The Use of Language Data in Comparative Research: A Note on Blust (2008) and Onvlee (1984)

Klamer, Marian

Squib

The Use of Language Data in Comparative Research: A Note on Blust (2008) and Onvlee (1984)

Marian Klamer

LEIDEN UNIVERSITY

This squib corrects and explain errors in the representation, interpretation, and analysis of Kambera data used in Blust (2008). By highlighting the problems with the Kambera data, some pitfalls in the comparativist's task of using others' descriptions of primary data are identified. The primary source of Blust's Kambera lexical data is Onvlee (1984), a dictionary containing about 6,000 entries. Some background information about this source is given in order to evaluate its usefulness for comparative research. More generally, this squib stresses the crucial importance of including detailed metadata in synchronic linguistic descriptions of primary data, as well as in comparative studies.

1. INTRODUCTION. An article by Robert Blust, entitled "Is there a Bima-Sumba subgroup?" appeared in *Oceanic Linguistics* 47 (1).[1] The main object of that paper was to present the diachronic phonology of Kambera, Hawu, Bimanese, and Manggarai in order to argue that the Bima-Sumba subgroup does not exist and that Kambera and Hawu form a Sumba-Hawu subgroup. I have no quibble with Blust's overall conclusions; and the article provides a very detailed and highly welcome analysis of the historical phonology of Kambera. My focus here is on the Kambera data as they appear in the article. Kambera is a (Central–)Malayo-Polynesian language spoken by about 200,000 people, who occupy about three-quarters of the 12,297 sq. km of Sumba Island. The Kambera data used by Blust come from Onvlee (1984), a Kambera–Dutch dictionary containing approximately 6,000 entries.

The first aim of this squib is to correct and explain errors in the representation, interpretation, and analysis of Kambera data in Blust's comparative study, as these are too numerous to go uncorrected. Also, by highlighting the ways in which the Kambera data were misrepresented, some of the pitfalls in the comparativist's task of using others' descriptions of primary data may be identified. The second and more general aim of this squib is thus to point out the crucial importance of including detailed metadata in synchronic linguistic descriptions of primary data, as well as in comparative studies.

As far as I could see, about 30 percent of the Kambera data cited in Blust (2008) are debatable, or misrepresented. Particularly unreliable are the data on Kambera vowels and their historical changes, for reasons that will be explained in section 2. Also, the informa-

1. I would like to thank two anonymous reviewers for their comments on an earlier draft of this squib.

tion on the morphology of Kambera words and roots appears to confound synchronic and diachronic structure. This issue will be addressed in section 3. Finally, there are differences between the extended Swadesh wordlist compiled by Blust and the one compiled by me during fieldwork that require an explanation. Some explanations are discussed in section 4, which also presents background information about Onvlee (1984). In section 5, I discuss some of the implications that the methodological concerns raised in this paper may have for diachronic and synchronic linguistic studies in general.

2. VOWEL LENGTH. Unlike the other three languages analyzed in Blust (2008)—Hawu, Manggarai, and Bimanese—the Kambera data used in the paper come from a single source, Onvlee (1984), a Kambera–Dutch dictionary containing approximately 6,000 entries. One other source is occasionally consulted (Klamer 1994),[2] but when discrepancies (appear to) exist between the synchronic analysis presented in the latter and the data as presented by Onvlee, Blust takes Onvlee as authoritative. Overall, Blust seems to assume that the orthography used by Onvlee is (a) phonemic, (b) not phonetic, and (c) does not mix phonetic and phonemic representations. Thus, for example, it is assumed that Onvlee's three representations of /a/ as *a, á,* and *à* represent three distinct vowel phonemes, though as the *á* is rare and "irrelevant" it is ignored (Blust 2008:50). Further, it is assumed that any vowel represented with an acute accent in Onvlee (1984) is phonemically long. Neither of these assumptions is explicitly mentioned in the paper, and unfortunately both turn out to be wrong. This has implications for the Kambera vowel changes proposed, in particular for the question on how the long vowels developed (Blust 2008:59–60, 61). The problem is that the primary data on vowel length on which Blust relies are less reliable than assumed, because the orthography used in Onvlee (1984) is not phonemic. Onvlee uses more vowel symbols than there are vowel phonemes, and mixes phonemic and phonetic representations of vowels by using acute accents to represent (a) phonemic vowel length, (b) phonetic (but not phonemic) vowel length, and (c) stress. I will explain this here in some detail, to stress the importance of proper annotation in sources that describe primary data: an inconsistent orthography places heavy restrictions on future research using these sources.

The Kambera vowels and diphthongs are given in table 1. All can occur in the initial, stressed syllable of the (synchronic) root.[3] In unstressed syllables, we only find the cardinal vowels /i a u/. The contrast between /u/ and /u:/ is always quantitative, but the contrast

TABLE 1. KAMBERA VOWELS AND DIPHTHONGS

	Front	Central	Back
high	i (i)		u (u)
	i: (i)		u: (ú)
low	e	a (à)	o
	ai	a: (a)	au

2. An extended and revised version of Klamer (1994), my PhD thesis, was published as Klamer (1998); this is more widely available, but for some unknown reasons not referred to by Blust.
3. Synchronic roots in Kambera are not necessarily isomorphic to diachronic root forms as, for example, the reconstructed Proto–Malayo-Polynesian forms; see section 3 below.

between /a/ and /a:/ and between /i/ and /i:/ may also be realized qualitatively, as a lax/tense distinction.

