Northern Language Studies 10: 61-82（Japan Association of NoLS, 2020）

[Special Feature: Linguistic Typology and Studies of Northern Languages]

How to Handle "Missing Values" in Linguistic Typology: A Pitfall in the Statistical Modelling Approach

Yohei ONO
(Graduate Student at the Open University of Japan)

Keywords: Computational Linguistics, Descriptive Linguistics, Probability Axioms, Statistical Imputation, The World Atlas of Language Structures

1. Introduction

The recent development of an invaluable large-scale database in linguistic typology, The World Atlas of Language Structures Online (Dryer and Haspelmath 2013; hereafter WALS) or Autotyp 0.1.0 (Bickel et al. 2017), has resulted in the promotion of interdisciplinary research among linguists, information scientists, and statisticians.

As a statistician, however, I consider the current research environment to be unsuitable for further developments in the realm. Since information scientists/ statisticians publish in their own journals, their papers are replete with numerical formulae, probability functions, and statistical model equations that even researchers in the same field have difficulty understanding. Thus, linguists cannot comprehend the substantive contribution and discuss the content based on their linguistic knowledge.

In the field of statistical typology, the papers by information scientists/statisticians are far removed from substantive knowledge in linguistics, while papers by linguists are far removed from substantive knowledge in information science/statistics. My motivation in conducting this study was to raise the alarm on this situation in statistical typology and to demonstrate what form interdisciplinary research should take, taking "missing values", the basic notion in information science/statistics, as an example.

There are two main streams in statistical typology: one, examining WALS data with probability distribution (e.g., Albu 2006, Daumé and Campbell 2007, Daumé 2009), the other elucidating WALS data without probability distribution (e.g., Tsunoda, Ueda, and Itoh 1995, Cysouw 2007, Ono, Yoshino, Hayashi, and Whitman 2017, 2018, Whitman and Ono 2017, 2020). These two streams of research differ in the following three points: (1) the selection of WALS data in the analysis; (2) the purpose of applying statistical methods to WALS data; (3) the selection of statistical methods based on the previous two points. This paper focuses on the former stream, which in this paper is called the "statistical modelling approach."

The statistical modelling approach requires as much WALS data as possible, even

though more than 70 percent of WALS data contain missing values (Murawaki 2019: 202). Furthermore, researchers in this stream consider WALS data to comprise one sample drawn from a true probabilistic structure and construct statistical modelling as its approximation (see Section 3). Thus, these researchers utilize probability function and probabilistic modelling (e.g., Bayesian statistical modelling) in accordance with the purpose of their analysis.

However, the other stream, research without probability distribution, selects a language sample to avoid biases from the viewpoints of both genealogy and areality. Moreover, their statistical analysis aims to obtain substantive knowledge not only for linguistic typology but also linguistics, visualizing the relationships among linguistic features and languages. Thus, these researchers utilize descriptive statistical methods, without probability function or probabilistic modelling.

As mentioned above, the two streams of research demonstrate a clear disposition regarding whether probability distribution is applicable for analysis of WALS data. Focusing on the statistical modelling approach, this paper discusses whether probability distribution can apply to "missing values" in WALS from the viewpoint of linguistic materials, taking Ainu, Chukchi, Khalkha, and Navajo as examples.

The main objective is to clarify the issue that the statistical modelling approach does not appropriately handle "missing values" in WALS from the viewpoint of either linguistics or statistics, and to demonstrate directions for further development in statistical typology. To my knowledge, previous studies have not dealt with the issue of probability distribution or presented an alternative. Thus, this paper is novel in proposing that the statistical modelling approach should primarily address the issue of missing values in the database.

The remainder of this paper is organized as follows. Section 2 demonstrates that the missing values in WALS data differ from those in information science/statistics. Since WALS does not describe the reason for the missing values in the languages, I show what the missing values demonstrate in Ainu, Chukchi, Khalkha, and Navajo, referencing the literature on these languages. I clarify the differences between the missing values in WALS and those in information science/statistics, which results in violation of the assumption that probability axioms are indispensable for probability function and probabilistic modelling.

Section 3 explains the notion of probability axioms with the example of dice, and indicates the problem of probability axioms in making the missing values in WALS conform to those in information science/statistics, taking statistical imputation in previous studies as examples.

Section 4 states the significance of this paper and demonstrates that the missing values in WALS data contain some similarity information among languages, which could potentially improve previous statistical results. Thus, descriptive research, which

examines the missing values in WALS data, is a promising direction for future developments in the realm of statistical typology.

2. Materials

Section 2.1 focuses on the missing values in WALS data, taking Ainu, Chukchi, Khalkha, and Navajo as examples. Since WALS does not explain why languages are classified into missing values in the features concerned, this paper focuses on four features and the corresponding missing values: 12A: Syllable Structure; 14A: Fixed Stress Locations; 30A: Number of Genders; 47A: Intensifiers and Reflexive Pronouns. Section 2.2 then demonstrates that probability axioms, the basic notion of probability function and probabilistic modelling, cannot be applied to the missing values in WALS.

2.1. Missing Values from a Linguistic Perspective

Table 1 shows the missing values focused on in this paper. It shows the four languages (Ainu, Chukchi, Khalkha, and Navajo) and the four features: 12A: Syllable Structure; 14A: Fixed Stress Locations; 30A: Number of Genders; 47A: Intensifiers and Reflexive Pronouns. While the original data in WALS represent the missing values as blanks, here, the missing values in each feature are tentatively represented as NA[1]. Since WALS does not reference the specific literature on NA in each language, the following section will examine the reason for NA by scrutinizing the reference list in WALS.

Table 1. Each feature and the corresponding missing values in Ainu, Chukchi, Khalkha, and Navajo. Note that original data in WALS represent the missing values (or NA in this table) as blanks.

Name	12A	14A	30A	47A
Ainu	2 Moderately complex	NA	1 None	NA
Chukchi	3 Complex	2 Initial	1 None	NA
Khalkha	NA	1 No fixed stress	1 None	1 Identical
Navajo	2 Moderately complex	1 No fixed stress	NA	2 Differentiated

2.1.1. 12A: Syllable Structure

Maddieson (2013) classifies feature 12A: Syllable Structure into three categories: 1: Simple syllable structure; 2: Moderately complex syllable structure; and 3: Complex syllable structure, and explains the combination of strings for consonant and vowel sound symbols as the criterion for complexity in this feature.

Maddieson (2013: Ch.2) states, “Languages which permit a single consonant after the vowel and/or allow two consonants to occur before the vowel, but obey a limitation to only the common two-consonant patterns described above, are counted as having

[1] NA corresponds to different meanings in each field: “Not Attestable”, “Not Applicable”, or “Not Available”. Here, I replace the blank spaces in WALS with NA for the sake of convenience.

moderately complex syllable structure". Thus, Ainu (Simeon 1969: 754) and Navajo (Sapir and Hoijer 1967: 3) correspond to 2: Moderately complex syllable structure.

Maddieson (2013: Ch.2) also states, "Languages which permit freer combinations of two consonants in the position before a vowel, or which allow three or more consonants in this onset position, and/or two or more consonants in the position after the vowel, are classified as having complex syllable structure".

