2009

Computing in the field: Automated elicitation & documentation

Andrew R. Hippisley
University of Kentucky, andrew.hippisley@uky.edu

Gregory Stump
University of Kentucky, gstump@uky.edu

Raphael Finkel
University of Kentucky, raphael.finkel@uky.edu

Computing in the field: Automated elicitation & documentation

Andrew Hippisley, Gregory Stump & Raphael Finkel

University of Kentucky

{a.hippisley, g.stump}@uky.edu, raphael@cs.uky.edu

1. Introduction

University of Kentucky

Amanda Barie
Stanley Brunn
Darya Bukhtoyarova
Raphael Finkel
Andrew Hippisley
Mark Lauersdorf
Jeanmarie Rouhier-Willoughby
Gregory Stump

Khorog State University,
***Institute of Humanities of the Tajik Academy of Sciences**

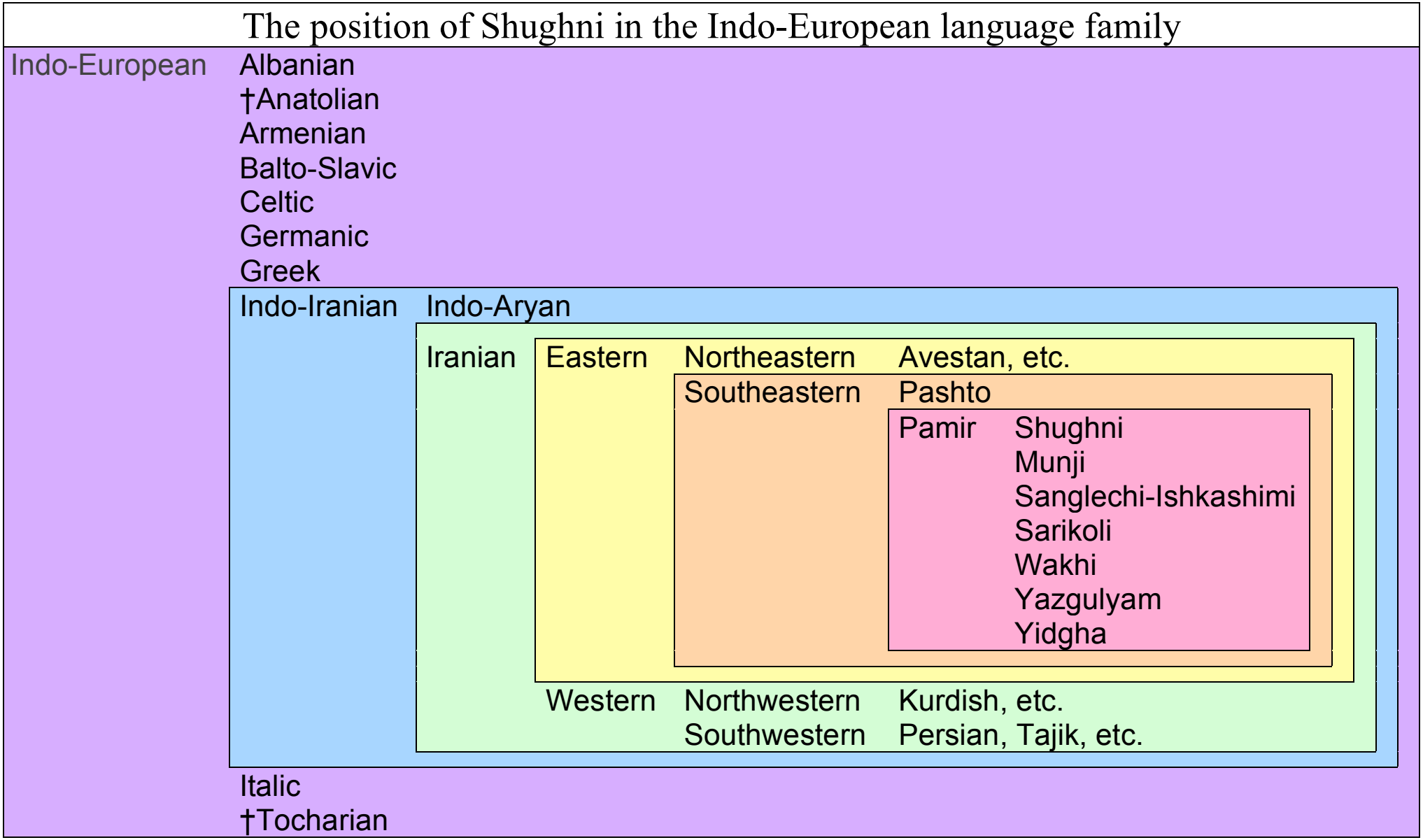
Muqbilsho Alamshoev*
Shoxnazar Mirzoev*
Gulnoro Mirzovafoeva
Shahlo Nekushoeva*