The fact that long vowels /a:/ and /i:/ may also be realized as tensed (nonlong) vowels, and short vowels may be realized as lax vowels, implies that the allophones of short /a/ are [a] and [ɑ]; short /i/ has allophones [i] and [ɪ]; long /a:/ has allophones [a:] and [a]; and long /i:/ has allophones [i:] and [i]. In other words, the allophones [a] and [i] are *shared* among the tense/long and the lax/short realizations of phonemes /a/ and /i/. These shared allophones are a major source of inconsistency in Onvlee's orthography. He uses three distinct symbols for /a/—*a* = [a] (plain), *à* = [ɑ] (lax/tense), and *á* = [a:] (long)—and as a result it becomes unclear whether an orthographic *a* represents a long/tense or a short/lax phoneme. As a consequence of the fact that Onvlee uses three distinct symbols, Blust assumes that his representations of /a/ (*a, á, à*) represent three distinct vowel phonemes, while in fact there are only two. To add to the confusion, Onvlee also uses acute accents to indicate stress, probably because the phonetic manifestation of stress in Kambera is by higher pitch as well as increased vowel length, and he was influenced by the orthography of Dutch, where acute accents also mark stress. However, using accents for different purposes has the practical consequence that in Onvlee (1984), vowels *with* acute accents can mark phonemically long *or* short vowels, while vowels *without* accent can also represent phonemically long *or* short vowels![4]

Clearly, on the basis of these data, it is virtually impossible to study the origins of synchronically long vowels, for the simple reason that the source does not provide reliable information about which words contain long vowels and which words don't.

Further, on the basis of a source like this, it is impossible to formulate generalizations on the synchronic phonotactics of heavy syllables, or to evaluate phonotactic generalizations made by others. Blust, who as mentioned above apparently assumes that all accented vowels in Onvlee (1984) represent phonemically long vowels, considers orthographic words such as *túya* 'mother's brother', *wúya* 'crocodile', *yíyipu* 'by bits, a little at a time', and *yúyungu* 'jerk, shudder' as evidence to "abandon" (Blust 2008:52) my earlier generalization that a long vowel cannot be followed by a vowel or consonant of equal height in Kambera (Klamer 1998:26). The empirical question is then: do these words indeed contain phonemically long vowels? Before publication, I tested the generalization by checking with Kambera-speakers the pronunciation of these and other words from Onvlee that contain an accented vowel followed by a segment of equal height. It turns out that they can be pronounced as either long or short, and there are no minimal pairs (e.g., **wɯya* versus *wúya*) showing a phonemic long/short contrast; so synchronically, there is no reason to believe these words contain a long vowel phoneme. Since stress in Kambera may be manifested as increased vowel length, it seems more plausible that the accented vowels in the words cited by Blust represent stress, as this is also how stress is represented elsewhere in Onvlee's dictionary.[5]

4. My own work on Kambera maintains a strictly phonemic orthography, using one symbol per vowel phoneme. For more information, see Klamer (1998:13–14, 396 note 18), where it is also noted that previous work on Kambera has not been consistent in the matter of vowel representation, a note of caution that apparently remained unnoticed by Blust.
5. For example, the loan word *rupí* (< Malay *rupiah*), also cited by Blust, appears to have (deviant) stress on the second root syllable. (Is this deviant because of its borrowed status? We do not know, as the word is no longer in use today, being replaced by *rupiah*.)

In sum, Onvlee (1984) is not a reliable source for research on Kambera vowel length or long vowel phonotactics, diachronic or synchronic, because the vowels are not represented with a consistent and phonemic orthography.[6]

3. SYNCHRONIC AND DIACHRONIC STRUCTURE OF WORDS AND ROOTS. In the word lists in the appendix to Blust 2008, some Kambera words have their morpheme boundaries indicated, e.g., *ka-hilu* 'ear', *ha-pui* 'blow'. These morpheme boundaries are (presumably) inferred from the Proto–Malayo-Polynesian (PMP) ancestor form, as they are not given in Onvlee (1984). Because Onvlee's entries do not contain morphological information, synchronic morphological structure is usually not represented in the data cited by Blust either. As a result, many Kambera words that are synchronically complex are listed by Blust as simple words, or as base/root forms. Also, some words are listed as trisyllabic bases with a final phonemic root vowel /u/, while synchronically, the final vowel in these forms would be analyzed as an epenthetic (paragogic, default) vowel, which is only phonetically present. In table 2, some illustrations are given of the discrepancies between Kambera words cited in Blust and how they have been documented and analyzed in Klamer (1998).

The synchronic morphological structure of Kambera words is almost completely isomorphic with their synchronic prosodic structure. For example, in a word like *ha-ŋahu* 'breathe', the synchronic morphological "root" is *ŋahu*, and this entity is identical to the prosodic category "trochaic foot"—a foot with two syllables that has stress on the initial syllable.[7] The synchronic prefix *ha-* is identical to the prosodic category "(pretonic) syllable", a syllable preceding the stressed initial syllable of the foot.[8] Apart from CVCV

TABLE 2. DISCREPANCIES IN WORDS CITED IN BLUST (2008) AND KLAMER (1998)

