Testing a Grammar Customization System with Sahaptin

Scott Drellishak
University of Washington
Seattle, WA, USA
sfd@u.washington.edu

Abstract

I briefly describe a system for automatically creating an implemented grammar of a natural language based on answers to a web-based questionnaire, then present a grammar of Sahaptin, a language of the Pacific Northwest with complex argument-marking and agreement patterns, that was developed to test the system. The development of this grammar has proved useful in three ways: (1) verifying the correct functioning of the grammar customization system, (2) motivating the addition of a new pattern of agreement to the system, and (3) making detailed predictions that uncovered gaps in the linguistic descriptions of Sahaptin.

1 Introduction

The **LinGO Grammar Matrix** (Bender et al., 2002) is a resource for building implemented precision HPSG (Pollard and Sag, 1994) grammars of natural languages. Grammars based on the Matrix are expressed in the Type Description Language (TDL) (Krieger and Schäfer, 1994), are interpretable by the Linguistic Knowledge Building system (LKB) (Copestake, 2002) (a software tool for developing constraint-based grammars), and have semantic representations that are compatible with Minimal Recursion Semantics (MRS) (Copestake et al., 2005). The Grammar Matrix project, in particular the customization system described below, has drawn on the linguistics and linguistic typology literature during its development; the system is now complex enough that it is capable making contributions back to linguistics.

1.1 Matrix Customization System

In its earliest form, the Matrix provided a set of predefined types intended to give grammar engineers a head start, allowing them to avoid duplicating the effort required to develop analyses of linguistic structures thought to occur in all languages. However, there exist many linguistic phenomena that are widespread, but not universal. If the Matrix were restricted to supporting only what is truly universal, it would be a much less useful resource for grammar-writers working on languages containing such non-universal phenomena. Our solution has been to provide the **Matrix customization system**, which presents a linguist with a web-based typological questionnaire designed to elicit a description of a target language and, based on it, automatically produce a grammar that parses and generates the target language.[1] The grammars produced are not encumbered by phenomena that do not occur in the target language; rather, they contain just enough complexity to model it as described. Although the grammars produced by the customization system are intended as a starting point for further grammar engineering, that starting point is now far enough along that even without enhancement the grammars can be used for interesting linguistic work.

The customization system is conceived of as consisting of a set of **libraries**, each of which supports a particular linguistic phenomenon, and includes a section of the questionnaire and a syntactic analysis of the target phenomenon that can be

[1] A frozen version of the customization system as described here can be found on the Web at depts.washington.edu/uwcl/matrix/sfddiss/.

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 254–262,
Los Angeles, California, June 2010.

customized and included in output grammars. Recently, I have added three new libraries to the system (Drellishak, 2009). A library for case-marking supports a variety of patterns for the marking of up to two mandatory verbal arguments, including the nominative-accusative, ergative-absolutive, and tripartite patterns, as well as various split-ergative systems and Austronesian alignment (see Blake (2001) for definitions of these terms). A library for agreement supports agreement in syntactic and semantic features between verbs and their arguments. Finally, a library for so-called direct-inverse argument marking supports languages in which the marking of verbs and verbal arguments is conditioned on a grammatical scale—for example, languages in which clauses with a first person subject and a second person object are marked differently than clauses with a second person subject and a first person object. Languages can contain none, some, or all of these phenomena, and the customization system must produce consistent grammars for every combination.

1.2 Testing the Customization System

Work to add new libraries to the customization system is ongoing. Since the grammatical analyses of different phenomena can interact in unexpected ways, we utilize a system of regression testing to verify that the implementation new libraries does not break older libraries.

A customization system regression test consists of three parts. First, each test includes a stored set of answers to the questionnaire describing a language that illustrates one or more linguistic phenomena; this can be fed into the customization system to create a grammar. Second, each test has a list of strings, some grammatical and some ungrammatical in the test's language, that probe the behavior of the grammar with respect to the phenomena in question. Third, each test has the expected results, including semantic representations in the format of Oepen (2001), that are produced by the grammar when it parses the test sentences.