Skorik (1961) states on Chukchi, "The syllabic unit of the Chukchi language is the vowel. As part of a word, a vowel sound can be preceded or followed by one or two consonants, which are always non-syllable and dependent elements" (64-65) and "By their construction, the Chukchi language syllables can be monosonic (vowel), open (consonant + vowel), closed (vowel + consonant) and closed (consonant + vowel + consonant)…In addition, as indicated above, as a result of a confluence of consonants when a short vowel [ə] occurs, an open syllable can have two consonants (consonant + consonant + vowel)…A closed three consonants (consonant + consonant + vowel + consonant)…The composition of the word can combine all the indicated types of syllables. However, not each of them has equal opportunities in relation to the place occupied by him in the word" (65-66) [author's translation from the original Russian]. Thus, Chukchi corresponds to 3: Complex syllable structure in WALS.

However, Svantesson (2003: 158) states on Khalkha, "The maximal syllable structure is CVVCCC, i.e., the vowel kernel may be preceded by at most one consonant and followed by a cluster of up to three consonants. The vowel can be short, long or diphthong. In non-initial syllable, it can also be a non-phonemic schwa vowel. Onsetless syllables occur only word-initially. Whether a consonant combination can form a syllable coda or not depends on the phonetic properties of the consonants. Permitted types of coda include: voiced + voiceless consonant, e.g. *daws* [taws] 'salt', *alt* [aɮ^{h}t] 'gold', *bügd* [pugt] 'all'; nasal + stop or affricate, e.g. *xünd* [xunt] 'heavy', *möngg* [moŋg] 'silver', *myanggh* [m^{j}aŋɢ] 'thousand'; fricative + stop or affricate, e.g. *tsast* [ts^{h}as^{h}t] 'snowy'. Three-consonant codas consist of a voiced consonant followed by a fricative + stop or affricate, e.g. *ilst* [iɮs^{h}t] 'sandy'.[2]" This description indicates that the syllable structure of Khalkha cannot be precisely understood with the string combination for consonant and vowel sound symbols. Rather, it should be analyzed in terms of the

[2] Svantesson (2003: 156) originally describes the strong stops *p py t ty*, the strong affricates *ts c*, the weak stops *b by d dy g gy gh*, the weak affricates *dz j*, the fricatives *s sh x xy*, the nasals *m my n ny ng*, the laterals *l ly lh*, the vibrants *r ry*, and the glides *w wy y.* Furthermore, Svantesson (2003: 157) states, "the strong stops are probably produced with tensed vocal cords; they give a tense voice quality to the surrounding vowels, both the following and preceding one, resulting in some pre- and postaspiration. In the phonetic transcription, the strong stops and affricates are probably best written as postaspirated [p^{h} p^{jh} t^{h} t^{ʲh} ts^{h} tʃ^{ʰ}] in initial position, and preaspirated [^{h}p ^{h}p^{j} ^{h}t ^{h}t^{ʲ} ^{h}ts ^{h}tʃ] in medial and final position". Therefore, *alt* [aɮ^{h}t] 'gold', *tsast* [ts^{h}as^{h}t] 'snowy', and *ilst* [iɮs^{h}t] 'sandy' may also reflect the notations above (i.e., the strong stops *t* is preaspirated in these three words). However, the pre- and postaspiration on the strong stops is controversial (cf. Ueta 2019).

phonetic properties of the consonants restricting possible syllable structures, which results in the value for syllable structure in Khalkha being marked as missing.

2.1.2. 14A: Fixed Stress Locations

Goedemans and van der Hulst (2013) classify 14A: Fixed Stress Locations into seven categories: 1: No fixed stress (mostly weight-sensitive stress); 2: Initial: stress on the first syllable; 3: Second: stress on the second syllable; 4: Third: stress on the third syllable; 5: Antepenultimate: stress on the antepenultimate (third from right) syllable; 6: Penultimate: stress on the penultimate syllable; 7: Ultimate: stress is on the ultimate syllable.

On Chukchi, WALS refers to Bogoras (1922/2002). Bogoras (1922/2002: 680) states, “In all three languages [i.e., Chukchee, Koryak, and Kamchadal] the accent usually recedes to the beginning of the word, even as far as the fourth or fifth syllable from the end.” Thus, Chukchi corresponds to 2: Initial in WALS.

However, note that Skorik (1961: 67-68) states, “In the case when the monosyllabic stem is formed by the suffix, which is a syllable, the stress is always on a vowel of the stem. Therefore, for two-syllable suffix words (monosyllabic + syllable suffix), the stress regularly falls on the first syllable, i.e., the stem on the word. For the stem with a syllable suffix, which has two or more syllables, the stress is always on the last syllable of the stem” [author’s translation from the original Russian]. Furthermore, Minoura (1989: 927) explains, “In the case when the stem has two or more syllables, the stress falls on the last vowel on the stem, if the suffix attached to the stem has the vowel or is inserted with /ə/. Otherwise, the stress falls on the penultimate vowel in the stem” [author’s translation from the original Japanese].

Here, the descriptions indicate that the stress locations in Chukchi should not analyzed as 2: Initial in WALS, from the viewpoints of current linguistic knowledge, which results in the values on the fixed stress locations in Chukchi being classified as missing.

Furthermore, Street (1963: 62) explains on Khalkha, “There is a binary contrast between emphatic and normal accent; every nucleus with which emphatic accent symbol /ˊ/ is not written is assumed to have normal accent. The phonetic details of the two accents are not at all clear, but a few general statements can be made” and “The nuclei of an accentual word with normal accent throughout differ in prominence. If the word has only single vowel nuclei, the first of these is most prominent; otherwise, the first geminate nucleus or diphthong is most prominent (cf. Poppe 1951: 13). By prominence here is meant a combination of stress and length: the most prominent nucleus of a word has slightly greater stress than other nuclei of the word and somewhat longer than less prominent but phonemically identical nuclei.” Thus, Khalkha is classified into 1: No fixed stress (mostly weight-sensitive stress) in WALS.

However, note that Street (1963: 63) continues, "A nucleus with the emphatic accent is the most prominent nucleus of its word: it is more strongly stressed than any other nucleus of the word and begins at a relatively higher pitch than an equivalently placed nucleus with normal accent" and demonstrates some relationships between emphatic accent and pitch. Here, the descriptions indicate that the accent system in Khalkha can be analyzed in terms of both stress and pitch accent, which results in the values on the fixed stress locations in Khalkha being classified as missing.

Sapir and Hoijer (1967: 12) state on Navajo, "Each syllable in a Navaho utterance, except only those which are syllabic sibilants, contains one tone phoneme, if it has a single syllabic, or two tone phonemes, if the syllabic is doubled. As we have noted, tone is represented by an acute accent for the high tone and by a grave accent for the low tone. Although the tone is actualized in every voiced portion of the syllable, it is obviously most marked on the syllabic." Thus, Navajo corresponds to 1: No fixed stress (mostly weight-sensitive stress) in WALS[3].

However, Tamura (2000: 21) states on Ainu, "Accent is distinctive in most dialects of Hokkaido Ainu. Like Japanese, it is pitch accent, but unlike the Tokyo dialect (and others), in which the fall from high to low is distinctive, in Ainu, the rise from low to high is distinctive. The syllables before this rise are all low, and the syllables following it gradually fall with a certain degree of regularity."

