

Towards a General Model for Linguistic Paradigms

David Penton, Catherine Bow, Steven Bird and Baden Hughes
Department of Computer Science and Software Engineering
University of Melbourne
{djpenton, sb, cbow, badenh}@cs.mu.oz.au

ABSTRACT

Linguistic forms are inherently multi-dimensional. They exhibit a variety of phonological, orthographic, morphosyntactic, semantic and pragmatic properties. Accordingly, linguistic analysis involves multi-dimensional exploration, a process in which the same collection of forms are laid out in many ways until clear patterns emerge. Equally, language documentation usually contains tabulations of linguistic forms to illustrate systematic patterns and variations. In all such cases, multi-dimensional data is projected onto a two-dimensional table known as a linguistic paradigm, the most widespread format for linguistic data presentation. In this paper we survey a representative sample of paradigms and develop a simple relational data model. We show how XML technologies can be used to store and render paradigms. The result is a flexible and extensible model for the storage, interchange and delivery of linguistic paradigms.

1. INTRODUCTION

The paradigm is the most widely used data presentation format in language documentation and description. Complex multi-dimensional information is frequently presented in such a manner, such as phoneme charts, inflectional forms of words, and so forth. As a structured collection of data arranged for efficient search and retrieval, paradigms can be considered a type of database. The goal of this paper is to develop an encoding model for linguistic paradigms through consideration of a range of linguistic paradigms as found in the descriptive literature.

Bird [3] adopted the following working definition: “a paradigm (broadly construed) is any kind of rational tabulation of words or phrases to illustrate contrasts and systematic variation.” This definition needs to be extended to include content below the level of the word, such as phones or morphs. Adopting this definition excludes other tabulations used by linguists, such as certain formats for rule-based derivations and Optimality Theory tableaux, in which the content and position of table cells is dependent on theory-driven notions of ordering. Exchanging the rows or columns of such displays can change their meaning, or render them incoherent. By contrast, ordering “singular” before “plural” in linguistic paradigms is a matter of convention, and no information is distorted or lost when the ordering is reversed.

Paradigms can be viewed as a two-dimensional arrangement of elements and attributes, with optional row and column labels. The example in Figure 1 shows a paradigm for the German definite article, with number and gender labelled across the top, and case listed down the left hand side [13, 60]. The content of each cell is a word-form identifiable by its co-ordinate position.

A significant advantage of the linguistic paradigm is its ability to present complex data in tabular form, so that

PARADIGM FOR GERMAN DEFINITE ARTICLE

	SINGULAR			PLURAL
	MASCULINE	FEMININE	NEUTER	ALL GENDERS
NOMINATIVE	der	die	das	die
ACCUSATIVE	den	die	das	die
GENITIVE	des	der	des	der
DATIVE	dem	der	dem	den

Figure 1: Paradigm for German definite article

multiple dimensions of information can be presented in a two-dimensional table. Paradigms displayed on the printed page incorporate a variety of devices to represent more than two dimensions. The range of presentations possible for the same data set indicate that the underlying structure of the paradigm can be rendered into a variety of visual formats. The constraints inherent in the two-dimensionality of the printed page obscure the complexity inherent in the underlying model. The challenge is to clearly express dynamic multi-dimensional paradigms in the static two dimensional format of the printed medium. The present work seeks to complement the traditional display functionality with the utility provided by a structural encoding model.

This paper proposes a simple relational data model for linguistic paradigms. We demonstrate how the model can be used to derive an XML representation, permitting the data to be manipulated in a variety of ways, or to be constructed from external sources such as lexicons and interlinear texts. The presentation of a linguistic paradigm then becomes a rendering problem. The XML representation becomes a canonical underlying form which can be reused in many ways: by rendering into many different visual formats, or by direct conversion to other linguistic data models. The present work should be viewed in the context of other attempts to model linguistic information using XML, such as the four-level model of interlinear text ([4], [17]) and the GOLD ontology ([12]).

The structure of this paper is as follows. We begin by surveying paradigms as they are used in the literature on language documentation and description. Next, we briefly review previous work on modelling tabular linguistic structures, before developing our own formal model. This model is used as the basis for an XML representation, and we explore the use of XSLT stylesheets for visualization. We conclude with a discussion of issues for further research.