Meaning	Blust (2008)	Klamer (1998)	Difference
'breathe'	haŋahu	ha-ŋahu	root/base vs. complex form with prefix *ha-*
'cold'	maringu	ma-ringu	root/base vs. complex word with prefix *ma-*
'cloud'	karumaŋu	ka-rumaŋ^{u}	root/base vs. complex form with prefix *ka-*; lexical versus epenthetic vowel [u]
'white'	bàrahu	bàra	root with final syllable *hu* vs. absence of this syllable
'penetrate'	nditikungu	nditik^{u}-ng^{u}	root/base vs. complex word with applicative suffix *-ng* and two epenthetic vowels, one following the root, and one following the suffix.
'bird'	mahawurung	ma-ha-wurung	root/base vs. complex form with prefixes *ma-* and *ha-*
'all'	mbúlu/ndàba	mbu ndàba	two alternative words vs. one compound word

6. Blust (2008:61) observes that a long high front vowel sometimes developed out of a schwa "contrary to what might be expected based on clear evidence of its historical shortness." Unexpected cases like these also suggest that the accented vowels in Onvlee (1984) do not always encode phonemically long vowels.
7. More precisely, a foot/root in Kambera is minimally bimoraic (Klamer 1998:17).
8. There are several such prefixes: apart from *ha-,* there are *pa-, ka-, ma-, la-,* and *ta-*. Observe that this particular syllable always contains a vowel /a/, so we can say that the prefix is just a consonant, and /a/ in this position is a default vowel that creates a pretonic syllable. This is the synchronic pattern that may have forced all "penultimate" vowels to become /a/, a process observed in Blust (2008:60).

roots, there are also words with CVCVC roots. An example is *ka-rumaŋ*^{u} 'cloud', with a root-final consonant *ŋ*. In Kambera, root-final consonants are always followed by the paragogic vowel [u]. The appearance of this paragogic vowel is due to the restriction that Kambera does not allow closed syllables on the surface—that is, all syllables must be (phonetically) open—and CVC syllables are only found lexically. In sum, a morphologically complex word in Kambera is isomorphic to a prosodic word, and prosody is an important clue to morphological structure. (For more discussion, see Van der Hulst and Klamer 1996, 1997, Klamer 1998:30–31).

Investigating the possible positions for heavy syllables (that is, syllables with long vowels), Blust (2008:52) refers to Klamer (1994:19), who states that syllables with long vowels only occur under main stress, and that "main stress is without exception on the initial syllable of the root" (Klamer 1994:25). As counterexamples to this generalization, he then cites words like *mangú* 'grope, feel for s.t. with the hand', *paní* 'flying fox, fruit bat', and *larí* 'young (of female animal)', which appear to have stress on the *second* syllable of the root. But, as discussed in the previous paragraph, the initial syllable of a Kambera root is not necessarily also the initial syllable of a Kambera *word*, since words often have prefixes, and prefixes in Kambera are never stressed.[9] Given the etymology of, for example, *paní* 'flying fox, fruit bat' < PMP *paniki, it is clear that this word is diachronically a simple base, and not a complex word with a prefix. However, diachronic structure is not necessarily identical to synchronic structure. Synchronically, *paní* is analyzed as prefix *pa* + root *ní-,* and as such it *confirms* that stress is on the initial (in this case, the only) syllable of its root. The same applies to *larí* and *mangú.*

In fact, there are quite a few Kambera words whose PMP ancestral forms were morphologically simple trisyllabic words, and whose structure has been morphologically reinterpreted. Interesting examples include *ta-linga* 'ear', *la-yia* (*laqia) 'ginger', and *ta-leli* (*tulali) 'flute'. In the synchronic morphology of Kambera, words like these have been reinterpreted as morphologically complex forms, containing a root and a prefix (that is often semantically empty). They have been assigned a structure that is analogous to morphologically complex words that are productively derived, because their prosodic structure is identical to such words. In other words, any trisyllabic Kambera word with stress on the penultimate syllable is assigned a morphological structure <prefix Ca + root>, *even if* this implies that the morphemes so assigned are not independent meaningful elements.[10] So, even though *talinga* 'ear', *layia* 'ginger', and *taleli* 'flute' are diachronically monomorphemic, they are synchronically (formally) complex, and the fact

9. The fact that Kambera affixation does not affect stress placement is described in Klamer (1998, 1994, chapter 2).

10. Reanalysis or reassignment of morphological structure is one of the possible processes of morphological change, just as reanalysis of syllable structure (e.g., when an original syllable coda is reanalyzed as the onset of the next syllable) is one of the processes of phonotactic change. Often, morphological reanalysis involves the loss of morpheme boundaries, while in the Kambera case, a new boundary is assigned in a word that is originally monomorphemic. The following is an example from Dutch, where originally monomorphemic words are also reanalyzed to have a bimorphemic structure. The words *floppy* 'floppy disk', *puppy* 'puppy', and *guppy* 'guppy' are reanalyzed as containing the Dutch diminutive suffix -[i] and are thus assigned the morphological structure *flop-i pup-i*, *gup-i,* where *-i* is 'Diminutive' (Booij and Van Santen 1998:280). In the Dutch case, a suffix is assigned in formal analogy with a productive suffix; in the Kambera case, prefixes are assigned in formal analogy with productive prefixes.

that prosodic structure is generally isomorphic to morphological structure in Kambera has enabled the reinterpretation of these words. They are put in line with the general and productive word structure where *Ca* prefixes precede the root, which has an initial stressed syllable. The synchronic derivational morphology of Kambera, including words that are "formally" complex, is discussed in detail in Klamer (1998:178–273).[11]

In the data cited, Blust (2008) appears to assume that Onvlee (1984) follows standard lexicographic conventions, in that entries are monomorphemic base forms, have no prefixes or suffixes (unless indicated), and are not compounds (unless stated explicitly). This implies that the initial letter of an entry in the dictionary is considered as identical to the initial segment of a root/base. So, if Blust investigates where stress is placed in a Kambera word, he applies the rule "stress the initial syllable of the root" to the initial syllable of any entry in the dictionary, and then (of course) observes that the stress pattern of many words in Onvlee (1984) does not seem to conform to this rule. However, as pointed out, his counterexamples are only apparent, because many of Onvlee's entries are prefixed words, not roots.