At the time of this writing, the regression test suite includes 112 tests that fall roughly into two categories. The first category contains small artificial languages that illustrate a single phenomenon (e.g. nominative-accusative case marking or a particular word order). The second category contains larger grammars based on natural languages that illustrate a wider range of phenomena, and therefore test the interaction of the associated libraries. The largest and most complex test in the latter category is the regression test for Sahaptin.

2 Sahaptin

Sahaptin [uma] (Penutian) is a family of closely related dialects spoken in Washington, Idaho, and Oregon. The details of Sahaptin grammar are drawn primarily from a description of the language by Rigsby and Rude (1996) (henceforth R&R). It happens that Sahaptin contains extremely complex argument marking and agreement patterns that illustrate, in a single grammar, a number of phenomena covered by my recently-implemented Matrix libraries, including:

- Case marking on verbal arguments.
- Argument marking sensitive to a grammatical scale, including patterns analyzed here as proximate and obviative marking on third-person nominals.
- Two loci of agreement (a verbal prefix and a second-position enclitic) with both the subject and the object.
- A distinction in number between singular, dual, and plural on nominals, but only between singular and plural on agreement morphology.
- An inclusive/exclusive distinction in person reflected only in the second-position enclitic.

2.1 Sahaptin Grammar

This section contains a brief sketch of the structure of Sahaptin sentences. Consider the following simple sentence:

(1) *ín=aš á-tuxnana yáamaš-na*
I=1SG 3ABS-shot mule.deer-OBJ
'I shot the mule deer.' [uma]
(Rigsby and Rude, 1996, 676)

In (1) the first word consists of the first person singular pronoun in its unmarked form, the nominative, followed by a second-position enclitic that agrees with the pronoun. The second word is the verb, consisting of a verbal prefix appropriate to the person and number of the subject and object (glossed by

R&R as 3ABS, but see §3.6 below for a different analysis) and the verb stem. The third word consists of the noun stem meaning 'mule deer' and a suffix marking the objective case.

R&R describe several cases in Sahaptin, including an unmarked "nominative" case, a marked "objective" case, an "inverse ergative" case, and an "obviative ergative" case. In spite of their use of the term "ergative", R&R make it clear that the subject generally appears in the nominative case in both transitive and intransitive clauses, and that the object consistently appears in the objective case in transitive clauses. The "inverse ergative" and "obviative ergative" forms only occur with third person singular nominals, both nouns and pronouns, in addition to the subject and object forms, and they are used to distinguish the subject from the object in transitive clauses.

In addition to case marking on nominals, Sahaptin has two ways to cross-reference the arguments of verbs: a verbal prefix and a second-position enclitic that attaches to whichever word comes first in the sentence. R&R characterize the prefixes and enclitics in two ways: first, they provide a general description of the distribution of each; second, they provide detailed paradigms of intransitive and transitive sentence patterns that cover most, but not all, of the logical combinations.

Enclitic	**Description**
=naš ∼ *=aš* ∼ *=š*	"first-person singular"
=na	"first-person plural inclusive"
=nataš ∼ *=ataš* ∼ *=taš*	"first-person plural exclusive"
=nam ∼ *=am* ∼ *=m*	"second-person singular"
=pam	"second-person plural"
=maš	"second-person object with first-person subject (both singular)"
=mataš	"second-person object with first-person subject (one or both plural)"

Table 1: Sahaptin enclitics (Rigsby and Rude, 1996, 675)

R&R describe Sahaptin's second-position enclitics as shown in Table 1. Notice in particular that several of the enclitics are associated with a person and number, but R&R do not mention whether those values are associated with the subject or the object. The reason for this becomes clear when we examine the full paradigm of clauses. The enclitic *=nataš*, for example, occurs with first person plural exclusive subjects in intransitive clauses; in transitive clauses, however, it occurs when one argument is first person plural exclusive and the other is third person, regardless of which is the subject and which is the object. A similar pattern can be observed for *=na* and *=naš*. This variant of scale-sensitive argument marking motivated an enhancement to the customization system described in §5 below.

