, when the reader is uninterested (perhaps because already expert enough), the examples go unread, or are only summarily parsed, to the effect that sloppy annotation is filtered out and neutralized. In this way the readers avoid being misinformed simply because they do not act in order to be informed. In other cases, in a given work, sloppy annotation is concentrated in areas of the annotated examples which are not relevant to the general topic under discussion. For example, the morphologic properties of Japanese case particles are specifically addressed in only one work in my sample.[13] This means that the readers use all the works in the

---

[13]   The work in question is Kageyama 2016, discussed in §5. In Kageyama's case the whole article is on the morphological status of Japanese postnominal particles, so that the information provided by the main text overrides the conventional meaning of the boundary symbols used in annotation, effectively depriving them of semantic function. The same occurs in Horie 2018, discussed in § 6.2.2.

548

Annali di Ca' Foscari. Serie orientale   e-ISSN 2385-3042
56, 2020, 501-562 | ISSN 1125-3789

sample but one for learning something different from the morphology of Japanese case particles. To them, how case-particle boundaries are marked is irrelevant and they will rather pay attention to other features of the annotated texts. To make a more specific example, in Narrog (2017, 345-6) the discussion of (23) is focused on the allative phrase *gakkoo=e*, written in bold. Readers who follow Narrog's line of reasoning will hardly notice the bad segmentation in the rest of the two lines. Again, misinformation does not occur because the agent does not process the misinforming material.

The second mechanism is an effortful inferential procedure that kicks in and functions in very much the same way as when a subject is coping with output errors in processing natural language. In this situation, the reader does notice sloppy annotation, and also considers the information which the text is supposed to convey as epistemically relevant. The problem is how to extract it. The reader is confronted with a jumble of signs wherein it is unclear which ones are used as symbols, and with what meaning, and what function meaningless signs might have. The reader therefore activates a procedure for restoring the epistemic value of the received information consisting in four parallel operations (adapted from Lewandowsky et al. 2012, 112-13):

- A cohesion check, based on a comparison of the association of meaning and symbols within each annotated example.
- A coherence check, carried out by comparing contextual instances of annotation.
- A consistency check, performed by comparing the data provided by the present instance of annotation with previous encyclopedic knowledge (both in linguistics and of specific languages, including Japanese). This check might not necessarily reveal invalid analysis.
- A validation check, to assess the source's credibility and expertise, based on previous acquaintance with the author's works.

By so doing, the reader notices patterns and regularities, identifies errors, and eventually becomes able to assign valid, viable and consistent values to symbols and labels. This restores communication; a sense emerges from which the reader can acquire useful information, albeit with occasional data loss. For example, whenever period signs (and tildes) are used improperly, as in (22-23), it is not difficult to understand that they generally have the same function of hyphens. Once ascertained this, and confronted with the confused glossing of *mie.ta* as 'come.HON-PST' in (22), the reader only has to discard the incohesive segmentation as meaningless and, on the basis of the accompanying discussion, retain the fact that the sense of *mieta* is the sum of three meaning components: 'honorific', 'past' and 'come'. Here bad segmentation causes a loss of information about the individual

549
Annali di Ca' Foscari. Serie orientale
56, 2020, 501-562

e-ISSN 2385-3042
ISSN 1125-3789

true value of the items, but the relevant message, the meaning of the whole construction, goes through. The same process leads to restoring the epistemic contents of *zyunbisareta* in (19), *hanasiteimasu* in (20), *dete.ko-nai* in (24) and *shin de iru* in (29). In (19-20) the lack of segmentation in Line 1 causes a partial loss of data; but this loss is intentional, and permitted by the notational rules themselves. Only, the authors signal it with the wrong symbols. In (24) and (29), discarding the segmentation as meaningless prevents from being misinformed by the wrong value assigned to Line 2 morphs (see discussion at the end of § 7).

When case-particle boundaries are marked in inconsistent fashion, a reader will restore the value of boundary markers by familiarity, adopting as valid (that is, as expressing the author's true theoretical persuasion) the most frequent use of symbols. This is what happens in Iwasaki (2013), where the boundaries of case particle are marked by hyphens in the first chapter and by blanks in the rest of the volume. The valid symbols will be taken to be the latter, even by those readers unaware of Iwasaki's warning on the matter (2013, xix). Additionally, for each segmented item or type of item, the most frequently used symbol will override all other symbols. For instance, again in Narrog (2017), three distinct boundary markers are used for particles in allative function. Rather than ignoring them because of lack of cohesion, the reader might just reckon blanks as valid, since they have the highest distributional value throughout the article:

30. Override by most frequent boundary symbol in (Narrog 2017)
    a. *gakkoo=e*
       school ALL
    b. *heya ni*
       room ALL
    c. *heya ni*
       room-ALL

In (30) the variation in boundary marking for of all particles glossed ALL in Narrog (2017) is shown. The exact figures are omitted; combination (c) only occurs once, whereas (a-b) are quite frequent. From the frequency of blanks, a reader takes the valid boundary marker for allative particles to be a blank space, thus dismissing the use of equals signs and hyphens as an error. In this case, that allative case markers are free word-forms may, or may not, correspond to Narrog's true opinion.