2. PARADIGM SURVEY

In this section the results of a survey of paradigms in linguistic description are presented, beginning with simple two-dimensional paradigms, where possible analyses are constrained, then simple three dimensional occurrences, and finally more complex paradigms.

2.1 Simple Paradigms

A simple paradigmatic form is found in a representation of personal pronouns in Hua [15], shown in Figure 2(a). The horizontal axis is labelled with the various cases, while the vertical axis indicates a combination of person and number, and each cell contains corresponding word-forms. This type of layout makes it easy both to retrieve specific information (e.g. that the 2sg benefactive pronoun in Hua is ‘gai-si’) and to identify certain patterns (e.g. that the citation forms are all suffixed by ‘-a’ and the 1pl forms are prefixed by ‘r’).

A similar presentation is found in the collection of data showing related word forms in four Polynesian languages shown in Figure 2(b) [9]. The languages themselves are labelled along the vertical axis, and the horizontal axis labels the items numerically. The content of the right-most column could be interpreted either as a ‘header’ giving the English gloss for the word-forms, or as a fifth language column which is not explicitly labelled.

The Diyari paradigm in Figure 2(c) shows a range of different stems and their inflected forms [2]. The layout is similar to that of the Hua paradigm, however some cells contain cross-references rather than word-forms, for example where the “ALL” form is represented by “=LOC” for certain groups, and the “DAT” form as “=ALL” in certain groups. Also, there is a combination of two forms for “woman’s name” from ‘NOM’ and ‘ACC’ to a single ‘ABS’ form. Rows are labelled with both numbering and glossing (e.g. “5. stick”). There is a separation of rows into three groups, where the third group shows only the locative, allative and ablative forms, and both grammatical (e.g. “Temp. Loc.”) and lexical (“today”) information is given in the row header.

The Cherokee syllabary in Figure 2(d) lays out the onset of each syllable on the horizontal axis and the nucleus in the vertical axis [11]. However, only the vertical axis is labelled with the nucleus forms (e.g. ‘-a’, ‘-e’, etc.), while the onset forms of the horizontal axis are not labelled, but rather in each column the Cherokee character is listed with its syllable form. The inclusion of this information is informative only when there are two representations for the syllable, such as ‘da’ and ‘ta’, where the voicing component needs to be specified. Some other representations of the same data label both axes, rendering the phonetic representation redundant. Besides these voicing distinctions, there are other non-standard cells, such as the ‘na’ form which has two alternate forms, ‘hna’ and ‘nah’, an extra form for ‘s’ with no nucleus, and a gap where ‘mv’ would be predicted.

2.2 Three-Dimensional Paradigms

The paradigms presented in Figure 3 go beyond the simple horizontal and vertical axes of the previous samples, yet they are still represented visually in similar ways. The Kanarese sample [21] shows the distinction between caste and regional dialects of this language. The paradigm shows two binary distinctions, with each of six word forms (labelled along the vertical axis) shown by caste (Brahmin or non-Brahmin) and by region (Dharwar or Bangalore).

The samples from Russian [20] and Qafar [16] both show simple two-dimensional paradigms but have two different verbs represented in the same tabular structure. In the Russian case, showing stress exchange in singular and plural forms of monosyllabic neuter noun stems, the ‘okno’ and ‘mesto’ forms are in separate “cells” within the table. In case of Qafar, showing mood inflections in two classes of verb, the two forms are paired within each co-ordinate point. The Qafar sample also contains empty cells, where the requestive forms are only presented in the first person, and the imperative and jussive forms are in complementary distribution with regard to the second person forms.

The consonant chart of the International Phonetic Association [1], labels the place of articulation along the horizontal axis and manner of articulation along the vertical axis. However, within the cells there is a voicing distinction, which is noted below the table in prose, and must be inferred from spacing or alignment within the cells where only one form is given. Also, empty cells indicate the absence of specific symbols in certain places, while shading is used to indicate ‘impossible’ articulations.

2.3 More Complex Paradigms

The phoneme chart of Warumungu in Figure 4(a) gives both phonetic (represented by square brackets and aligned left within the cell) and orthographic (represented by boldface type and aligned right within the cell) forms [19]. Both place of articulation (conventionally labelled along the horizontal) and manner (vertical) are explicitly labelled, and non-existing forms are left blank. There is a sub-class of the category ‘stop’ which distinguishes three different types (long voiceless, short voiceless, short voiced), however none of the other manners of articulation have such distinctions.