Shughni Summer Workshop, University of Kentucky, July 2008

Shughni Summer Workshop, University of Kentucky, July 2008

2. The Shughni language


KAZAKHSTAN
ELEVATION:
4 000 m
2 000 m
1 000 m
200 m
Ozero Aydarkul
Syr-Dar'ya
Khodzhent
Kanibadam
UZBEKISTAN
KYRGYZ REPUBLIC
Ura-
Tuybe
Kyzylsu
Samarkand
Pendzhikent
Ayni
Zeravshan
Surkhob
Muksu
CHINA
Novabad
▲Pik Kommunizma
7495 m
Pik Revolyutsii
6974 m ▲
P A M I R
Dushanbe
TAJIKISTAN
Tursunzade
Bartang
Kurgan-
Tyube
Shazud
Kulyab
Gunt
Khorog
Amudar'ya
Kafirnigan
Vakhsh
Pyandzh
AFGHANISTAN
PAKISTAN
0
100
200
300 km
GRID-Arendal 1998


The position of Shughni in the Indo-European language family
Indo-European
Albanian
†Anatolian
Armenian
Balto-Slavic
Celtic
Germanic
Greek
Indo-Iranian
Indo-Aryan
Iranian
Eastern
Northeastern
Avestan, etc.
Southeastern
Pashto
Pamir
Shughni
Munji
Sanglechi-Ishkashimi
Sarikoli
Wakhi
Yazgulyam
Yidgha
Western
Northwestern
Kurdish, etc.
Southwestern
Persian, Tajik, etc.
Italic
†Tocharian

Grammatical descriptions

Bahtibekov, T. 1979. *Grammatikai Zaboni Šuγnoni* [Grammar of the Shughni language]. Dushanbe.

Dodykhudoeva, Lelia R. 1988. Shugnanskiĭ glagol v istoricheskom osveshchenii [Shughni verbs in historical perspective]. Dushanbe.

Karamshoev, Dodkhudo. 1986. *Kategorija roda v pamirskih jazykah (shugnano-rushanskaja gruppa)* [The category of gender in the Pāmir Languages, Shughni-Roshani group]. Dushanbe.

Nawata, Tetsuo. 1979. *Shughni* (Asian and African grammatical manual 17s). Tokyo.

History

Payne, John. 1980. The decay of ergativity in Pamir languages. *Lingua* 51, 147-186.

Payne, John. 1981. Iranian Languages. In Bernard Comrie (ed.), *The Languages of the Soviet Union*. Cambridge. 158-179.

Payne, John. 1989. Pāmir languages. In Rüdiger Schmitt (ed.), *Compendium linguarum Iranicarum*, 417-444. Wiesbaden.

Dictionaries

Karamshoev, Dodkhudo. 1988-1999. *Shugnansko-russkiĭ slovar'* [Shughni-Russian Dictionary], 3 vols. Moscow.

Zarubin, Ivan Ivanovich, ed. 1960. *Shugnanskie teksty i slovar'* [Shughni texts and dictionary]. Moscow & Leningrad.

3. Morphological elicitation

Inflection of *wiftow* ‘knit’

Nonpast	wuz	wāf-um	‘I knit’
	tu	wāf-i	‘you (sg.) knit’
	yu / yā	wof-t	‘he / she knits’
	māš	wāf-am	‘we knit’
	tama	wāf-et	‘you (pl.) knit’
	wāδ	wāf-en	‘they knit’
Past	wuz = um	wīft	‘I knitted’
	tu = t	wīft	‘you (sg.) knitted’
	yu = yi / yā = yi	wīft	‘he / she knitted’
	māš = ām	wīft	‘we knitted’
	tam = et	wīft	‘you (pl.) knitted’
	wāδ = en	wīft	‘they knitted’

Inflection of *wirīvdow* ‘stand’

