Sora Combining Forms and Pseudo-compounding[1]

Stanley STAROSTA
University of Hawaii

Abstract

This paper is a lexicase[2] analysis of a Sora word-formation process which will be referred to as 'pseudo-compounding'. The paper provides evidence and arguments in support of three claims:

1) Sora nominal 'compounds' must be treated as cases of lexical derivation, rather than as transformational or lexical compounding;

2) The shorter 'Combining Forms' of nouns are in most cases of greater antiquity than the disyllabic 'Free Forms'; and

3) The Combining Forms must be listed in the dictionary of a generative grammar, rather than being generated on demand by productive grammatical rules.

Comparative evidence from Khmer is adduced in support of the second claim, and data from word formation in Mandarin Chinese is referred to in order to demon-strate the naturalness and external applicability of the formal analysis adopted in this paper.

Contents

1 Introduction
    1.1 Sora
    1.2 Lexicase
2 Sora pseudo-compounding
    2.1 Pseudo-compounding and lexicase
    2.2 Sora pseudo-compounding analyzed
3 The historical priority of the Combining Form
    3.1 External cognates
    3.2 Possessive pseudo-compounds
    3.3 The conspiracy theory
4 Formal analysis
    4.1 Transformational solutions
    4.2 Lexical solutions and lexicase

---

[1]This paper is a revision of a paper written while I was a visiting researcher at the Institut für deutsche Sprache in Mannheim, Germany, under a fellowship from the Alexander von Humboldt Foundation. It is an expansion of the final section of my 'Sora nouns: possession, compounding, and the lexical status of Combining Forms', presented at the Second International Congress on Austroasiatic Linguistics, Central Institute of Indian Languages, Mysore, 1978. The first part of that paper, not including the final section on compounding and Combining Forms, will be published in the proceedings of the Congress as 'Sora noun inflection'.

[2]cf.Starosta 1988.

5 The long-to short-analysis
5.1 Analogy and the direction of derivation
5.2 Truncation rules
Appendix: Mandarin Chinese pseudo-compounds

1 Introduction

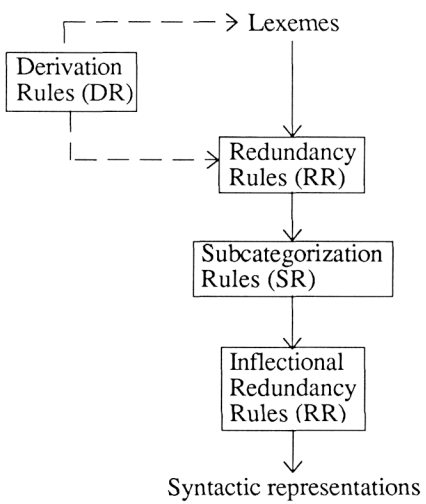
1.1 Sora

The Sora language is a member of the Southern Munda subgroup of the Munda branch of the Austroasiatic language family. It is spoken by about 300,000 people in the border district of Andhra Pradesh and Orissa in India. The first linguistic description of Sora was G.V. Ramamurti's *A manual of the Sora (or Savara) language*, published in 1931. Subsequent work includes a generative syntactic description (Starosta 1967), and more restricted studies of phonology (Stampe 1963), verbs (Biligiri 1965), verb derivation and case (Starosta 1971, 1976), noun inflection and classification (Starosta forth.), and nominal Combining Forms (Zide 1976).

1.2 Lexicase

Lexicase is a generative but non-transformational grammatical model which was developed at the University of Hawaii, and has been applied in the description of a number of primarily Asian and Pacific languages since 1970. In its current conception (cf. Starosta 1978, 1979a, 1988), it is composed of the following components:

1-1.



In the lexicase framework, words are pivotal. Morphology is the study of the internal structure of words, and syntax is an account of the distributions of words in sentences. Grammar is the set of all general statements that can be made about the internal structure and external distribution of words in sentences, and the lexicon is a list of lexemes, words which have been stripped of all aspects of their sound-meaning-form correspondences which can be predicted by means of redundancy and subcategorization rules. Syntactic representations are hierarchically structured strings of fully specified words in which every construction has at least one lexical head, and in which no contextual feature on a word is violated by a sisterhead (cf. Starosta 1979a).

The most important rules for the purposes of this paper are the Derivation Rules (DR's). These are the patterns by which a language adds new words to its lexicon based on analogies with words already present. These rules differ from the subcategorization and redundancy rules in that the items generated by the Derivation Rule component are potential rather than actual words (cf. Halle 1973). This distinction between potential words and actual words is characterized in the grammar by listing all the actual words in the lexicon. That is, each application of a DR is a separate historical event, and its result is recorded in the lexicon. Thus DR-0 is a Derivation Rule of English that states that manner adverbs can be formed from adjectives by adding the suffix *-ly:*

1-2. DR-0

$$\begin{bmatrix} +\text{Adj} \\ \alpha\text{Fi} \end{bmatrix} \longrightarrow \begin{bmatrix} +\text{Adv} \\ +\text{manr} \\ \alpha\text{Fi} \end{bmatrix}$$

$$\text{]} \longrightarrow \text{ly]}$$

However, the only way we can tell that *freshly* is an English manner adverb and **stalely* isn't but could be, is to look in the lexicon and find an entry for *freshly* but not for **stalely*.

Except for completely productive cases, DR's only approximate the relationship between lexical sets. This is because, first of all, there may be gaps on either side: there will normally be many words to which the rule has not yet applied, such as the example of *stale* above, which has no corresponding manner adverb **stalely,* at least in my lexicon; and there may be some derived forms whose historical sources have been lost from the lexicon. (The famous example of *aggression* versus **aggress* may be an instance of this situation.) Second, the actual semantic representation of the derived form is established at the point of derivation, in accordance with requirements of the situation for which it is coined, and may differ somewhat from the exact prediction of the DR. Thus *dully* for me can refer to emotions, but the source adjective *dull* can't. And since new forms are full-fledged independent lexical entries once they have been derived, they and their sources are subject to independent semantic and phonological shifts which may increase the distance between them, sometimes to the point of non-recognition of their relatedness.

In addition to serving as the pattern by which new words are formed, DR's have a parallel function: they are the part of the synchronic grammar which

characterizes the speaker's knowledge that the lexicon contains two sets of words which are related to each other in a certain way. Thus DR-0 accounts for the fact that when a speaker encounters an Adverb which ends in *-ly* and which is marked for the semantic feature [+manr], she expects to find another word in the lexicon which is a member of the class of adjectives and which differs from the Adverb semantically and phonologically only in that it lacks the semantic feature [+manr] and the *-ly* ending.

2. Sora pseudo-compounding

2.1 Pseudo-compounding and lexicase

This paper presupposes the analysis of Sora noun classes and inflectional categories proposed in *Sora noun inflection* (Starosta forthcoming). It provides evidence and arguments in support of two claims:

1) Sora nominal 'compounds' must be treated as cases of lexical derivation, rather than as transformational or lexical compounding; and

2) Sora 'Combining Forms' must be individually listed in the derivational component of the lexicon.

The paper considers and rejects a possible alternative approach, the derivation of Combining Forms from Free Forms by productive grammatical rules (cf. Zide 1976), and demonstrates how the lexicase framework forces the adoption of a lexical approach which, although initially somewhat counterintuitive, turns out to be historically and synchronically well motivated. This in turn provides empirical confirmation for the strongly constrained lexicase model.

2.2 Sora pseudo-compounds analyzed

One way to 'modify' a noun in Sora, that is to restrict the range of possible reference, is to mark it as possessed. This signals to the hearer that the intended referents are not all the things that are compatible with the semantic definition of the noun, but rather a restricted subset of of these referents, the ones that are placed in abstract juxtaposition with the 'possessor,' the external entity which can be optionally specified as a syntactic attribute to the head noun (cf. Starosta forthcoming).

In this paper, I will describe 'pseudo-compounding,' a second strategy used in Sora to restrict the range of reference of a noun. Pseudo-compounding derives morphologically complex nouns by combining a modifying noun or verb with a bound monosyllabic 'Combining Form.' Each Combining Form corresponds in meaning and usually to some extent in form with a normally disyllabic independent Full Form noun.

The term 'pseudo-compound' is taken from Huffman 1970:329, who uses it to refer to Khmer words which contain bound morphemes which recur with the same meaning in a number of complex words. (Laurence Thompson 1965 uses the same term in a similar sense to describe a comparable phenomenon in Vietnamese.) A similar class of words exists in English as well; for example:

2-1. meningitis
appendicitis
tonsilitis

Each of these words contains a morpheme *-itis,* bearing a meaning of 'inflammation', but although the word looks like a compound, the *-itis* almost never occurs as an independent noun.

