Journée ATALA: TAL et langues anciennes (Paris, 21-may-05)

Using Optimality Theory to "Learn" Elamite Phonology

Eric J. M. Smith

<eric.smith@utoronto.ca>

Dept. of Linguistics, University of Toronto

§1 Introduction

- Elamite: extinct language, spoken in Iran, with no known affiliations.
- Language is attested in cuneiform texts from 2300-360 BCE.
- Conflicting reconstructions of phonology (Paper 1955; Reiner 1969; McAlpin 1982; Grillot-Susini & Roche 1988; Khačikjan 1998).
- Strategy: To treat phonological reconstruction as a learning problem, and adapt existing tools from Optimality Theory.

§2 Optimality Theory

- Optimality Theory (Prince & Smolensky 1993) has become the dominant framework for modelling phonology.
- OT describes phonological processes as a set of competing ranked constraints.
- OT is a learning model (Tesar & Smolensky 2001): presented only with surface forms, language learners can derive their language's underlying forms and constraint rankings.
- For orthography, the relationship between surface (orthographic) and underlying (phonological) forms can be described as a set of constraints.
- A learning algorithm which is powerful enough to drive language acquisition should be able to "learn" the underlying forms and constraint rankings for written Elamite.
- Less ambitiously, such a learning algorithm should be able to evaluate the hypotheses about Elamite phonology presented by previous scholars.

§3 Data

- 16000 entries from *Elamisches Wörterbuch* (Hinz & Koch 1987). Includes words, personal names, and geographical names.
- Transcribed manually and lemmatised, with aid of a custom Mac OSX front-end.
- XML storage format. Attributes of <word> tag include: variant spellings, chronology, foreign-language cognates, corpus frequency, and morphology.

§4 Gradual Learning Algorithm

- Concrete implementation of OT learning model based on Gradual Learning Algorithm (Boersma & Hayes 2001):
  1) Choose an observed (orthographic) form from the lexicon.
  2) If there is an existing estimate of the underlying form (e.g. an Old Persian loanword), use that. Otherwise, calculate an estimate using Lexicon Optimization.
  3) Generate a set of "slightly incorrect" rivals which are similar to the observed form.
  4) Compare the incorrect and observed forms against the estimated underlying form using the current constraint rankings.
  5) If the constraint system selects the wrong winner, penalise the constraints which favoured the wrong winner, and reward the constraints which favoured the right winner.
  6) Repeat several thousand times, until constraint rankings have stabilised.

§5 Implementation details

- The set of constraints to be used is based on a set of 30 hypotheses proposed over the preceding 50 years by Paper (1955), Reiner (1969), McAlpin (1982), Grillot-Susini & Roche (1988), and Khačikjan (1998). (Listed in appendix)
- Sample constraint implementation:

  Hypothesis H1a: In a <CV$_1$-V$_2$C> sequence, V$_1$ and V$_2$ are articulated as separate vowels.

  Rule: Score a violation whenever the orthography contains a <CV$_1$-V$_2$C> sequence if:

  1) the underlying phonology for /V$_1$/ equals /V$_2$/ or
  2) either /V$_1$/ or /V$_2$/ is not a vowel.

  Example violations: /daʃ/ → <da-iš>, /dajʃ/ → <da-iš>

  Example non-violations: /daiʃ/ → <da-iš>
- During processing, internal representation of entries as annotation graphs (Bird & Liberman 1999; Sproat 2000).
- Forms are presented to the GLA according to an approximation of their frequency within the corpus. Earlier implementations failed to take frequency into account, and gave too much weight to Old Persian personal names.
- To avoid diachronic complications, restricted data to entries from Achæmenid Elamite period.
- Generation of "slightly incorrect" candidates using a constraint-driven Gen algorithm based on Heiberg (1999).
- Lexicon Optimization also implemented using constraint-driven "anti-Gen" algorithm.

