An Electronic English-Tamil Dictionary

Harold F. Schiffman

South Asia Regional Studies

University of Pennsylvania

The Tamil language is well-studied in its written form, and Tamil-to-English dictionaries are plentiful and adequate, but English-to-Tamil dictionaries lack information necessary for non-natives, such as the complexities of verbal morphology, syntactic frames, and in particular, any information about spoken forms. The complex phonetics and verbal morphology of the modern (spoken) dialects of this radically diglossic language are not accessible to researchers wishing to study its phonetics, phonology, or grammar. In particular, there are no searchable databases for linguistic information on Tamil, and especially none for spoken Tamil.

This project takes an extant database for a pedagogical reference dictionary (An English Dictionary of the Tamil Verb, now approximately 85% done) to enhance, extend, and electronically tag the sentence examples, both in their readable (visual) form as well as in digitized sound files (to be provided). The project will eventually involve recording sound files for all examples, and organizing the data for searchability by providing indexes and tools, and for publication by CD-ROM.

The extant database consists of approximately 17,500 records, English to Tamil, each English record having a different Tamil verb given in its Literary Tamil (LT) form, its spoken (ST) form, verb-class specification (i.e. the complex morphology of past-tense formation), with synonyms, and with example sentences in LT (in Tamil script), ST (in roman transliteration and in audio format), and with English glosses. The final product will be made available on CD-ROM, or via other electronic databases (such as Hyperlex) and networks (such as TalkBank), and made available at cost to interested researchers. The final CD version will give researchers unfamiliar with this morphologically and syntactically complex language access to phonetic, grammatical and syntactic information that can be mined for purposes not usually accessible to non-specialists, thus expanding the data available to researchers interested in these phenomena.

1. History of Tamil Dictionaries

Tamil is the Dravidian language with the longest written tradition in India, dating from the early centuries of the Common Era or before. Because it was spoken along the trade routes to the East Indies, it was one of the earliest South Asian languages learned by Europeans and is the first Indian language to appear in (western-style moveable-type) print-

-the VOCABULARIO TAMULICO COM A SIGNIFICAൺAM PORTUGUEZA [D255] of da Proen2a in 1679. Because of its ancient literature and its spread both in ancient and recent times into Sri

Lanka and southeast Asia, Tamil is important as a historical language in these areas and is studied by non-Tamils to a degree that is out of proportion to the size of its population of speakers. Because of its long history, Tamil developed a very large lexicon independent of Sanskritic and Indo-Aryan influences, and this vocabulary is available in print dictionaries.

Today, computing resources for Tamil are in better shape than many other South Asian languages, so that there are now reported to be more Tamil speakers with access to computing and the Web than for any other Indian language, including Hindi. This means that databases have developed, networks of researchers exchange data, and the possibilities of research in and on this language exceed the resources available for any other South Asian language, and promise to continue to do so.

2. Research on the modern living language.

Nevertheless, researchers who are unfamiliar with Tamil script but who would like information about the complex morphology and syntax of this language face many barriers. As noted, Tamil-to-English dictionaries are plentiful, but they do not give information about the spoken language; since Tamil is severely diglossic (its spoken forms cannot be understood by those only knowing the written form, and vice versa) most dictionaries of the language are useful only for the written version of the language. No electronic searchable databases exist for modern Tamil except for an on-line version of the Cre-A Tamil Dictionary restricted to select users at the U. of Chicago, and none exist whatsoever that could access any information about the phonetics, phonology, or morphology of modern spoken Tamil. Since the database of the pedagogical dictionary is now about 85% complete, we are modifying it to allow it to serve as a searchable database for researchers interested in linguistic issues such as its complex phonetics, phonology (see Schiffman 1993) and morphology.

In Dhamotharan's 1978 bibliography of Tamil dictionaries there are actually some 55 English-Tamil dictionaries or glossaries listed. All of these suffer from various faults, such as being intended for Tamil speakers only, for students (or children or tea planters) only, are extremely brief, or are simply out of print. Many of them list rare English words but do not give simpler or more colloquial items such as come' or go', or verb-particle combinations such as come off,' burn down', etc. None of them gives information on Tamil spoken usage and pronunciation. The most modern and scholarly attempt, the three-volume Madras University English-Tamil Dictionary (Chidambaranatha Chettiar, 1965), while containing much more information than the others, still does not list verb classes, transitivity status, or any spoken forms. None give examples of aspectually-marked complex verbs (a distinctive feature of modern spoken Dravidian languages in particular) and of course none contain sound files.

