The use of formal language models in the typology of the morphology of Amerindian languages

Andrés Osvaldo Porta
Universidad de Buenos Aires
hugporta@yahoo.com.ar

Abstract

The aim of this work is to present some preliminary results of an investigation in course on the typology of the morphology of the native South American languages from the point of view of the formal language theory. With this object, we give two contrasting examples of descriptions of two Aboriginal languages finite verb forms morphology: Argentinean Quechua (quichua santiagueño) and Toba. The description of the morphology of the finite verb forms of Argentinean quechua, uses finite automata and finite transducers. In this case the construction is straightforward using two level morphology and then, describes in a very natural way the Argentinean Quechua morphology using a regular language. On the contrary, the Toba verbs morphology, with a system that simultaneously uses prefixes and suffixes, has not a natural description as regular language. Toba has a complex system of causative suffixes, whose successive applications determinate the use of prefixes belonging different person marking prefix sets. We adopt the solution of Creider et al. (1995) to naturally deal with this and other similar morphological processes which involve interactions between prefixes and suffixes and then we describe the toba morphology using linear context-free languages.[1].

1 Introduction

It has been proved (Johnson, 1972; Kaplan and Kay, 1994) that regular models have an expresive power equal to the noncyclic components of generative grammmars representing the morphophonology of natural languages. However, these works make no considerations about what class of formal languages is the natural for describing the morphology of one particular language. On the other hand, the criteria of classification of Amerindian languages, do not involve complexity criteria. In order to establish criteria that take into account the complexity of the description we present two contrasting examples in two Argentinean native languages: toba and quichua santiagueño. While the quichua has a natural representation in terms of a regular language using two level morphology, we will show that the Toba morphology has a more natural representation in terms of linear context-free languages.

2 Quichua Santiagueño

The quichua santiagueño is a language of the Quechua language family. It is spoken in the Santiago del Estero state, Argentina. Typologically is an agglutinative language and its structure is almost exclusively based on the use of suffixes and is extremely regular. The morphology takes a dominant part in this language with a rich set of validation suffixes. The quichua santiagueño has a much simpler phonologic system that other languages of this family: for example it has no series of aspirated or glottalized stops.

Since the description of the verbal morphology is rich enough for our aim to expose the regular nature of quichua santiagueño morphology, we have restricted our study to the morphology of finite verbs forms. We use the two level morphology paradigm to express with finite regular transducers the rules that clearly illustrate how naturally this language phonology is regular. The construction uses the descriptive works of Alderetes (2001) and Nardi (Albarracín et. al, 2002)

[1]This work is part of the undergraduate thesis Finite state morphology: The Koskenniemi's two level morphology model and its application to describing the morphosyntaxis of two native Argentinean languages

Proceedings of the ACL 2010 Student Research Workshop, pages 109–114,
Uppsala, Sweden, 13 July 2010.

2.1 Phonological two level rules for the quichua santiagueño

In this section we present the alphabet of the quichua santiagueño with which we have implemented the quichua phonological rules in the paradigm of two level morphology. The subsets abbreviations are: V (vowel), Vlng (underlying vowel).Valt (high vowel),VMed (median vowel), Vbaj (bass vowel) , Ftr (trasparent to medialization phonema), Cpos (posterior consonant).

```
ALPHABET
  a e i o u p t ch k q s sh ll m
  n  l r w y b d g gg f h x r rr
  A E I O U W N Q Y + ’
NULL 0
ANY @
BOUNDARY #
SUBSET C       p t ch k q s sh ll m
               n  l r w y b d f h
               x r rr h Q
SUBSET V       i e a o u A E I O U
SUBSET Vlng    I E A O U
SUBSET Valt    u i U I
SUBSET VMed    e o E O
SUBSET Vbaj    a A
SUBSET Ftr     n y r Y N
SUBSET Cpos    gg q Q
```

With the aim of showing the simplicity of the phonologic rules we transcribe the two-level rules we have implemeted with the transducers in the thesis. R1-R4 model the medialization vowels procesess, R5-R7 are elision and ephentesis procesess with very specific contexts and R7 represents a diachornic phonological process with a subjacent form present in others quechua dialects.

```
Rules

R1 i:i /<= CPos:@ __

R2 i:i /<=  __ Ftr:@ CPos:@

R3 u:u /<= CPos:@ __

R4 u:u /<=  __ Ftr:@ CPos:@

R5  W:w <=> a:a a:a +:0__a:a +:0

R6 U:0 <=> m:m __+:0 p:p u:u +:0

R7 N:0 <=> ___+:0 r:@ Q:@ a:a +:0
```

