Converging Transition Networks and Sub-Morphemic Regularities in Latin Noun Inflection

Michael A. Covington
Artificial Intelligence Center
The University of Georgia
Athens, GA 30602-7415

DRAFT, 1999 February 15

Abstract

Many Latin inflectional endings share material constituting less than a whole morpheme. Conventional linguistic analysis must either ignore the shared material (treating it as historical relics) or make it into morphemes through abstract morphophonemics, poorly motivated in Latin. The shared material contributes to economy of representation when the inflectional system is stored as a transition network (character tree), a representation that is computationally efficient and may be psychologically realistic. For example, all the genitive plural endings can converge on "—u—m—⟨GenPl⟩" although the material preceding this depends on the inflectional class.

1 The problem

Table 1 shows the inflectional ending system for Latin nouns. Within this system, there are some obvious syncretisms:

- All neuter accusatives are the same as the corresponding nominatives.
- All ablative plurals are the same as the corresponding datives.
- Classes 1 and 2 fall together in the dative/ablative plural.

But there are many more regularities that can only be expressed by referring to units smaller than the complete suffix. For example:

Class	1	2	2	3c	3c	3i	3i	4	4	5
Gender	*f*	*m*	*n*	*m,f*	*n*	*m,f*	*n*	*m*	*n*	*f*
Singular										
Nominative	a	us	um	∅/s	∅	is	e	us	ū	ēs
Accusative	am	um	um	em	∅	em	e	um	ū	em
Dative	ae	ō	ō	ī	ī	ī	ī	uī	ū	eī
Ablative	ā	ō	ō	e	e	ī	ī	ū	ū	ē
Genitive	ae	ī	ī	is	is	is	is	ūs	ūs	eī
Plural										
Nominative	ae	ī	a	ēs	a	ēs	ia	ūs	ua	eī
Accusative	ās	ōs	a	ēs	a	ēs	ia	ūs	ua	ēs
Dative	īs	īs	īs	ibus	ibus	ibus	ibus	ibus	ibus	ēbus
Ablative	īs	īs	īs	ibus	ibus	ibus	ibus	ibus	ibus	ēbus
Genitive	ārum	ōrum	ōrum	um	um	ium	ium	uum	uum	ērum

Table 1: Latin noun endings. Some details of classes 3 and 4 are omitted, as is the vocative case, which is almost always identical to the nominative.

- All nonneuter accusative singulars end in *-m*.
- All genitive plurals end in *-um*.
- In classes 3 and 4, the genitive singular ends in *-s*.
- In classes 3, 4, and 5, the dative/ablative plural ends in *-bus*.

The modern-day language learner generally feels that these submorphemic regularities simplify the language and make it easier to remember. Native speakers of Latin apparently felt the same way, since the regularities were preserved and actually increased during the development of Latin; for example, the pattern of ablative endings in *-$\bar{V}d$*, later *-$\bar{V}$*, was generalized from class 2 to classes 1 and 4, and the *-$\bar{V}rum$* genitive plural spread from class 1 to class 2.

Confronted by these sub-morphemic regularities, present-day linguistic analysis has two alternatives: simply ignore them, treating them as historical relics with no synchronic significance, or postulate an abstract phonology in which, for instance, *(-$\bar{V}$)rum* rather than *ārum* is a morpheme, and *a* is a separate element present in some cases in Class 1 and absent in others.

Neither alternative is entirely satisfactory. To ignore the regularities is to miss obvious generalizations. However, a highly abstract analysis is hard to motivate. On the surface, the Latin noun endings appear to attach directly to the stems. Class 3 does motivate some morphophonemic rules, since it makes sense to derive forms such as *nox* (stem *noct-*) from underlying **nocts*. Apart from that, though, Latin has no obvious morphophonemic processes comparable to, for instance, the Greek contractions (*agapā+eis* → *agapāis,* *mere+a* → *merē,* and so forth); the analysis of Latin will have to be either very abstract or very shallow, with no middle ground.