Blust's assumption that all Onvlee's entries are bases/roots also explains his claim that "24 percent of the entries in Onvlee" are "*bases* that begin with *p* or *b*" (Blust 2008:51; my italics). In fact, the dictionary gives both bases *and* derived words as entries without representing the morpheme boundaries that distinguish them. As a result, words with a prefix *pa-* sit side-by-side with forms with a root-initial syllable *pa* without morpheme boundaries indicated.

The dictionary has 19 pages with entries starting with *pa,* among which are roots whose initial syllable is *pa*, though most of these entries are morphologically complex items with a prefix *pa-*. Note that dictionary entries starting with *pe, pi, pu,* or *po* are all root forms (because Kambera has no prefixes **pe-, *pi-, *pu-,* or **po-*) and these four types together make up 16 pages of the dictionary, against 19 with just *pa*. Words with other prefixes show even more heavily skewed patterns: there are 54 pages of entries starting with *ma* (among which are many words with prefix *ma-*), while entries with initial *mi, mo, me,* and *mu* (not prefixes) together fill only 10 pages. Entries with initial *ka* (among which are many words with prefix *ka-*) fill 85 pages, but words with initial *ki, ko, ke, ku* (not prefixes) only fill 15 pages. Initial *ta* fills 37 pages (among which are many items with prefix *ta-*), but words with initial *ti, to, te, tu* fill 20 pages. Initial *ha* fills 40 pages (as many words have prefix *ha-)*, while words with initial *hi, ho, he, hu* fill only 17 pages.

11. Also related to Kambera morphology, Blust (2008) suggests that the prenasalized consonants in Kambera are the product of prefixation with *ma- 'stative' followed by syncope of the vowel of the prefix (*ma-panas > *mbanahu* 'hot'), and that this process is generally confined to bases with an initial labial stop, so that it may have been motivated by an inherited constraint against dissimilar labials separated by a single vowel. The major problem with this suggestion is that it fails to explain how Kambera developed the prenasalized segments that do not involve a labial (*nj*, *nd*, *ny*, *ŋg*).

Further, note that PMP *ma- has various synchronic reflexes in Kambera: (i) the clitic *ma=*, which marks subjective and possessor relative clauses and nominalizations (Klamer 1998:261, 318–21), (ii) the prefix *ma-* found in nominal and verbal derivations (Klamer 1998:261–62), and (iii) the (unproductive) nasal prefix that derived intransitive achievement verbs from transitive ones (Klamer 1998:262–65). How these reflexes might be connected to a protoprefix that also is supposed to have created the prenasalized consonants (or, better, one of them) remains to be investigated.

These striking differences in frequency between words starting with *Ca* on the one hand and *Ci/o/u/e* on the other can be observed by simply glancing through the dictionary, and cry out for an explanation if indeed all entries had been bases, as Blust seems to assume.

The assumption that Onvlee's entries are (monomorphemic) roots also explains why Blust (2008:52–53) attests so many entries that are CVCVCV roots and yet do not end in [u], in apparent contradiction to Klamer's (1998:17–18) "odd" claim that roots like *rimuna* and *puita* do not exist in Kambera. Among the words Blust cites as counterevidence are *ka-lau, ka-pita, ka-reni, ka-wana, la-ngoda, la-yia, ma-hàna, ma-nila, pa-tola, ta-leli,* and *ta-nai,* etcetera. All of these are apparently considered to be roots, because they are presented as counterexamples to a claim about root structure. But, in fact, all of them have a (synchronic) prefix, and hence exemplify disyllabic roots with CVCV structure, not trisyllabic CVCVCV roots.

Blust further notes that a few words have a trisyllabic PMP ancestral form while other forms reflect prefixed PMP forms, but adds that "there is no evidence in Onvlee that they still contain a morpheme boundary" (Blust 2008:53). However, as I have pointed out, the fact that Onvlee does not indicate morphological boundaries in his dictionary does not mean these boundaries do not exist. The words cited by Blust as counterexamples to a generalization about synchronic Kambera root structure are in fact synchronically prefixed forms, so that none of them constitutes counterevidence to my "odd claim"; rather, all of them confirm it.[12] Synchronically, Kambera does not have CVCVCV roots like *rimuna* and *puita.* This is not a strange or unexpected pattern, as "trisyllabic bases are rare … in most Austronesian languages" (Blust 2008:82).