§6 Results

- After 40000 iterations, the algorithm produced the constraint rankings shown in the appendix (number in lower left of each cell).
- H1: Broken <CV$_1$-V$_2$C> writings most likely represent /CV$_1$C/ (as Paper 1955, Reiner 1969).
- H2: No conclusion regarding voicing. Hinz' hypothesis does seem likely.
- H3: Geminate stops in orthography represent voicelessness in phonology (as Reiner 1969). Gemination of liquids in orthography is also significant (as McAlpin 1982).
- H4: No conclusion. There were problems with the alignment algorithm.
- H5: Word-final vowels are generally significant.
- H6: There is a /tʃ/. Largely driven by Old Persian loanword data.
- H7: No conclusion on status of /h/.
- H8: No evidence to support existence of /f/ and /v/ phonemes (contra Khačikjan 1998).
- H9: No conclusion on status of /j/.
- H10: <u$_2$> is not being used to write /w/.
- H11: There may be an /e/ phoneme, but it is confined to initial syllables of words.

§7 Conclusions

- First application of Optimality Theory to problems of mapping between phonology and orthography. General approach shows promise.
- Lack of known phonology against which to check results. Elamite may not have been the wisest choice of languages.
- Digital version of *Elamisches Wörterbuch* could have future utility for students of Elamite.

§8 Appendix: Hypotheses to be evaluated

H1) Interpretation of broken <CV$_1$-V$_2$C> writings

H1a) The written vowels of the V_1-V_2 sequence are articulated as two separate spoken vowels, possibly separated by a glottal stop (e.g. /daˀiʃ/ → <da-iš>).	H1b) The V_1-V_2 sequence is being used to represent a diphthong (e.g. /dajʃ/ → <da-iš>).	H1c) The combination of V_1 and V_2 in the orthography is being used to represent an intermediate vowel which could not otherwise be written in cuneiform (e.g. /dɛʃ/ → <da-iš>).	H1d) The V_1-V_2 sequence in the orthography is simply being used to indicate an underlying phonology of V_1; the presence of V_2 in the orthography is merely a scribal convention (e.g. /daʃ/ → <da-iš>).
-2282.89	-2431.23	-2853.94	-1217.99

H2) Voicing of stops

H2a) The language's phonology includes a true voicing distinction, and this distinction is reflected in the choice of graphemes with voiced or unvoiced values (Grillot-Susini & Roche 1988).	H2b) The choice of graphemes with voiced or voiceless values is significant, but the opposition being represented is tense/lax, or some other distinction than voicing.	H2c) The choice of graphemes using voiced and voiceless grapheme values does not reflect a distinction in the phonology. The choice of graphemes is merely an orthographic convention.	H2d) Voicelessness is indicated using the orthographic mechanism suggested by Hinz & Koch (1987). The voicing feature is supplied by one grapheme and the place of articulation by another (e.g. /upa/ → <uk-ba>).
-1018.38	-1496.04		-357.78

H3) Geminate consonants

H3a) Geminate orthographies represent underlying geminate phonologies.	H3b) Geminate orthographies are being used to indicate voicelessness, as suggested by Reiner (1969).	H3c) Certain geminate spellings are used to indicate a distinction other than voicing, such as retroflex/alveolar.
-3569.45	-551.42	-625.69

H4) Nasal vowels

H4a) The observed alternations in the writing of nasals indicate the presence of nasal vowels (e.g. /hũban/ → <hu-um-ban>, <hu-ban>).	H4b) The observed alternations in the writing of nasals can be explained by underlying nasal consonants which are deleted through some phonological process (e.g. /humban/ → [huban] → <hu-um-ban>, <hu-ban>).
-1235.16	-543.03

H5) Word-final vowels

H5a) The alternations in the writing of word-final vowels indicate an attempt to render a word-final consonant cluster which could not otherwise be written in cuneiform.	H5b) The alternations in the writing of word-final vowels indicate the presence of a /ə/ or other underspecified vowel.	H5c) The alternations in the writing of word-final vowels indicate an actual alternation in the phonological vowel.
-2541.74	-3371.33	-781.40