2 Transition networks

In this paper I propose an alternative phonological representation in which sub-morphemic generalizations are captured, and serve to simplify the system, without requiring abstract rules. Further, this representation leads to efficient computer recognition of inflected forms and has a reasonable chance of turning out to be psychologically real.

The representation that I propose is a character-by-character (or rather phoneme-by-phoneme) transition network of a kind known in computer science as a CHARACTER TREE or TRIE (for *retrieval*; de la Briandais 1959; Fredkin 1960; Knuth 1973:481-505; Sedgewick 1998:623–668). Transition networks are a solution to the problem of how to store a large lexicon compactly and search it quickly.[1]

To understand how transition networks work, consider the following small lexicon:

nauta	(class 1, masc.) 'sailor'
patria	(class 1, fem.) 'homeland'
puella	(class 1, fem.) 'girl'
pupa	(class 1, fem.) 'doll'

One way to use a lexicon would be to accept an entire word from the input source, then compare it with each of the words in the lexicon until a match is found. A much more efficient way is to accept the input one phoneme at a time and follow the transitions in the transition network shown in Figure 1. In that way, all available information is exploited as soon as possible and the correct lexical entry is found with a minimum number

[1] Tries of this type are one of the two things that are also known as DAWGs (Directed Acyclic Word Graphs), as in Sgarbas, Fakotakis, and Kokkinakis 1995. Elsewhere in the literature, a DAWG is a trie that combines all suffixes of a string, e.g., the DAWG for *abc* would comprise *abc*, *bc*, and *c*.

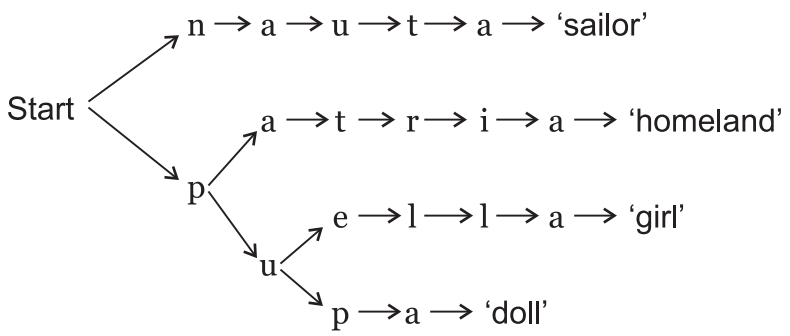

Figure 1: Four-word vocabulary stored in a transition network.

of comparisons. Transition networks are widely used in lexical analysis of programming languages and in speech recognition and are coming into use in natural-language parsing (Roche and Schabes 1997).

Applying transition networks to Latin is especially convenient because of the lack of morphophonemic rules and the nearly phonemic writing system; nothing of significance is lost by using ordinary Latin spelling in place of a phonemic transcription, as I shall do. (For convenience, I shall even treat *ae* as two letters, which reflects its history though not its synchronic status in classical Latin.) Bubenheimer (1995) has implemented a Latin morphological analyzer based on transition networks.[2]

Although not needed in Latin, morphophonemic processes can be incorporated into character-by-character transition networks using the "two-level morphology" of Koskienniemi (1983) and related techniques. Further, transition networks can be built by a mechanical process from sets of parsed individual forms. This makes them handy for computer implementation and plausible as regards psychological reality.

3 Adding inflection to transition networks

Each of the four words in Figure 1 has nine inflected forms, some of which are ambiguous. It is of course unnecessary to list all nine forms in the lexicon. Instead, the inflection (in this case, class 1) can be incorporated into the transition network as shown in Figure 2.

[2]I particularly thank Uli Bubenheimer for showing me his working system just when I was starting to think about analyzing Latin in this way, and Richard A. O'Keefe for discussing trie-based morphological analysis with me as long ago as 1994.