Nonpast	wuz	wirāfc-um	‘I am standing’
	tu	wirāfc-i	‘you (sg.) are standing’
	yu / yā	wirofc-t	‘he / she is standing’
	māš	wirāfc-am	‘we are standing’
	tama	wirāfc-et	‘you (pl.) are standing’
	wāδ	wirāfc-en	‘they are standing’
Past	wuz = um	wirūvd	‘I (masc.) stood’
	wuz = um	wirovd	‘I (fem.) stood’
	tu = t	wirūvd	‘you (masc. sg.) stood’
	tu = t	wirovd	‘you (fem. sg.) stood’
	yu	wirūvd	‘he stood’
	yā	wirovd	‘she stood’
	māš = ām	wirovd	‘we stood’
	tam = et	wirovd	‘you (pl.) stood’
	wāδ = en	wirovd	‘they stood’

4. Default inheritance & morphological generation

In order to investigate verb morphology in a heavily inflected language, it is necessary to postulate not just individual word forms, but rather ***entire paradigms***. A computer program for ***morphological generation*** is well suited to this purpose.

In our research, we have integrated automatic morphological generation ***into the elicitation process***: A native speaker evaluates the generated paradigms; where necessary, we revise the generation program and confirm the validity of its subsequent output.

The most suitable morphological generation program for use in the elicitation process is one which models morphology as a ***default inheritance hierarchy***: a program of this sort allows the morphology of a language to be modelled very succinctly and allows revisions (with potentially far-reaching consequences) to be made quickly and easily.

DATR and KATR

Evans, Roger & Gerald Gazdar. 1996. DATR: A language for lexical knowledge representation. *Computational Linguistics* 22, 167-216.

Raphael Finkel, Lei Shen, Gregory Stump & Suresh Thesayi. 2002. 'KATR: A Set-Based Extension of DATR', Technical Report No. 346-02, Department of Computer Science, University of Kentucky.

Realizational approaches to morphology

Corbett, Greville G. & Norman M. Fraser. 1993. Network Morphology: A DATR account of Russian nominal inflection. *Journal of Linguistics* 29, 113-142.

Hippisley, Andrew. 1997. Declarative Derivation: A Network Morphology Account of Russian Word Formation with Reference to Nouns Denoting `Person', Unpublished PhD thesis, University of Surrey.

Stump, Gregory T. 2001. *Inflectional Morphology*. Cambridge University Press.

The verb hierarchy

```
Verb:
  {} == SubjectPronoun Adverb , "<stemPresent>" Agreement eow
  {past} == SubjectPronoun "<auxiliary>" , "<wordformPast>"
  {perfect} == SubjectPronoun "<auxiliary>" , "<wordformPerfect>"
  {auxiliary} == Agreement
  {perfectSuffix} == - č
  {wordformPast} == "<stemPast>" - t
  {wordformPerfect} == "<stemPerfect>" "<perfectSuffix>"
  {stemPerfect} == "<stemPast>"
  {stemPast} == "<stemPresent>"
  .
```

```
MiddleVerb:
  {auxiliary 3 sg} ==
  {perfectSuffix fem sg} == - c
  {} == Verb
  .

ActiveVerb:
  {auxiliary 3 sg} == - i
  {} == Verb
  .
```

```
Agreement:
  {1 sg} == - u m
  {2 sg} == - i
  {3 sg} == - t
  {auxiliary 2 sg} == - a t
  {1 pl} == - ā m
  {2 pl} == - e t
  {3 pl} == - e n
  .
```

```
SubjectPronoun:
  {1 sg} == w u z
  {2 sg} == t u
  {3 sg masc} == y u
  {3 sg fem} == y ā
  {1 pl} == m ā š
  {2 pl} == t a m a
  {3 pl} == w ā ð
  .
```

A regular lexical entry

```
Disturb:
    {stemPresent} == wiš
    {} == ActiveVerb
    .
```