In both Khmer and English, this process of lexeme construction involves loans from classical languages, and it has only a marginal status in the word-forming repertoires of these two languages. In Sora, however, pseudo-compounding is highly productive, and plays a very important role in noun modification. Every independent noun has a corresponding but distinct Combining Form which is semantically indentical to it but which differs phonologically sometimes to the point of suppletion.

In previous studies beginning with Ramamurti's *Manual* (Ramamurti 1931), this process of word-formation was described as a type of syntactic attribution in which the head noun of a noun phrase had to be drawn from a set of 'Combining Form' nouns (*ibid.*, p. 48). To take a fairly extreme example, the Sora word for 'woman' is *ənsɛlɔ* and the word for 'beauty' (female) is *laŋgi.* The word for 'beautiful woman', however, is not**laŋgi -ənsɛlɔ*, but *laŋgi-bɔj.* The form *-bɔj* is the conceptual head of the pseudo-compound *laŋgi-bɔj*, and of many other such compounds involving a referential restriction of the concept 'woman', but it never occurs in my data as a free noun meaning 'woman,' and the free form *ənsɛlɔ* never occurs in compounds.

Sora Combining Forms occur in at least five classes of complex words, as illustrated by the following five examples:

2-2. 1) kɨnar-siŋ-ən 'mother-in-law's house'; cf. *kɨnar-ən* 'mother-in-law', *suʔuŋ-ən* 'house'
2) suʔuŋ-leŋ-ən 'in the house; the house's inside'
3) soṛa-siŋ-ən 'big house'; cf. *soṛa* [+V] 'big'
4) ə-siŋ-bu-n 'red ants' nest'; cf. *əbu-n* 'species of red ant'
5) gu-siŋ- [+V] 'build a house'; cf. *gu-* [+V] 'sow, plant, set up in the ground'

Type 1 derivatives can be characterized by rules such as the following:

2-3. DR-1 $\begin{bmatrix} +N \\ aF_i \end{bmatrix} \rightarrowtail \begin{bmatrix} +N \\ +habitation \\ aF_i \end{bmatrix}$

$] \longrightarrow$ siŋ]

This rule states that given a noun with the features [αF_i], it is potentially possible to derive a new noun differing from the original one in a) having a suffix *-siŋ,* and b) having a new semantic representation composed of the semantic features of the original input noun plus additional features introduced separately by each such rule, in this case something like [+habitation]. In the event that the new features conflict with old features carried over from the input noun, the new features take precedence.

The native speaker's knowledge that the forms derived by this rule are related to two corresponding Full Form nouns, i.e. *kinar* 'mother-in-law' and *su?uŋ* 'house' in the example given for Type 1 above, is shown by the DR: the matrix to the left of the fletched arrow matches a Full Form noun from the lexicon, in this case *kinar.* The feature [+habitation] on the right side of the arrow is a cover symbol for the semantic increment differentiating *kinar* 'mother-in-law' from *kinar-siŋ* 'mother-in-law's house', and matches the semantic representation of the noun *su?uŋ* 'house':

2-4. su?uŋ

$$\begin{bmatrix} +N \\ +habitation \end{bmatrix}$$

That is, the speaker's knowledge that *su?uŋ* and *kinar-siŋ* are related is a matter of semantics rather than phonology, which explains how a speaker is able to associate the suffix with the corresponding Full Form noun *su?uŋ* via their shared semantic features even if the phonological resemblance is not very great.

The following Derivation Rule exemplifies rules for the derivation of Type 2 pseudo-compounds:

2-5. DR-2

$$\begin{bmatrix} +N \\ \alpha F_i \end{bmatrix} \rightarrowtail \begin{bmatrix} +N \\ +lctn \\ +ntrr \\ \alpha F_i \end{bmatrix}$$

$$] \longrightarrow leŋ]$$

Type 2 is formally a special case of Type 1. It is treated separately here because of the special status of its outputs in Sora syntax. I referred to forms such as *-leŋ* in my dissertation (Starosta 1967) as 'noun auxiliaries', nouns combined with other nouns to add certain localistic features needed in the case-marking system. Later lexicase grammars, starting with Marybeth Clark's *Coverbs and case in Vietnamese* (Clark 1978) adopted the term 'relator noun', following Laurence Thompson's earlier usage (Thompson 1965: 200-202).

The fact that Type 2 derivatives really are derived noun lexemes and not postpositional phrases (cf. Starosta 1976:1086-1100) is supported by the fact that the definite suffix *-(ə)n* follows rather than precedes them, e.g.

2.6 su?uŋ-leŋ-ən 'in the house'
•su?uŋ-ən-leŋ
su?uŋ-ba-n 'to, at the house'
•su?uŋ-ən-ba

and by the fact that some lexemes derived by this rule have undergone semantic shifts to the extent that they are no longer synchronically related, e.g.

2.7 sərɔ-ba-n 'paddy field'; cf. *sərɔ* 'paddy'

In addition to its role in creating nouns that have a special function in the case-marking system, rather than just creating new lexemes to fit with new concepts when needed, this derivation type differs in another way from Type 1 pseudo-compounds: unlike Type 1 nouns, Type 2 noun suffixes typically have no synchronic noun sources. In the example above, *-leŋ* might conceivably be historically related to *luʔuŋ-ən* 'pit' (CF *-luŋ*), but other Type 2 suffixes such as *-ba* 'place at which' do not have any obvious synchronic sources.

Type 1 and Type 2 derivatives are very common in Sora, and are superficially difficult to distinguish from Type 3 forms, which can be derived by similar rules, e.g.

2-8. DR-3

$$\begin{bmatrix} +V \\ +sttv \\ \alpha F_i \end{bmatrix} \rightarrowtail \begin{bmatrix} +N \\ +sttv \\ +habitation \\ \alpha F_i \end{bmatrix}$$

$$] \longrightarrow siŋ]$$

Instead of nouns, this rule applies to stative verbs such as *soṛa* 'big' and *bəŋsa* 'good', with the semantic properties of the derived form being essentially a combination of the features of the input stative verb plus the noun-type semantic features introduced by the rule. This rule, and others of Type 3, tend to be more productive and semantically regular than DR-1, and there is seldom any doubt about the meaning of forms derived in this way.

The feature [+sttv] in the input to DR-3 may turn out to be too restrictive, since the existence of forms such as:

2-9. ñəŋñəŋ-mər 'teach-man', i.e. 'teacher'
tañ-kab-mər 'weave-cloth-man', i.e. 'weaver'

would seem to be composed of a CF in construction with a non-stative verb, unless the first verb is nominalized before entering into the derivation (cf. 4-10, section 4.2).

Type 4 pseudo-compounds are analyzed in accordance with rules such as the following:

2-10. DR-4

$$\begin{bmatrix} +N \\ \alpha F_i \end{bmatrix} \rightarrowtail \begin{bmatrix} +N \\ +pssd \\ -[+adrs] \\ -[+N]\_\_ \\ +fierce \\ +formic \\ \alpha F_i \end{bmatrix}$$

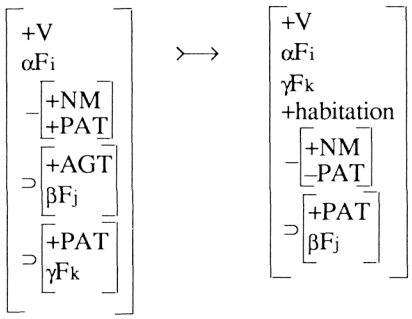
$$] \longrightarrow bu]$$

Words derived by this pattern are relatively rare in my field data, though Ramamurti lists a number of them in his *Manual* and dictionaries. They are superficially similar to Type 1 forms in that they are composed of two nominal morphemes. They differ syntactically and semantically from these forms, however. Semantically, words formed by this process refer to a relation in which the first element of the pseudo-compound is the 'head', the semantic center of the compound, and the second member is the attribute or 'possessor', an element which restricts the reference of the first member. Thus *ə-siŋ-bu-n* 'the red ants' nest' is not a 'house ant', an ant which lives in the house, but rather a 'house' (nest) built, inhabited, and fanatically defended by these merciless creatures. This difference between Type 1 and Type 4 words is reflected syntactically and morphologically in the specification of Type 4 nouns as inflected for third person possession ([+pssd], –[+adrs]___) and as disallowing overt external syntactic possessors (–[+N]___). The prefix *ə-* which is characteristic of these compounds is thus simply the regular inflectional affix for nouns of this category (cf. Starosta forthcoming).