The process of retrieving relevant content from poor boundary marking and sloppy annotation is within the reach of every linguist and is customarily applied by all members of the community. Being first of all a method for restoring cohesion, that is, for retrieving consistent associations of meanings and symbols, this process is used

in every daily-life instance of linguistic communication. Moreover, as happens in games and natural languages, a linguist knows that rules may have loopholes or might be difficult to understand and follow properly, and that when writing a paper many authors, including herself, prefer to invest time and mental energy in expounding their ideas than in polishing up their annotation. Authors know that readers would understand, readers know that authors know, and so on in a regression which is typical of the working of conventions. Indeed, a certain acceptance of sloppiness is in itself a deeper convention forming an integral part of the ethos of the profession; the trust in the expertise of some authors – and the deference towards editors – may decrease as an effect of bad annotation, but again theories and ideas are considered more important than technical accuracy.

It should be remembered, though, that the above process is sustained by a trust in the encoders' intention of informing truthfully about their theoretical persuasions. The process leads to recovering the integrity of the original message regardless of its truth content (that is, whether it represents valid linguistic analysis). As seen, when the use of symbols is too confused, it is just ignored as meaningless, but when it maintains, or is restored to, a certain cohesion, then it may be taken at face value, decoded as theory-consistent and hence convey false information. Without mobilizing previous linguistic knowledge of Japanese, which they may not have, readers cannot validate the decoded information and decide if it may be a source of new knowledge.

From the examples discussed above, it is clear that a misinformation effect has the greatest chance of happening when the use of boundary symbols implies that Japanese case markers are split into different morphological classes (independent words, clitics, affixes), while in fact they are not. Several minor cases of inconsistency (§ 6.2.1) have this misleading effect, for the marking pattern they display is logical enough that can be taken for the result of valid morphological analysis (see discussion of 11, 12, 13, 14 = 24 etc.). For example, among the major cases of inconsistency (§ 6.2.2), the over-hyphenation in (15), (16) and (17) implies that particles like the genitive marker *no* and the focus marker *mo* become affix-like in certain constructions.

Finally, the atheoretical use of blanks is the least damaging instance of sloppy annotation. In extreme cases, like the intrusion of blanks into morpheme breaks shown in (29), it is easily detected and ignored as an instance of wrong encoding. In all other cases it is undetectable, yet it involuntarily conveys a meaning 'case particles are free word-forms' which is not inconsistent with morphological theory. Therefore, the atheoretical use of blanks may not correspond to an author's true theoretical persuasion – so representing a betrayal of the reader's second type of trust and a lack of observance of Grice's

551
Annali di Ca' Foscari. Serie orientale
56, 2020, 501-562

e-ISSN 2385-3042
ISSN 1125-3789

maxim of quality – but does not otherwise misinform the reader by conveying incorrect analysis.

11   Conclusions

Matters of relevance, under which the attention and cognitive resources of the reader are channeled to filter out misinformation and recover meaning selectively, coupled with an ethos that tolerates sloppy annotation, help limit the disruption caused by poor boundary marking and allow it to persist in the English-language literature of Japanese linguistics. What follows is a conclusive assessment of the inconsistent treatment of Japanese case-particle boundaries under such conditions.

At the level of single works, be they monographs, journal articles or lone chapters in collections not specifically dealing with Japanese language, it is reasonable that the reductionist use of boundary symbols discussed in § 5 be let to intrude into Japanese annotation. Given the cognitive constraint working against the adoption of dedicated *fuzokugo* break symbol, and being Japanese case markers closer to free word-forms than to suffixes, the most accurate way of showing case-particle boundaries should be with blanks, reserving hyphens for the segmentation of word-internal morphemes. This two-marker-only strategy is indeed the one currently followed by Shibatani, in the latest stage of this author's long search for optimality. But there would be no harm in hyphenating case-particle boundary either. At most, such a choice of symbols could be criticized as being inaccurate but, within a single work, would not misinform.

In single works, for the sake of simplicity I also deem it acceptable for syntactic analysis to be stated by way of the glosses themselves and not in words. The readers will just assume that boundary marking is backed by a theory, whichever it is, without necessarily knowing of alternate theories, and of the validity thereof.

Care should be applied though when letting the pre-theoretical boundary marking of the original Latin orthography in Line 1 colonize Line 2 and grammatical annotation in general. Those authors who give no theoretical significance to the graphic way they show word boundaries should be aware of the fact that they are doing so and declare it explicitly.