Since the accent system in Ainu should be analyzed in terms of pitch accent but not stress accent, WALS classifies the values on the fixed stress locations in Ainu as missing.

2.1.3. 30A: Number of Genders

Corbett (2013) classifies this feature (i.e., 30A: Number of Genders) into five categories: 1: None; 2: Two; 3: Three; 4: Four; 5: Five or more. Corbett (2013: Ch.1) explains, "The defining characteristic of gender is agreement: a language has a gender system only if we find different agreements ultimately dependent on nouns of different types. In other words, there must be evidence for gender outside the nouns themselves."

Refsing (1986: 81) states on Ainu, "General nouns are independent words; free forms which may by themselves perform the role of sentence subject. Number is marked in a few cases by the suffix, *-utar* (as in *menokoutar*, 'women') or by reduplication (as in *ramram* 'fish scales'), but otherwise number is unmarked. So is gender, and as for case, only a small group among the general nouns form a kind of genitive, namely 'the

[3] WALS refers to Baskakov (1966: 107) in this feature (i.e., 14A: Fixed Stress Locations). However, Baskakov (1966: 107) does not exist in practice and should be corrected as Azimov, Amansaryev, and Saryev (1966: 107). Moreover, Azimov, Amansaryev, and Saryev (1966) deal with Turkmen language but not with Navajo. Throughout this study, I was confronted with several mistakes on the references in WALS. Thus, the improvement on the references will matter in future, as discussed in Section 4.

belonging form’. This form is used only to express inalienable possession.” Thus, Ainu corresponds to 1: None in WALS.

Dunn (1999: 61-63) does not explain gender agreement in Chukchi. Thus, Chukchi also corresponds to 1: None in WALS. Note that Dunn (2000) demonstrates that gender difference is expressed by lexical choice, pronunciation, etc. However, Corbett (2013: Ch.1) notes, “A further consequence of the definition is that differences in use of language which depend on the sex of the speaker (lexical choice, voice quality and so on) are not treated here; an example is the difference between men’s and women’s pronunciation in Chukchi (Dunn 2000).”

Poppe (1951: 33) explains on Khalkha, “A grammatical gender does not exist in Khalkha-Mongolian. The natural gender of humans or animals is expressed by special words: *adzarga* ‘stallion’ and *güü* ‘mare’” [author’s translation from the original German]. Thus, Khalkha corresponds to 1: None in WALS[4].

On Navajo, Young and Morgan (1972: 1) state, “Many of us still think of the English as one expressing gender, although actually there exists only a vestige of such a grammatical distinction, and this only in the forms of the 3rd person pronoun. Thus we use he, his, him to designate the masculine; she, hers, her for the feminine, and it, its, it for the neuter. In Navaho, gender is not expressed by special forms, even of the 3rd person pronoun. Its expression is as unimportant for clarity in the Navaho pronoun as it is in English nouns. Of course, in both languages, there are nouns such as man, woman, girl etc. which, by virtue of their meaning, are concerned with gender; but there is not a special word form, as in Latin or Spanish, to make such distinction. When, in Navaho, it is necessary to expressly designate the sex of an animal, this is accomplished by means of the counterparts of English male, female, or by specific names distinguishing the male or the female of the species.”

Here, the descriptions indicate that the gender system of Navajo cannot be precisely understood with the agreement, which results in the value for number of genders in Navajo being marked as missing.

2.1.4. 47A: Intensifiers and Reflexive Pronouns

König, Siegmund, and Töpper (2013) classify 47A: Intensifiers and Reflexive Pronouns into two categories: 1: Intensifiers and reflexive pronouns are formally identical; 2: Intensifiers and reflexive pronouns are formally differentiated. König, Siegmund, and Töpper (2013: Ch.1) explain these two categories as follows.

“By intensifiers we mean expressions like German *selbst*, Russian *sam*, Turkish *kendi*, Mandarin *zìjǐ*, English *x-self*, which can be adjoined to either NPs or VPs, are

[4] In this feature (i.e., 30A: Number of Gender), WALS originally refers to Orlovskaja (1961) in Khalkha. Since Orlovskaja (1961) is difficult to obtain, I refer to Poppe (1951), instead of Orlovskaja (1961).

invariably focused and thus are prosodically prominent. The main function of intensifiers can be seen in the evoking of alternatives to the referent of the NP they relate to: a. (adnominal) *The director himself opened the letter*; b. (adverbial) *The director opened the letter himself*" (König, Siegmund, and Töpper 2013: Ch.1);

"Reflexive pronouns are expressions which are prototypically used to indicate that a non-subject argument of a transitive predicate is coreferential with (or bound by) the subject, i.e. expressions like German *sich*, Russian *sebja*, Turkish *kendi*, Mandarin *zìjǐ*, English *x-self*: *John$_i$ saw himself$_i$ in the mirror*" (König, Siegmund, and Töpper 2013: Ch.1).

On Khalkha, Poppe (1951: 73) states, "Reflexive Pronouns is *ööröö* 'self'. Genitive *öörīŋ* is used in the meaning of 'own'. Dative and Locative, *öörtöö* or *öörtööŋ*, with reflexive-possessive ending, is also commonly used in the meaning of 'sich selber' and 'sibi' in German" [author's translation from the original German].

Furthermore, Kazama (2018) supports *ööröö* as the intensifiers by König, Siegmund, and Töpper (2013: Ch.1), referencing to Kullmann and Tserenpil (1996: 261-263). Therefore, Khalkha corresponds to 1: Intensifiers and reflexive pronouns are formally identical with respect to *ööröö*. Note that the interested reader should refer to Kazama (2018) for further discussions in this topic.

On Navajo, Young and Morgan (1972: 4) state, "A variant and particularized form of the above subjectives [i.e., independent subjective pronouns] is obtained by suffixation thereto of relatival enclitic *-í*, and prepounding of the particle *t'áá* (just). Thus *t'ááshíhí, t'áánihí, t'áábíhí, t'ááhóhí* etc., the meaning being thus modified to 'I myself', 'you yourself', 'he himself', etc."

Furthermore, Young and Morgan (1980: 23) explains, "The particle *t'áá*, just, plus the enclitics *-í-* the one, combine to produce a variant of the independent subjective pronoun comparable to the more emphatic 'I myself', 'you yourself' of English. Thus, '*ásdzaa*, 'I did it', more emphatic *shí 'ásdzaa*, 'I did it', and still more emphatic *t'áá shíhí 'ásdzaa*, 'I myself did it'."

Therefore, the form, which combines independent subjective pronouns (e.g., *shí* in this case) with the particle *t'áá* plus the enclitics *-í-*, can be considered as the intensifiers by König, Siegmund, and Töpper (2013: Ch.1).

Also, Young and Morgan (1980: 183) state[5,6], "The reflexive pronoun component -

[5] WALS originally refers to Young and Morgan (1980: 8-9, 721) in this feature. However, I confirmed that Young and Morgan (1980: 8-9, 721) only dealt with the intensifiers in Navajo but not with the reflexive pronouns. Thus, the improvement on the references will also matter in future.