Related to the claim that, synchronically, Kambera has trisyllabic root forms is Blust's remark that it is "difficult to see what synchronic evidence there is for treating *-u* … as a paragogic vowel," and that in Klamer (1994) "no evidence is given to support the claim that the 'paragogic u' is nonphonemic, apart from its insertion in loanwords" (Blust 2008:53). Yet, this latter argument is only one of five pieces of synchronic evidence presented in Klamer (1998 [1994], chapter 2) where I argue that these "epenthetic" syllables do not play any role in the prosodic structure of Kambera.[13] The remaining four are as follows. First, the presence of a syllable with a paragogic vowel does not alter the stress pattern: stress remains where it is, on the initial syllable of the root (*akat* [ˈakatᵘ] 'be bad'). Second, in a word game that involves permutations of the final syllable to the beginning of the word, a sequence of two syllables, one of which contains epenthetic [u], count together as a single syllable: *àulung(u)* 'snatch away' [ˈaw | lu | ŋᵘ] > [**lu** | **ŋu** | ˈaw]), while this is not the case for syllables with a lexical vowel /u/: *ka-modu* 'yesterday' [ka | ˈmo | du] > [ˌ**du** | ka | ˈmo:], not *[**mo** | **du** | ˈka:] (Klamer 1998: 32–33). Third, epenthetic [u] may be added productively and iteratively in a single word. For example, the root form *uhuk* 'sit' is pronounced as [ˈu | hu | kᵘ] 'sit', with one epenthetic vowel; with an applicative suffix *-ŋ* it becomes *ukuk-ŋ* 'sit on something/someone', which is pronounced as [ˈu | hu | kᵘ | ŋᵘ], with two epenthetic vowels (Klamer 1998:20). Fourth, the syllable with an epenthetic vowel does not take part in foot reduplication. Examples of foot reduplication

12. It would indeed have been quite "odd" to investigate a language for several years and then come up with a generalization to which hundreds of counterexamples could be found by reading through Onvlee (1984) for a few minutes.

13. See also Van der Hulst and Klamer (1996, 1997) and Klamer (2002, 2005).

are *rama* 'work' > *rama-rama,* or *kaunda* 'stalk away'> *kaunda-kaunda.* In the reduplication of *wunang^{u}* 'priest'> *wuna-wunang^{u}* (**wunang^{u}-wunang^{u}*), the syllable with the epenthetic vowel is not reduplicated, and hence is considered not to be part of the foot (cf. Klamer 1998:37–38). These five reasons brought me to analyze the vowel *u* in phonetically trisyllabic roots as an epenthetic vowel that is distinct from the other, lexical, vowels. The trisyllabic roots listed in Onvlee (1984) are all phonologically disyllabic; all of them end with an epenthetic vowel [u] and have initial stress. Generally, Onvlee represents the epenthetic vowel as identical to a lexical vowel /u/; only occasionally is it put between brackets. This is unfortunate. An orthography that mixes phonetics and phonemics in this way is a source of confusion, and becomes an obstacle for later research on the phonemic and morphological analysis of Kambera entries, as Blust (2008) illustrates.

4. LEXICAL DIFFERENCES. Apart from the errors in representation and interpretation of the Kambera data discussed in the previous sections, there are also lexical differences between the Kambera data used in Blust (2008) and my own Kambera field notes. Of the 200 Kambera words in the extended Swadesh list as given in the appendix of Blust (2008), 17 have different roots in my files. Some illustrations are given in table 3.

The 17 different word forms do not have serious consequences for the overall pattern of similarity between Kambera and the other three languages in the comparison, nor do they change the relation between PMP and Kambera. Some cognate words in Blust's data set are not cognates in mine, others are not cognate in his set, but cognate in mine, but they generally outnumber each other. Two differences between Blust (2008) and my own data that are worth mentioning are that Kambera/Hawu has 67/196 (34.2%) cognates (instead of 70/196 or 35.7 percent in Blust), and that Kambera/Tetun has 58/200 (29 percent) cognates (instead of 63/200 or 31.5 percent in Blust). While these figures suggest fewer links between Kambera and Tetun or Hawu than Blust suggests, they do not change the validity of his proposal to reject the "Sumba-Bima" group, nor do they affect his suggestion of a Sumba-Hawu subgroup.

TABLE 3. DIFFERENT KAMBERA WORDS IN BLUST (2008) AND MY FIELDNOTES

Blust's (2008) appendix	My fieldnotes	Meaning	Comment
tamu	ŋara	'name'	
kawiŋgu	ka-jia	'back'	
diŋi	kei	'buy'	*diŋi* 'request', *kei* 'buy'
kadipu	kau	'cut'	*ka-dipu* attested as root in *ha-ka-dipu* 'a piece'
mbàlaru	ka-longga, ma-lau	'wide'	*kalongga* 'wide, spacious', *ma-lau* 'with wide opening'
tímbi	ma-nàba	'thick'	
yia	ni-na	'this'	*yia* < *ye* in *ye-na* 'this one'
kànja	wà-ngu	'say'	
alahu	omang	'forest'	
kadumba	ka-nduba	'dull'	

However, for future research that intends to use Onvlee (1984) as a source of primary data, it is important to investigate why the basic wordlists compiled by Blust and myself turn out to be different. One explanation may be sought in the fact that Onvlee (1984) represents different speech styles and genres, and in particular combines archaic, ritualistic, and colloquial words. Such data are important to historical-comparative research. But a comparative study based on such data would want to include information on their nature, thus making more explicit the basis of the diachronic analysis that is proposed. Onvlee (1984) does not contain much information on why, when, and how it was compiled, or on which varieties and genres were included and which ones were not.