Sub-classes are also evident in the Anejom pronoun paradigm in Figure 4(c) [18], however in this case each subset (singular, dual, etc.) is repeated in each category, giving a third dimension. Where forms are not possible (i.e. singular forms of 1.INC), this is indicated with a dash.

A more complex paradigm is found with the French example in Figure 4(b). Here, only gender and number are labelled on the horizontal and vertical axes, yet each ‘cell’ contains example phrases showing three different cases and two different languages (French and English). The result is a four-dimensional paradigm represented visually as a two-dimensional table [10].

2.4 Discussion

The preceding survey has covered several issues. We review the major ones here: providing a complete description of the paradigm; describing the assumptions of cell interpretations; parameterising the presentation of the model; and extending the model to perform more complex operations.

To provide a complete description of the paradigm, a model must handle multi-dimensionality. Examples of four dimensional paradigms exist (as in the French example given above) and a greater number of dimensions are possible. Within each dimension there must be allowance for sub-classes, such as those shown in the Warumungu and Anejom examples.

Inherent assumptions about the interpretation of cell content may be complex, for example the inclusion of both the syllable and character in the Cherokee example, as well as multiple units within certain cells. The Russian and Qafar

Table 27.1 Personal pronouns

	Citation	Benefactive	Ergative	Genitive
1sg.	dgai-a	dgai-si'	dgaivi'bamu'da	dgai-'
2sg.	kgai-a	kgai-si'	kgaivi'bamuga	kgai-'
3sg.	gai-a	gai-si'	gaivi'bamu'	gai-'
1du.	ra'agai-a	ra'agaisi'	ra'agaimuta'a	ra'agai-'
2/3du.	pa'agai-a	pa'agaisi'	pa'agaimuta'a	pa'agai-'
1pl.	rgai-a	rgaisi'	rgaimuta	rgai-'
2/3pl.	pgai-a	pgaisi'	pgaimuta	pgai-'

(a) Hua personal pronouns (Haiman 1998:544).

	Tongan	Samoan	Rarotongan	Hawaiian	
1.	tapu	tapu	tapu	kapu	'forbidden'
2.	pito	pute	pito	piko	'navel'
3.	puhi	feula	pu'ʻi	puhi	'blow'
4.	tafaʻaki	tafa	taʻa	kaha	'side'
5.	taʻe	tae	tae	kae	'faeces'
6.	taʻata	taʻata	taʻata	kanaka	'man'
7.	tahi	tai	tai	kai	'sea'
8.	malohi	malosi	kaʻa	ʻaha	'strong'
9.	kalo	ʻalo	karo	ʻalo	'dodge'
10.	aka	aʻa	aka	aʻa	'root'
11.	ʻahu	au	au	au	'gall'

(b) Polynesian cognate forms (Crowley 1992:91).

Table 3.3. Diyari case forms

Stem	ERG	NOM	ACC	LOC	ALL	ABL	DAT
1. person-DL	kaṇawuḷali	kaṇawuḷu	kaṇawuḷaṇa	kaṇawuḷaṇu	=LOC	kaṇawuḷaṇundu	kaṇawuḷaṇi
2. person-PL	kaṇawaṛali	kaṇawaṛa	kaṇawaṛaṇa	kaṇawaṛaṇu	=LOC	kaṇawaṛaṇundu	kaṇawaṛaṇi
3. woman's name	ṭirimirindu	ṭirimirini	ṭirimiriṇa	ṭirimiriṇaṇu	=LOC	ṭirimiriṇundu	ṭirimiriṇaṅka
4. man's name	waṭamaṅkali	ABS waṭamaṅkaṇa		waṭamaṅkaṇu	=LOC	waṭamaṅkaṇundu	waṭamaṅkaṇi
5. stick	piṭali	piṭa	piṭaṇi	piṭaya	piṭandu	=ALL	
6. young man	ṭariyali	ṭari	ṭariṇi	ṭariya	ṭarindu	=ALL	
7. boy	kankuyali	kanku	kankuṇi	kankuya	kankundu	=ALL	
8. man	maṭarali	maṭari	maṭaraṇi	maṭaraya	maṭarandu	=ALL	
9. Place Name 'Farina'				widawaṭaṇi	widawaṭaya	widawaṭandu	
10. Temp. Loc. 'today'				karaṇi	karaṇaya	karaṇandu	
11. Spatial Loc. 'there'				ṇaka	ṇakaṇi	ṇakandu	

(c) Diyari case forms (Austin 1981:51).