Although this paper is concerned with nouns rather than verbs, it should be pointed out that 'object incorporation' in verbs can be handled in exactly the same way that pseudo-compounding for nouns is handled: by means of a derivation rule, exemplified by DR-5 below, which treats the incorporated noun as a derivational affix. (Cf. Starosta 1971 for an earlier case treatment of this phenomenon.)

This rule takes as input a transitive verb [+V, –[+NM,+PAT]] implying an agent with the semantic features [βF_j] and a Patient with the features [γF_k], and derives an intransitive verb [+V, –[+NM, –PAT]] with the inherent semantic features [αF_i, γF_k, +habitation] and implying a Patient (in this situation, the performer of the action; cf. Starosta 1978:474) with the features [βFj]. Phonologically, it adds the CF suffix *-siŋ* as an overt marker of the derivation. In pseudo-compounds of this type, as in Type 4 pseudo-compounds, the first element of the word is the intuitive head of the construction. The problems of representing this aspect of the derived semantic representation using a binary feature notation have not yet been explored within a lexicase framework (see however Starosta 1980).

2-11. DR-5



] ⟶ siŋ]

3 The historical priority of Combining Forms

In this paper, I am proposing a direction of derivation going from Combining Form to Full Form, which contradicts the long-to-short analysis proposed in Biligiri 1965, Stampe 1965, and in Zide 1976. In the present section, I will point out some historical considerations which favour treating Sora Combining Forms as being more basic than Full Form independent nouns. In Section 4, I will provide a theoretical justification for the short-to-long direction of derivation, and in Section 5, I will discuss the most detailed and explicit example of the long-to short analysis, Arlene Zide's 'Nominal combining forms in Sora and Gorum' (Zide 1976), and attempt to point out problems which make it unacceptable in its present form as an alternative to my short-to-long approach.

3.1 External cognates

The first kind of evidence one might consider in discussing the historical priority of a form would of course be comparative. I have only just begun to look in this direction, but so far, almost all of the likely Sora-Khmer pairs I have come up with in scanning the first third of Jenner and Pou's *A lexicon of Khmer morphology* (Jenner and Pou, 1980-81) involve Sora Combining Forms, not Full Forms:

3-1.	*Sora*			*Khmer*	
	FF	CF			
	əsu	su	'illness, fever'	chýy	'suffer, be ill'
				tuu	'moan'
	aŋgaj	gaj	'moon'	khaaɛ	'moon'
	–	ba-n	'place; at'	ban	'group together' ?
	daŋgo	daŋ	'stick'	dɑɑŋ	'shaft, haft'
				dɑmbɑɑŋ	'staff, stick'
	daʔa	da	'water'	týk	'water' (< OKh. *dyk*)

jaʔaŋ	jaŋ	'bone'	cqỳŋ	'bone'
jiʔi	ji	'tooth'	tkiiəm	'molar tooth' (cf. *kiĭəm* 'bite')
kɨnad	kad	'crab'	kdaam	'crab'
kondi	kon	'knife'	kandìíəw	'sickle' (< *kiĭəw*)
–	leŋ	'inside'	knoŋ	'inside, interior'
	luŋ	'inside space'	kamloŋ	'space within, interior'
loŋər	loŋ	'cave'	luŋ	'fall in a hole, pit'
luʔuŋ	loŋ			
təŋɔr	gɔd	'way, road'	tnal	'raised road'
ubban		'brother'	baaŋ	'older sibling'
usal	sal	'skin'	saak	'skin'

Until we can establish some regular correspondences, of course, this doesn't mean much, but the fact that it is generally the CF rather than the FF as a whole which has the probable cognate is certainly support for the historical priority of the CF rather than the FF.

3.2 Possessive pseudo-compounding

Type 4 pseudo-compounds, discussed in the preceding section, differ from the other types in that they are obligatorily possessed, although no syntactic possessor may be attached, and in that the Combining Form suffix corresponds to the possessor rather than the conceptual head, as is true of the other pseudo-compound classes. It is very interesting to note the parallels between the construction of these compounds and the structure of possessively inflected nouns as described in my Sora noun inflection paper (Starosta forthcoming):

3-2			suʔuŋ	-ñɛn		'my house'
			suʔuŋ	-nəm		'your (sg.) house'
		ə-	suʔuŋ		-ən	'his/her/its house'
	(ənsɛlɔn	ə-	suʔuŋ		-ən	'the woman's house')
		ə-	sɨŋ	-bu	-n	'red ants' nest' (Type 4 pseudo-compound)

The parallels between possessive suffixation and Type 4 pseudo-compounds could be accounted for if both reflect lexicalizations of syntactic constructions that existed at an earlier stage, when possessors followed their heads, as they do in modern Southeast Asian and Austronesian languages, instead of preceding them, as they do in modern South Asian languages. Once these possessive constructions had contracted and become lexicalized as single complex words, they would have been exempt from the syntactic change in typology that must have occurred later, possibly through the influence of surrounding non-Austroasiatic languages. If this word class does reflect an older stage of the language, then the fact that the form of the noun which gets lexicalized into the new complex word is the short one rather than the long one suggests that the short form is the older one, and may have been

the free form in an earlier SVO stage.[3]

The relative antiquity of the possessive pseudo-compound rule in Sora is further confirmed by the fact that it is only marginally productive for non-pronominal possession, and that the 'head' forms sometimes also occur in what I take to be the historically earlier monosyllabic Combining Forms, e.g. *siŋ* 'house' in *ə-siŋ-bu-n* 'red ants' nest', rather than the later and more common disyllabic Full Forms, e.g. *jɛlu* 'meat' in *ə-jɛlu-bɔŋ-ən* 'water buffalo meat' (cf. *bɔŋtɛl* 'buffalo', CF *bɔŋ*, and *jɛlu* 'meat', CF *jɛl*).

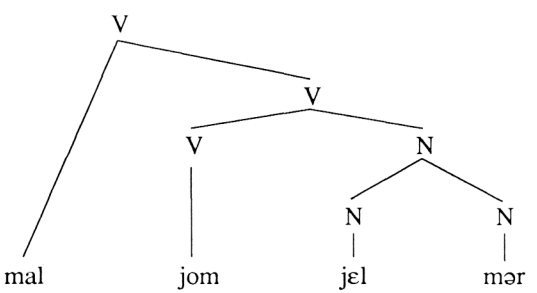
A form occurring in one of Zide's examples also appears to contain an instance of this old CF-CF head-possessor compounding functioning as an incorporated object (Zide 1976:1261):

3-3. kəṛib-ñɛn mal-jom-jɛl-mər-te
1 2 3 4 5 6 7

My sword longs to eat human flesh.
2 1 3 7 4 6 5

The incorporated-object verb *mal-jom-jɛl-mər-* here seems to be a three-stage derivation, as shown by the IC tree below, with the innermost derivation layer composed of a Type 4 head-possessor lexicalization, *jɛl-mər* 'flesh of man':

3-4.



(The prefix *ə-* on Type 4 pseudo-compounds, as mentioned before, is an inflectional affix, and would not be expected to carry over in derivation.)

Further evidence that Type 4 derivatives are quite old is the existence of words apparently derived by this pattern in which an otherwise unused morpheme shape is preserved, or which speakers no longer recognize as being morphologically complex, e.g.

---

[3]Note in this connection that pseudo-compounds in Khmer also show an attribute-head order which is counter to the order prevailing in syntactic constructions (Huffman 1970:329), although this is presumably to be explained as a carry-over from their Sanskrit and Pali sources rather than as a historical retention. The differences Huffman notes between independent words and their combining forms occurring in pseudo-compounds would presumably also be explicable in terms of the different phonological changes that would affect a form as an independent word and as an integrated part of a larger complex word.

3-5. adre-(s)im-ən 'chicken's egg'; cf. *adre-n* 'egg', *kənsim-ən* 'chicken', CF *-im*, not *-sim*.

əkəndar-ən 'tree branch'; cf. *ara-n* 'tree' CF *-neb; kən-* (frequent noun prefix)

The word for 'chicken's egg' is a Type 4 derivative, although the initial *ə-* is regularly absorbed by initial *a* or *ə* and thus is not available to help in identifying the derivation class. The form *-sim* as a CF appears only in the FF *kənsim* and in this derivative; the productive CF for 'chicken' is *-im*. For *əkəndarən*, my consultant did not recognize any connection betweeen this word and the word for 'tree', nor did he recognize a morpheme division after the initial *ə-*. The freezing of this form may be fairly recent, though, since Ramamurti's dictionary lists it without the initial *ə-*, suggesting that its derived status may have been more transparent twenty-five years earlier.