Prefix	**Description**
i-	"third-person nominative"
pa-	"third-person plural nominative"
á- ∼ *áw-*	"third-person absolutive"
pá-	"inverse"
patá- ∼ *patáw-*	"third-person plural subject with third-person object"

Table 2: Sahaptin prefixes (Rigsby and Rude, 1996, 675)

As for Sahaptin's verbal prefixes, R&R describe them as shown in Table 2.[2] These descriptions are less straightforward than those for the enclitics. In particular, the description of *á-* ∼ *áw-* as "absolutive" is misleading. Regarding that prefix, R&R write, "...this pronominal marks subjects in intransitive clauses when they are possessors, and objects in transitive clauses when the subject is first or second person." (675) In other words, it does not occur in all transitive clauses, and only in those intransitive clauses where the subject is possessive. Furthermore, all the prefixes above appear on the verb, not the nominal arguments, as one might expect for an "absolutive" affix. In spite of the use of the term "absolutive", the distribution of the prefix *á-* ∼ *áw-* does not give evidence of ergative alignment in Sahaptin. Similarly, although there is evidence of argument marking sensitive to a grammatical scale, the description of *pá-* as "inverse" is misleading, since that prefix does not appear if and only if the object outranks the subject.

[2]There are three further verbal prefixes in Sahaptin that mark reflexives and reciprocals, but there is currently no support for these phenomena in the customization system.

3 Sahaptin Test Case

The phenomena described above make Sahaptin an excellent test case for demonstrating the flexibility and expressive power of the customization system. In this section, I will show how a significant fragment of Sahaptin can be described in the customization system questionnaire, producing a grammar that correctly models some of the complexity of Sahaptin morphosyntax.

It should be noted that some aspects of Sahaptin are beyond the current capabilities of the customization system, so some simplifying assumptions were necessary. For instance, the customization system models complex morphosyntax but not complex morphophonology. In effect, the grammars it outputs expect a morpheme-by-morpheme gloss as input rather than orthography, leaving the problem of morphological analysis to other systems.[3] The Sahaptin test grammar therefore uses only a single spelling for each stem and morpheme, and the morphemes are separated by '-' or '=' characters. The facts of Sahaptin word order are also too complex for the customization system; in particular, it cannot model truly free word order (i.e., discontinuous noun phrases), and the attachment behavior of the second-position enclitic is similarly beyond its capability. However, given these simplifying assumptions, the customization system is capable of modeling all the agreement and marking patterns of Sahaptin intransitive and transitive clauses shown in Tables 7 and 8 in R&R (1996, 676).

After the design and implementation of the libraries for case, direct-inverse languages, and agreement were completed, the construction of the Sahaptin test case took only about 80 hours of work, including the creation of test sentences (described in more detail in §4 below), a linguistic analysis of Sahaptin, filling out the questionnaire to reflect that analysis, and debugging the answers to the questionnaire.

3.1 Word Order

In the test grammar, I treat Sahaptin as a VSO language, and the enclitic as a suffix on verbs. This means that the sentences recognized and generated by the grammar are in a legal word order—VSO sentences with the verb followed by the second-position enclitic are grammatical in Sahaptin—but there are other legal word orders that the test grammar will not accept. The analysis of the enclitic is therefore limited by the current capabilities of the customization system's word order library; however, if that library is enhanced in the future to support second-position clitics, the analysis presented below should transfer straightforwardly.

3.2 Number

I analyze Sahaptin as having three values of number: singular (sg), dual (du), and plural (pl). All three values are distinguished on pronouns, as shown in Table 3; however, agreement with enclitics and verbal prefixes only shows a singular/plural distinction (with dual pronouns agreeing with the plural morpheme). It will be necessary in several places for the grammar to refer to a non-singular category covering *du* and *pl*. The questionnaire allows the explicit description of such a category; however, it also allows the user to select multiple values for a feature, and from those values infers the existence of categories like *non-singular*. I have made use of the latter mechanism in this grammar.

Table 3 shows the Sahaptin pronoun forms that distinguish singular, dual, and plural; in the questionnaire, therefore, I specified a number value on each. So-called plural agreement morphemes, on the other hand, do not distinguish between the dual and plural so are simply specified as covering both values.