2.1 The pedagogical dictionary D-base

Because of the serious need for an adequate English-Tamil dictionary, the P.I. has been compiling since 1972 a corpus that is intended to be eventually published as an English Dictionary of the Tamil Verb.

The project has gone through many stages, and received funding from a number of sources (NEH, Smithsonian, Consortium for Language Teaching and Learning) and the database has

been restructured a number of times as computer technology changed, and especially as database management became more sophisticated. The d-base now consists of approximately 17,500 basic English verbs (such as 'break'), with verb-plus-particle variants (such as 'break up, break out, break in') treated as separate items. Verbs were chosen based on the following criteria:

- Currency or frequency in Modern (American) English
- Equivalency in modern Tamil
- Use in modern literary and journalistic Tamil, i.e. short stories, novels, newspapers, and in spoken sources.

Tamil verbs are a finite set; the Tamil language does not neologize or innovate new lexical verbs except through borrowing (usually from English or another Indian language) or by reusing something from an older stage of Tamil. English borrowings are done by compounding the English item with a Tamil verb meaning do', i.e. cey, pannu, e.g. draiv pannu drive (a car).' In general, English entries have been taken from the commonest verbs in use in standard sources such as Websters and Fowler's Dictionary of Modern English Usage. In other words, the most important criteria here is whether an English entry will get the user to a Tamil verb. If the English verb is archaic or has no equivalent in Tamil (such as the English verb to burr' [i.e. make a burring or trilling sound, as in Scottish pronunciations of English], or smite' [which has a Tamil equivalent, but is not likely to be the starting point for a search for a Tamil verb]), we do not include it. Modern English neologisms that are likely to not have Tamil equivalents other than direct translation, such as download' (daunlod pannu), are also excluded.

2.2 Finiteness of the Verbal Corpus

A reason for focusing on verbs in this corpus is that Tamil verbs are a finite set. Tamil does not currently borrow verbs from other languages, nor does it invent new ones except by compounding existing items or by the use of a loan word plus a Tamil verb, as mentioned in the previous section (e.g. English drive' may be rendered by draiv plus pannu do', although this process is not permitted in LT). Tamil also has recourse to borrowing' verbs from older stages of the language (Old Tamil) and occasionally from its own dialects, but for the most part any listing of the verbs of Tamil now in use would remain fairly constant and not become obsolescent for some time. Nouns, the other major part of speech' in the Tamil lexicon, on the other hand, are an infinite set and therefore somewhat unmanageable given the present circumstances. Noun morphology is also much less complex and interesting in Tamil than verbal morphology. Thus this verbal corpus is about as complete as it can be for the modern language, and constitutes a fairly stable database that will not change much in the future, except for borrowing. Borrowing of this type can then be compared with restrictions on borrow in the literary language, for interesting cross-comparisons.

3. Computerization

Computerization has allowed us to incorporate all useful material from previous dictionaries, and to edit and phototypeset in English and Tamil. Computerization also allows us to update the database periodically and to reprint revised editions if and when necessary, and eventually

should also allow us to add nouns if that is warranted. The development of the web has allowed us to create an on-line version (still in the testing stage) with sound files; it also allows us to edit on-line from remote locations, and to share data with Tamil scholars world-wide, who consult our proposed entries and make suggestions for improvements. This version can be viewed on line at http://ccat.sas.upenn.edu/plc/tamilweb/dictionary

4. Sources of data

Since the work is not an etymological or historical dictionary, we concentrate primarily on modern materials. We consult monolingual dictionaries, bilingual dictionaries, and spoken materials such as radio plays, and other more modern reference works such as the Cre-A Dictionary of Contemporary Tamil. Modern contemporary journalistic usage can be obtained from web versions of Tamil newspapers, such as Dinamani, which is close to, but not identical with, spoken usage. Otherwise, there are very few modern sources, since Tamil culture does not dignify the spoken language by using it in print; some spoken examples can be obtained from the conversational portions of novels and short-stories, but these are few and far between, and writers are inconsistent in their usage. In fact for this we rely on the native-speaker intuition of our various consultants, who can generate (as can all educated Tamils) spoken equivalents of literary Tamil on demand.

4.1 The problem of standardization of Spoken Tamil.

The question of whether there exists a variety of spoken Tamil that is standard' is a somewhat difficult, but not intractible issue. Many linguistic scholars have approached the issue and have various conclusions to offer; the concensus seems to be that a standard spoken Tamil, if it does not already exist, is at least emerging' and can be described as that variety that one hears used in the Tamil social' film, and on the radio and in the production of social' dramas, both live and on radio and television. We are of the opinion that a restandardization' process is taking place for spoken Tamil, and that a consensus exists as to what this consists of.I have dealt with the issue extensively in my 1998 article listed in the bibliography.