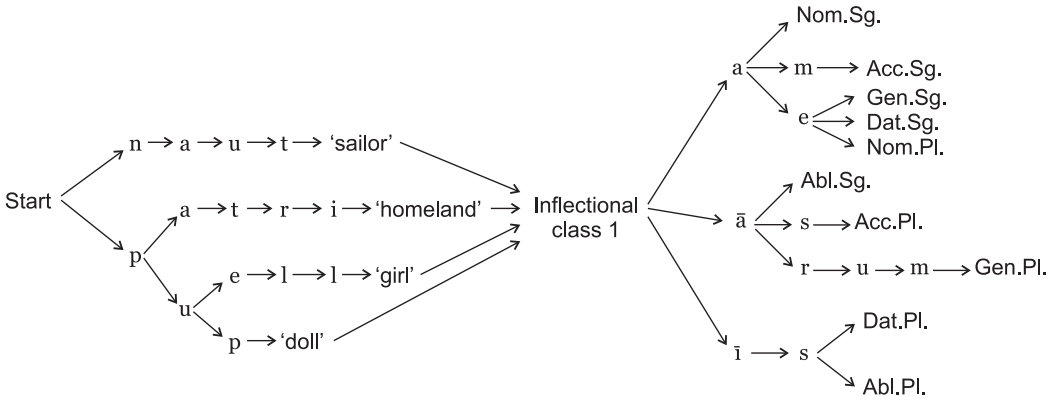

Figure 2: Transition network with inflectional class 1 added.

Purely for expository convenience, I have mixed together two kinds of nodes in the transition network – nodes that correspond to phonemes and nodes that reflect identification of a word, class, or case ending. A purer approach would treat the latter not as nodes, but as annotations to arcs. Nothing of significance is lost by putting all the information into nodes as long as it is understood that not all nodes correspond to phonemes.

Now – and at last I get to unveil my main point – consider what happens when inflectional class 2 and part of class 3 are added to the transition network for class 1. Figure 3 shows the result, and, crucially, large parts of the existing structure can be shared between inflectional classes. Looking at Figure 3, note that every piece of shared material corresponds to a convergence in the tree. So do shared ambiguities, however trivial they seem. For instance, there is a node in the tree reflecting the fact that in classes 1 and 2 (masc.), the genitive singular matches the nominative plural.

4 Two technical points

Note also that the tree in Figure 3 is nondeterministic: when a class-2 suffix begins with *ī*, there are two arrows to be followed, not just one. This was done in order to capture the fact that the dative/ablative plural ending -*īs* is shared by classes 1 and 2, but there is another class-2 ending that begins with *ī*.

The implication is that when there are two paths to be followed, both will be pursued until at least one of them runs into a dead end. In terms of psychological reality, this is justifiable on the ground that phonemes heard