The introduction to Onvlee (1984) is only seven pages, of which four pages describe orthography, pronunciation, the structure of the dictionary entries, and word derivations, while three pages provide sketchy morphosyntactic and lexical information. Following the Kambera–Dutch dictionary, a 59-page Dutch–Kambera finderlist that contains approximately 4,800 words is provided (compiled by P. J. Luijendijk, a missionary who had been posted on Sumba and knew Kambera). Most of the entries in the finderlist refer to several (usually more than two, sometimes more than 10) different Kambera words, without explanation of their similarities or differences. In order to establish semantic contrasts, the reader must look up all the words under one entry to see how each of them translates into Dutch, and how they are used in the example sentences. These example sentences can only be understood by readers with a fair knowledge of Kambera lexicon and morphosyntax.

In Blust (2008) it is not explained why certain words were selected from Onvlee (1984) to become part of the Swadesh basic wordlist, over others that are given as more or less synonymous in the finderlist. For example, Onvlee's finderlist gives two words for *naam* 'name': *ngara* and *tamu*. *Tamu* is selected for the Swadesh list, not *ngara*, and the question is: why? Interestingly, the word *ngara* is the one generally used for 'name' in my field site. Similarly, the finderlist gives three words for *kopen* 'buy': *dingi, kadingi,* and *kei.* What determined the selection of *dingi* for the item 'buy' in the Swadesh list? In my fieldnotes and in Onvlee, *kei* translates as 'buy', *dingi* as 'to request', so what determined the selection of *dingi* for Blust? The paper does not contain information on the motivation of these lexical choices and the procedures that were applied in the selection of items. Such information is relevant, as comparative work like Blust (2008) often becomes a secondary source of language data for further research that needs to be able to evaluate how, and why, the data were selected.

Having worked with Onvlee (1984) for about a decade, I found it an invaluable resource on Kambera, as acknowledged in Klamer (1998:4). However, many entries have to remain a mystery, because speakers do not recognize them as part of their language (anymore?): this was the case for approximately one out of five items I checked with speakers. I am unsure what linguistic researchers (diachronic or synchronic) can do with words that are not recognized by a community of speakers, except to put them to one side as "maybe from another dialect, maybe obsolete, maybe an error."

In the evaluation of the nature of the primary data in Onvlee (1984), it is important to consider the sociohistorical context in which the dictionary was compiled. The work itself provides only very sketchy, if any, information on this, so I will elaborate on it here. Dr.

Louis Onvlee (1893–1986) worked on the island of Sumba as a Dutch missionary and Bible translator from 1926 through 1947 and again 1951–55. In 1926 he first arrived on Sumba, and after spending a few years in Waikabubak in West Sumba, he moved to East Sumba in 1929 to start work on the Kambera language with his local language assistant Umbu (previously spelled as Oemboe) Hina Kapita. After his return to the Netherlands in 1955, he was offered a professorship in Cultural Anthropology at the Vrije Universiteit te Amsterdam, from which he retired a few years later. Apart from the dictionary, published two years before his death, Onvlee also published a grammar of Kambera in 1925. This work was written before Onvlee set foot on Sumba, and is entirely based on earlier written sources about Kambera. It contains 76 pages of grammatical notes and 131 pages of translated (but not glossed) Kambera texts. Besides this grammar, Onvlee published a few short papers on particular topics in Kambera (Onvlee 1927, 1936a–d, 1950). His linguistic publications remained limited in number, probably because his employer (the Dutch Bible Society) had assigned him the tasks of translating the Bible in Weyewa and Kambera and developing literacy programs in the local languages. Onvlee also wrote two pedagogical grammars that were used in the 1950s to teach Kambera and Weyewa to the Dutch missionaries who came to the island then (Onvlee n.d.a, n.d.b). His Sumbanese language assistant, Umbu Hina Kapita, published a 296-page Kambera–Indonesian dictionary (Kapita 1982), a short 90-page grammar of Kambera in Indonesian (Kapita 1983), and several books with Kambera traditional ritual texts, songs, stories, and sayings (Kapita 1977, 1979, 1985, 1987). The latter four books, together with the Kambera New Testament (1961) and a Kambera Hymn book, make up the Kambera written literature.

In the introduction to the Kambera dictionary, Onvlee remarks that during World War II, his own data collection on Kambera had been destroyed, but that the copies of his consultant Kapita had survived the war. From this I infer that it is Kapita's collection that formed the basis of Onvlee (1984). Although the dictionary does not inform us about it, checking the data with local speakers seems to indicate that it reflects mostly the Kambera language as spoken in the area around the capital (port) town Waingapu. The data date roughly from the 1920s till the second half of the 1950s, as Onvlee returned to the Netherlands at that time, and Kapita became occupied elsewhere. Of course, later decades must have brought additions, but the bulk of the material must have been collected between 1920 and 1950. In other words, although the book has 1984 as its year of publication, it contains data collected two or three generations earlier.

This explains, for example, why the dictionary contains relatively few Malay loan words, and also why these loans are adapted into Kambera in ways that are different from the way Indonesian loans are incorporated into the language today. In the context of borrowing from Malay, Blust (2008:48) notes that "the great majority" [of loanwords] appear to have entered "the local languages … centuries" before the foundation of the Republic of Indonesia. For Kambera, this is probably not correct: all evidence points to a limited Malay influence on the island until the arrival of the Dutch administration at the end of the nineteenth century. For example, Kontroleur S. Roos (1872:1) observed that outside of the "kampong" of Waing-apoe (the later capital Waingapu) "op Soemba, geen Maleisch ... wordt gesproken" ("on Sumba, no Malay ... is spoken"). (See also the historical overview in Forth 1981.) On Sumba, the major influx of Indonesian loans started in

the mid-1970s, when primary education in Indonesian became available at the village level throughout the island.[14] It might also explain why the meanings of some words are rather different from their meaning today: since the time they were collected two or three generations ago, the semantics of words may have shifted. An example of such a shift could be *riu,* which is translated as 'ten thousand' in Onvlee (1984) but is nowadays used to mean 'thousand', perhaps by analogy with Indonesian *ribu*.