A. Cherokee

	-a	-e	-i	-o	-u	-v = [ə]					
a	D	e	R	i	T	o	ḍ	u	Ḟ	v	i
ga	Ṣ	ka	Ḟ	ge	Ṣ	hi	Ḟ	go	Ḟ	gu	J
ha	Ḟ	he	Ḟ	hi	Ḟ	ho	Ḟ	hu	Ḟ	hv	Ḟ
la	W	le	Ḟ	li	Ṣ	lo	G	lu	M	lv	Ḟ
ma	Ḟ	me	Ḟ	mi	H	mo	Ḟ	mu	Ḟ		
na	Ḟ	hna	t	nah	G	ne	Ḟ	ni	h	no	Z
qua	Ḟ	que	Ḟ	qui	Ḟ	quo	Ḟ	quu	Ḟ	quv	Ḟ
s	Ḟ	sa	Ḟ	se	Ḟ	si	B	so	Ḟ	su	Ḟ
da	L	ta	W	de	Ṣ	te	Ḟ	di	J	ti	J
dla	Ḟ	tla	Ḟ	tli	L	tlo	Ḟ	tlu	Ḟ	tlv	P
tse	G	tse	Ḟ	tse	Ḟ	tsi	Ḟ	tso	K	tsu	J
wa	G	we	Ḟ	wi	Ḟ	wo	Ḟ	wu	Ḟ	wv	Ḟ
ya	Ḟ	ye	Ḟ	yi	Ḟ	yo	Ḟ	yu	G	yv	B

(d) Cherokee syllabary (Daniels 2001:65).

Figure 2: Examples of simple paradigms.

Table 2. Regional and caste differences in Kanarese

	Brahmin		non-Brahmin	
	Dharwar	Bangalore	Dharwar	Bangalore
'it is'	əðə	ide	ayti	ayti
'inside'	-oɭage	-alli	-āga	-āga
infinitive affix	-likke	-ōk	-āk	-āk
participle affix	-ō	-ō	-ā	-ā
'sit'	kūt-	kūt-	kunt-	kunt-
reflexive	kō	kō	kont-	kont-

(a) Regional and caste differences in Kanarese (Trudgill 1974:36).

(20) Case	Singular	Plural	Singular	Plural
nominative	oknó	ókna	méstó	mestá
accusative	oknó	ókna	méstó	mestá
genitive	okná	ókon	méstá	mest
dative	okné	óknam	méste	mestám
instrumental	oknó	óknami	méstom	mestámi
locative	okné	óknax	méste	mestáx

(b) Stress exchange patterns in Russian (Spencer 1998:137).

4

(12)	1sg.	2sg.	3m. sg.	3f. sg.	1pl.	2pl.	3pl.
requestive	ardóð				nardóð		
	nakóð				naknóð		
imperative		eréd			eréda		
		nák			náka		
jussive	árday	yárday	tárday	nárday			yardoónay
	nákay	nákay	náktay	náknay			nakoónay

(c) Mood in Qafar (Hayward 1998:638).

THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993)
CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill				r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

(d) IPA consonant chart (IPA 1993).

Figure 3: Samples of three-dimensional paradigms.

Table 32.1 Phonetic chart

	bilabial		apico-alveolar		apico-post alveolar		lamino-palatal ^a		dorso-velar	
stop ~										
long voiceless	[p:]	pp	[t:]	tt	[t:]	rtt	[c~t:]	jj	[k:]	kk
short voiceless	[p]	pp	[t]	tt	[t]	rtt	[c~t]	jj	[k]	kk
short voiced	[p~b]	p	[t~d]	t	[t~d]	rt	[c~t~j~d]	j	[k~g]	k
nasal	[m: ~m]	m	[n:~n~dn ^b]	n	[n:~n]	mn	[ɲ:~ɲ:~j~ɲ]	ny	[ŋ:~ŋ]	ng
lateral			[l:~l~dl]	l	[l:~l]	rl	[ʎ:~ʎ:~ʎ~j]	ly		
flap			[ɾ~r]	rr						
semivowel	[w]	w			[ɹ]	r	[y]	y		

Notes:

^a Lamino-dentals are sometimes used instead of these.