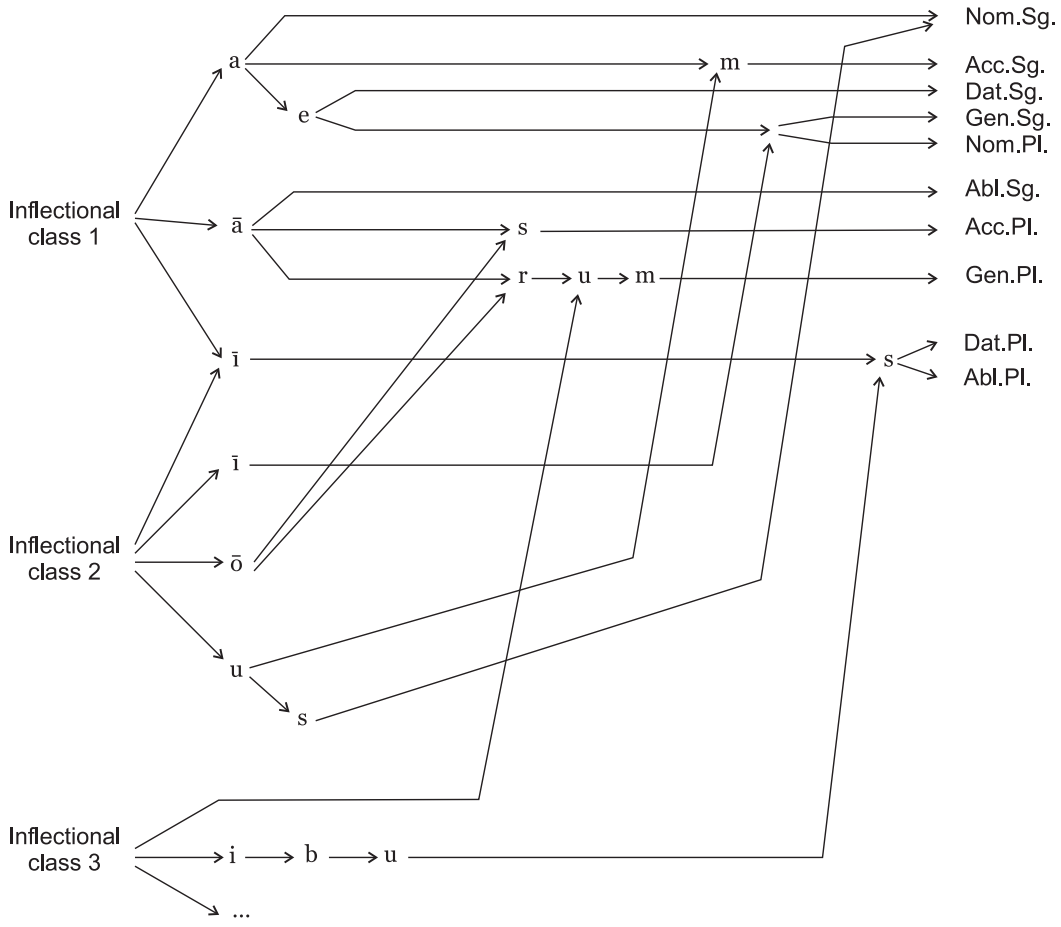

Figure 3: Several inflectional classes can share much of the same structure.

by the listener are often ambiguous. If a word is not heard clearly, so that (for example) one of the vowels is uncertain, the practical thing to do is follow all possible paths until (hopefully) all but one of them run into dead ends.

Finally, an argument can be made for an "elsewhere condition" in transition-network morphology. Consider the dative/ablative plural *fīliābus* 'daughters,' a word that is otherwise a regular member of class 1. Where do we put it in the transition network? Accounting for *-ābus* by means of an arrow joining the *-bus* of class 3 is no problem; but how thetn to account for the rest of the forms? One possibility is to jump to just the parts of class 1 that do not include the *-īs* dative/ablative plural ending. The other possibility is to jump to class 1 entire – that is, link *fīlia* to class 1 in the normal manner, adding a special link to *-ābus* and a stipulation that the special link prohibits use of the ordinary link to the same case and number.

References

Bubenheimer, Uli (1995) *YALL: eine morphologische Analysekomponente für das Lateinische zum Einsatz in einem lehrunterstützende System.* Studienarbeit, University of Koblenz-Landau.

Covington, Michael A. (1994) *Natural Language Processing for Prolog Programmers.* Englewood Cliffs, N.J.: Prentice-Hall.

de la Briandais, R. (1959) File searching by using variable length keys. *Proceedings of the Western Joint Computer Conference* 15:295–298. New York: Institute of Radio Engineers.

Fredkin, E. (1960) Trie memory. *Communications of the ACM* 3:490-499.

Keenan, E. L., and Comrie, B. (1977) Noun phrase accessibility and Universal Grammar. *Linguistic Inquiry* 8:63–99.

Knuth, D. E. (1973) *The Art of Computer Programming,* vol. 3: *Sorting and Searching.* Reading, Mass.: Addison-Wesley.

Koskenniemi, K. (1983) Two-level morphology: a general compuational model for word-form recognition and production. Publication 11, Department of General Linguistics, University of Helsinki.

Ringe, Don (1995) Nominative-accusative syncretism. Manuscript, University of Pennsylvania.

Sgarbas, K; Fakotakis, N.; and Kokkinakis, G. (1995) Two alternative algorithms for incremental construction of directed acyclic word graphs. *Inter-*

national Journal on Artificial Intelligence Tools 4:369 ff. VERIFY

Sihler, Andrew L. (1995) *New Comparative Grammar of Greek and Latin.* New York: Oxford University Press.