In sum, then, in works that are not part of collections with a general introduction on Japanese morphology, the problem of how to interpret boundary symbols could be neutralized by simply adding some clarifying remarks. That authors hardly ever do so is another problem altogether.

In collections, my judgment is that inconsistence is ethically bad. An innocent reader, with no previous knowledge of Japanese, unaware of the morphologic nature of Japanese case markers, is confused

by seeing them treated now as affixes, now as independent words, now as clitics. Eventually, such a reader will most probably dismiss the conflicting pieces of information as inaccurate, leading to no misinformation, or conclude that Japanese case markers are of a controversial nature. Still, that reader will be no wiser as to what nature they are, with no new instrument to form a personal opinion about the matter. This type of inconsistency actually deprives boundary signs of meaning, and forces the morphology of Japanese case marker into irrelevance, to the effect that the reader acquires no information about it. It is a surrender of grammatical annotation.

The best possible solution is to require all contributors to comply with a general boundary marking strategy decided at editorial level, even if the chosen symbols do not reflect the personal opinions of all authors. In presence of uniformity, no explicit statement of the enforced policy would be needed in a preface or general introduction – but of course some explanatory remarks would do no harm either. The second best possible solution is to require all contributors to explicitly state the rationale of their boundary marking. This would allow all authors to freely express their theoretical stances and adopt the preferred marking method. Again, I could find no evidence that either of the above solutions has ever being adopted. I identified the cause in the little power that editors have over contributors.

A mix of several case-particle boundary marking strategies should never occur in one and the same work, for it may only generate uninformative or misinformative outputs. The most outstanding case of uninformative symbol-mixing in my sample is Horie (2018), discussed in § 6.2.2. It is true that readers have at their disposal a powerful cognitive procedure for disambiguating between conflicting boundary symbols, but after adopting one forced interpretation they will still be unsure about the true morphological status of Japanese case markers. In linguistic typology, it often happens that authors are not specifically competent in Japanese. When they use other sources' Japanese examples, in order to avoid using conflicting boundary symbols they should seek advice, or arbitrarily choose and enforce one method only. Then, assuming the choice is limited to hyphens and blanks, which one they use is ultimately unimportant – after all, if both can be found in original sources, both have some theoretical validity. On how to actually implement the chosen strategy, thereby altering the annotation of a language the author has little or no competence whereof, again advice should be sought. As an expert, though, a typologist is required to have enough knowledge to identify the relevant morphemes and separate them correctly, to the extent that Lehmann says:

> If the author knows the number and order of morphs in an L1 form, then he should indicate them. If the author does not even know so much, he should probably not use the example. (2004b, 1854)

553
Annali di Ca' Foscari. Serie orientale   e-ISSN 2385-3042
56, 2020, 501-562   ISSN 1125-3789

This is why I consider the cases of Malchukov, Haspelmath, Comrie (2010) and of Malchukov (2016) particularly serious, given the authors' high standing[14] and the scope of the volumes hosting their chapters.

The differential treatment of case-marker boundaries, which implicitly splits Japanese postnominal particles into several distinct morphological classes, is absolutely inacceptable. The role of hyphens is to mark word-internal morpheme breaks, not some vague, affix-like morphological relation that sprouts out when certain words are in certain sequences, as all the instances of over-hyphenation discussed above imply. This view of Japanese morphology is aberrant. Such a free use of hyphens has no morphological basis and no place in the notational conventions. It must be avoided at all costs because of the misinformative effect on the innocent reader.

As for the mix of hyphens, blanks, tildes, occasional equals and period signs, freely alternating between the two lines of individual examples, often to mark the same type of boundary, and across the annotated examples of entire articles, it is nothing but a rude mockery of annotation. It has no epistemic value whatsoever, no matter how easily it may be ignored by an expert reader. It should not be allowed in the professional literature of linguistics.

What brings then many leading scholars in the field of Japanese linguistics to be so careless in their annotation? As I anticipated in the introduction, there is no way of giving a satisfactory answer to this question. However, I do have an opinion on the matter, and I will express it here, with no further analysis.

I think that the bad annotation of Japanese is not actually caused by sloppiness, but by a biased belief in the inutility of annotating Japanese. The line of reasoning underlying such a belief may be explained in the following way. Annotating Japanese is useless because no annotation, no matter how accurate, will ever be able to effectively encode Japanese grammar or bring about the 'spirit of the language' – which is, after all, the ultimate purpose of interlinear morphemic glossing (Lehmann 2004b, 1834). This lack of power does not lie in some technical weakness of the notational tools themselves. If this were the case, the tools would only have to be better tuned to become efficient. The problem is that they are based on Western linguistics, whose notions reflect Western categories of thought and can only represent languages to Western minds. Japanese, however, is different, because it originates in the mind of Japanese people, or perhaps it existed in nature and then penetrated into their brains.