[6] One of reviewers indicated that *'ádi-* is the affix in Navajo. However, Young and Morgan (1980: 448) juxtapose several meanings on *'ádi-*, including reflexive pronoun as (indirect or) direct object and reflexive possessive prefix with certain nouns. Furthermore, Young and Morgan (1980: 15) state on *'ádi-*, "a composite prefix representing self as the direct object of the verb. It will be noted in the more complex constructions involving the reflexive pronoun that the position of *'á-* is variable".

di- is normally in Position IV, while *’á-*, self, is in Position Ic. The possessive pronoun prefix *di-* in Sarcri, meaning ‘own, a person’s own’ suggests that Navajo *di-* as well as *’á-* (*’á-di-*) may have once functioned in a reflexive sense. At present, the compound prefix *’ádi-* serves as a reflexive with verbs and postpositions, and *’á-* is occasionally found as a reflexive possessive prefix, but *di-* does not appear without *’á-*.”

Thus, Navajo corresponds to 2: Intensifiers and reflexive pronouns are formally differentiated in WALS.

On Ainu, Tamura (2000: 204) and Satō (2008: 249-253) explain *yay-* and *si-* as reflexive prefix. Satō (2008: 249) shows *yay-pusu* and *si-pusu* in Chitose Ainu dialect, which may correspond to ‘emerge from ground or water consciously’ and ‘emerge from ground or water spontaneously’ in English, respectively[7].

On Chukchi, Kämpfe and Volodin (1995: 33-36) demonstrate the derivational affixes in Chukchi as a list, and state that *=эт* is the suffix expressing the reflexive, taking *эюпы=к* and *эюп=эты=к* as an example[8]. Furthermore, Dunn (1999: 255-257) also explains *=эт* (and *=эв*) as “ubiquitous verb derivational suffixes” (Dunn 1999: 255).

The descriptions demonstrate that *yay-* and *si-* in Ainu and *=эт* in Chukchi should be considered as reflexive affix; in other words, *yay-* and *si-* in Ainu and *=эт* in Chukchi satisfy the condition on affix in Carstairs-McCarthy (2006: 83): a bound morph that (1) is not a root and (2) is a constituent of a word rather than of a phrase or sentence[9].

Thus, the reflexive in Ainu and Chukchi cannot be precisely understood using the notion of pronoun, which results in the value of the intensifiers and reflexive pronouns in Ainu and Chukchi being marked as missing.

2.2. Missing Values from a Statistical Perspective

This section states the formal definition of probability axioms in Kolmogorov (1950) and explains how the missing values in Section 2.1 are dealt with in the context of statistics, taking “rolling a die” as an example. Kolmogorov (1950) defines probability axioms, the basic notion of probability space and probability function, as follows:

“Let *E* be a collection of elements $\xi, \eta, \zeta, \ldots$, which we shall call *elementary events*, and *F* a set of subsets of *E*; the elements of the set *F* will be called *random events.*

1. *F is a field of sets.*
2. *F contains the set E.*

Thus, the reflexive pronouns in WALS can be reconsidered from the current substantive linguistic knowledge.

[7] I referred to Satō (2007) and Kobayashi (2010) and translated *pusu* as ‘emerge from ground or water’.

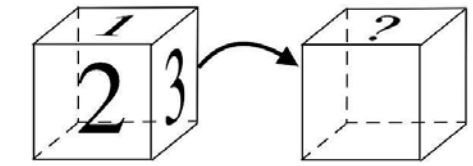
[8] Note that *эюпы=к* corresponds to ‘stechen’ and *эюп=этыы=к* to ‘sich stechen’ in German.

[9] Different interpretations could stand on the intensifiers and reflexive pronouns, depending on parts of speech proposed by the professionals on Ainu. The interested reader should refer to Izutsu (2006).

3. *To each set A in F is assigned a non-negative real number* P(A).
   This number P(A) *is called the probability of the event A.*
4. P(E) *equals* 1.
5. *If A and B have no elements in common, then*

$$\mathrm{P}(A + B) = \mathrm{P}(A) + \mathrm{P}(B)$$

A system of sets, *F*, together with a definite assignment of numbers P(A), satisfying Axioms 1-5, is called a *field of probability*" (Kolmogorov 1950: 2)[10]. Since this formal definition is not intuitive for linguists, the following sentences clarify the idea of probability axioms, taking "rolling a die" as an example.



A six-sided die **The result of rolling a die: which pip appeared on the top?**

Figure 1. Rolling a six-sided die.

Suppose that we once roll a die and define all results as *E*, then *E* can be decomposed into six subsets: C_1: one pip appears on top, when the die is rolled once; C_2: two pips appear on top, when the die is rolled once; C_3: three pips appear on top, when the die is rolled once; C_4: four pips appear on top, when the die is rolled once; C_5: five pips appear on top, when the die is rolled once; C_6: six pips appear on top, when the die is rolled once.

These six subsets have no elements in common and do not occur simultaneously. Therefore, the following equation holds[11]:

$$\begin{aligned} P(E) &= P(C_1 \cup C_2 \cup C_3 \cup C_4 \cup C_5 \cup C_6) \\ &= P(C_1) + P(C_2) + P(C_3) + P(C_4) + P(C_5) + P(C_6) \\ &= 1 \end{aligned}$$

Thus, we usually assign $1/6$ to $P(C_1)$, $P(C_2)$, $P(C_3)$, $P(C_4)$, $P(C_5)$, and $P(C_6)$, considering a die as unbiased.

The same analysis can be applied to the features in WALS. First, consider 12A: Syllable Structure as one die with all results of rolling this die defined as *E*. Then, *E* can be decomposed into three subsets: C_1: (1) Simple syllable structure appears on top, when the die is rolled once; C_2: (2) Moderately complex syllable structure appears on top, when the die is rolled once; C_3: (3) Complex syllable structure appears on top, when the die is rolled once. We observe that these three subsets have no elements in common and do not occur simultaneously. Therefore, 12A: Syllable Structure can be considered

[10] The interested reader can refer to the original definition of probability axioms in Kolmogorow (1933) and the Japanese literature in Itō (1951), respectively.
[11] Here I apply ∪, instead of + in the original notation in Kolmogorov (1950).

as a three-sided die.

Second, suppose that we consider 14A: Fixed Stress Locations as one die and define all results of rolling this die as *E*. Then, *E* can be decomposed into seven subsets: C_1: (1) No fixed stress (mostly weight-sensitive stress) appears on top, when the die is rolled once; C_2: (2) Initial (i.e., stress on the first syllable) appears on top, when the die is rolled once; C_3: (3) Second (i.e., stress on the second syllable) appears on top, when the die is rolled once; C_4: (4) Third (i.e., stress is on the third syllable) appears on top, when the die is rolled once; C_5: (5) Antepenultimate (i.e., stress is on the antepenultimate syllable) appears on top, when the die is rolled once; C_6: (6) Penultimate (i.e., stress is on the penultimate syllable) appears on top, when the die is rolled once; C_7: (7) Ultimate (i.e., stress is on the ultimate syllable) appears on top, when the die is rolled once. These seven subsets have no elements in common and do not occur simultaneously. Therefore, 14A: Fixed Stress Locations can be considered as a seven-sided die.