If I am correct in assuming that the head-possessor order is historically earlier, then the sequence of historical changes I suggested for Munda languages in my Sora noun inflection paper (Starosta, forthcoming, section 4.3) was backwards. In that paper, I suggested that preposed pronominal possessors came first in time, and that they were later replaced by postposed pronoun-derived possessive suffixes for kinship terms in Mundari and for all noun classes in Sora. In that scenario, though, I made no proposals about where the possessive suffixes might have come from. If it is in fact true that the suffixed possessors are retentions from an earlier head-modifier stage in Munda languages, and if it is also true that these languages did undergo a typological change after the possessive pronouns and a few non-pronoun possessors had been lexicalized as suffixes, then it would be easy to see how the Munda languages could have introduced a new modifier-head construction in accordance with the new syntactic regulations, and how this construction could have won out in competition with the older possessive constructions in the other Munda languages, being retained only in Sora possessive noun inflection and in Mundari kinship terms.

3.3 The conspiracy theory

One very interesting point which Arlene Zide raises in her paper is that there seems to be a 'conspiracy' in Sora grammar to produce CVC combining forms regardless of the shape of the initial Full Form noun from which she considers them to be derived (Zide 1976: 1269, 1271, 1278). On p. 1278, she states that 'the monosyllabic shape of the CF takes precedence over all other considerations'.

I think that Zide is right about the conspiracy, but not about the direction. Since it seems from other evidence that the monosyllabic forms came first, there must instead have been a conspiracy to derive polysyllabic forms from monosyllables, and this conspiracy would probably have been an external anti-Munda plot hatched by the same sinister forces that brought about the differences in syntactic typology that distinguish Munda languages from cousin languages such as Khmer. So, we still have a conspiracy, but this time we have a more plausible scenario for the crime: the various derivation rules presented in the previous section, and the extensive and pervasive system of independent polysyllabic FF nouns and

corresponding monosyllabic CF suffixes can be viewed as the result of a monosyllabic language (pre-Munda?) finding itself in an area where it is not at all fashionable to be monosyllabic, and desperately groping about itself for mechanisms like prefixation, reduplication, infixation, glottal breaking, suffixation, and compounding to cover up its suddenly revealed canonical nakedness. Thus, the short-to-long conspiracy theory has something going for it that the long-to-short theory doesn't: a motive.

I have recently begun looking outside Munda for examples of pseudo-compounding and evidence for a short-to-long conspiracy. It is easy to find evidence for the historical priority of the monosyllabic morpheme in Khmer, since as Jenner and Pou have shown (ms: xv-xviii), native wordbases are almost all monosyllabic. The extensiveness of the affixation system, the difficulty in pinning down the functions of some of the affixes, and Jenner's comment that 'the distinguishing phonological feature of this set [of complex prefixes]. . . is no more than their syllabicity' (Jenner and Pou 1980-81: xxx) does suggest a short-to-long conspiracy in progress, but if such a conspiracy existed, it was not as successful as the Sora counterpart, since independent monosyllabic words are still normal in Khmer. Pseudo-compounding does of course exist in Khmer, but there seems to be no reason to consider it to be part of a word-lengthening strategy. Instead, it has much closer parallels to the English derivation process which produces words such as *tonsilitis* and *hemisphere,* since in both cases the results are of marginal importance in the overall word-formation system, and involve only non-native morphemes as 'combining forms.'

It turns out, however, that there *is* a basically disyllabic language which has a system of Full Form and corresponding Combining Form nouns which almost exactly parallels the Sora system, and in which there is ample evidence for the historical priority of monosyllabic morphemes. The language is Mandarin Chinese. Thus it is very easy to find sets of Mandarin compounds sharing a single monosyllabic morpheme which seldom if ever occurs independently. To cite one such set (IO = 'included object', AC = 'attributive compound'):

3-6.	CF:	事	shì	
	FF:	事情	shìqing	'matter, affair, thing, business'
	IO:	做事	zuòshì	'to do work; to work'
	AC:	壞事	huàishì	'bad deeds'

A large subset of the disyllabic Full Forms end in 子 *-zi,* a character glossed as 'seed' when it is not acting as the second element of a pseudo-compound. In pseudo-compounds, it is an empty morph whose sole function seems to be to add a phonological syllable to a Combining Form to produce an acceptable disyllabic independent FF noun when the CF is not itself part of some other disyllabic compound, e.g.:

3-7.	háizi	'child'	xiǎoháizi	'small child; child'
			xiǎohái	'small child; child'

		xiǎohár	'small child; child'
pízi	'leather'	píxié	'(leather) shoes'
xiézi	'shoes'		
chūzi	'a cook'	chūfáng	'kitchen'
fángzi	'house, room'		

The list of CF's and compound sets given as an Appendix to this paper should serve to indicate the extensiveness of this process in modern Mandarin.

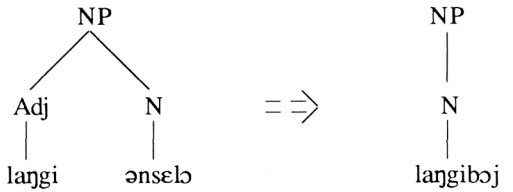
I am not about to suggest that Sora and Chinese are related (though I am also not about to rule it out either; in fact, I have been collecting possible cognates for years). Rather, I think the remarkable similarity in morphological structure of noun stems and included object verb stems is more likely to be the result of a parallel historical development: in both cases, it looks as if the original language was monosyllabic, and then changed to a typologically disyllabic structure, necessitating the development of strategies for making disyllabic nouns out of monosyllabic ones. Part of the motivation for this change in Mandarin may have been the need to compensate for the ambiguity introduced by the historical loss of final stops (Benjamin T'sou, personal communication). However, this would not account for the fact that it is so often the same syllable *-zi* which is added, since pairs of words which are homophonous before the addition of *-zi* are still homophonous after it is attached. This explanation might yet turn out to be the primary motivation for the change, if it is found that in case of homophony, one member of a homophonous pair was selected for the addition of *-zi* and the other was left unmarked or was suffixed with some other form. However, this still would not be enough to account for the very strong tendency to disallow the occurrence of monosyllabic nouns as the sole constituents of Noun Phrases in Mandarin. I would like to suggest that at least part of the motivation for the change may have been the same for both languages: a shift in structure brought about through the influence of contacts with another language family with disyllabic root structure, such as Austronesian. (See Ballard 1979 for further discussion of the notion that the modern Chinese languages have been strongly influenced by close contact with their neighbours.)

4 Formal analysis

4.1 Transformational solutions

It has been common practice since Lees' *Grammar of English nominalizations* (Lees 1968) to derive compound nouns from clauses or phrases. However, if one regards Sora pseudo-compounds as phrases at some level of analysis, there may be problems in assigning an appropriate derived constituent structure. Consider for example a word such as *laŋgi-bɔj* 'beautiful woman'. At one level of analysis, this is represented perhaps as a Noun Phrase, a syntactic construction composed of an Adjective and a Noun dominated jointly by a single NP node. Subsequently, this is converted into a single unbranching node, as follows:

4-1.



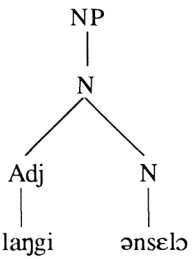
It would be fairly easy to write a transformation that would produce the correct external syntactic structure, e.g.:

4-2. T-1 SD: X – Adj – N – Y
1 2 3 4

SC: 1, 2, 3, 4 ⇒
1, Ø, 2 # 3, 4

which would produce the following derived constituent structure:

4-3.



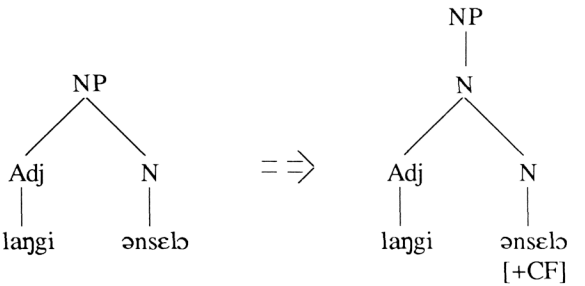
The problem, though, is to replace the sequence *ənsɛlɔ* in this derived structure by the Combining Form *-bɔj*, and this is not a simple matter even in a transformational grammar. As a first approximation, we might try something like:

4-4. T-2 SD: X – Adj – N – Y
1 2 3 4

SC: 1, 2, 3, 4 ⇒
1, Ø, 2 # $\left[\begin{matrix}3\\+CF\end{matrix}\right]$, 4

This rule would perform the following operation:

4-5.