---

14 Malchukov, Haspelmath, Comrie 2010 is by Andrej Malchukov, Martin Haspelmath and Bernard Comrie. Haspelmath and Comrie are coauthors of the LGR (Bickel, Comrie, Haspelmath 2008).

554

Annali di Ca' Foscari. Serie orientale
56, 2020, 501-562

Either way, Japanese people think differently from all the other people in the world. Since Japanese does not follow the same logic as any other human language, the tools for representing them are powerless for Japanese. This is the same bias that originates the many theories on the unicity of the Japanese people (*nihonjinron* 日本人論) and of Japanese culture (*nihonbunkaron* 日本文化論) (Befu 2001). In the domain of language, this bias goes under the name of *nihongoron* 日本語論, 'the question of Japanese language', and manifests itself in a system of myths ascribing to Japanese interlaced properties such as being ancient, natural, prelogical, pure, unique, difficult, untranslatable (Miller 1982, 10-11; 1977).

Under this attitude, there is no reason to waste time and effort in trying to accurately represent a language that so defies Western categorization: no annotation could ever be accurate for Japanese. Since annotation is expected of them, authors do annotate, using symbols and applying rules. But many just do it for the sake of it, without acknowledging the true purpose of the convention, to the effect that their annotation is chaotic, sometimes unintelligible, often misleading. But under the discourse of *nihongoron*, it does not matter, because those readers who know Japanese will understand the contents even without annotation, whereas to the others the annotated text will be useless anyway. This is of course one of the many Japanese expressions of what is called 'native-speakerism' (Holliday 2005). Such a denial of the role of metalanguage is the only explanation I could find for the existence of the many instances of poor annotation discussed above. I am persuaded that all authors mentioned in this paper are fine scientists who are contributing with their works to the progress of the science of linguistics and of human knowledge in general. I am also positive that, if asked, they would categorically deny they believe in some nonsense like the incommensurability of Japanese. This is the very reason they should pay more attention to the way they annotate.

List of abbreviations

ACC, ACC, Acc	accusative
all	allative
CAUS	causative
CL	classifier
COMP	complementizer
CON	converb
cop	copula
cvb	converb
dat	dative
EMPH	emphatic
evid	evidential
FOC	focus
\ft	free translation tier
GEN, GEN	genitive
ger	gerund
\gr	grammatical tagging tier
\it	gloss tier
HJLL	Handbooks of Japanese Language and Linguistics
IMG	interlinear morphemic gloss
INS	instrumental
ISO	International Organization for Standardization
hon	honorific
honor	honorific
L1	object language
LGR	Leipzig Glossing Rules
\mo	morpheme break tier
neg	negative
NOM, NOM, Nom	nominative
NONPST, NONPST	nonpast
NMLZM	nominalization marker
nmlzr	nominalizer
NP	noun phrase
OBL	oblique
\orth	orthographic transcription tier
Per	perfective
\phonem	phonemic transcription tier
\phonet	phonetic transcription tier
PART, part	particle
PASS, PASS	passive
PAST	past
POL, POL	polite

556

Annali di Ca' Foscari. Serie orientale   e-ISSN 2385-3042
56, 2020, 501-562   ISSN 1125-3789

PP	postpositional phrase
PRES	present
PROG	progressive
prs	present
PST, PST	past
Q	question
\ref	reference tier
RES	resultative
TOP, TOP, top	topic

Bibliography

Bakker, Dik et al. (1994). *EUROTYP Guidelines*. Strasbourg: European Science Foundation, Programme in Language Typology.

Befu Harumi (2001). *Hegemony of Homogeneity. An Anthropological Analysis of Nihonjinron*. Melbourne: Trans Pacific Press.

Bickel, Balthasar; Comrie, Bernard; Haspelmath, Martin (2004). *The Leipzig Glossing Rules. Conventions for Interlinear Morpheme-by-Morpheme Glosses*. Leipzig: Max-Planck-Institut für Evolutionäre Anthropologie.

Bickel, Balthasar; Comrie, Bernard; Haspelmath, Martin (2008). *The Leipzig Glossing Rules. Conventions for Interlinear Morpheme-by-Morpheme Glosses*. 2nd ed. Leipzig: Max-Planck-Institut für Evolutionäre Anthropologie.

Bird, Steven; Liberman, Mark (2001). "A Formal Framework for Linguistic Annotation". *Speech Communication*, 33, 23-60. https://doi.org/10.1016/S0167-6393(00)00068-6.

Creissels, Denis (2014). "Functive Phrases in Typological and Diachronic Perspective". *Studies in Language*, 38(3), 605-47. https://doi.org/10.1075/sl.38.3.07cre.