3.3 Person

Sahaptin distinguishes three values of person: first, second, and third. The enclitics (but, interestingly, not the pronouns) further distinguish a first person inclusive and first person exclusive. I filled out the person section of the questionnaire with answers reflecting the presence of an inclusive/exclusive distinction.

3.4 Case

As described above, Sahaptin has a nominative case that marks intransitive and transitive subjects and an objective case that marks transitive objects. This

[3] The construction of such systems is well-understood (Beesley and Karttunen, 2003), as is the method for hooking up such a system to the LKB.

	Singular		Dual		Plural	
	Subject	**Object**	**Subject**	**Object**	**Subject**	**Object**
1	*ín*	*ináy*	*napiiní*	*napiinamanáy*	*náma*	*naamanáy*
2	*ím*	*imanáy*	*imiiní*	*imiinamanáy*	*imáy*	*imaamanáy*
3	*pín*	*paanáy*	*piiní*	*piinamanáy*	*pmáy*	*paamanáy*
3 obv erg	*piiní*					
3 inv erg	*pním*					

Table 3: Umatilla Sahaptin Pronouns (Rigsby and Rude, 1996, 682–683)

is the common nominative-accusative pattern, so in the case section of the questionnaire I describe it as such. Note that I do *not* analyze the inverse ergative and obviative ergative as case; see §3.6 for details.

3.5 Direct-Inverse

I analyze Sahaptin as a direct-inverse language—that is, a language whose argument marking is sensitive to a grammatical scale—though one that lacks clear direct or inverse forms of the verb, with the exception of the *pá-* prefix. The scale I propose for Sahaptin is:

(2) 1P $>$ 2P $>$ 3P topic $>$ 3P non-topic

The customization system interprets this scale, creating a series of rules that constrain the value of a feature DIRECTION on verbs. This feature takes the values *direct* and *inverse* and can be used to constrain the form either of verbs themselves or of their arguments.

3.6 Other Features

I use two additional features in my analysis of Sahaptin: a semantic TOPICALITY feature and a syntactic PROXIMITY feature, both on nominals.

Marking of Sahaptin transitive clauses distinguishes between topical and non-topical third person arguments. There is no overt marking of topicality on nominals, but clausal marking is conditioned on pragmatic distinctions that influence the felicity of the sentence in different discourse contexts. In order to systematically test this aspect of Sahaptin grammar in terms of string grammaticality, I introduced an artificial mark on topical noun phrases, the suffix *-TOP*. This suffix constrains the value of the TOPICALITY feature on nominal indices.

I use the syntactic PROXIMITY feature to model the "inverse ergative" and "obviative ergative" forms of nominals. In Sahaptin transitive clauses, the inverse ergative occurs precisely when the subject is third person singular and the clause is inverse (that is, the object is higher on the scale). The obviative ergative occurs in exactly one case: when the subject is third person singular and the object is a topical third person singular. These "ergative" forms function very much like the so-called proximate and obviative forms in Algonquian languages. However, in contrast to those languages, I analyze Sahaptin as having three values of the PROXIMITY feature rather than two: *proximate*, corresponding to the inverse ergative *-ním*, which promotes the marked nominal up the scale; *obviative*, corresponding to the obviative ergative *-in*, which demotes the marked nominal down the scale; and *neutral*, the unmarked form, which does not affect the nominal's position on the scale.[4]

3.7 Lexicon

Having defined the necessary features and values, we can now describe the lexicon of the Sahaptin grammar, which includes lexical types and inflectional morphemes. In the questionnaire, inflectional morphology is described as a series of slots, each attaching to one or more lexical types or other slots, and each containing one or more morphemes, each of which in turn specifies features. In order to prevent spurious ambiguity, the features on each set of morphemes are specified in such a way that no morpheme overlaps another, but also so that no legal combination of features goes unexpressed.