Third, consider 30A: Number of Genders as one die and define all results of rolling this die as *E*. Then, *E* can be decomposed into five subsets: C_1: (1) None appears on top, when the die is rolled once; C_2: (2) Two appears on top, when the die is rolled once; C_3: (3) Three appears on top, when the die is rolled once; C_4: (4) Four appears on top, when the die is rolled once; C_5: (5) Five or more appears on the top, when the die is rolled once. These five subsets have no elements in common and do not occur simultaneously. Therefore, 30A: Number of Genders can be considered as a five-sided die.

Finally, consider 47A: Intensifiers and Reflexive Pronouns as one die and define all results of rolling this die as *E*. Then, *E* can be decomposed into two subsets: C_1: (1) Intensifiers and reflexive pronouns are formally identical appears on top, when the die is rolled once; C_2: (2) Intensifiers and reflexive pronouns are formally differentiated appears on top, when the die is rolled once. These two subsets have no elements in common and do not occur simultaneously. Therefore, 47A: Intensifiers and Reflexive Pronouns can be considered as a two-sided die.

Section 2.1 demonstrated that the missing values in WALS represent that the language possesses some linguistic characteristic that the features in WALS cannot precisely describe with the given values in the corresponding feature. They do not represent that the language possesses some linguistic characteristic that the features in WALS can describe with the given values in the corresponding feature but that researchers have not yet observed for some reason (e.g., lack of descriptive research on the language or some environment around the language). Thus, the missing values in WALS raise a question regarding whether researchers can apply probability axioms to the corresponding features; in other words, whether researchers can consider the concerned features as an *n*-sided die.

Although we can consider an *n*-sided die when *n* is known or even known as infinite, the missing values in WALS indicate that the die (i.e., the feature with missing values)

has unknown *n* sides. Therefore, every attempt to handle the feature with missing values as a probability function contradicts probability axioms, especially axiom 4, with no exception.

For example, if researchers deal with 12A: Syllable Structure as a probability function, they need to fix the values in the feature (e.g., the three-sided die as explained above) and assign a probability to each value (e.g., $P(C_i)$) in accordance with axiom 4. However, the missing values in WALS demonstrate other sides overlooked by researchers. Thus, there are no guarantees in research that uses probability distribution for WALS data that their statistical modelling approach analyzes what WALS data represent.

This section has demonstrated that previous studies have dealt with "missing values" using statistical rather than linguistic methods, which enables information scientists/statisticians to apply probability function and probabilistic modelling. Again, the missing values in WALS represent that the language possesses some linguistic characteristic that the features in WALS cannot precisely describe, while the missing values in statistical manners represent that the languages possess some linguistic characteristic that the features in WALS can describe but that researchers have not yet observed for some reason (e.g., lack of descriptive research on the language or some environment around the language).

Next, I focus on the differences between the missing values in linguistics and those in statistics, which have resulted in confusion in previous statistical typology.

3. Missing Values in Previous Statistical Typology

This section explains the application of statistical methods to the missing values in WALS and clarifies the issues surrounding missing values in previous statistical typology. Section 3.1 focuses on (1) how the research on learning WALS data with probability distribution understands the nature of WALS data, (2) what "learning WALS data with probability distribution" indicates in the view of (1), and (3) what the "probability function cannot appropriately handle with the missing values in WALS" suggests about the view of (2).

Section 3.2 confirms the issues surrounding the missing values in Section 3.1, taking numerical formulae, probability functions, and statistical model equations in previous studies as examples. Note that the research on learning WALS data with probability distribution is paraphrased as the "Statistical Modelling Approach" in the following section.

3.1. The Statistical Modelling Approach and Its Understanding of WALS Data

As demonstrated in Figure 2, the Statistical Modelling Approach assumes one "true" statistical structure M, from which our data A in hand are generated. Therefore, the

Statistical Modelling Approach and the researchers who utilize it understand the nature of WALS data as one sample (or outcome) drawn from the "true" statistical structure M (i.e., our data A in hand could vary as A', A'', A''' etc.).

Notably, they usually consider the "true" statistical structure as non-specifiable from the data in hand. Instead, they construct some statistical model M' for the purpose T, and let statistical model M' learn from the data A in hand (see Figure 3). Note that the purpose T differs in each research: typological implications (Daumé and Campbell 2007), phylogeny in linguistic typology (Murawaki 2015), areality in linguistic typology (Daumé 2009), and both areality and phylogeny in linguistic typology (Murawaki and Yamauchi 2018).

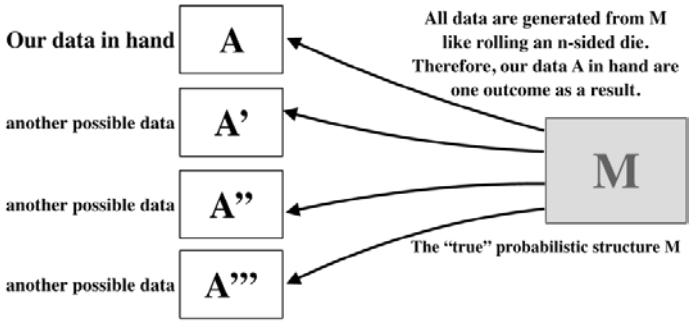


Figure 2. Data in the Statistical Modelling Approach.

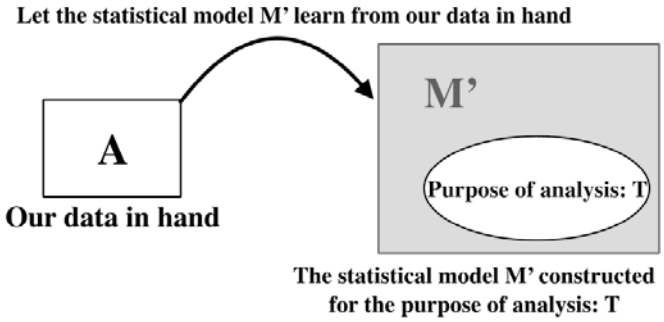


Figure 3. Statistical Learning from Data.

However, researchers are soon confronted with the problem that learning data A with statistical model M' requires data A to be able to be "appropriately" handled by the probability function. As demonstrated in Section 2, the missing values, more than 70 percent of WALS data (Murawaki 2019: 202), should be considered in the linguistic context but not in the statistical.

Again, the former represents that the language possesses some linguistic characteristic that the features in WALS cannot precisely describe, while the latter represents that the language possesses some linguistic characteristic that the features in

WALS can describe but that researchers have not yet observed for some reason (e.g., lack of descriptive research on the language or some environment around the language).

These facts indicate that previous statistical typology has analyzed something that does not reflect the nature of linguistic typological data at all. Thus, as the wise saying in information science, "garbage in, garbage out", goes, there is no guarantee that the results obtained are meaningful from the perspective of substantive knowledge.

Thus, the Statistical Modelling Approach should primarily address the issue of missing values in linguistic typology, as the Statistical Modelling Approach itself cannot avoid utilizing the probability function. To my knowledge, however, information science does not propose an alternative to the probability function on this issue and research in the field of information science has not yet addressed this problem. My hope, therefore, is that information scientists and statisticians who apply the Statistical Modelling Approach to linguistic typology will address this fundamental problem in the future.