Croft, William (2003). *Typology and Universals*. 2nd ed. Cambridge: Cambridge University Press.

Duranti, Alessandro (1997). "Transcription. From Writing to Digitized Images". Duranti, Alessandro, *Linguistic Anthropology*. Cambridge: Cambridge University Press, 122-61. https://doi.org/10.1017/cbo9780511810190.006.

Edwards, Jane A. (2001). "The Transcription of Discourse". Schiffrin, Deborah; Tannen, Deborah; Hamilton, Heidi E. (eds), *The Handbook of Discourse Analysis*. Oxford: Blackwell, 321-48. https://doi.org/10.1002/9780470753460.ch18.

Fetzer, James H. (2004). "Information. Does It Have to Be True?". *Minds and Machines*, 14(2), 223-9. https://doi.org/10.1023/b:mind.0000021682.61365.56.

Floridi, Luciano (2011). *The Philosophy of Information*. Oxford: Oxford University Press.

Frellesvig, Bjarke; Sells, Peter (eds) (2013). *Japanese/Korean Linguistics*, vol. 20. Stanford: Centre for the Study of Language and Information (CSLI) Publications.

Frellesvig, Bjarke; Shibatani Masayoshi; Smith, John Charles (eds) (2007). *Current Issues in the History and Structure of Japanese*. Tokyo: Kurosio.

Fukuda Shin (2006). "Japanese Passives, External Arguments, and Structural Case". *San Diego Linguistic Papers*, 2, 86-133.

Giriko Mikio et al. (eds.) (2014). *Japanese/Korean Linguistics*, vol. 22. Stanford: Centre for the Study of Language and Information Publications.

Grice, Herbert Paul (1975). "Logic and Conversation". Cole, Peter; Morgan, Jerry L. (eds), *Syntax and Semantics*. Vol. 3: *Speech Acts*. New York: Academic Press, 41-58. `https://doi.org/10.1163/9789004368811_003`.

Harada Naomi; Larson, Richard K. (2009). "Datives in Japanese". Shibagaki, Reiko; Vermeulen, Ryosuke (eds), *Proceedings of the 5th Workshop on Altaic Formal Linguistics* (London, 23-5 May 2008). *MIT Working Papers in Linguistics*, 54, 3-17.

Hasegawa Yoko. (2016). "Translation. A Theoretical Perspective". Minami 2016, 415-39.

Hasegawa Yoko (2018a). "Introduction". Hasegawa 2018b, 3-14.

Hasegawa Yoko (ed.) (2018b). *The Cambridge Handbook of Japanese Linguistics*. Cambridge: Cambridge University Press.

Hattori Shirō 服部四郎 [1949] (1960). "Gutaiteki gengo tan'i to chūshōteki gengo tan'i" 具体的言語単位と抽象的言語単位 (Concrete and Abstract Linguistic Items). Hattori Shirō (ed.), *Gengogaku no hōhō* 言語学の方法 (Methods in Linguistics). Tōkyō: Iwanami shoten, 447-60.

Hattori Shirō 服部四郎 (1950). "Fuzokugo to fuzoku keishiki" 付属語と付属形式 (Subordinate Words and Subordinate Forms). *Gengo kenkyū* 言語研究, 15, 1-26; 103-4. Reprinted in Hattori 1960, 461-90.

Hellan, Lars; Malchukov, Andrej L.; Cennamo, Michela (eds) (2017). *Contrastive Studies in Verbal Valency*. Amsterdam: John Benjamins.

Heycock, Caroline (2008). "Japanese -wa, -ga and Information Structure". Miyagawa, Saito 2008, 53-83. `https://doi.org/10.1093/oxfordhb/9780195307344.013.0003`.

Himmelmann, Nikolaus P. (2006). "The Challenges of Segmenting Spoken Language". Gippert, Jost; Himmelmann, Nikolaus P.; Mosel, Ulrike (eds), *Essentials of Language Documentation*. Berlin: de Gruyter, 253-74. `https://doi.org/10.1515/9783110197730.253`.

Holliday, Adrian R. (2005). *The Struggle to Teach English as an International Language*. Oxford: Oxford University Press.

Horie Kaoru (2018). "Linguistic Typology and the Japanese Language". Hasegawa 2018b, 65-86.

Hume, David [1751] (1983). *An Enquiry Concerning the Principles of Morals*. Indianapolis: Hackett.

Hutton, Christopher (1993). "Analysis and Notation. The Case for a Non-Realist Linguistics". Harré, Rom; Harris, Roy (eds), *Linguistics and Philosophy. The Controversial Interface*. Oxford: Pergamon, 165-78. `https://doi.org/10.1016/b978-0-08-041937-4.50011-8`.