The feature [+CF] would be considered a signal that the original noun represented by term 3 in the Structural Description is to be be replaced in a second lexical pass by the corresponding Combining Form; but how is the concept of 'corresponding CF' represented in such a grammar? If we could modify the transformational rule above so that it would automatically change the full form into the appropriate Combining Form, or if we could maybe add some regular morphophonemic rules which would be triggered by the [+CF] feature and would produce the right form, this analysis would be rather attractive, even if there were a few suppletive forms left around that had to be listed lexically. However, as I will attempt to show in section 5, the only serious attempt so far that I know of to do this (Zide 1976) is only partially successful. In most cases, it turns out not to be possible to derive CF's from FF's by general rules. That means that the 'corresponding CF's' will have to all be listed in the lexicon, as in the lexicase solution, and the correspondence between FF's like *ənsɛlɔ* and CF's like *-bɔj* will have to be shown either by indexing them or juxtaposing them in a list, which would be completely ad hoc, or by choosing the CF on the basis of shared semantic features as is done in the lexicase approach, though it isn't clear how this would be formalized in the transformational approach.

Another problem in trying to do lexical derivation transformationally has often been mentioned in arguments supporting the lexicalist hypothesis: transformations can't change meaning, but the meanings of some of the Sora pseudo-compounds are not predictable from the meanings of the input forms. Thus the word *kɨnar-siŋ-ən* 'mother-in-law's house' refers to a regular house, while *ə-siŋ-bu-n* 'red ants' nest' does not. Even if one decided to relax this restriction, one would need to use Rule Features and a lot of transformations which applied to only one form each to make the system work, and that is the same thing as listing all the forms in the lexicon instead of deriving them by rules.

A third problem involves Type 2 pseudo-compounds, which are formally like Type 1 words but involve CF's with no synchronically extant corresponding FF's. This would be a serious problem for an analysis in which Combining Forms are considered to be derived synchronically from corresponding Full Forms. In such an analysis, Combining Forms such as *-leŋ* would have to be derived from some unattested underlying Full Form or treated as either exceptional or completely unrelated to the other pseudo-compounding constructions. In the analysis proposed in this paper, of course, no Full Form of *-leŋ* is required as an input for DR-2, and Type 2 pseudo-compounds are derived in exactly the same way as Type 1

constructions. In both cases, the grammar is able to represent bound 'relator noun' structures as derived by the same process as Type 1 derivatives, and thus capture an important synchronic and diachronic generalization.

A transformational description of Sora pseudo-compounding which tried to work in the correct short-to-long direction would not be much better off than the one which works the other way, since I can see no way of writing a single general rule which takes a monosyllabic CF as input and says, 'When this form occurs in isolation, make a disyllable out of it in any way possible.' As mentioned in the previous section, there are a number of strategies available for making FF's from CF's, and there seems to be no general way to predict which strategy is going to be used on which form, though tendencies can be seen in Zide's analysis. This means that each CF will have to be lexically marked with a Rule Feature indicating which of a battery of expansion rules is to apply; but a Rule Feature is just a notational variant for a separate lexical entry: it doesn't indicate any internal property of the entry on which it is marked, but rather is a signal that another form having a certain set of properties exists somewhere else in the lexicon. Thus we are back to lexical listing again.

4.2 Lexical solutions and lexicase

In an unconstrained descriptive framework such as Ramamurti's traditional approach or the transformational framework assumed by Lees, one can of course start with a syntactic input and come out with a single lexical unit. In a grammatical model such as lexicase, however, where transformations have been rejected to restore empirical content to the associated metatheory, this analysis is not available. In such a framework, all word-formation takes place in the lexicon, and in a lexicase grammar, the only available mechanism for describing such a process is a Derivation Rule, since inflection is clearly not involved. The apparent formal problem this creates comes from the fact that

1) lexical Derivation Rules in a lexicase grammar take lexemes as their entries, and

2) lexical entries are words, or stems which differ from words only in the absence of inflectional affixes. However, Combining Forms like *-bɔj* are neither words *nor* stems, since they never appear as independent words in Sora sentences either with or without inflectional affixes. This means that they are not lexemes, and so can't serve as the input to Derivation Rules. Thus while another non-transformational linguist's first impulse might be to treat this kind of word formation as compounding, using a rule such as DR-A:

4-6. DR-A

$$\begin{bmatrix} +N \\ \alpha F_i \end{bmatrix} + \begin{bmatrix} +N \\ +CF \\ \beta F_j \end{bmatrix} \longrightarrow \begin{bmatrix} +N \\ \beta F_j \\ \alpha F_i \end{bmatrix}$$

this is not possible in a lexicase grammar as currently conceived, since only one of the elements of the 'compound' would be present in the lexicon to serve as the input to the compounding rule.

The solution adopted in this study is the only one that appears to be compatible with lexicase assumptions. It requires complex words such as *laŋgibɔj* 'beautiful woman' to be treated as affixed forms, with the Combining Forms considered to be the affixes added in the process of derivation. Taking English *tonsilitis* as an illustration, the corresponding Derivation Rule would have the following form:

4-7. DR-B

$$\begin{bmatrix} +N \\ \alpha F_i \end{bmatrix} \rightarrow \begin{bmatrix} +N \\ +\text{inflammation} \\ \alpha F_i \end{bmatrix}$$

$$] \rightarrow \text{itis}]$$

This rule takes a noun like *tonsil* whose semantic representation contains a subset of features [αF_i], and derives a new noun, one which carries the original features but which differs semantically and phonologically from the input in having an extra semantic increment symbolized here by [+inflammation] in its matrix, and an additional final sequence *-itis* in its phonological representation.

In the event of any conflict between the new features added by the rule and the features carried over from the input, the new features take precedence.

Since DR-B is a Derivation Rule, we know that it describes a *potential* word-formation process (cf. Starosta 1974: 294). Of the nouns characterized as eligible to undergo the rule by the specification of the features [αF_i], there is no way to predict in an individual case which noun will undergo the rule. Thus looking at this rule alone, there is no way to know that *tonsilitis* and *appendicitis* are English words, but that **earlobitis* is not, though it could be.

A linguist encountering this analysis for the first time may find it rather awkward and counterintuitive, but I think that the awkward parts turn out to be places where the lexicase analysis faithfully represents aspects of the data which are glossed over or covered up in other analyses. For instance, the fact that every Combining Form needs a separate Derivation Rule may seen wrong, since it fails to capture the generality of the process of pseudo-compounding. This can be partly remedied through the use of square brackets to bring together all the individual subrules for each derivation type, with the overall rule then in effect incorporating a list of all the CF's involved in the word-formation process. The fact that a separate rule or sub-rule is required for each CF is regrettable but unavoidable: since there is no way of uniformly predicting FF's from CF's *or* vice versa, both sets have to be listed somewhere. Including the CF's as part of a lexical Derivation Rule is no more expensive than listing each of them in the lexicon, and it avoids the problem of matching up corresponding forms discussed in 4.1.

Another apparent flaw in the lexical derivation approach is that every derived form must also be listed in the lexicon, in addition to having a separate rule for each CF (cf. Halle 1973). Although this seems uneconomical, it simply reflects the fact

that the precise semantic representations of lexically derived words are determined at the point of derivation in accordance with the requirements of the moment, and thus may not be exactly characterized by the semantic part of the rule; and since the newly derived words are listed independently, they can shift semantically and phonologically in unpredictable ways. Thus the rule serves to relate pairs of items in the lexicon, but listing all the derivatives separately makes allowances for differences. The lexical approach to derivation accepts and accounts for something that transformational solutions normally camouflage with Rule Features: the derivatives have to be lexically listed.

Lexically listing related forms is of course something that has to be done sometimes anyway for cases of suppletion, and in fact the present analysis requires the relation between CF's and FF's in Sora to be analyzed as a kind of suppletion. Thus by this analysis there is no direct connection established between the Combining Form and the corresponding Full Form. The phonological part of the DR and the semantic feature matrix [αF_i] serve to characterize the speaker's knowledge that a complex word like *tonsilitis* is related to the word *tonsil*, but only the added semantic features symbolized by [+inflammation] here show any connection between the *itis*-suffixed noun and independent Full Form nouns such as *inflammation* or *illness*, which have similar semantic representations. That is, the relation between the Full Form and the corresponding Combining Form is treated like suppletion in a lexicase grammar: two forms like *go* and *went* are listed separately in the lexicon, each with its own phonological representation and semantic features, and only the complementary distribution and shared semantic features show the two words to be related:

4-8.

go	went
+V	+V
+traverse	+traverse
	+finite
	+past

It is possible in Sora and Gorum to have pseudo-compounds incorporating more than one nominal CF, e.g.