Onvlee is a collection of words from different Kambera dialects. As has been mentioned, the majority seem to come from the dialect spoken in the Kambera region, in and around Waingapu, where the main port of Sumba is located and where Kapita and Onvlee resided. However, consultants informed me that there are also many entries in the book from other dialects, for example, from the dialect of Mangili, spoken on the far eastern tip of the island. In general, we cannot be sure about any of the items' origins, because the regions where they have been collected were not systematically catalogued.

The dictionary also contains words from different speech styles and genres. As evidenced by his publications, Kapita must have been fascinated by Kambera ritual speech, songs, stories, and sayings. (When I visited him as a student in 1988, his main interest was in how linguists should document the rich oral traditions of Kambera.) In the dictionary, everyday colloquial words sit next to the literary and archaic words used in prayers, ceremonial offerings, songs, and ancestor narratives. Moreover, as Kambera ritual speech is characterized, among other things, by lexical parallelism (documented in Kapita 1987), which is neither used nor understood by common speakers, many parallel words are included in the dictionary that are only known, used, and understood by a selective few speakers. Unfortunately the dictionary does not indicate whether an entry is colloquial or not, archaic or current, obsolete or still in use, if it is a semantically empty element only used in parallelisms, or one that has a meaning known by a few people. (Sometimes, however, the genres can be reconstructed by looking at the example sentences, for which, as mentioned, a fair knowledge of Kambera grammar and lexicon is a prerequisite.)

As a result of this mixed content, the Kambera words used in Blust are also a mix of unknown words (archaic or obsolete, or from an unknown dialect), as well as colloquial words that were in use between 1930 and 1960. Some items have shifted their meaning since they were collected by Onvlee two or three generations ago, while some have changed their form (for example, some have lost their consonantal suffixes *-k* or *-ng*). In contrast, my data are more homogeneous, being collected from native speakers living in one village in the 1990s, and containing words from everyday language as spoken in that period only. This is another explanation for the lexical differences between the data in Blust (2008) and mine.

5. DISCUSSION. Comparative studies commonly work with data reported in word lists or dictionaries that have been collected by others and not by the comparativist. Such word lists often turn out to be highly heterogeneous, and contain rather "noisy" data. In

14. Blust (2008:51) also claims that some of the loans cited in Onvlee (1984) contradict the generalization in Klamer (1998:11, 1994:13) that "in loan substitutions, /b, d, g, ʤ/ are always prenasalized." These works describe how Indonesian loans are being adapted into the Kambera system in the 1990s, when prenasalization is systematic. In contrast, Onvlee (1984) presents data from earlier times (1930–60), and the process may have been less systematic then.

order to evaluate the reliability and comparability of word lists, questions like the following may be asked: When were the lists collected: recently, or several generations ago? Who were the compilers: missionaries, anthropologists, teachers, linguists, or local speakers? What kind of data did they compile: spoken data or data from written sources? Do the items in the word list represent one ideolect, or are they from different speakers? If different speakers contributed, did they speak the same dialect? Does the compilation represent colloquial language, or does it also include items from other registers? Does it include only words as they are used today, or also archaic or obsolete words? And last but not least: how does the orthography used in the compilation relate to the phonemic and phonetic representation of the items? Is the orthography based on the collector's native language, on the national language, or a "colonial" language? Not asking questions like these explicitly is one of the pitfalls of comparative work, as primary data may be misinterpreted, misrepresented, and analyzed erroneously, and apples may be compared with pears.

In this squib I have argued that it is useful to address such issues for *every* language in a comparative study, and why I believe that these questions have not been adequately addressed by Blust for the Kambera data he cites. These contain many errors and misrepresented facts that could have been avoided if Blust had not relied on a single source, and had made an effort to check whether the assumptions he had about that source were actually correct. For efficiency's sake, comparative researchers obviously have to make choices about the data they use and the sources they consult. However, if, as Blust noted for Kambera, some primary data appear to be problematic or contradictory, it is never a bad idea to consult additional published sources on the language,[15] to ask native speakers what they think, or to check with language experts. Especially before publishing results that can become secondary sources on a language, an effort must be made to present data that are as reliable as possible.

It is also crucial to include with a comparison very explicit information on how the data used for that comparison were compiled, and which assumptions steered the research and use of a particular source. If such information is lacking, there is no way future research can build further on the comparison fruitfully.

Of course, this in turn implies that the primary data source should also contain explicit information on all these questions and issues, that it should have a consistent and phonemic orthography, and that it must comply with standard lexicographic conventions in providing unambiguous information on morphological structure and stress placement. In this squib I have shown how Onvlee (1984) fails on many of these critical points, as I have discovered working with this book for many years. This is important information to share, because it helps future researchers to see its restrictions.[16] While Onvlee (1984) is a tremendously rich source, the data must be approached with caution, especially if they are used for phonological and morphological research. In such cases, the data should preferably be double-checked with native speakers.

15. Other existing sources on Kambera include Wielenga (1909, 1913), Onvlee (1925, 1950, n.d.), Klamer (1998, 2002, 2005), and the nineteenth-century sources given in the list of references.
16. For this reason I now regret that I did not include a review of Onvlee (1984) in my Kambera grammar. But at the time of writing the grammar it did not feel "right" for me to include a critical evaluation of the final publication of someone whose deep knowledge of the language and culture deserves so much respect.