Iannàccaro, Gabriele (2000). "Per una semantica più puntuale del concetto di 'dato linguistico'. Un tentativo di sistematizzazione epistemologica". *Quaderni di Semantica*, 21(1), 51-79.

Ide, Nancy et al. (2017). "Designing Annotation Schemes. From Model to Representation". Ide, Pustejovsky 2017, 73-111.

Ide, Nancy; Pustejovsky, James (eds) (2017). *The Hanbook of Linguistic Annotation*. Dordrecht: Springer.

ISO, International Organization for Standardization (1989). *ISO3602. Documentation – Romanization of Japanese (kana script)*. Genève: International Organization for Standardization. `https://www.iso.org/standard/9029.html`.

558

Annali di Ca' Foscari. Serie orientale | e-ISSN 2385-3042
56, 2020, 501-562 | ISSN 1125-3789

Ito Junko; Kubozono, Haruo; Mester, Armin (2017). "A Prosodic Account of Consonant Gemination in Japanese Loanwords". Kubozono Haruo (ed.), *The Phonetics and Phonology of Geminate Consonants*. Oxford: Oxford University Press, 283-320. `https://doi.org/10.1093/oso/9780198754930.003.0013`.

Iwasaki Shoichi (2013). *Japanese. Revised Edition*. Amsterdam: John Benjamins.

Iwata Seizi (2017). "Locative Alternation. A Lexical-Constructional Approach". Shibatani, Miyagawa, Noda 2017b, 235-69. `https://doi.org/10.1075/cal.6`.

Kageyama Taro (2016). "Lexical Integrity and the Morphology-Syntax Interface". Kageyama, Kishimoto 2016, 490-528.

Kageyama Taro; Jacobsen, Wesley (eds) (2016). *Transitivity and Valency Alternations. Studies on Japanese and Beyond*. Berlin: de Gruyter.

Kageyama Taro; Kishimoto Hideki (eds) (2016). *Handbook of Japanese Lexicon and Word Formation*. Berlin: de Gruyter.

Kishimoto Hideki (2017). "Case Marking". Shibatani, Miyagawa, Noda 2017b, 447-96.

Kishimoto Hideki; Kageyama Taro; Sasaki Kan (2015). "Valency Classes in Japanese". Comrie, Bernard; Malchukov, Andrej (eds), *Valency Classes in the World's Languages*. Vol. 1: *Introducing the Framework, and Case Studies from Africa and Eurasia*. Berlin: de Gruyter, 765-805.

Kitagawa Yoshihisa (2017). "Interfacing Syntax with Sounds and Meanings". Shibatani, Miyagawa, Noda 2017b, 497-552.

Kubozono Haruo (ed.) (2015). *Handbook of Japanese Phonetics and Phonology*. Berlin: de Gruyter.

Lehmann, Christian (1982). "Directions for Interlinear Morphemic Translations". *Folia Linguistica*, 16(1-4), 199-224. `https://doi.org/10.1515/flin.1982.16.1–4.199`.

Lehmann, Christian (2004a). "Data in Linguistics". *The Linguistic Review*, 21(3-4), 175-210. `https://doi.org/10.1515/tlir.2004.21.3–4.175`.

Lehmann, Christian (2004b). "Interlinear Morphemic Glossing". Booij, Geert et al. (eds), *Morphology. An International Handbook on Inflection and Word-Formation*, vol. 2. Berlin: de Gruyter, 1834-57.

Lewandowsky, Stephan et al. (2012). "Misinformation and Its Correction. Continued Influence and Successful Debiasing". *Psychological Science in the Public Interest*, 13(3), 106-31. `https://doi.org/10.1177/1529100612451018`.

Lewis, David [1969] (2002). *Convention. A Philosophical Study*. Oxford: Blackwell.

Malchukov, Andrej L. (2016). "'Ambivalent Voice'. Markedness Effects in Valency Change". Kageyama. Jacobsen 2016, 389-422.

Malchukov, Andrej; Haspelmath, Martin; Comrie, Bernard (2010). "Ditransitive Constructions. A Typological Overview". Malchukov, Andrej; Haspelmath, Martin; Comrie, Bernard (eds), *Studies in Ditransitive Constructions. A Comparative Handbook*. Berlin, de Gruyter, 1-64. `https://doi.org/10.1515/9783110220377.1`.

Malchukov, Andrej L.; Spencer, Andrew (eds) (2009). *The Oxford Handbook of Case*. Oxford: Oxford University Press.

Marmor, Andrei (2009). *Social Conventions. From Language to Law*. Princeton: Princeton University Press.

Miller, Roy Andrew (1977). *The Japanese Language in Contemporary Japan*. Washington: American Enterprise Institute for Public Policy Research.

Miller, Roy Andrew (1982). *Japan's Modern Myth. The Language and Beyond*. New York: Weatherhill.