Next, I focus on how previous statistical typology has addressed the issue of missing values, taking descriptions, numerical formulae, probability functions, and statistical model equations as examples.

3.2. Examples in Previous Statistical Typology

This section demonstrates that previous statistical typology has not yet addressed the fundamental issue of missing values. Daumé (2009) focuses on areality in WALS and states, "We will consider a data set consisting of N languages and F typological features. We denote the value of feature f in language n as $X_{n,f}$. For simplicity of exposition, we will assume two things: (1) there is no unobserved data and (2) all features are binary. In practice, for the data we use (described in Section 4), neither of these is true. However, both extensions are straightforward" (Daumé 2009: 596). However, the first assumption that there are no unobserved data clearly violates the nature of missing values in WALS, as discussed in Section 2, while the second assumption that all features are binary also violates that the values in the corresponding feature in WALS are mutually exclusive.

Murawaki (2015) handled the missing values in WALS using the imputation method with multiple correspondence analysis. However, statistical imputation needs to fix the number of values in the corresponding feature (e.g., the number of sides of a die in Section 2.2) in the case of WALS data. Thus, the statistical imputation also has the missing values in the linguistic context conform to those in statistics, resulting in the same crucial problem as discussed in Section 3.1.

For example, I applied the statistical imputation method to WALS data, utilizing the missMDA packages (Josse and Husson 2016) and imputeMCA command in R language (R Core Team 2019), the same statistical imputation method using Multiple correspondence analysis (Josse, Chavent, Liquet, and Husson 2012) adopted by Murawaki (2015).

Table 2. Result of statistical imputation by missMDA packages in Ainu, Chukchi, Khalkha, and Navajo. Note that the grey cells represent the imputed value by imputeMCA command.

Name	12A	14A	30A	47A
Ainu	2 Moderately complex	1 No fixed stress	1 None	2 Differentiated
Chukchi	3 Complex	2 Initial	1 None	2 Differentiated
Khalkha	2 Moderately complex	1 No fixed stress	1 None	1 Identical
Navajo	2 Moderately complex	1 No fixed stress	1 None	2 Differentiated

Table 2 demonstrates that the missing values in each language are imputed by the values in the corresponding features; that is, the missing value in Khalkha on 12A: Syllable Structure by 2: Moderately complex, the missing value in Ainu on 14A: Fixed Stress Locations by 1: No fixed stress (mostly weight-sensitive stress), the missing value in Navajo on 30A: Number of Gender by 1: None, the missing value in Ainu on 47A: Intensifiers and Reflexive Pronouns by 2: Differentiated, and the missing value in Chukchi on 47A: Intensifiers and Reflexive Pronouns by 2: Differentiated, respectively. However, as discussed in Section 2, the missing values in each language do not correspond to the values in the corresponding feature but demonstrate that the language possesses some linguistic characteristic that the features in WALS cannot precisely describe with the given values in the corresponding feature.

Thus, the data, imputed by statistical methods, are far from what the linguistic data in WALS represent in their substantive aspect, and are not appropriate for statistical analysis.

Furthermore, Daumé and Campbell (2007), Murawaki and Yamauchi (2018), and Murawaki (2019) utilized the statistical imputation as the performance criteria of their statistical modelling. However, these studies have the same problem as Murawaki (2015). For example, the statistical imputation precludes the missing values in WALS, as explained in Section 2.1, from the domain and range of the value, predicted by the statistical imputation in the corresponding feature.

Table 3 demonstrates the values, predicted by the statistical imputation (i.e., imputeMCA adopted by Murawaki [2015]), in grey cells. Note that the statistical imputation method using Multiple correspondence analysis (Josse et al. 2012) is considered as a statistical learning model in this case and is performed by treating the grey cells as NA.

We observe that the statistical imputation method fails to predict the values in Chukchi (i.e., 12A: Syllable Structure and 14A: Fixed Stress Locations) and succeeds in predicting the other values. This approach adopts, as the performance criteria of their statistical modelling, the ratio of cells, on which the statistical modelling succeeds in the prediction, to the total cells without NA (e.g., accuracy of the modelling).

Table 3. Result of statistical imputation by missMDA packages in Ainu, Chukchi, Khalkha, and Navajo. Note that the grey cells are dealt with as NA and inserted by the value, which is predicted by imputeMCA command. The prediction fails in 12A and 14A in Chukchi.

Name	12A	14A	30A	47A
Ainu	2 Moderately complex	NA	1 None	NA
Chukchi	2 Moderately complex	1 No fixed stress	1 None	NA
Khalkha	NA	1 No fixed stress	1 None	1 Identical
Navajo	2 Moderately complex	1 No fixed stress	NA	2 Differentiated

However, the property whereby the statistical imputation precludes the missing values in the linguistic context from its domain and range raises a question on the fundamentals of these studies as discussed in Section 3.1 (i.e., what the statistical modelling learns from does not correspond to what WALS data represent in the linguistic context). We observed that, as a result, previous statistical typology (e.g., Daumé and Campbell 2007; Daumé 2009; Murawaki 2015; Murawaki and Yamauchi 2018; Murawaki 2019) has not yet responded to the issue of missing values in linguistic typology.

4. Discussion and Conclusions

This section discusses the significance of this paper from the perspectives of both linguistics and statistics. From the viewpoint of linguistics, the main results can be summarized in five points. First, the paper demonstrates the shortcomings of the WALS reference list. As explained in Section 2.1, WALS explains that Baskakov (1966: 107) has written on the Navajo language, but Baskakov (1966: 107) does not exist. Furthermore, the corrected reference (i.e., Azimov, Amansaryev, and Saryev [1966: 107]) covers Turkmen but not Navajo. Therefore, the WALS reference list needs to be thoroughly reviewed.

Second, WALS should add page references for the works in the reference list. The lack of such page references resulted in difficulties when I examined the values in Table 1. Third, WALS should present the reference list and specific pages of references, based on which the value in the corresponding feature is classified as missing. Furthermore, as explained, these missing values are present in the linguistic context but not in the statistical. Again, the former represents that the language possesses some linguistic characteristic that the features in WALS cannot precisely describe, while the latter represents that the language possesses some linguistic characteristic that the features in WALS can describe with the given values in the corresponding feature but that researchers have not yet observed for some reason.

Fourth, WALS should describe the specific content of each missing value from the substantive linguistic knowledge. As discussed in Section 3, the missing values in WALS

do not clarify the different notions in linguistics, which partly results in confusion about statistical typology by information scientists/statisticians.

Finally, researchers need to pay more attention to the problem that even the values in the corresponding feature, as categorized by WALS, are disputed among the specialists on the concerned language. For example, Street (1963: 63) demonstrates that the accent system in Khalkha is related to both stress accent and pitch accent. Also, Skorik (1961: 67-68) and Minoura (1989: 927) demonstrate that the stress locations in Chukchi should not be classified as initial stress or categorized with the values on the concerned feature in WALS. Furthermore, at the 2nd annual conference of The Japan Association of Northern Language Studies, some participants explained that the specialists on each language were also dissatisfied with the data from the viewpoint of their own linguistic knowledge.

These facts suggest a reconsideration of previous statistical typology by information scientists/statisticians, mainly focusing on examination of the dice in statistical modelling approach, to bring new insight to this realm[12].