2.2 Quichua Santigueño morphology

The grammar that models the agglutination order is showed with a non deterministic finite automata. This implemented automata is presented in Figure 1. This description of the morphophonology was implemented using PC-KIMMO (Antworth, 1990)

3 The Toba morphology

The Toba language belongs, with the languages pilaga, mocovi and kaduveo, to the guaycuru language family (Messineo, 2003; Klein, 1978). The toba is spoken in the Gran Chaco region (which is comprised between Argentina, Bolivia and Paraguay) and in some reduced settlements near Buenos Aires, Argentina. From the point of view of the morphologic typology it presents characteristics of a polysynthetic agglutinative language. In this language the verb is the morphologically more complex wordclass. The grammatical category of person is prefixed to the verbal theme. There are suffixes to indicate plurals and other grammatical categories as aspect, location-direction, reflexive and reciprocal and desiderative mode. The verb has no mark of time. As an example of a typical verb we can considerate the $sanadatema$:

Example 1 .

s-	*anat(a)*	*-d*	*-em*	*-a*
1Act-	*advice*	*2 dat*	*ben*	[2]
” I	*advice*	*you”*		

One of the characteristics of the toba verb morphology is a system of markation active-inactive on the verbal prefixes (Messineo, 2003; Klein, 1978). There are in this language two sets or verbal prefixes that mark action:

1. Class I (In):codifies inactive participants, objects of transitive verbs and pacients of intransitive verbs. .
2. Class II(Act): codifies active participants, subjects of transitive and intransitive verbs.

[2] abrev: Act:active, ben:benefactive, dat:dative,inst: intrumental,Med: Median voice, pos: Posessor, refl: reflexive