The simplest grammars are those that do not resort to homophony—that is, they do not have multiple lexical items or morphemes with the same

[4]Note that, for consistency with R&R's description, I nonetheless continue to refer to the marked forms as the "inverse ergative" and "obviative ergative".

spelling but different semantics or features. It is often possible to avoid homophony by adding complexity to feature hierarchies, but overly complex hierarchies can be as difficult to manage as extensive homophony. In the Sahaptin grammar, I have attempted to strike a balance between homophony and hierarchy complexity. For example, to make the grammar easier for users to understand, I segregated verbal prefixes and enclitics each into two classes: those attaching to intransitive stems and those attaching to transitive stems. This produced two homophonous variants of the prefixes *i-* and *pa-*, and of the enclitics *=naš*, *=na*, *=nataš*, *=nam*, and *=pam*. Furthermore, the distributions of two transitive prefixes (*pá-* and the null variant) and of three transitive enclitics (*=nam*, *=pam*, and *=mataš*) were easier to model using homophonous variants. Finally, the third person singular obviative pronoun and the third person dual subject pronoun are both *piiní* (as shown in Table 3) and it seemed simplest to represent these using two separate lexical entries. The grammar, then, contains 22 lexical items, of which only two are homophonous, and 24 non-null inflectional morphemes representing 12 distinctly spelled prefixes and enclitics.

A full description of the morphosyntactic details of the Sahaptin test grammar would be too long for this paper; instead, I will provide a summary.[5] The lexicon of the test grammar contains six inflectional slots: a slot for the topic morpheme described above that attaches to nominals; a slot for verbal prefixes that attach to intransitive verbs; a slot for verbal prefixes that attach to transitive verbs; a slot for enclitics that attach to intransitive verbs; a slot for enclitics that attach to transitive verbs; and a slot that contains no overt morphemes, but is used to produce lexical rules that constrain the appearance of topic, proximate, and obviative on a verb's nominal arguments. Each of these slots contains morphemes, on which are specified values for one or more features. To give an idea of what this looks like, Table 4 shows the features that are defined for the most complex of these slots, the one that contains transitive prefixes.

[5]The full details of the Sahaptin grammar can be found in my dissertation (Drellishak, 2009). How the questionnaire can be filled out to model Sahaptin can be seen by visiting the customization system web site at `depts.washington.edu/uwcl/matrix/sfddiss/` and clicking the Umatilla Sahaptin link at the bottom of the main page, which fills out the questionnaire automatically.

4 Testing the Sahaptin Grammar

In order to test the correctness of the Sahaptin grammar, it was necessary to create a suite of test sentences, some grammatical and some not, that probe its expected lexical and grammatical coverage. I started with the sentence patterns in R&R's Tables 7 and 8 (Rigsby and Rude, 1996, 676); from each, I created a sentence with the appropriate prefix, verb, enclitic, subject, and object. In every case where a plural argument was called for, I actually created two sentences, one with a dual argument—and in cases with two plural arguments, I created four: *du/du*, *du/pl*, *pl/du*, and *pl/pl*.

All these sentences were expected to be grammatical based on the descriptions in R&R. To generate ungrammatical sentences, I initially permuted the grammatical sentences in the following ways:

1. For each grammatical sentence with a prefix, I created an ungrammatical variant with the prefix missing.
2. For each grammatical sentence with an enclitic, I created an ungrammatical variant with the enclitic missing.
3. For each grammatical sentence, I created variants that contained every incorrect prefix and variants that contained every incorrect enclitic.

After duplicates were removed, this produced a list of 89 grammatical and 220 ungrammatical sentences, for a total of 309.

The permutation of the grammatical sentences as described above was sufficient to test the phenomena of interest for intransitive sentences, producing ungrammatical sentences consisting of correctly-formed words in the correct basic word order but with an ungrammatical agreement pattern, and this permutation was a small enough job to perform by hand. For transitive sentences, though, there is a much larger space of sentences with the right word order but wrong agreement, so in order to test the grammar thoroughly, I decided to supplement the ungrammatical sentences I created by hand by writing a small program to generate every sentence containing the verb *ɖínun* 'see' that followed the pattern:

Transitive prefix	**Subject** PERNUM	**Subject** TOPICALITY	**Object** PERNUM	**Object** TOPICALITY
i-	*3sg*			*non-topic*
pa-	*3du, 3pl*			*non-topic*
á-	*1st, 2nd*		*3rd*	
pá-	*2sg*		*1sg*	
pá-	*3sg*	*non-topic*	*3sg*	*topic*
patá-	*3du, 3pl*	*non-topic*	*3sg*	*topic*
∅	*1st*		*2nd*	
∅	*2du, 2pl*		*1st*	
∅	*2sg*		*1du, 1pl*	