4-9. kid-suŋ-kui 'die-house-woman', i.e. 'widow' (Gorum)
əsɔŋ-bəb-mər 'feces-head-man', i.e. 'insolent man'
tañ-kab-mər-ən 'weave-cloth-man', i.e. 'weaver'

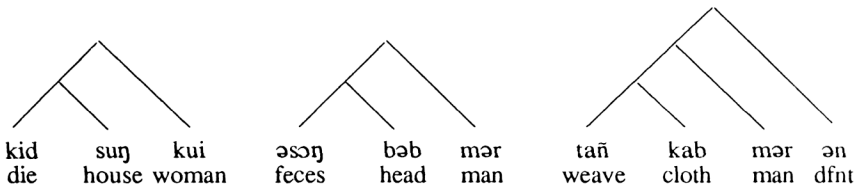
(The first two examples are taken from Zide 1976:1260 and the third from Ramamurti 1931:44.) The analysis proposed in this paper requires such forms to be derived in stages, with each stage adding a nominal CF at the right end of the input form. For instance, the third example would be derived in the following stages:

4-10. Stage I: tañ- [+V] ⟶ tañkab- [+V] by object-incorporation (Type 5 pseudo-compounding)

Stage II: tañkab- [+V] ⟶ tañkabmər [+N] by Type 3 pseudo-compounding

(The *-ən* on 'weaver' is a definite inflexional affix; see Starosta forthcoming). Clearly the present analysis makes the correct prediction for such forms: they are left-branching structures containing only binary branchings, so that a *kid-suŋ-kui* is a woman (*kui*) associated with someone having died (*kid*) in a house (*suŋ*), etc.:

4-11.



Finally, there is the problem mentioned earlier that the Combining Forms in pseudo-compounds of Types 1-3 are conceptually the heads of their words, but that my rules treat the CF's in all five derivation types formally as affixes. It is interesting to note that this is one place where Zide's analysis seems to agree with mine, since she refers to some 'complex nominal forms consisting of a series of CF's or created by the addition of a CF to a FF or by the addition of one or more CF's to a verb root' (Zide 1976:1259; cf. also p. 1261). I don't think she is right about the IC analysis implied by the terms 'series' and 'one or more', for the reasons above, but her statement that the CF is added to the Full Form seems to fit with my treatment of CF's as suffixes.

I'm not sure how serious a drawback this is, since it is not clear that the concept of internal constituency in words has any synchronic significance anyway. Presumably, a complex word couldn't be lexicalized until its semantic representation had become an integrated unit, and once this has happened there is no semantic reason to refer in a grammar to the morphological structure of a word; the whole phonological representation could be replaced by some arbitrary sequence without affecting the syntax at all. (This is of course the situation with suppletion, as mentioned earlier.)

5 The long-to-short analysis

5.1 Analogy and the direction of derivation

Previous linguists working on Sora who have considered the question have all concluded that the Combining Forms of Sora nouns are derived from the Full Forms by a process of truncation. So far, no one seems to have seriously considered the other logical possibility, the one proposed in this paper: that CF's are primary and that the so-called Full Form nouns are secondary derivations from the Combining Forms. One wonders if this might not be a case of terminology influencing analysis, since intuitively it seems sensible to take 'Full Forms' as the starting point, the independent basic entities from which the stripped-down special-purpose 'Combining Form' can be derived. In this section, though, I will try to

show that the long-to-short analysis, as exemplified especially in Arlene Zide's paper on nominal Combining Forms in Sora and Gorum (Zide 1976), presents serious formal and conceptual problems that make it unacceptable.

Zide considers her analysis of Sora Combining Forms to be superior to Biligiri's 1965 description in that Biligiri presented only a classification based on arbitrarily chosen formal features of CF's and their corresponding FF's. The basis for her claim to have produced an explanation rather than a description (*ibid.*, p. 1263) is that she has taken other productive morphological processes in the language into consideration, whereas Biligiri's description did not.

To some extent, Zide's claim seems justified, except in one major respect: the processes she cites do in fact in general seem to be the ones involved in the relation between CF's and FF's but her evidence does not seem to support the long-to-short *direction* of derivation, since the processes she cites all *add* elements rather than subtract them as Zide's truncation procedure assumes. For example, the reduplication and infixation processes which apply to verbs would certainly take the unaffixed and unreduplicated forms as basic, since they have the widest distribution and are formally unmarked as compared with the reduplicated and affixed forms.

This is especially clear in the case of nominalizations, which by anyone's analysis would be derived from the corresponding verbs by the infixation of elements such as *-ən-* and *-ər-* or by the suffixation of elements such as *-an*, rather than by deriving the shorter form from the longer one by a subtractive process, which is literally what Zide proposes (*ibid.*, p. 1266) as the way to derive the CF's of nominalized verbs.

What Zide proposes as a strategy for finding the correct CF for a FF is to first find out if the FF is derived from some monosyllabic form elsewhere in the language, and if so, derive it back again. As she states on p. 1264:

> It has already been suggested that certain of the morphological properties of FF nominals must be assessed before prediction of derived CF's can be undertaken with any hope of accuracy ... information can be retrieved in some cases thanks to the existence in the language of a related form or forms, such as a verb, from which the FF nominal is itself derived.

And on p. 1266:

> From this much it can be seen that the first step in accounting for the derivation of CF's must be the identification of FF's which are themselves derived from monosyllabic morphemic nominal stems ... For nominals derived from verbs, reduction is effected by deinfixation, ... in known instances

which I take to mean, 'in those instances in which we have already determined that the monosyllabic unaffixed form is the underlying and basic one', that is, where a short-to-long direction of derivation is implicitly assumed.

Zide's approach seems even more implausible when applied to reduplication. Certainly there is plenty of evidence in Sora morphology for positing reduplication processes, especially in verb derivation; however, I know of no justification anywhere in the grammar for assuming that CF's are derived from reduplicated FF's by a rule of dereduplication. In fact, writing such a rule in a general form would be very clumsy if it is possible at all. Zide has an ingenious solution to this particular problem, however; instead of deleting phonological material, the dereduplication rule simply deletes 'R', which is essentially a Rule Feature invoking the Reduplication Rule. It is generally acknowledged that Rule Features are already ad hoc devices indicating that no generalization has been achieved. If so, an analysis that treats them as phonological segments and deletes them by pseudo-phonological rules seems to compound the ad hociery.

Zide's rule of deprefixation, unlike the disinfixation and dereduplication rules, does not have any counterpart elsewhere in the morphology. For disinfixation, we can cite the reverse process of infixation in nominalization, but for 'deprefixation,' the only reverse counterpart one might cite would be one which has the sole function of making FF's from CF's. Yet if we already have such a process for relating these forms, why do we need another one that does the same thing but in the opposite direction? Even the use of the term 'prefix' makes it clear what is involved here: a prefix is something added on to something else more basic, which would mean that the CF is basic and the (prefixed) FF is derived. Similarly when Zide states on p. 1266 that 'For nominals formed by concatenation of several potentially independent morphemes, truncation is effected by deleting the second syllable in most cases,' the process of 'concatenating potentially independent morphemes' would be a means of deriving an independent disyllabic word from a canonically inadequate monosyllabic morpheme, and if this process is assumed anyway, it is totally redundant to have another process of truncation to go back and undo everything again.[4]

Shakiest of Zide's disinfixation rules is the one which deletes the 'infix' /ʔ/. This has no support from anywhere else in Sora morphology, as far as I know, and fails to explain why this 'infix' is only deleted between identical vowels. On the other hand, an analysis which assumed that FF's of the form $C_1V_1ʔV_1C_2$ were historically derived from CF's of the shape $C_1V_1C_2$ by a process of 'breaking' (perhaps related to consonant checking, or to the register phenomenon found in other Austroasiatic languages) would be superior in accounting for the non-occurrence of sequences of non-identical vowels interrupted by glottal stop which are related to CF's without the glottal stop.