From the viewpoint of statistics, the missing values in WALS do not correspond to what information science and statistics generally consider “missing values.” Rather, they represent that the language possesses some linguistic characteristic that the features in WALS cannot precisely describe. Thus, the statistical modelling approach does not learn what WALS data represent from the substantive linguistics knowledge but what WALS data conform to in the statistical context. This raises a question on the fundamentals of statistical typology using the statistical modelling approach. Statistical typology should address how to deal with the missing values in WALS using the probability function.

In conclusion, researchers should focus on the well-attested languages in WALS. WALS provides 100 or 200 languages that are selected to avoid areal and genetic biases. To produce a more reliable typological database, researchers should correct the corresponding WALS data as attested by professionals in each language. For example, Japanese and Ainu demonstrate the missing value in 14A: Fixed Stress Locations, which corresponds to the pitch accent system in both languages. Therefore, Japanese and Ainu are more similar in this respect. Note that *Nobori kaku* ‘rising kernel’ and *Sage kaku* ‘lowering kernel’ in pitch accent will matter in this case.

The direction will enable researchers to avoid the problem of missing values caused by the statistical modelling approach, that is, research with probability distribution requires as much WALS data as possible. Furthermore, the attested typological database

[12] Furthermore, one of reviewers indicated that different interpretations could stand on linguistic structures (e.g., phonemes, syllable structures, or accent system) in the same language or dialect. For example, on Khalkha, the syllable structures are still disputed among the specialists on Khalkha. Note that the interested reader should refer to Ueta (2018) in detail. Thus, this topic will also matter in applying statistical methods to WALS data.

has the potential to revive previous statistical typology: research without probability distribution could be improved with more reliable data, and research with probability distribution (i.e., Statistical Modelling Approach) could analyze the data consistent to the linguistic content.

The results indicate the necessity for the humanities and information science/statistics fields to be comprehensible to each other. This will enable each side to respond to the other's findings, with appropriate feedback from substantive knowledge, and will also allow for constructive complementary studies.

Conversely, as this paper demonstrates, any interdisciplinary research without the above conditions will be unfruitful. I end this paper in the hope that interdisciplinary research among the humanities and information science/statistics will be established as a discipline and prove invaluable in the future.

Acknowledgments

Parts of this paper were presented at the 2nd annual conference of The Japan Association of Northern Language Studies, November 9. I am very grateful to the conference organizers and participants for their invaluable questions and comments. Also, I would like to thank the editors and two highly conscientious reviewers for their constructive and invaluable comments; all errors are of course my own.

References

Albu, Mihai (2006) *Quantitative Analyses of Typological Data.* Doctoral Dissertation. Leipzig: Leipzig University. Retrieved from https://pure.mpg.de/rest/items/item_405604/component/file_405603/content

Azimov, Pigam A., Amansaryev Dzumamurad and Saryev K. (1966) Turkmenskij jazyk. In Baskakov, Nikolaj A. (ed.), *Jazyki narodov SSSR. Volume 2: Tjurkskie jazyki,* 91-111. Moscow: Nauka.

Baskakov, Nikolaj A. (1966) Altajskij jazyk. In Baskakov, Nikolaj A. (ed.), *Jazyki narodov SSSR. Volume 2: Tjurkskie jazyki*, 506-522. Moscow: Nauka.

Bickel, Balthasar, Nichols Johanna, Zakharko Taras, Witzlack-Makarevich Alena, Hildebrandt Kristine, Rießler Michael, Bierkandt Lennart, Zúñiga Fernando and Lowe John B. (2017) The AUTOTYP typological databases. Version 0.1.0

Bogoras, Waldemar (2002) Chukchee. *Series: Classics in American Thought, Handbook of American Indian Languages*. *Volume 4, part 2*, 637-903. Bristol: Thoemmes Press. (Reprinted from *Handbook of American Indian Languages*. *Volume 4, part 2*, 637-903, In Boas, Franz [ed.], 1922, Washington: Government Printing Office)

Carstairs-McCarthy, Andrew (2006) Affixation. In Brown, Keith (ed.), *Encyclopedia of Language and Linguistics. Volume 1, Second Edition,* 83-88. Oxford: Elsevier Ltd.

Corbett, Greville G. (2013) Number of Genders. In Dryer, Matthew S. and Haspelmath

Martin (eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology (Available online at http://wals.info/chapter/30, Accessed on 2019-10-17.)

Cysouw, Michael (2007) New Approaches to Cluster Analysis of Typological Indices. *Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday,* 61-76, Berlin: De Gruyter Mouton.

Daumé III, Hal (2009) Non-Parametric Bayesian Areal Linguistics. In Johnston, Michael and Popowich Fred (eds.), *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics,* 593-601. Boulder: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N091067.pdf

Daumé III, Hal and Campbell Lyle (2007) A Bayesian Model for Discovering Typological Implications. In Zaenen, Annie and van den Bosch Antal (eds.), *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics,* 65-72. Prague: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P07-1009.pdf

Dryer, Matthew S. and Haspelmath Martin (eds.) (2013) *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.(Available online at http://wals.info, Accessed on 2019-10-17.)

Dunn, Michael J. (1999) *A Grammar of Chukchi.* Doctoral dissertation. Canberra: The Australian National University. Retrieved from https://openresearch-repository.anu.edu.au/handle/1885/10769/

Dunn, Michael J. (2000) Chukchi Women’s Language: A Historical-Comparative Perspective. *Anthropological Linguistics*, 42(3), 305-328.

Goedemans, Rob and van der Hulst Harry (2013) Fixed Stress Locations. In Dryer, Matthew S. and Haspelmath Martin (eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/14, Accessed on 2019-10-17.)

Itō, Kiyoshi (1951) *Kakuritsu Ron [Probability Theory]*. *Volume 1*. Tokyo: Iwanami Shoten.

Izutsu, Katsunobu (2006) Ainugo no hinshi bunrui saikou [Ainu Parts of Speech Revisited: with Special Reference to So-called Personal Pronouns]. *Journal of Hokkaido University of Education (Humanities and Social Sciences)*, 56(2), 13-27.

Josse, Julie, Chavent Marie, Liquet Benot and Husson François (2012) Handling Missing Values with Regularized Iterative Multiple Correspondence Analysis. *Journal of Classification*, 29(1), 91–116.

Josse, Julie and Husson François (2016) missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *Journal of Statistical Software*, 70(1), 1-31.

Kazama, Shinjiro (2018) Arutai shogengo to chousengo, nihongo niokeru iwayuru “saiki

daimeishi" no taishou kenkyū [A Contrastive Study of the "Reflexive Pronouns" among the Altaic Languages, Korean and Japanese]. *Northern Language Studies*, 8, 23-58.

Kämpfe, Hans-Rainer and Volodin Alexander P. (1995) *Abriß der Tschuktschischen Grammatik auf der Basis der Schriftsprache*. Wiesbaden: Harrassowitz Verlag.

Kobayashi, Miki (2010) Ainugo no doushi settouji si- to yay- no kou doutei kinou [The Function of Argument Identification of Ainu Verbal Prefixes *si-* and *yay-*]. *Studies on Humanities and Social Sciences of Chiba University,* 21, 140-159.