4.2 Transliteration.

The Roman transliteration chosen represents a fairly phonetic attempt at rendering spoken Tamil without getting into fine phonetic detail that is actually predictable from a general knowledge of Tamil. Unlike some Indian languages, Tamil does not have a single standard transliteration system. Authoritative sources such as the Madras University ENGLISH-TAMIL DICTIONARY (Chidambaranatha Chettiar 1965), the Madras University TAMIL LEXICON, and Burrow and Emeneau's DRAVIDIAN ETYMOLOGICAL DICTIONARY (1961) use different transliterations, especially for some of the laterals and rhotics, where true confusion reigns. To make matters worse, popular transcriptions, such as those used in public signing, transliterations of personal names, etc. typically do not mark differences in vowel length, retroflexion, and other distinctions. This is unfortunate, but scholars and others have not been able or willing to agree on a standard transliteration, so we have chosen one that can be used by lay persons as well as scholars.

Since Tamil has a number of stop consonants series that western languages do not exhibit, i.e.

the retroflex and alveolar stop and nasal contrasts, we have chosen a transliteration system for these series that utilize lower-ascii characters unambiguously and mnemonically.The special" alveolar and retroflex consonants are represented with numerals, which are mnemonic for Tamil speakers since 2 represents the two-loop n", 3 represents the three-loop n," 6' stands for the alveolar r in the Tamil word for six,' 7' stands for the retroflex frictionless continant in the Tamil word for seven eeru, etc. The schema works like this:

Transliteration of Special Consonants

Tamil Character	Description	Phonetic Symbol	Input Symbol	Output Symbol
ன2ன	alveolar nasal	[n]	2	൹
ன2ன	retroflex nasal	[ø]	3	’
ன4ன	retroflex shibilant	[♣]	4	൹൹
ன5ன	palatal nasal	[ ]	5	‘
ன6ன	alveolar rhotic/stop	[ρ]	6	ൿ
ன7ன	retroflex rhotic	[è]	7	
ன8ன	retroflex stop	[ÿ ≤]	8	“
ன9ன	retroflex lateral	[→]	9	ൽ
ன ng ன	velar nasal	[÷]	ng	3

With this system combined with intuitively phonetic ascii symbols such as p for a labial stop, m for a labial nasal, aa etc. for long vowels and u etc. for short vowels, both spoken and written Tamil can be represented simply in an almost one-to-one correspondence; except for ng as a digraph for velar nasal (and the double-vowel representation of long vowels), all other Tamil sounds can be represented with one symbol for any one consonant or vowel. In the output symbol column, these are for ST only, and since the distinction between alveolar and non-alveolar rhotics and nasals is not phonemic in ST, we do not represent these differences.

5. References

1 Burrow, T. and M. B. Emeneau 1961. A Dravidian Etymological Dictionary. Oxford: the Clarendon Press.

2 Chidambaranatha Chettiar, A. (ed.) 1961 English-Tamil Dictionary. Madras: University of Madras.

3 Dhamotharan, A. 1978. Tamil Dictionaries: A Bibliography

Beitr€œge zur S1000dasien-Forschung, S1000dasien-Institut der Universit€œt Heidelberg, vol. 50. Wiesbaden: Franz Steiner Verlag.

4 da Proen2a, Antam. Vocabulario Tamulico com a significa2am Portugueza. Na impremssa Tamulica da Provincia do Malabar, por Ignacio Aichamoni impressor della. 1679. 247 fo. New edition, Ant€œo da Proen2a's Tamil-Portuguese dictionary, A.D. 1679. Kuala Lumpur: University of Malaya, 1966. Xavier S. Thani-Nayagam, ed.

6 Schiffman, Harold. 1974. ``Causativity and the Tamil Verbal Base."International Journal of Dravidian Linguistics, V:238-48.

7 .....1993. ``Intervocalic V-deletion in Tamil: Its Domains and its Constraints." 1993. Journal of the American Oriental Society 113(4)513-528.

8 .....1996. ``Desiderata for an English Dictionary of Tamil." In Studies in Tamil Lexicography, Gregory James (ed.) Hong Kong: University of Science and Technology.

9 .....1997. ``Diglossia as a Sociolinguistic Situation." In Florian Coulmas (ed.), The Handbook of Sociolinguistics. London: Basil Blackwell, Ltd.

10. ..... 1998. Standardization and Restandardization: the case of Spoken Tamil." Language in Society, Vol. 27 (3) 359-385.

11 .....1999. A Reference Grammar of Spoken Tamil. Cambridge: Cambridge University Press.