Minami Masahiko (ed.) (2016). *Handbook of Japanese Applied Linguistics*. Berlin: de Gruyter.

Miyagawa Shigeru (2017). "Numeral Quantifiers". Shibatani, Miyagawa, Noda 2017b, 581-610.

Miyagawa Shigeru; Tsujioka Takae (2004). "Argument Structure and Ditransitive Verbs in Japanese". *Journal of East Asian Linguistics*, 13(1), 1-38. https://doi.org/10.1023/b:jeal.0000007345.64336.84.

Miyagawa Shigeru; Saito Mamoru (eds) (2008). *The Oxford Handbook of Japanese Linguistics*. Oxford: Oxford University Press.

Motomura Mitsue (2013). "An Analysis of Inalienable Possession Constructions in Japanese". *Osaka keidai ronshū*, 64(1), 103-26.

Murao Haruhiko (2009). *Cognitive Domains and Prototypes in Constructions*. Tokyo: Kurosio.

Murasugi Keiko (2015). "Root Infinitive Analogues in Child Japanese". Nakayama 2015, 117-47.

Nakamura Wataru (2018). "Case". Hasegawa 2018b, 249-75.

Nakayama Mineharu (ed.) (2015). *Handbook of Japanese Psycholinguistics*. Berlin: de Gruyter.

Nakazawa Tsuneko (2012). "On Interpretation of Resultative Phrases in Japanese". *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation* (Bali, 7-10 November 2012). S.l.: Universitas Indonesia, 259-601.

Narrog, Heiko (2017). "The Morphosyntax of Grammaticalization in Japanese". Shibatani, Miyagawa, Noda 2017b, 333-70.

Narrog, Heiko (2018). "Modality". Hasegawa 2018b, 357-78.

Nishikawa Ichiro (1983). "Traslitterazione della lingua giapponese e questioni varie". *Annali della Facoltà di Scienze Politiche dell'Università di Milano*, 3, 469-505.

Ogawa Akio (2009). "Case in Topic-Prominent Languages. Pragmatic and Syntactic Functions of Cases in Japanese". Malchukov, Spencer 2009, 779-88.

Ono Naoyuki (2013). "Two Modes of Argument Selection in Nominals". Frellesvig; Sells 2013, 413-20.

Otoguro Ryo (2003). "Focus Clitics and Discourse Information Spreading". Butt, Miriam; Holloway King, Tracy (eds), *Proceedings of the LFG03 Lexical-Functional Grammar Conference* (Saratoga Springs, 16-18 July 2003). Stanford: Centre for the Study of Language and Information Publications, 367-86.

Otoguro Ryo (2006). *Morphosyntax of Case. A Theoretical Investigation of the Concept* [PhD Dissertation]. Colchester: University of Essex.

Pardeshi, Prashant; Kageyama Taro (eds) (2018). *Handbook of Japanese Contrastive Linguistics*. Berlin: de Gruyter.

Rehbein, Jochen; Hohenstein, Christiane; Pietsch, Lukas (eds) (2007). *Connectivity in Grammar and Discourse*. Amsterdam: John Benjamins.

Saito Mamoru; Lin, Jonah T.-H.; Murasugi, Keiko (2014). "N'-Ellipsis and the Structure of Noun Phrases in Chinese and Japanese". Saito 2014, 1-24. https://doi.org/10.1093/acprof:oso/9780199945207.003.0001.

Saito Mamoru (ed.) (2014). *Japanese Syntax in Comparative Perspective*. Oxford: Oxford University Press.

Saito Mamoru (2017). "Ellipsis". Shibatani, Miyagawa, Noda 2017b, 701-50.

Sakamoto Tsutomu (2015). "Processing of Syntactic and Semantic Information in the Human Brain. Evidence from ERP Studies in Japanese". Nakayama 2015, 457-510.

Sakoda Kumiko (2016). "Errors and Learning Strategies by Learners of Japanese as a Second Language". Minami 2016, 129-50.

Sato Yutaka (2014). "Japanese Passives with Verbal Nouns". Giriko et al. 2014, 307-22.

Schultze-Berndt, Eva (2006). "Linguistic Annotation". Gippert, Himmelmann, Mosel 2006, 213-52.

Shibatani Masayoshi (2003). "Directional Verbs in Japanese". Shay, Erin; Seibert, Uwe (eds), *Motion, Direction and Location in Languages. In Honor of Zygmunt Frajzyngier*. Amsterdam: John Benjamins, 259-86.

Shibatani Masayoshi (2004). "Voice". Booij, Geert et al. 2004, 1154-65.

Shibatani Masayoshi (2006). "Japanese". Brown, Keith (ed.), *Encyclopedia of Language and Linguistics*, vol. 6. 2nd ed. Amsterdam: Elsevier, 102-5.