Theorem of ‘disturb’

	sg			pl		
present	sg	masc	fem	pl	masc	fem
	1	wuz wiš-um	wuz wiš-um	1	māš wiš-ām	māš wiš-ām
	2	tu wiš-i	tu wiš-i	2	tama wiš-et	tama wiš-et
	3	yu wiš-t	yā wiš-t	3	wāð wiš-en	wāð wiš-en
past	sg	masc	fem	pl	masc	fem
	1	wuz-um wiš-t	wuz-um wiš-t	1	māš-ām wiš-t	māš-ām wiš-t
	2	tu-yat wiš-t	tu-yat wiš-t	2	tama-yet wiš-t	tama-yet wiš-t
	3	yu-yi wiš-t	yā-yi wiš-t	3	wāð-en wiš-t	wāð-en wiš-t
perfect	sg	masc	fem	pl	masc	fem
	1	wuz-um wiš-č	wuz-um wiš-č	1	māš-ām wiš-č	māš-ām wiš-č
	2	tu-yat wiš-č	tu-yat wiš-č	2	tama-yet wiš-č	tama-yet wiš-č
	3	yu-yi wiš-č	yā-yi wiš-č	3	wāð-en wiš-č	wāð-en wiš-č
future	sg	masc	fem	pl	masc	fem
	1	wuz-ta wiš-um	wuz-ta wiš-um	1	māš-ta wiš-ām	māš-ta wiš-ām
	2	tu-ta wiš-i	tu-ta wiš-i	2	tama-ta wiš-et	tama-ta wiš-et
	3	yu-ta wiš-t	yā-ta wiš-t	3	wāð-ta wiš-en	wāð-ta wiš-en

5. Elicitation query generation: A demonstration

Cycle 1: Start with Theory 1

1. Computational model based on standard lexical entries to produce theorem consistent with language consultant
2. Non-standard lexical entry of type 1 plugged into model, produces theorem inconsistent with language consultant
3. Model constrained to produce all theorems consistent with language consultant – result is Theory 2

Cycle 2: Start with Theory 2

1. Computational model based on standard + non-standard type 1 lexical entries
2. Non-standard lexical entry of type 2 plugged into model, produces theorem inconsistent with language consultant
3. Model constrained to produce all theorems consistent with language consultant – result is Theory 3.

Cycle n results in Theory $n + 1$, and may lead to the further Cycle $n + 1$.

Example 1: Morphonological overgeneralization

```
Buzz:
    {stemPresent} == b ā ɣ
    {} == ActiveVerb
    .
```

Overgeneralized theorem for ‘buzz’

	sg			pl		
present	sg	masc	fem	pl	masc	fem
	1	wuz bāɣ-um	wuz bāɣ-um	1	māš bāɣ-ām	māš bāɣ-ām
	2	tu bāɣ-i	tu bāɣ-i	2	tama bāɣ-et	tama bāɣ-et
	3	yu bāɣ-t	yā bāɣ-t	3	wāð bāɣ-en	wāð bāɣ-en
past	sg	masc	fem	pl	masc	fem
	1	wuz-um bāɣ-t	wuz-um bāɣ-t	1	māš-ām bāɣ-t	māš-ām bāɣ-t
	2	tu-yat bāɣ-t	tu-yat bāɣ-t	2	tama-yet bāɣ-t	tama-yet bāɣ-t
	3	yu-yi bāɣ-t	yā-yi bāɣ-t	3	wāð-en bāɣ-t	wāð-en bāɣ-t
perfect	sg	masc	fem	pl	masc	fem
	1	wuz-um bāɣ-č	wuz-um bāɣ-č	1	māš-ām bāɣ-č	māš-ām bāɣ-č
	2	tu-yat bāɣ-č	tu-yat bāɣ-č	2	tama-yet bāɣ-č	tama-yet bāɣ-č
	3	yu-yi bāɣ-č	yā-yi bāɣ-č	3	wāð-en bāɣ-č	wāð-en bāɣ-č
future	sg	masc	fem	pl	masc	fem
	1	wuz-ta bāɣ-um	wuz-ta bāɣ-um	1	māš-ta bāɣ-ām	māš-ta bāɣ-ām
	2	tu-ta bāɣ-i	tu-ta bāɣ-i	2	tama-ta bāɣ-et	tama-ta bāɣ-et
	3	yu-ta bāɣ-t	yā-ta bāɣ-t	3	wāð-ta bāɣ-en	wāð-ta bāɣ-en

Fixing ‘buzz’ by fixing the model

```
#sandhi $voicedObstruent - č => $1 - ǰ .
#sandhi $voicedObstruent - t => $1 - d .

#vars $voicedObstruent: b ž z ɣ v ʒ g d ǰ ð.
```