^b The very oldest speakers occasionally pre-stopped alveolar nasals and laterals instead of lengthening them.

(a) Warumungu phonetic chart (Simpson 1998:710)

	Masculine Nouns	Feminine Nouns
Singular	<i>le bon livre</i> 'the good book'	<i>la bonne maison</i> 'the good house'
	<i>ce livre vert</i> 'this green book'	<i>cette maison verte</i> 'this green house'
	<i>mon grand frère</i> 'my big brother'	<i>ma grande soeur</i> 'my big sister'
Plural	<i>les bons livres</i> 'the good books'	<i>les bonnes maisons</i> 'the good houses'
	<i>ces livres verts</i> 'these green books'	<i>ces maisons vertes</i> 'these green houses'
	<i>mes grands frères</i> 'my big brothers'	<i>mes grandes soeurs</i> 'my big sisters'

(b) French concord (Crowley, Siegel et al 1995:322)

Table 5. Anejom Pronouns

	1. INC	1. EXC	2.	3.
Independent				
Singular	—	<i>añak</i>	<i>aek, aak</i>	<i>aen, aan</i>
Dual	<i>akajau</i>	<i>ajamrau</i>	<i>ajourau</i>	<i>aarau</i>
Trial	<i>akataj</i>	<i>ajamtaj</i>	<i>ajoutaj</i>	<i>aattaj</i>
Plural	<i>akaja</i>	<i>ajama</i>	<i>ajowa</i>	<i>aara</i>
Object				
Singular	—	<i>ñak</i>	<i>yic, -c</i>	<i>yin, -n</i>
Dual	<i>cajau</i>	<i>camrau</i>	<i>courau</i>	<i>rau</i>
Trial	<i>cataj</i>	<i>camtaj</i>	<i>coutaj</i>	<i>ettaj</i>
Plural	<i>caja</i>	<i>cama</i>	<i>cowa</i>	<i>ra</i>
Possessive				
Singular	—	<i>-k</i>	<i>-ñ</i>	<i>-n</i>
Dual	<i>-jau</i>	<i>-mrau</i>	<i>-mirau</i>	<i>-rau</i>
Trial	<i>-taj</i>	<i>-mtaj</i>	<i>-mitaj</i>	<i>-ttaj</i>
Plural	<i>-ja</i>	<i>-ma</i>	<i>-mia</i>	<i>-ra</i>
Subject (aorist)				
Singular	—	<i>ek</i>	<i>na</i>	<i>et</i>
Dual	<i>tau</i>	<i>ekrau</i>	<i>erau</i>	<i>erau</i>
Trial	<i>taj</i>	<i>ettaj</i>	<i>ettaj</i>	<i>ettaj</i>
Plural	<i>ta</i>	<i>ekra</i>	<i>eka</i>	<i>era</i>
Subject (past)				
Singular	—	<i>kis</i>	<i>as</i>	<i>is</i>
Dual	<i>tus</i>	<i>eris</i>	<i>arus</i>	<i>erus</i>
Trial	<i>tijis</i>	<i>eris</i>	<i>atijis</i>	<i>etijis</i>
Plural	<i>eris</i>	<i>ekris</i>	<i>akis</i>	<i>eris</i>
Subject (inceptive)				
Singular	—	<i>ki</i>	<i>an</i>	<i>iñiyi</i>
Dual	<i>tu</i>	<i>ekru</i>	<i>aru</i>	<i>eru</i>
Trial	<i>tiji</i>	<i>etiji</i>	<i>atiji</i>	<i>etiji</i>
Plural	<i>ti</i>	<i>ekri</i>	<i>aki</i>	<i>eri</i>

(c) Anejom pronouns (Lynch 1998:106)

Figure 4: Examples of more complex paradigms.

examples show single paradigms including more than one lexical item. The Diyari example has some cell contents given as word forms and others as cross-references. The interpretation of empty cells is also variable, and a model must encode the linguistic intuition regarding systematic gaps. Empty cells may be the result of either incomplete data, non-existing values (as in Warumungu, Qafar) or impossible values (as in IPA, Anejom). The model must distinguish between each of these scenarios. The model must also handle the various non-Roman scripts common to linguistic description.