Shibatani Masayoshi (2007a). "Grammaticalization of Converb Constructions. The Case of Japanese *-te* Conjunctive Constructions". Rehbein, Hohenstein, Pietsch 2007, 21-49.

Shibatani Masayoshi (2007b). "Grammaticalization of Motion Verbs". Frellesvig, Shibatani, Smith 2007, 107-34.

Shibatani Masayoshi (2009a). "Case and Voice. Case in Derived Constructions". Malchukov, Spencer 2009, 323-38.

Shibatani Masayoshi (2009b). "Elements of Complex Structures, where Recursion Isn't. The Case of Relativization". Givón, Talmy; Shibatani Masayoshi (eds), *Syntactic Complexity*. Amsterdam: John Benjamins, 163-98.

Shibatani Masayoshi (2009c). "Japanese". Comrie, Bernard (ed.), *The World's Major Languages*. 2nd ed. London: Routledge, 741-63.

Shibatani Masayoshi (2013). "What Can Japanese Dialects Tell Us about the Function and Development of the Nominalization Particle 'No'?". Frellesvig, Sells 2013, 429-51.

Shibatani Masayoshi (2016). "The Role of Morphology in Valency Alternation Phenomena". Kageyama, Jacobsen 2016, 445-78.

Shibatai Masayoshi (2017). "Nominalization". Shibatani, Miyagawa, Noda 2017b, 271-331.

Shibatani Masayoshi (2018a). "Nominalization". Hasegawa 2018b, 461-2.

Shibatani Masayoshi (2018b). "Nominalization in Crosslinguistic Perspective". Pardeshi, Kageyama 2018, 346-410.

Shibatani Masayoshi; Kageyama Taro (2015). "Introduction to the *Handbooks of Japanese Language and Linguistics*". Kubozono 2015, vii-xxxiv.

Shibatani Masayoshi et al. (2014). "Genitive Modifiers. *Ga/no* Conversion Revisited". Giriko et al. 2014, 355-94.

Shibatani Masayoshi; Miyagawa Shigeru; Noda Hisashi (2017a). "Introduction". Shibatani, Miyagawa, Noda 2017b, 1-26.

Shibatani Masayoshi; Miyagawa Shigeru; Noda Hisashi (eds) (2017b). *Handbook of Japanese Syntax*. Berlin: de Gruyter.

Shimojo Mitsuaki (2018). "Word Order and Extraction. A Functional Approach". Hasegawa 2018b, 404-28.

Shirai Yasuhiro (2015). "L2 Acquisition of Japanese". Nakayama 2015, 217-34.

Sperber, Dan; Wilson, Deirdre (1995). *Relevance. Communication and Cognition*. Second Edition. Oxford: Blackwell.

Stassen, Leon (2009). *Predicative Possession*. Oxford: Oxford University Press.

Takahashi Daiko (2008). "Noun Phrase Ellipsis". Miyagawa, Saito 2008, 394-422.

Takezawa Koichi (2016). "Inflection". Kageyama, Kishimoto 2016, 459-87.

Tamaoka Katsu (2015). "Processing of the Japanese Language by Native Chinese Speakers". Nakayama 2015, 583-632.

Tomioka Satoshi (2007). "The Japanese Existential Possession. A Case Study of Pragmatic Disambiguation". *Lingua*, 117(6), 881-902. `https://doi.org/10.1016/j.lingua.2006.04.004`.

Tsuboi Eijirō (2010). "Malefactivity in Japanese". Zúñiga, Fernando; Kittilä, Seppo (eds), *Benefactives and Malefactives. Typological Perspectives and Case Studies*. Amsterdam: John Benjamins, 419-35.

Tsujimura Natsuko (2007). *An Introduction to Japanese Linguistics. Second Edition*. Oxford: Blackwell.

Tsujimura Natsuko (2013). *An Introduction to Japanese Linguistics. Third Edition*. Hoboken (NJ): John Wiley & Sons.

Vance, Timothy J. (1993). "Are Japanese Particles Clitics?". *The Journal of the Association of Teachers of Japanese*, 7(1), 3-33. `https://doi.org/10.2307/489122`.

Wellisch, Hans H. (1978). *The Conversion of Scripts. Its Nature, History and Utilization*. New York: Wiley & Sons.

Wilcock, Graham (2009). *Introduction to Linguistic Annotation and Text Analytics*. San Rafael (CA): Morgan & Claypool.

Zwicky, Arnold (1985). "Clitics and Particles". *Language*, 61(2), 283-305. `https://doi.org/10.2307/414146`.

Zwicky, Arnold; Pullum, Geoffrey K. (1983). "Cliticization vs. Inflection. English N'T". *Language*, 59(3), 502-13. `https://doi.org/10.2307/413900`.