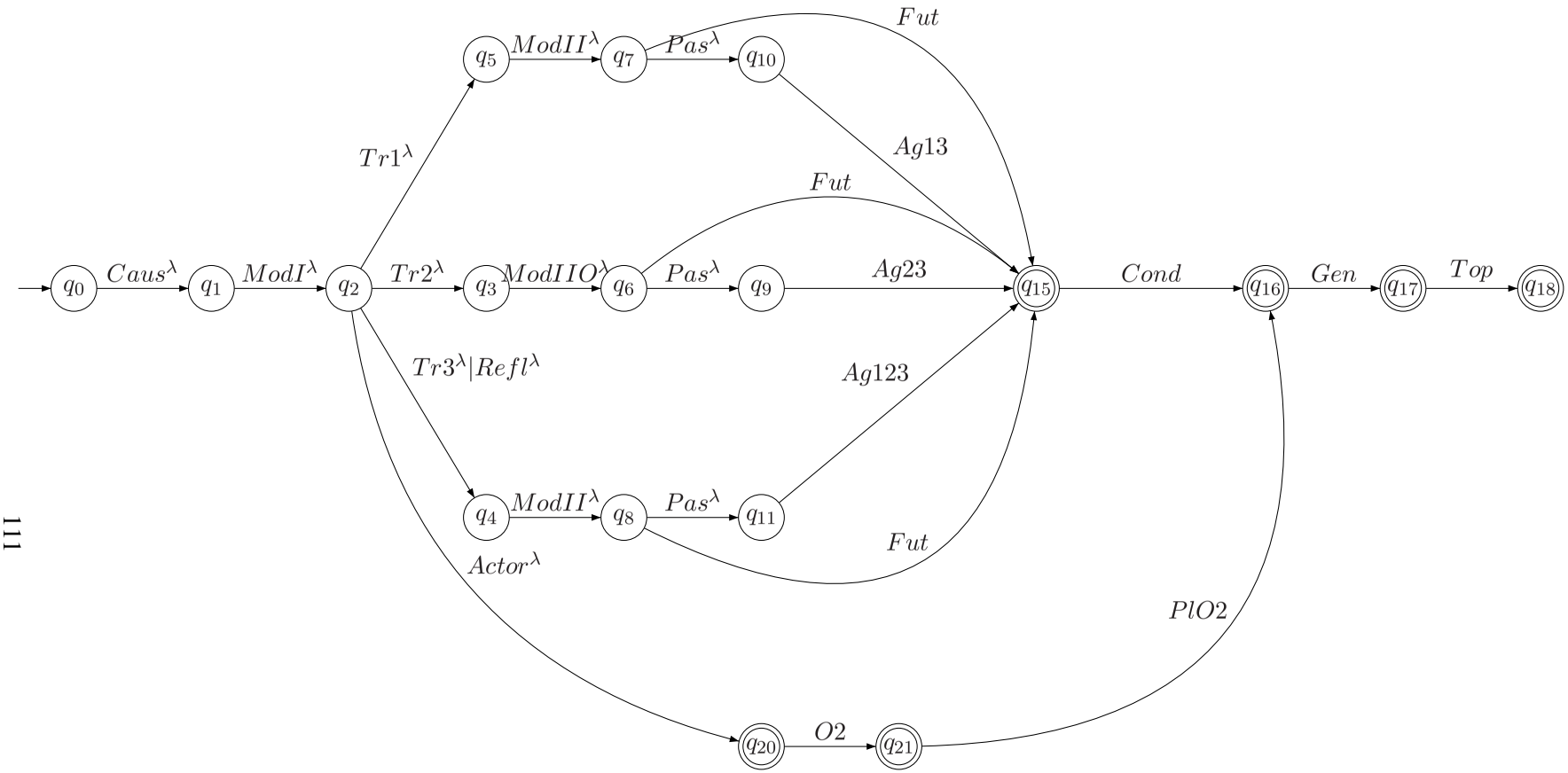

Figure 1: Schema of the verbal morphology of the quichua santiagueño. The supra indices λ indicate possible null transitions.
Abrev.: Caus: Causative suffixes.ModI: Set I of Modal Suffixes.Tri : i th. person trasition Suffixes.ModII: Set II of modal suffixes .Pas: Past suffixes. Ag: AgentSuffix , in this case, for example, Ag1, indicates the agent suffix for the 1st person. Ag12 is an abbreviator for $A1 \cup A2$. Cond : Conditional Suffixes. Gen : General Suffixes. Fut: future suffixes.Top : Topicaliser suffixes. O2: Object 2nd person Suffixes. PLO2: Plural of Object 2nd person Suffixes.

Active affected(Medium voice, Med): codifies the presence of an active participant affected by the action that the verb codifies. .

The toba has a great quantity of morphological processes that involve interactions between suffixes and prefixes. In the next example the suffixation of the reflexive ($-l'at$) forces the use of the active person with prefixes of the voice medium class because the agent is affected by the action.

Example 2 .

(a)	*y-*	*alawat*	
	3Activa	*-kill*	
	" He	*kills"*	
(b)	*n-*	*alawat*	*-l'at*
	3Med-	*kill*	*-refl*
	" He kills	*himself"*	

The agglutination of this suffix occurs in the last suffix box (after locatives, directional and other derivational suffixes). Then, if we model this process using finite automata we will add many items to the lexicon (Sproat, 1992). The derivation of names from verbs is very productive. There are many nominalizer suffixes. The resulting names use obligatory possessing person nominal prefixes.

Example 3 .

l-	*edaGan*	*-at*
3pos	*write*	*instr*
	"his pencil"	

The toba language also presents a complex system of causative suffixes that act as switching the transitivity of the verb. Transitivity is specially appreciable in the switching of the 3rd person prefix mark. In section 3.2 we will use this process to show how linear context free grammars are a better than regular grammars for modeling agglutination in this language, but first we will present the former class of languages and its acceptor automata.

3.1 Linear context free languages and two-taped nondeterministic finite-state automata

A linear context-free language is a language generated by a grammar context-free grammars G in which every production has one of the three forms (Creider et al., 1995):

1. $A \rightarrow a$, with a terminal symbol
2. $A \rightarrow aB$, with B a non terminal symbol and a a terminal symbol.
3. $A \rightarrow Ba$, with B a non terminal symbol and a a terminal symbol.

Linear context-free grammars have been studied by Rosenberg (1967) who showed that there is an equivalence between them and two-taped nondeterministic finite-state automata. Informally, a two-head nondeterministic finite-state automata could be thought as a generalization of a usual nondeterministic finite-state automata which has two read heads that independently reads in two different tapes, and at each transition only one tape moves. When both tapes have been processed, if the automata is at a final state, the parsing is successful. In the ambit that we are studying we can think that if a word is a string of prefixes, a stem and suffixes, one automata head will read will the prefixes and the other the suffixes. Taking into account that linear grammars are in Rosenberg's terms: " The lowest class of nonregular context-free grammars", Creider et al. (1995) have taken this formal language class as the optimal to model morphological processes that involve interaction between prefixes and suffixes.