---

[4]An interesting problem arises in connection with Zide's use of the term 'morpheme' for the second element of, for example, the word for 'buffalo', *bɔŋtɛl*, CF *-bɔŋ*. If *-tɛl* is a 'morpheme,' what does it mean? Since *bɔŋ* carries the full meaning by itself in all occurrences, then *tɛl* presumably has no content at all, that is, it is an empty morph. Alternatively, it could be considered to be one of the many and varied segmental and processual allomorphs for the stem-forming morpheme {FF}. By my analysis, this formal difficulty doesn't arise, since CF's and FF's are treated like suppletive allomorphs of the same morpheme related synchronically only by common meaning and complementary distribution, so that there would be no internal segmentation of forms like *bɔŋtɛl*.

Looking at the various morphological processes Zide has cited in her discussion, then, there seems to be ample evidence for a conspiracy theory (Zide 1976:1271), but if we look at the direction in which these processes operate, the conspiracy seems to be to avoid monosyllabic nouns and verbs under any circumstances. By Zide's own criterion, a short-to-long analysis of Full Forms would have a better claim to being an explanation than one going in the other direction, since it would be more in concert with other morphological processes operating in the language.

5.2 Truncation rules

Throughout her paper, Arlene Zide assumes that Sora CF's are derived from FF's rather than vice versa, and that it is possible to state this relationship by means of generative rules. We have already seen that the long-to-short direction of derivation is probably wrong. In this section, I will examine Zide's generative rules to show that her claim to be able to generate CF's from FF's is not well founded either.

There are a number of difficulties in evaluating Zide's claim to have achieved some kind of explanatory adequacy, since the rules as she formalizes them don't always correspond to the rules discussed in prose (*ibid.*, p. 1268, 1270). This is especially notable in the sample derivations (*ibid.*, p. 1274-5), where the rules have been rearranged, collapsed, and abbreviated in such a way that it is quite difficult to figure out which of the rules on these pages matches up with which rules discussed in the text. This appears to have confused even the author, since she seems to have inadvertantly left out one rule (dereduplication), and introduced another completely ad hoc one (A-c) without going back and inserting some justification for it in the main text.

The second problem is that the form of the rules seems to be of a new type, created especially for this occasion. This means that it is impossible to determine whether such rules can be fitted into the kind of full-fledged theory that is presupposed by Zide's claims to have produced an explanation rather than a mere description. To cite several odd features of the rules,

1) B-d on p. 1269 seems to have an unnecessary subscript; in fact, since the environment is clearly specified at the right of the rule, it isn't clear why either of the subscripts is needed. (Crucial stress marks have also been omitted in the same rule.)

2) The notation $[\ ]_{syll}$, by the usual convention of generative grammar, implies a hierarchical structure containing a node labeled 'syll.' I am unaware of any generative framework that posits such a word-internal labelled hierarchy. An apparently related convention is Zide's use of the notation $\_\_|_{CF}$. Again, this seems to imply some sort of labelled hierarchical structure within a word, this time with grammatical node labels such as 'CF' (Combining Form), again something that I have not seen in other grammatical frameworks. This label can't refer to something generated in Deep Structure, since CF's in Zide's framework are not Deep Structure categories, so it must have been introduced by a later rule which is never mentioned here.

3) The deprefixing and disinfixing rules are stated without adequate specification of their domain of application, so that as stated, they could detach elements indiscriminately from anywhere in the word, not just from the beginning or post-initial-consonant position respectively.

4) Leaving aside the question of the metatheoretical status of Zide's rules, and giving her the benefit of the doubt in instances of overly casual formalization, it seems that her rules do in fact work, at least when they apply to the items they are supposed to apply to and not to those they aren't. Thus it is possible to take the lexical forms of the FF's as she gives them, and apply her operations in sequence to produce the appropriate Combining Forms. However, this brings us to a fourth problem: the status of the input FF's. This is a problem because almost all of Zide's rules are crucially dependent on stress: in order for the rules to work, each vowel in the Full Form must be marked for either primary, secondary, or no stress in almost all of Zide's entries. The strange thing is that Zide is the first of the modern linguists working on Sora to notice that Sora has stress. As she mentions in a footnote (p. 1293), Stampe, Starosta, Biligiri, and Mahapatra all regard stress as subphonemic in Sora. In another footnote, she uses the term 'nonphonemic' (p. 1294): '... stress is nonphonemic, it is used here only as suggestive raw data yet to be assessed.' I consider this latter term to be more appropriate, since I personally was unable, after a total of ten months' field work on Sora, to discover any consistent pattern of stress at all.

The problem then should be clear: Zide has based her analysis and her claim to having achieved a higher degree of adequacy on a feature which no one else can hear, but which she can hear consistently, and in fact can hear three different kinds of. Something which has the status of 'suggestive raw data yet to be assessed' has been found to be sufficient grounds for the claim that she has produced an analysis superior in explanatory power to, for example, Biligiri's previous work (*ibid.*, p. 1263). I am more than ready to grant that Zide may have a better ear for stress than the rest of us, and that she can really hear three degrees of non-phonemic stress. This raises an interesting point, however: can the Soras hear it? If it is 'subphonemic,' then presumably they can't. And if they can't hear it, how can they use it to derive Combining Forms from Full Forms? If 'subphonemic' means that stress varies freely, then obviously it can't serve as a basis for consistently deriving the correct CF. If, on the other hand, Zide means that stress is predictable in terms of some other factor which *is* phonemic, then what is that factor, and why wasn't *it* used in writing the CF-deriving rules?

Unless the above questions can be answered satisfactorily, the possibility must be considered that the Three Sora Stresses are really notational variants for Rule Features. Evidence for this hypothesis, in addition to the considerations in the preceding paragraph, include inconsistencies between the statements made *about* stress and the stress actually marked on the underlying Full Forms. The most glaring example of this is Zide's claim (p. 1293) that /ə/ and /i/ are inherently unstressed. I have found 34 examples of primary-stressed /ə/ and 12 examples of primary-stressed /i/ in Zide's data, and the examples are always cases where the stress is needed in order to condition the application of one of her rules. The actual stress patterns themselves are also sometimes strange-looking, as in the case where single disyllabic words have two primary stresses when this is needed for the application of a particular rule, and other cases where no stress at all is marked on

the whole FF. With three exceptions, all of the completely unstressed words in Zide's list of 285 FF's are cases in which stress is not needed to condition the application of the truncation rule, either because the CF is suppletive or because the rule is conditioned by the presence of a glottal stop.

If Zide's stress markers are really Rule Features in disguise, and I see no alternative to this assumption at the moment, then her claims to have produced an explanatorily superior description of the relation between Sora FF's and CF's is vacuous. As she herself stated (p. 1263):

> One can indeed account for all types of contraction in Sora simply by recognizing a sufficient number of classes ... But, those classes are unsatisfactory in terms of linguistic motivation, or explanation, for what are otherwise merely arbitrary classifications.

Rule Features of course are an exponent of an arbitrary classification in its clearest form.

I think that it is not surprising that Zide's attempt to provide an explanatorily adequate analysis of Sora pseudo-compounding was unsuccessful, because she attempted to do a synchronic explanation of the phenomenon, whereas it seems to me that synchronically there is no explanation. As far as the speakers of Sora are concerned, the CF's, FF's, and their correspondence with each other are simply lists that have to be memorized, and this fact is reflected in the analysis presented in this paper. The real explanation, I am forced to conclude, must be a historical one. It is extremely awkward to try to derive Combining Forms of nouns from the corresponding Full Forms because the Combining Forms were there first, and because historically the corresponding Full Forms were derived at different points and in accordance with differing and unpredictable strategies.

Zide has in fact recognized the historical nature of at least some of the FF-CF relationships, but has failed to draw all the consequences. Thus she has pointed out (*ibid.*, pp.1270-3) that the relation of FF's to CF's may sometimes be unpredictable due to historical phonological changes. Further remarks to this point:

> If one works exclusively with synchronic Sora data the required information is in many instances not retrievable from surface forms. (*ibid.*, p. 1264)

> Working synchronically without recourse to comparative data, we can only set such exceptions aside without explanation. (*ibid.*, pp. 1272-3)

> ... unexpected vowel alternations ... must be accounted for in terms of vowel reconstruction rather than of deviation from more general derivation rules. (*ibid.*, pp. 1270-71)

> From this it can be seen that some CF's of Sora-Gorum, or at least of pre-Sora, must have been derived at a time when either an initial *s* was present (and later lost) or when the effects of an initial *s* ... were in force. (*ibid.*, p. 1273)

What all of these observations entail, of course, is that CF's and FF's could only have diverged from each other in this way if they were all separate and independent lexical entries from the beginning, with no general synchronic rules deriving one from the other at any stage. That is, they have always been related via suppletion, and that situation continues to the present.