H6) Sibilants		
H6a) The sibilant inventory includes a /tʃ/ which is written using the Akkadian <ṣV> and <Vṣ> graphemes (Paper 1955).	H6b) The sibilant inventory includes a /z/ which is written using the Akkadian <ṣV> and <Vṣ> graphemes (Grillot-Susini & Roche, 1988).	H6c) The sibilant inventory includes a /ts/ which is written using the <sV>, <šV>, <tV>, <Vs>, <Vš>, or <Vt> graphemes (Khačikjan 1998).
-944.56	-1135.58	-2149.88

H7) The phonemic inventory includes an /h/.	
H7a) The <hV> and <Vh> graphemes are being used to write the phoneme /h/.	H7b) The <hV> and <Vh> graphemes are purely orthographic variants of the equivalent <V> graphemes (Paper, 1955).
-487.53	-473.603

H8) The phonemic inventory includes an /f/ or a /v/.	
H8a) The <pir$_2$> grapheme is being used to indicate a /fr/ or /vr/ sequence (Khačikjan 1998).	H8b) The <pir$_2$> grapheme is being used to indicate an ordinary /pr/ or /pir/ sequence.
-457.38	-226.38

H9) The phonemic inventory includes a /j/, written with the <ya> grapheme.	
H9a) The <ya> grapheme is being used to write the phoneme /j/.	H9b) The <ya> grapheme is being used to write a non-syllabic allophone of /i/.
-518.90	-520.39

H10) The phonemic inventory includes a /w/, written with the <u$_2$> grapheme.	
H10a) The grapheme <u$_2$> is being used to indicate a /w/ (McAlpin 1982).	H10b) The grapheme <u$_2$> is being used to indicate a /u/.
-508.96	-133.80

H11) The phonemic inventory includes an /e/.		
H11a) There is an /e/ vowel, distinct from /i/.	H11b) There is an /e/ vowel, but it is distinct from /i/ only in the first syllable of a word (McAlpin 1982).	H11c) The <e>, <eC> and <Ce> graphemes are purely orthographic variants of the equivalent <i>, <iC>, and <Ci> graphemes (Paper 1955).
-490.84	-363.41	-473.88

References

Bird, Steven, and Liberman, Mark. 1999. A Formal Framework for Linguistic Annotation. Philadelphia: Department of Computer and Information Science, University of Pennsylvania.

Boersma, Paul, and Hayes, Bruce. 2001. Empirical Tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32:45-86.

Grillot-Susini, Françoise, and Roche, Claude. 1988. *Eléments de grammaire élamite*: Etudes elamites. Paris: Editions Recherche sur les civilisations.

Heiberg, Andrea. 1999. Features in Optimality Theory: A computational model, University of Arizona: Doctoral dissertation.

Hinz, Walther, and Koch, Heidemarie. 1987. *Elamisches Wörterbuch*: Archäologische Mitteilungen aus Iran. Ergänzungsband ; 17. Berlin: D. Reimer.

Khačikjan, M. L. 1998. *The Elamite Language*: Documenta Asiana, v. 4. Roma: Consiglio Nazionale delle Ricerche Istituto per gli Studi Micenei ed Egeo-anatolici.

McAlpin, David W. 1982. Proto-Elamo-Dravidian: The Evidence and its Implications. *Transactions of the American Philosophical Society* 71:1-155.

Paper, Herbert H. 1955. *The Phonology and Morphology of Royal Achæmenid Elamite*. Ann Arbor: University of Michigan Press.

Prince, Alan, and Smolensky, Paul. 1993. Optimality Theory. *Rutgers Optimality Archive* #537.

Reiner, Erica. 1969. The Elamite Language. *Handbuch der Orientalistik I/II/1/2/2*:54-118.

Sproat, Richard. 2000. *A Computational Theory of Writing Systems*. Cambridge: Cambridge University Press.

Steve, M.-J. 1992. *Syllabaire élamite: Histoire et paléographie*: Série II, Philologie. Neuchâtel: Civilisations du Proche-Orient.

Tesar, Bruce, and Smolensky, Paul. 2000. *Learnability in Optimality Theory*. Cambridge, Mass.: MIT Press.