3.2 Analysis of the third person verbal paradigm

In this section we model the morphology of the third person of transitive verbs using two-taped finite nondeterministics automata. The modeling of this person is enough to show this description advantages with respect to others in terms of regular languages. The transitivity of the verb plays an important role in the selection of the person marker Class. The person markers are (Messineo, 2003):

1. i-/y- for transitive verbs y and some intransitive subjects (Pr AcT).
2. d(Vowel) for verbs typically intransitives (Pr ActI).
3. n: subjets of medium voice (Pr ActM).

The successive application of the causative seems to act here, as was interpreted by Buckwalter (2001), like making the switch in the original verb transitivity as is shown en Example 4 in the next page.

Example 4 . .

IV	*de-*	*que'e*		*he eats*
TV	*i-*	*qui'*	*-aGan*	*he eats(something)*
IV	*de-*	*qui'*	*-aGanataGan*	*he feeds*
TV	*i-*	*qui'*	*-aGanataGanaGan*	*he feeds(a person)*
IV	*de*	*qui'*	*-aGanaGanataGan*	*he command to feed*

If we want to model this morphological process using finite automata again we must enlarge the lexicon size. The resulting grammar, althought capable of modeling the morphology of the toba, would not work effectively. The effectiveness of a grammar is a measure of their productivity (Heintz, 1991). Taking into account the productivity of causative and reflexive verbal derivation we will prefer a description in terms of a context-free linear grammar with high effectivity than another using regular languages with low effectivity.

To model the behavior of causative agglutination and the interaction with person prefixes using the two-head automata, we define two paths determined by the parity of the causative suffixes wich have been agglutinated to the verb. We have also to take into consideration the optative posterior aglutination of reflexive and reciprocal suffixes wich forces the use of medium voice person prefix. From the third person is also formed the third person indefinite actor from a prefix, qa -, which is at left and adjacent to the usual mark of the third person and after the mark of negation sa-. Therefore, their agglutination is reserved to the last transitions. The resulting two-typed automata showed in Figure 2 also takes into account the relative order of the boxes and so the mutual restrictions between them (Klein, 1978).

4 Future Research

It is interesting to note that phonological rules in toba can be naturally expressed by regular Finite Transducers. There are, however, many South American native languages that presents morphological processes analogous to the Toba and some can present phonological processes that will have a more natural expression using Linear Finite Transducers. For example the Guarani language presents nasal harmony which expands from the root to both suffixes and prefixes (Krivoshein, 1994). This kind of characterization can have some value in language classification and the modeling of the great diversity of South American languages morphology can allow to obtain a formal concept of natural description of a language.

References

Lelia Albarracín, Mario Tebes y Jorge Alderetes(eds.) 2002. *Introducción al quichua santiagueño por Ricardo L.J. Nardi*. Editorial DUNKEN: Buenos Aires, Argentina.

Jorge Ricardo Alderetes 2002. *El quichua de Santiago del Estero. Gramática y vocabulario.*. Tucumán: Facultad de Filosofía y Letras, UNT:Buenos Aires, Argentina.

Evan L. Antworth 1990. *PC-KIMMO: a two-level processor for morphological analysis.No. 16 in Occasional publications in academic computing*. No. 16 in Occasional publications in academic computing. Dallas: Summer Institute of Linguistics.

Alberto Buckwalter 2001. *Vocabulario toba*. Formosa / Indiana, Equipo Menonita.

Chet Creider, Jorge Hankamer, and Derick Wood. 1995. Preset two-head automata and morphological analysis of natural language . *International Journal of Computer Mathematics*, Volume 58, Issue 1, pp. 1-18.

Joos Heintz y Claus Schönig 1991. Turcic Morphology as Regular Language. *Central Asiatic Journal*, 1-2, pp 96-122.

C. Douglas Johnson 1972. *Formal Aspects of Phonological Description*. The Hague:Mouton.

Ronald M. Kaplan and Martin Kay. 1994. Regular models of phonological rule systems . *Computational Linguistics*, 20(3):331-378.

Harriet Manelis Klein 1978. *Una gramática de la lengua toba: morfología verbal y nominal*. Universidad de la República, Montevideo, Uruguay.

Natalia Krivoshein de Canese 1994. *Gramática de la lengua guaraní*. Colección Nemity, Asunción, Paraguay.

María Cristina Messineo 2003. *Lengua toba (guaycurú). Aspectos gramaticales y discursivos*. LINCOM Studies in Native American Linguistics 48. München: LINCOM EUROPA Academic Publisher.

A.L Rosenberg 1967 A Machine Realization of the linear Context-Free Languages. *Information and Control*, 10: 177-188.