Finally, although Zide's long-to-short derivation isn't diachronically or synchronically valid, there are some instances where the disinfixation and second-syllable deletion approaches do seem to be indicated. Although the alternation of monosyllabic CF's and polysyllabic FF's is quite old and is basically a retention, Zide has found evidence that in at least some cases speakers have created CF's for nouns which didn't have any historically retained ones through a process of deletion of what she calls, following Mahapatra and Zide 1972:2, 'phonological infixes' (Zide 1976:1266). She tries to incorporate this process in her rule system, but I would consider it a matter of back-formation which applies sporadically to fill in gaps in an analogical pattern. This is supported by the fact that it tends to involve loan words of more than one syllable (*ibid.*, p. 1283), that it can be blocked by the existence of 'preserved verbal relations' (*ibid.*, p. 1268), and that it sometimes produces two alternative CF's for the same FF (*ibid.*, p. 1281-3, 1289), presumably through the application of two different analogies, although this shouldn't be possible if the rules were regular and phonologically conditioned as Zide claims. Lexicase derivation rules are, of course, ideally suited to describing such semi-productive word-formation processes.

Appendix: Mandarin Chinese pseudo-compounds

CF	FF		IO		AC	
ài	àiqing	'love'	zuò'ài	'make love'	rén'ài	'benevolence'
bào	bàozhǐ	'newspaper'	màibào	'sell papers'	rìbào	'daily paper'
bīng	bīngshì	'soldier'	dāngbīng	'be a soldier'	bùbīng	'infantry'
bìng	bìngzhèng	'disease'	shēngbìng	'get sick'	fèibìng	'lung disease'
chē	chēzi	'car, vehicle'	shàngchē	'get in the car'	huǒchē	'train'
chéng	chéngshì	'city'	shàngchéng	'go to town'	dàchéng	'large city'
dǎo	hǎidǎo	'island'			bàndǎo	'peninsula'
dǐ	dǐxià	'bottom, floor'	dàodǐ	'thoroughly'	haǐdǐ	'seabed'
fǎ	fǎlǜ	'law'	fànfǎ	'break the law'	xiànfǎ	'constitution'
gōng	gōngfu	'work'	zuògōng	'do work'	réngōng	'manual labour'
guó	guójiā	'country'	chūguó	'leave the country'	měiguó	'America'
huì	huìyì	'meeting'	kāihuì	'have a meeting'	dàhuì	'big meeting'
hūn	hūnyīn	'marriage'	jiéhūn	'marry'	zǎohūn	'early marriage'

CF	FF		IO		AC	
jiā	jiālǐ	'home'	bānjiā	'move house'	dàjiā	'everybody'
jià	jiàzhí	'price, value'	jiǎngjià	'bargain'	dìngjià	'fixed price'
láo	láodòng	'labour'	shǎng láo	'reward labor'	qíangláo	'strong laborer'
lì	lìqi, lìliang	'strength' -"-	yònglì	'use strength'	kǔlì	'hard work'
liào	cáiliào	'materials'	beìlìao	'prepare materials'	yuánliào	'source materials'
lóu	lóufáng	'building'	shànglóu	'go upstairs'	jiǔlóu	'restaurant'
lù	dàolù	'road, street'	zǒulù	'walk'	xiǎolù	'path'
mén	ménkǒu	'gate'	chūmén	'go out'	dàmén	'main gate'
mín	mínrén	'people'	zhímín	'colonize'	guómín	'nation's people'
míng	míngzi	'name'	qiānmíng	'sign one's name'	nǎimíng	'milk name'
mìng	mìngyùn	'fate'	yàomìng	'terrible'	duǎnmìng	'short-lived'
shēn	shēntǐ	'body'	dòngshēn	'move on'	quánshēn	'all over the body'
shí	shíhou	'time'	shǒushí	'punctual'	zhànshí	'temporary'
shì	shìqing	'affair, matter'	zuòshì	'work'	huàishì	'bad deeds'
tú	túhuà	'drawing'	huàtú	'draw a design'	dìtú	'map'
tú	túdì	'disciple'	dāngtú	'apprentice'	jiàotú	'follower of a religion'
tǔ	dìtǔ	'ground, earth'	dòngtǔ	'get busy'	huángtǔ	'löss'
yán	yǔyán	'language'	fāyán	'make a speech'	fāngyán	'dialect'
yī	yīfu	'clothing'	cǐ yī	'wash clothes'	wàiyī	'outer clothes'
yì	yìyi	'meaning'	yuànyì	'willing'	zhúyì	'plan'
yǔ	yǔyán	'language'			déyǔ	'German language'
zhàn	zhànzhēng	'war'	zuòzhàn	'make war'	dàzhàn	'major war'

IO = Included object, AC = Attributive compound.

REFERENCES

Annamalai, E., Gérard Diffloth, Bijoy P. Mahapatra, and David L. Stampe (eds). [forthcoming]. *Proceedings of the Second International Congress on Austroasiatic Linguistics*. Mysore: Central Institute of Indian Languages.

Ballard, W. L. 1979. 'Chinese: a bastard at the Sino-Tibetan family reunion?' *12th International Conference on Sino–Tibetan Languages and Linguistics*, Paris.

Biligiri, H. S. 1965. 'The Sora verb, a restricted study'. In *Indopacific Linguistic Studies, Part II: descriptive linguistics,* ed. G. B. Milner and E. J. A. Henderson, 231-50. Amsterdam: North-Holland.

Clark, Marybeth. 1978. *Coverbs and case in Vietnamese*. (Pacific Linguistics B-48). Canberra: Australian National University.

Halle, Morris. 1973. 'Prolegomena to a theory of word formation'. *Linguistic Inquiry* 4, 3-16.

Huffman, Franklin. 1970. *Modern spoken Cambodian*. New Haven: Yale University Press.

Jenner, Philip and Saveros Pou. 1980-81. *A lexicon of Khmer morphology. MKS* IX-X.

Lees, Robert. 1968. *The grammar of English nominalizations*. Fifth printing. Mouton: The Hague.

Mahapatra, Khageshwar, and Norman H. Zide. 1972. 'Gta? nominal combining forms'. *Indian Linguistics* 33:3, 179-202.

Ramamurti, G. V. 1931. *A manual of the Sora (or Savara) language*. Madras: Government Press.

Ramamurti, G. V. 1933. *English–Sora dictionary*. Madras: Government Press.

Ramamurti, G. V. 1938. *Sora–English dictionary*. Madras: Government Press.

Stampe, David. 1963. *Proto–Sora–Parengi phonology*. Master's Thesis, University of Chicago.

Stampe, David. 1965. 'On the Sora noun'. Unpublished ms, Chicago. Abstracted in *IJAL* 31:4, 336-62.

Starosta, Stanley. 1967. *Sora syntax: a generative approach to a Munda language*. Ph.D. dissertation, University of Wisconsin.

Starosta, Stanley. 1971. 'Derivation and case in Sora verbs'. *Indian Linguistics* 32: 3, 194-206.

Starosta, Stanley. 1974. 'Causative verbs in Formosan languages'. *Oceanic Linguistics* 13:1-2, 279-369.

Starosta, Stanley. 1976. 'Case forms and case relations in Sora'. In *Austroasiatic Studies*, ed. P. Jenner, L. Thompson, and S. Starosta, 1069-1108. Honolulu: University Press of Hawaii.

Starosta, Stanley. 1978. 'The one per Sent solution'. In *Valence, semantic case and grammatical relations,* ed. W. Abraham, 459-576. Amsterdam: John Benjamins.

Starosta, Stanley. 1979a. 'The end of Phrase Structure as we know it'. *University of Hawaii Working Papers in Linguistics* 11:1, 59-76.

Starosta, Stanley. 1979b. 'Lexicase references'. *University of Hawaii Working Papers in Linguistics* 11:3, 79-85.

Starosta, Stanley. 1982. 'Lexical decomposition: features or atomic predicates? *Linguistic Analysis* 9.4: 379-93.

Starosta, Stanley. 1988. The case for lexicase. London: Pinter Publishers.

Starosta, Stanley. forthc. 'Lexicase: Versuch einer generativen Reformation'. *Deutsche Sprache.*

Thompson, Laurence. 1965. *A Vietnamese grammar*. Seattle: University of Washington Press.
Zide, Arlene. 1976. 'Nominal combining forms in Sora and Gorum'. In *Austroasiatic Studies*, ed. P. Jenner, L. Thompson and S. Starosta, 1259-94. Honolulu: University Press of Hawaii.