Table 4: Morphemes appearing in the transitive prefix slot

(3) **prefix-*ɖínun*=enclitic subject object**

The possible fillers for each position in (3) are shown in Table 5:

prefix	*i-, pa-, á-, pá-, patá-,* and ∅
enclitic	*=naš, =na, =nataš, =nam, =pam, =maš, =mataš,* and ∅
subject	subject forms in Table 3
object	object forms in Table 3

Table 5: Fillers for positions in (3)

As mentioned above, the lexicon of the Sahaptin grammar, and consequently the test sentences, uses the various forms of the personal pronoun to represent the possible person, number, case, and proximity values of subject and object noun phrases. In addition to plain case-marked pronouns, the subject and object positions may also contain third person pronouns marked as the topic with *-TOP*.

Generating every sentence that followed the pattern in (3) produced 6048 sentences, but some additional filtering was required. First, since it appears that topic marking is only relevant when disambiguating third person arguments, I removed all sentences where the *-TOP* suffix appeared with a first or second person pronoun. Second, 192 of the permutations of (3) are actually duplicates of the ungrammatical transitive test sentences created by hand above, so I removed those as well. After filtering, a total of 5856 programmatically-generated sentences remained. Added to the aforementioned 309 examples, this made 6165 unique test sentences.

After using the customization system to generate a grammar of Sahaptin, I used that grammar to attempt to parse every test sentence. All 89 sentences corresponding to R&R's grammatical transitive and intransitive patterns parsed and were assigned exactly one analysis.[6] Among the ungrammatical sentences, 5848 out of 5856 failed to parse, as expected. To my surprise, however, eight of the sentences did parse. These sentences were:

(4) a. *i-ɖínun pm̀-TOP piinamanáy*
3SG-see 3SG.NOM-TOP 3DU.OBJ
'He saw them (DU).'
b. *i-ɖínun pm̀-TOP paamanáy*
3SG-see 3SG.NOM-TOP 3PL.OBJ
'He saw them.'
c. *pa-ɖínun piiní paanáy*
3NONSG-see 3DU.NOM 3SG.OBJ
'They (DU) saw him.'
d. *pa-ɖínun pmáy paanáy*
3NONSG-see 3PL.NOM 3SG.OBJ
'They saw him.'
e. *pa-ɖínun piiní-TOP piinamanáy*
3NONSG-see 3DU.NOM-TOP 3DU.OBJ
'They (DU) saw them (DU).'
f. *pa-ɖínun piiní-TOP paamanáy*
3NONSG-see 3DU.NOM-TOP 3PL.OBJ
'They (DU) saw them.'
g. *pa-ɖínun pmáy-TOP piinamanáy*
3NONSG-see 3PL.NOM-TOP 3DU.OBJ
'They saw them (DU).'
h. *pa-ɖínun pmáy-TOP paamanáy*
3NONSG-see 3PL-TOP.NOM 3PL.OBJ
'They saw them.'

[6] Multiple analyses would not necessarily have been wrong—some sentences in some languages are structurally ambiguous—but the grammatical Sahaptin sentences in the test suite are marked explicitly enough for agreement that none was ambiguous.

Notice that the eight sentences fall into three patterns. The first two sentences have a third person singular topical subject and a third person non-singular non-topical object, the next two have a third person non-singular non-topical subject and a third person singular non-topical object, and the last four have a third person non-singular topical subject and a third person non-topical object. These are precisely the patterns that are absent from R&R's Table 8; corresponding sentences were therefore not included in the list of 89 grammatical sentences. In developing the Sahaptin grammar, I had, without considering these eight patterns, defined the prefixes in such a way that the grammar expected *i-* to appear in the first two sentences and *pa-* in the last six.