Richard Sproat 1992. *Morphology and Computation*. The MIT Press.

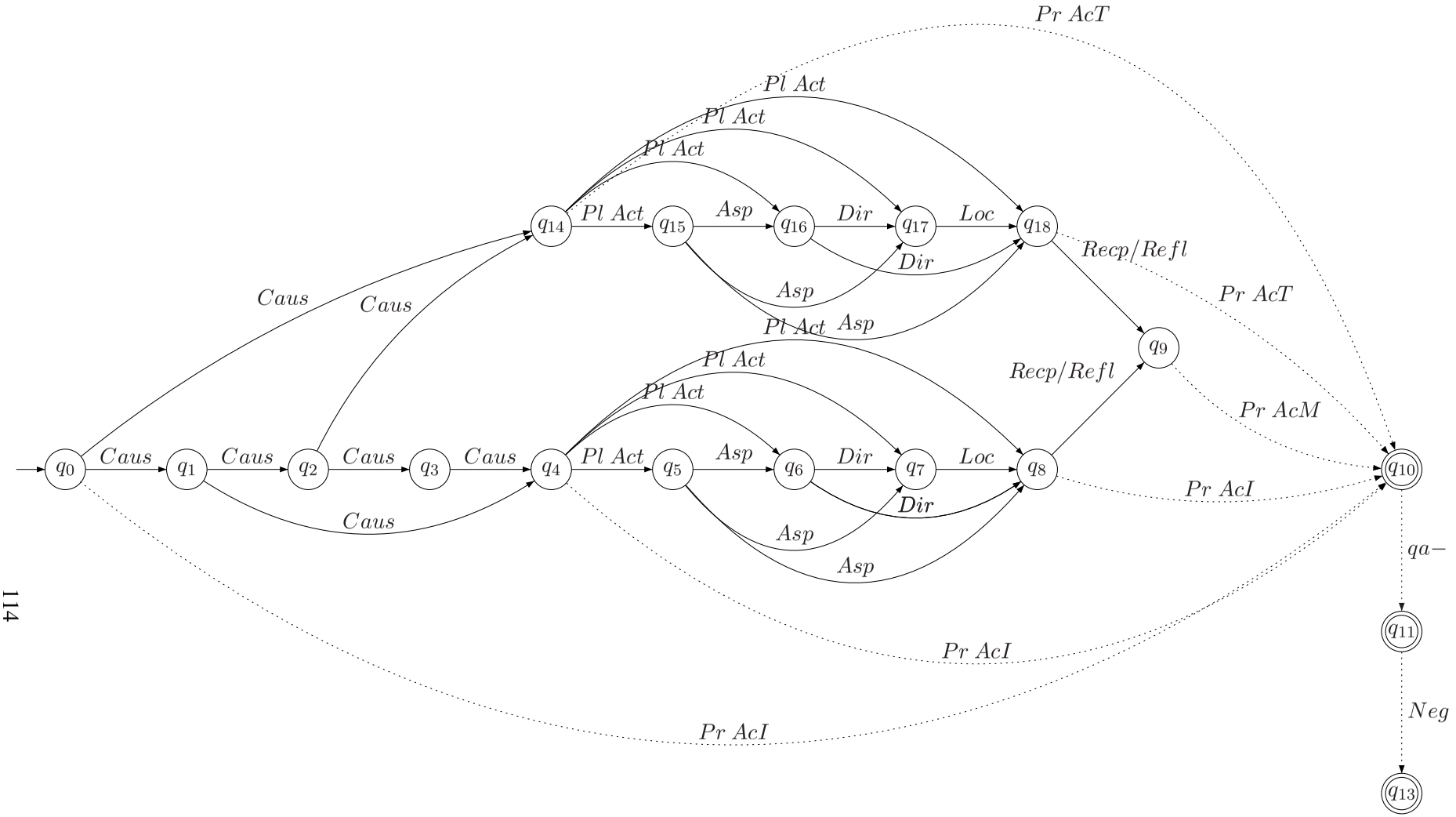

Figure 2: Schema of the 3rd person intransitive verb morphology of the toba .
The entire and dotted lines indicating transitions of the sufix and preffix tape, respectively
Abrev: Caus: Causative suffix. Pl Act: plural actors suffix. Asp: aspectual suffix. Dir: directive suffix. Loc: locative suffix. Recp: reciprocal action suffix. Refl: reflexive suffix. Pr.Ac: acting person prefix(T: transitive, I: intransitive, M: medium) qa-: indeterminate person prefix. Neg: negation prefix