Kolmogorow, Andrei N. (1933) *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Berlin: Springer Verlag.

Kolmogorov, Andrei N. (1950) *Foundations of the Theory of Probability. Second English Edition.* New York: Chelsea Publishing Company.

König, Ekkehard, Siegmund Peter and Töpper Stephan (2013) Intensifiers and Reflexive Pronouns. In Dryer, Matthew S. and Haspelmath Martin (eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/47, Accessed on 2019-10-22.)

Kullmann, Rita and Tserenpil Dandii-Yadam (2005) *Mongolian Grammars,* Hong Kong: Jensco ltd.

Maddieson, Ian (2013) Syllable Structure. In Dryer, Matthew S. and Haspelmath Martin (eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/12, Accessed on 2019-10-17.)

Minoura, Nobukatsu (1989) *Chukuchigo [Chukchee]*. In Kamei, Takashi, Kouno Rokurou and Chino Eiichi (eds.), *The Sanseido Encyclopedia of Linguistics. Volume 2,* 925-932. Tokyo: Sanseidō.

Murawaki, Yugo (2015) Continuous Space Representations of Linguistic Typology and Their Application to Phylogenetic Inference. In Mihalcea, Rada, Chai Joyce and Sarkar Anoop (eds.), *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 324-334. Denver: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N15-1036.pdf

Murawaki, Yugo (2019) Bayesian Learning of Latent Representations of Language Structures. *Computational Linguistics*, 45(2), 199-228.

Murawaki, Yugo and Yamauchi Kenji (2018) A Statistical Model for the Joint Inference of Vertical Stability and Horizontal Diffusibility of Typological Features. *Journal of Language Evolution*, 3(1), 13-25.

Ono, Yohei, Yoshino Ryozo, Hayashi Fumi and Whitman John (2017) A Multiple Correspondence Analysis of the Latent Structure in Linguistic Typology (1): A

Statistical Reanalysis of Tsunoda, Ueda, and Itoh (1995a). *Mathematical Linguistics*, 31(3), 189-204.

Ono, Yohei, Yoshino Ryozo, Hayashi Fumi and Whitman John (2018) A Multiple Correspondence Analysis of the Latent Structure in Linguistic Typology (2): A Statistical Reanalysis of Tsunoda, Ueda, and Itoh (1995a). *Mathematical Linguistics*, 31(4), 261-280.

Orlovskaja, Marija N. (1961) *Imena sushchestvitel’nye i prilagatel’nye v sovremennom mongol’skom jazyke.* Moscow: Izdatel’stvo vostochnoj literatury.

Poppe, Nicholas [Nikolaus] (1951) *Khalkha-Mongolische Grammatik mit Bibliographie, Sprachproben und Glossar.* Wiesbaden: Franz Steiner Verlag GmbH.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. URL: https://www.R-project.org/.

Refsing, Kirsten (1986) *The Ainu language: the morphology and syntax of the Shizunai dialect.* Aarhus: Aarhus University Press.

Sapir, Edward and Hoijer Harry (1967) *The Phonology and Morphology of the Navaho Language.* Berkeley: University of California Press.

Satō, Tomomi (2007) Ainugo chitose hōgen no saiki setsuji yay- to si- ni tsuite [The Reflexive Prefixes yay- and si- in the Chitose Dialect of Ainu]. *Journal of Cognitive Science,* 5, 31-39.

Satō, Tomomi (2008) *Ainugo bunpō no kiso [Fundamentals of Ainu Grammars]*. Tokyo: Daigaku Shorin.

Simeon, George (1969) Hokkaido Ainu phonemics. *Journal of the American Oriental Society*, 89(4), 751-757.

Skorik, Pjotr Ja. (1961) *Grammatika chukotskogo jazyka. Chast’ pervaja. Fonetika i morfologiya imennykh chastey rechi*. Moscow/Leningrad: Akademija Nauk SSSR.

Street, John C. (1963) *Khalkha Structure*. Bloomington: Indiana University Press.

Svantesson, Jan-Olof (2003) Khalkha. In Janhunen, Juha (ed.), *The Mongolic Languages,* 154-176. London: Routledge.

Tamura, Suzuko (2000) *The Ainu Language* (ICHEL Linguistic Studies 29). Tokyo: Sanseidō.

Tsunoda, Tasaku, Ueda Sumie and Itoh Yoshiaki (1995) Adpositions in word-order typology. *Linguistics*, 33(4), 741-762.

Ueta, Naoki (2018) *Mongorugo no boin nikansuru sougouteki kenkyū [Comprehensive Studies on Vowels in Khalkha Mongolian]*. Doctoral dissertation. Kyoto: Kyoto University. Retrieved from https://repository.kulib.kyotou.ac.jp/dspace/ bitstream/ 2433/232172/2/dbunk00767.pdf

Ueta, Naoki (2019) *Mongorugo haruha hōgen no zenkion no onseiteki tokuchou [A Phonetic Characteristic on Preaspiration in Khalkha Mongolian]*. Research presentation at the 2nd annual conference of The Japan Association of Northern

Language Studies, Toyama, 9 November 2019.
Whitman, John and Ono Yohei (2017) Diachronic interpretation of word order cohesion. In Mathieu, Éric and Truswell Robert (eds.), *Micro-change and Macro-change in Diachronic Syntax,* 43-60, Oxford: Oxford University Press.
Whitman, John and Ono Yohei (2020) Applying Multiple Correspondence Analysis to Non-Word Order Typological Features Reveals Areal and Genetic Groupings. Manuscript in preparation.
Young, Robert W. and Morgan William (1972) *The Navaho Language.* Salt Lake City: Deseret Book Company. Retrieved from https://archive.org/details/TheNavaho Language
Young, Robert W. and Morgan William (1980) *The Navajo Language: A Grammar and Colloquial Dictionary.* Albuquerque: University of New Mexico Press.

Summary

There are two main streams in statistical typology: one that learns WALS data with probability distribution, the other that elucidates WALS data without probability distribution. These two streams differ in the following three points: (1) the selection of WALS data in the analysis; (2) the purpose of applying statistical methods to WALS data; (3) the selection of statistical methods based on the previous two points.

This paper focuses on the first stream, called the "statistical modelling approach" in this paper, and discusses whether probability distribution can apply to "missing values" in WALS from the viewpoint of linguistic materials, taking Ainu, Chukchi, Khalkha, and Navajo as examples. The results demonstrate that "missing values" are not dealt with in the context of linguistic materials but conform to statistical notions, which enables information scientists/statisticians to apply probability function and probabilistic modelling. Thus, the statistical modelling approach does not learn what WALS data represent in terms of substantive linguistics knowledge and distorts WALS data in the statistical context. This raises a question regarding the fundamentals of statistical typology with the statistical modelling approach. Statistical typology should primarily address how the missing values in WALS are dealt with using the probability function.

The findings indicate that interdisciplinary research among the humanities and information science/statistics necessitates that information scientists/statisticians explain their research using linguistics concepts and that linguists explain their research using concepts from information science/statistics. This will enable mutual responses from both fields, with appropriate feedback from substantive knowledge, as well as constructive complementary studies.

(linguistics.dialectometry@gmail.com)