In order to determine whether this analysis was correct, Sharon Hargus presented the Yakima Sahaptin equivalents of the sentences in (4) by telephone to Virginia Beavert, a native speaker of that dialect, who accepted all eight of them with the readings shown in (4). Note that, in order for these sentences to be acceptable, they had to be cast in the past tense, a feature not modeled in my Sahaptin grammar fragment. Note also that Dr. Beavert considered sentence (4c) somewhat less acceptable, saying that it is "[a] little awkward, but has meaning."

The Sahaptin grammar, then, which was created using the customization system and based on its support for case, direct-inverse languages, and agreement, correctly analysed all 6165 of the test sentences, including eight that fell outside of the patterns described in the linguistic literature.

5 Summary and Discussion

Based on these results, I conclude that even Sahaptin, a language with extremely complex argument marking morphology, can be modeled using the customization system. Note that the system was not designed with the facts of Sahaptin in mind, and with two exceptions, the system did not need to be modified to enable it to handle Sahaptin.

One of the exceptions was trivial: formerly, grammars produced by the system treated '=' as punctuation, stripping it out and breaking words containing it. The other exception concerns an unusual agreement pattern I first encountered in Sahaptin: morphemes that agree, not with the subject or the object of a verb, but with the nominal argument that is more highly ranked on the direct-inverse scale. Supporting this agreement pattern proved worthwhile later, when it was used again in a test grammar for Plains Cree [crk] (Algonquian), another direct-inverse language. Although this latter change was a substantive one that allows grammars to be described more compactly, it did not increase the descriptive power of the system—languages showing that pattern of agreement could still be modeled using duplicated, homophonous morphemes. Such an enhancement to the system is an example of the feedback loop between grammar engineering and customization system development, where new languages with new phenomena (or new variations of old phenomena) inform the design and, in some cases, the descriptive power of the system.

After constructing the Sahaptin grammar and test suite described here, it was natural to include it in two places in the customization system. First, it is now one of the regression tests that is regularly run to ensure that future enhancement of the system does not break earlier features. Second, Sahaptin has been added to the list of sample grammars accessible from the main page of the questionnaire—by clicking on links in this list, users can see detailed examples of how the questionnaire can be filled out to model a target language.

The Sahaptin grammar, developed using the customization system, has proved itself useful—not only to the Grammar Matrix project, where it inspired the addition of support for scale-sensitive agreement and serves as a regression test of the correct functioning of the system, but also to the field of linguistics. By analyzing Sahaptin in the precise detail required by the customization system, I found unnoticed gaps in linguistic descriptions of the language, and in collaboration with linguists studying the language was able to help resolve those gaps.

Acknowledgments

My thanks go to Emily Bender and the Matrix team, Sharon Hargus, and Virginia Beavert. This work was supported by a gift to the Turing Center from the Utilika Foundation, by the Max Planck Institute for Evolutionary Anthropology, and by the National Science Foundation under Grant No. 0644097.

References

[Beesley and Karttunen2003] Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI, Stanford.

[Bender et al.2002] Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix. In *Proceedings of COLING 2002 Workshop on Grammar Engineering and Evaluation*, Taipei, Taiwan.

[Blake2001] Barry J. Blake. 2001. *Case, Second Edition*. Cambridge University Press, Cambridge.

[Copestake et al.2005] Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language & Computation*, 3(2–3):281–332.

[Copestake2002] Ann Copestake. 2002. *Implementing Typed Feature Structure Grammars*. CSLI, Stanford.

[Drellishak2009] Scott Drellishak. 2009. *Widespread, but Not Universal: Improving the Typological Coverage of the Grammar Matrix*. Ph.D. thesis, University of Washington.

[Krieger and Schäfer1994] Hans-Ulrich Krieger and Ulrich Schäfer. 1994. Tdl – a type description language for constraint-based grammars. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 893–899, Kyoto, Japan.

[Oepen2001] Stephan Oepen. 2001. [incr tsdb()] — Competence and performance laboratory. User manual. Technical report, Saarbrücken, Germany.

[Pollard and Sag1994] Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. CSLI, Stanford.

[Rigsby and Rude1996] Bruce Rigsby and Noel Rude. 1996. Sketch of sahaptin, a sahaptian language. In Ives Goddard, editor, *Languages*, pages 666–92. Smithsonian Institution, Washington DC.