Parameterising the presentation of the data is important in paradigms which are constrained by convention (as in IPA and Warumungu) as well as those with more flexibility in layout (as in the reversal of horizontal and vertical axes between Hua and Qafar). The model should allow extensions such as templating of conventional or standard layouts. Selection and ordering of elements and inclusion or exclusion of row and column headings should be definable by the user, and a general model needs to allow for a wide range of practice, such as the use of labels or numbering of items in the Polynesian and Diyari examples. The ability to choose an appropriate ordering or to customise the layout allows the user to draw contrast between different properties of the paradigm. For example, the Kanarese paradigm shows the caste distinction above the regional distinction, where the data could be visualised differently to make a contrasting point. The model must allow operations which constrain display to subsets of the given data model, such as constraining the presentation of the French paradigm to show just one phrase or one language.

As has been seen there is enormous diversity in the nature and presentation of linguistic paradigms. The versatility and extendibility of the model is essential because the display of exceptional or incomplete sets is as important or more so than the display of standard material. Therefore, the model must both have a solid theoretical basis and a strong technical architecture. In an ideal case, a formalism would facilitate algebraically-expressed transformations. A model would not constrain preferences for structural navigation, allowing a seamless transition between depth-first and breadth-first traversals, and support multiple levels of recursion. Furthermore, a model must by default support the expression of conventional paradigmatic forms using a logical ordering (e.g. greatest number of dimensions first).

3. PREVIOUS WORK

Previous work on modelling linguistic paradigms is scant, a striking fact given the prevalence of this information type in language documentation and description. Few researchers have considered the requirements for multi-dimensional tabular representations, encoding models and algebraic formalisms, and the most significant work is reviewed here.

Tufis and Barbu [22] proposed a flat attribute-value representation independent of linguistic formalism, grounded in inflectional morphology, which allows flexibility of selection and manipulation of systematised linguistic information without structural impediments. Motivated by the need for computational lexicons to support natural language generation, these authors focus largely on the abstraction of tabular representations into formal computational models. While there are affinities between this approach and the present work, this contribution is differentiated by its grounding in descriptive, rather than computational linguistics.

Gyssens et al [14] proposed a tabular algebra for transformations of non-normative and semistructured data. It is of interest to note the focus on semi-structured data as this situation is typical of linguistic descriptions with incomplete analyses. While the present work does not explicitly seek an algebraic formalism for linguistic paradigms, there a number of similarities here, notably techniques to robustly handle incomplete data.

Yu et al [23], motivated by the need to integrate a range of disparate data sources in tabular formats and subsequently render these according to new transformational syntax, developed an XML algebra for diverse tabular representations. It should be noted that this research focuses on the need to handle validation as a precursor to transformation, a notion also adopted here. The present work has an affinity with this earlier contribution, although here tabular representations are approached as a presentational form which requires exploration, rather than the inverse.

Bird [3] reported on a Perl/CGI system called HyperLex which could generate complex linguistic paradigms from data stored in SIL's Shoebox format. The motivation for this earlier work in deriving a high degree of flexibility in visualisation, and efficiency gains through single data entry are common themes we also adopt here. The current paper extends and generalizes that work, replacing Shoebox format with XML, and Perl/CGI processing with XSLT transforms.

4. REPRESENTING PARADIGMS

In spite of their variety, linguistic paradigms simply represent an association between linguistic forms and linguistic categories. For example, in the German definite article paradigm in Figure 1, the form *den* is categorized as masculine singular accusative and as dative plural. Systematic changes in layout, such as interchanging rows and columns, or flipping axes, do not affect the associations between forms and categories. Accordingly, we can view a paradigm as a function mapping a vector of properties to a form as follows:

$$f : \langle \text{masc, sg, acc} \rangle \mapsto \text{den}$$

Generalizing, let $D_0 \dots D_n$ be a set of linguistic properties (or domains). Then a paradigm is a function:

$$f : D_1 \times \dots \times D_n \rightarrow D_0$$

Let $D_1 = \{\text{masc, fem, neut}\}$, $D_2 = \{\text{sg, pl}\}$, and $D_3 = \{\text{nom, acc, gen, dat}\}$. Also, let $D_0 = \{\text{der, die, das, \dots}\}$. We can now write down the functional representation of the German paradigm as shown in Figure 5.