REFERENCES

Blust, Robert. 2008. Is there a Sumba-Bima subgroup? *Oceanic Linguistics* 47:45–113.

Booij, Geert, and Ariane van Santen. 1998. *Morfologie: de woordstructuur van het Nederlands.* Amsterdam: Amsterdam University Press.

De Roo van Alderwerelt, J. 1891. Soembaneesch–Hollandsche woordenlijst met een schets eener grammatika. *Tijdschrift voor Indische Taal-, Land- en Volkenkunde* 34:234–82.

Forth, G. 1981. *Rindi: An ethnographic study of traditional domain in Eastern Sumba,* The Hague: Martinus Nijhoff.

Heijmering, G. 1846. Bijdrage tot de kennis van de taal der Z. W. Eilanden, benevens een proeve van vergelijking derzelve met acht andere inlandsche talen. *Tijdschrift voor Nederlandsch-Indië* 8(3):1–81.

Kapita, Oe. H. 1976. *Sumba di dalam jankauan jaman*. Waingapu: Gereja Kristen Sumba.

———. 1977. *Ludu Humba: Pakangutuna*. Waingapu: Gereja Kristen Sumba.

———. 1979. *Lii Ndai: Rukuda da Kabihu dangu la Pahunga Lodu (Sejara Suku-suku di Sumba Timur)*. Waingapu: Gereja Kristen Sumba.

———. 1982. *Kamus Sumba/Kambera-Indonesia*. Waingapu: Gereja Kristen Sumba.

———. 1983. *Tatabahasa Sumba Timur dalam dialek Kambera*. Ende: Arnoldus.

———. 1986. *Pamangu ndewa (Perjamuan dewa)*. Ende: Arnoldus.

———. 1987. *Lawiti luluku Humba (Pola peribahasa Sumba)*. Ende: Arnoldus.

Kambera Hymn Book [*Ludu Pamalangu*]. 1979. Ende: Percetakan Offset Arnoldus.

Kambera New Testament [*Na Paràndingu Bidi*]. 1961. Ende: Percetakan Offset Arnoldus.

Klamer, Marian. 1994. *Kambera: A language of Eastern Indonesia.* PhD diss., Vrije Universiteit Amsterdam. The Hague: Holland Academic Graphics.

———. 1998. *A grammar of Kambera.* Berlin/New York: Mouton de Gruyter.

———. 2002. Semantically motivated lexical patterns: A study of Dutch and Kambera expressives. *Language* 78(2):258–86.

———. 2005. Kambera. In *The Austronesian languages of Asia and Madagascar,* ed. by K. Alexander Adelaar and Nikolaus P. Himmelmann, 709–34. London: Routledge.

Onvlee, Louis. 1925. *Eenige Soembaasche Vertellingen*. Leiden: Brill.

———. 1950. Over de mediae in het Soembanees en Sawoenees. In *Bingkisan budi: een bundel opstellen aan Dr Philippus Samuel van Ronkel door vrienden en leerlingen aangeboden op zijn tachtigste verjaardag 1 augustus 1950*, [no editor listed], 215–34. Leiden: Sijthoff.

———. 1984. *Kamberaas (Oost-Soembaas)–Nederlands Woordenboek.* Dordrecht: Foris.

———. N.d.a. *Lessen Kamberaas.* Unpublished typescript.

———. N.d.b. *Lessen Weweas.* Unpublished typescript.

Pos, W. 1901. Sumbaneesche woordenlijst. *Bijdragen tot de Taal-, Land- en Volkenkunde van Nederlandsch-Indië* 58:184–284.

Roos, S. 1872. Bijdrage tot de kennis van taal, land en volk op het eiland Sumba. *Verhandelingen van het Bataviaasch Genootschap van Kunsten en Wetenschappen* 36:1–125.

Sneddon, J. N. 1993. The drift towards final open syllables in Sulawesi languages. *Oceanic Linguistics* 32:1–44.

Van der Hulst, Harry, and Marian Klamer. 1996. The uneven trochee and the structure of Kambera roots. In *Dam Phonology,* ed. by M. Nespor and N. Smith, 39–57, The Hague: Holland Academic Graphics.

———. 1997. The prosodic structure of Kambera roots and words. In *Proceedings of the Seventh International Conference on Austronesian Linguistics,* ed. by C. Odé and W. Stokhof, 105–23, Amsterdam: Rodopi.

Vermast, A. M. 1895. Lijst van Soembaneesche woorden en uitdrukkingen, alphabetisch gerangschikt. *Veeartsenijkundige Bladen voor Nederlandsch-Indië* 9:122–42.

Wielenga, D. K. 1909. *Schets van een Soembaneesche spraakkunst (naar 't dialect van Kambera)*. Batavia: Landsdrukkerij.
———. 1913. Soembaneesche verhalen in het dialect van Kambera, met vertaling en aanteekeningen. *Bijdragen tot de Taal-, Land- en Volkenkunde van Nederlandsch Indië* 68:1–287.
———. 1917. Vergelijkende woordenlijst der verschillende dialecten op het eiland Soemba en eenige Soembaneesche spreekwijzen. *Verhandelingen van het Bataviaasch Genootschap van Kunsten en Wetenschappen* 61:1–96.

M.A.F.Klamer@hum.leidenuniv.nl