Correct theorem for ‘buzz’

	sg		pl	
present	sg	masc	fem	
	1	wuz bāɣ-um	wuz bāɣ-um	
	2	tu bāɣ-i	tu bāɣ-i	
	3	yu bāɣ-d	yā bāɣ-d	
	pl	masc	fem	
	1	māš bāɣ-ām	māš bāɣ-ām	
	2	tama bāɣ-et	tama bāɣ-et	
	3	wāð bāɣ-en	wāð bāɣ-en	
past	sg	masc	fem	
	1	wuz-um bāɣ-d	wuz-um bāɣ-d	
	2	tu-yat bāɣ-d	tu-yat bāɣ-d	
	3	yu-yi bāɣ-d	yā-yi bāɣ-d	
	pl	masc	fem	
	1	māš-ām bāɣ-d	māš-ām bāɣ-d	
	2	tama-yet bāɣ-d	tama-yet bāɣ-d	
	3	wāð-en bāɣ-d	wāð-en bāɣ-d	
perfect	sg	masc	fem	
	1	wuz-um bāɣ-ǰ	wuz-um bāɣ-ǰ	
	2	tu-yat bāɣ-ǰ	tu-yat bāɣ-ǰ	
	3	yu-yi bāɣ-ǰ	yā-yi bāɣ-ǰ	
	pl	masc	fem	
	1	māš-ām bāɣ-ǰ	māš-ām bāɣ-ǰ	
	2	tama-yet bāɣ-ǰ	tama-yet bāɣ-ǰ	
	3	wāð-en bāɣ-ǰ	wāð-en bāɣ-ǰ	
future	sg	masc	fem	
	1	wuz-ta bāɣ-um	wuz-ta bāɣ-um	
	2	tu-ta bāɣ-i	tu-ta bāɣ-i	
	3	yu-ta bāɣ-d	yā-ta bāɣ-d	
	pl	masc	fem	
	1	māš-ta bāɣ-ām	māš-ta bāɣ-ām	
	2	tama-ta bāɣ-et	tama-ta bāɣ-et	
	3	wāð-ta bāɣ-en	wāð-ta bāɣ-en	

Example 2: Stem overgeneralization

```
See:
    {stemPresent} == w i n
    {} == ActiveVerb
    .
```

Overgeneralized theorem for ‘see’

	sg			pl		
present	sg	masc	fem	pl	masc	fem
	1	wuz win-um	wuz win-um	1	māš win-ām	māš win-ām
	2	tu win-i	tu win-i	2	tama win-et	tama win-et
	3	yu win-t	yā win-t	3	wāð win-en	wāð win-en
past	sg	masc	fem	pl	masc	fem
	1	wuz-um win-t	wuz-um win-t	1	māš-ām win-t	māš-ām win-t
	2	tu-yat win-t	tu-yat win-t	2	tama-yet win-t	tama-yet win-t
	3	yu-yi win-t	yā-yi win-t	3	wāð-en win-t	wāð-en win-t
perfect	sg	masc	fem	pl	masc	fem
	1	wuz-um win-č	wuz-um win-č	1	māš-ām win-č	māš-ām win-č
	2	tu-yat win-č	tu-yat win-č	2	tama-yet win-č	tama-yet win-č
	3	yu-yi win-č	yā-yi win-č	3	wāð-en win-č	wāð-en win-č
future	sg	masc	fem	pl	masc	fem
	1	wuz-ta win-um	wuz-ta win-um	1	māš-ta win-ām	māš-ta win-ām
	2	tu-ta win-i	tu-ta win-i	2	tama-ta win-et	tama-ta win-et
	3	yu-ta win-t	yā-ta win-t	3	wāð-ta win-en	wāð-ta win-en

Fixing ‘see’ by fixing lexical entry

```
See:
    {stemPresent} == w i n
    {} == ActiveVerb
    {stemPast} == w ī n
    {stemPresent 3 sg} == w ī n
    .
```

Fixing ‘see’ by introducing a generalization

```
Verb:
%    {stemPast} == “<stemPresent>”
     {stemPast} == “<stemPresent 3 sg>”
     . . .
```