Observe that the original paradigm display in Figure 1 is a compact view of this table. It shows the domain values just once, and dispenses with the gender property for the plural forms.

Now, the above representation is just a relational table with schema `GermanParadigm`(gender, number, case, form). We can use relational algebra to extract the columns of the original paradigm display, e.g.:

D_1	D_2	D_3	D_0
gender	number	case	form
masc	sg	nom	der
masc	sg	acc	den
masc	sg	gen	des
masc	sg	dat	dem
masc	pl	nom	die
masc	pl	acc	die
masc	pl	gen	der
masc	pl	dat	den
fem	sg	nom	die
fem	sg	acc	die
fem	sg	gen	der
fem	sg	dat	der
fem	pl	nom	die
fem	pl	acc	die
fem	pl	gen	der
fem	pl	dat	den
neut	sg	nom	das
neut	sg	acc	das
neut	sg	gen	des
neut	sg	dat	dem
neut	pl	nom	die
neut	pl	acc	die
neut	pl	gen	der
neut	pl	dat	den

Figure 5: Function for the German Paradigm

$$\begin{aligned}
& \{s \mid t \in \text{GermanParadigm} \wedge t[\text{number}] = \text{'sg'} \\
& \wedge t[\text{gender}] = \text{'masc'} \wedge t[\text{case}] = s[\text{case}] \wedge t[\text{form}] = s[\text{form}]\} \\
= & \{ \langle \text{nom}, \text{der} \rangle, \langle \text{acc}, \text{den} \rangle, \langle \text{gen}, \text{des} \rangle, \langle \text{dat}, \text{dem} \rangle \}
\end{aligned}$$

The same query is expressed in SQL as follows:

```

SELECT case, form
FROM GermanParadigm
WHERE number = "sg"
AND gender = "masc".

```

```

nom, der
acc, den
gen, des
dat, dem

```

A more convenient way to map from this abstract representation to the range of visualizations is to use standard XML technologies. The relational table can be trivially represented in XML as follows:

```

<paradigm>
  <form>
    <attribute name="gender" value="masc"/>
    <attribute name="number" value="sg"/>
    <attribute name="case" value="nom"/>
    <attribute name="content" value="der"/>
  </form>
  ...
</paradigm>

```

XSLT transforms can then be used to convert the material to HTML, or some other presentational markup language,

for delivery to users. Using this approach we will accomplish a round-trip: from existing visualizations (surveyed in §2); to an abstract underlying form (discussed in this section); and back to visualizations. It remains for us to provide this final step. This is the topic of the next section.

5. IMPLEMENTATION

In this section, an encoding model developed for representing and displaying linguistic paradigms is presented here, and its utility demonstrated. To illustrate, the Kanarese paradigm discussed earlier is used as a case study to demonstrate some of the issues involved in implementing a general model for paradigms.

An initial distinction is useful here. We differentiate between the encoded base data; the memory-resident abstract representation of the paradigm; and the browser-rendered output. The first section describes the memory-resident, or DOM[5], representation of the paradigm. The second section discusses an intermediate form of the XML[8] data which is used to simplify the presentational rendering. The third section includes a discussion of the transformational process which generates a browser-rendered XHTML[7] document from the intermediate representation. The fourth section describes the integration of these documents and transforms into a software system on the web, and the final section discusses the software architecture which provides the machinery in previous sections.

5.1 An XML Representation of the Paradigm

An adequate model for representing linguistic paradigms must preserve the underlying properties of the paradigm. That is, the model must preserve relationships between each cell and each heading. Consider an interpretation of the cell containing 'ide' in Figure 3(a), which represents the word-form corresponding to 'it is' for a Kanarese speaker of Brahmin caste in Bangalore. The headings related to the cell (Kanarese, Brahmin and Bangalore) constrain the possible interpretations of the cell. Describing each cell with row or column headings as coordinates uniquely identifies that cell and preserves every constraint placed on its content. A complete model must also represent the implicit relationships between headings such as the association between Dharwar and Bangalore in Figure 3(a). The DOM representation visualised in Figure 6 expresses completely both the content and relationships of the associated Kanarese paradigm.