‘see’ generalization (nearly) predicting stem for ‘stand’

	sg			pl		
present	sg	masc	fem	pl	masc	fem
	1	wuz wirāfc-um	wuz wirāfc-um	1	māš wirāfc-ām	māš wirāfc-ām
	2	tu wirāfc-i	tu wirāfc-i	2	tama wirāfc-et	tama wirāfc-et
	3	yu wirofs-t	yā wirofs-t	3	wāð wirāfc-en	wāð wirāfc-en
past	sg	masc	fem	pl	masc	fem
	1	wuz-um wirūv-d	wuz-um wirofs-t	1	māš-ām wirofs-t	māš-ām wirofs-t
	2	tu-yat wirūv-d	tu-yat wirofs-t	2	tama-yet wirofs-t	tama-yet wirofs-t
	3	yu wirūv-d	yā wirofs-t	3	wāð-en wirofs-t	wāð-en wirofs-t
perfect	sg	masc	fem	pl	masc	fem
	1	wuz-um wirūv-ǰ	wuz-um wirofs-c	1	māš-ām wirofs-č	māš-ām wirofs-č
	2	tu-yat wirūv-ǰ	tu-yat wirofs-c	2	tama-yet wirofs-č	tama-yet wirofs-č
	3	yu wirūv-ǰ	yā wirofs-c	3	wāð-en wirofs-č	wāð-en wirofs-č
future	sg	masc	fem	pl	masc	fem
	1	wuz-ta wirāfc-um	wuz-ta wirāfc-um	1	māš-ta wirāfc-ām	māš-ta wirāfc-ām
	2	tu-ta wirāfc-i	tu-ta wirāfc-i	2	tama-ta wirāfc-et	tama-ta wirāfc-et
	3	yu-ta wirofs-t	yā-ta wirofs-t	3	wāð-ta wirāfc-en	wāð-ta wirāfc-en

(semi) fixing ‘stand’ through lexical specification

```
Stand:
    {stemPresent} == w i r ā f c
    {stemPresent 3 sg} == w i r o f s
    {stemPast sg masc} == w i r ū v
    {stemPerfect sg fem} == w i r ī v
    {} == MiddleVerb
    .

Verb:
     {stemPast} == “<stemPresent 3 sg>”
     {stemPerfect} == “<stemPast>”
      . . .
```

(nearly) correct theorem for ‘stand’

	sg		pl	
present	sg	masc	fem	
	1	wuz wirāfc-um	wuz wirāfc-um	
	2	tu wirāfc-i	tu wirāfc-i	
	3	yu wirofs-t	yā wirofs-t	
	pl	masc	fem	
	1	māš wirāfc-ām	māš wirāfc-ām	
	2	tama wirāfc-et	tama wirāfc-et	
	3	wāð wirāfc-en	wāð wirāfc-en	
past	sg	masc	fem	
	1	wuz-um wirūv-d	wuz-um wirofs-t	
	2	tu-yat wirūv-d	tu-yat wirofs-t	
	3	yu wirūv-d	yā wirofs-t	
	pl	masc	fem	
	1	māš-ām wirofs-t	māš-ām wirofs-t	
	2	tama-yet wirofs-t	tama-yet wirofs-t	
	3	wāð-en wirofs-t	wāð-en wirofs-t	
perfect	sg	masc	fem	
	1	wuz-um wirūv-ǰ	wuz-um wirīv-ʒ	
	2	tu-yat wirūv-ǰ	tu-yat wirīv-ʒ	
	3	yu wirūv-ǰ	yā wirīv-ʒ	
	pl	masc	fem	
	1	māš-ām wirofs-č	māš-ām wirofs-č	
	2	tama-yet wirofs-č	tama-yet wirofs-č	
	3	wāð-en wirofs-č	wāð-en wirofs-č	
future	sg	masc	fem	
	1	wuz-ta wirāfc-um	wuz-ta wirāfc-um	
	2	tu-ta wirāfc-i	tu-ta wirāfc-i	
	3	yu-ta wirofs-t	yā-ta wirofs-t	
	pl	masc	fem	
	1	māš-ta wirāfc-ām	māš-ta wirāfc-ām	
	2	tama-ta wirāfc-et	tama-ta wirāfc-et	
	3	wāð-ta wirāfc-en	wāð-ta wirāfc-en	

Overgeneralization Type	*Theory Refinement*	*Example*
Rule is completely accurate but incomplete	Add a complementary rule	rule of voicing assimilation affecting past tense suffix *-t*
rule is sometimes accurate, sometimes not	introduce overrides to rule	rule overriding default identity of a verb's present stems
rule is only superficially accurate	replace the rule	past stem = present stem replaced by past stem = 3 sg present stem

6. Conclusion

- technology of morphological generation is a quick and accurate hypothesis tester for data elicitation verification
- hypothesis by default, cyclical hypothesis refinement through extension, overrides and substitution
- consultant as system evaluator
- outcome is formal and informed description of the language

- compact theory generating exhaustive set of theorems