The XML element named "attributes" contains a vocabulary of terms for each heading in the name and value elements. As such, the category 'caste' which was not present in the original paradigm (though it was in the label), has values 'Brahmin' and 'non-Brahmin'. The "form" element of the XML representation describes a set of constraints that uniquely identifies each cell in the paradigm. Each attribute element in the form section has a name-value pair which correlates to an element in the attributes section.

Typically, a paradigm such as the Kanarese example as found in the literature represents only a single view of the underlying data. An essential requirement for the XML representation is that it must not constrain the possible ways of displaying the data.

5.2 A Simplified Intermediate Form

The XML encoding model is useless unless there is a viable technique for visualising and manipulating the paradigm. The underlying XML representation structure is not suitable for tabular display. Use of an XSLT[6] transform is

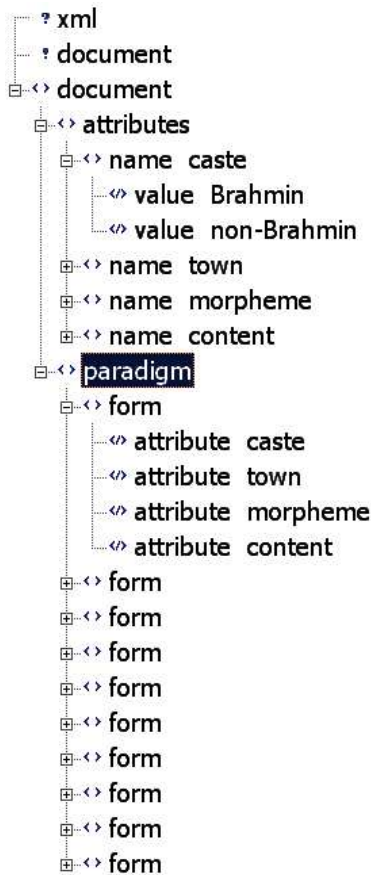


Figure 6: A partial DOM representation of the Kanarese paradigm.

proposed, which maps the original structure to a hierarchical structure while maintaining the integrity and completeness of the original data. The new hierarchical structure is conceptually equivalent to a decision tree. Figure 7 shows a partial decision tree representation for Kanarese and figure 8(b) and 8(a) two views of the equivalent XML representation.¹ Figure 8(b) shows how the tree is balanced like the decision tree in Figure 7. Figure 8(a) provides a depth first traversal of the tree, which shows that the structure is capable of handling unlimited depths of recursion.

5.3 Paradigm Visualisation

Presentation of the hierarchy is non-trivial as the nuances of implicit design decisions quickly become apparent. The display of headings and choice of axes (which produce a portrait or landscape orientation in tables) are purely arbitrary, although for ease of recognition, we present examples which are similar to conventional linguistic paradigms in the literature. Essentially another XSLT document translates the hierarchical structure into XHTML for a web browser to display.

One of the most challenging tasks for displaying an n -dimensional hierarchy as a table is choosing the number of dimensions to encode at each level. A three-dimensional table has two equivalent representations of its data; a two-dimensional table with a one-dimensional vector in each cells (Figure 9(a)); and a one-dimensional vector with two-dimensional tables in each cell (Figure 9(b)). A two-dimensional table is also equivalent to a one-dimensional vector of one-dimensional vectors as shown in Figures 9(c) and 9(d). The XSLT model developed starts at the root and attempts to fit the highest dimension table possible. The process continues recursively until every leaf is encoded. Figure 10(b) and 10(a) show the underlying XHTML structure of one such representation. A review of Figures 8(b) and 8(a) show the intermediate representation is almost mirrored in the XHTML output. There are two significant differences, the order of the nodes are transformed as discussed in the next section and the table markup is added at least every two steps down the tree (See Figure 10(a)).

Two views of the Kanarese paradigm visualised in a standard web browser using the aforementioned transformations are shown in Figures 11(a) and 11(b). The complete system allows arbitrary arrangement of axes by the user.

To display a two dimensional table in XHTML requires each row to be generated independently as shown in Figure 10(a). This creates problems when generating a table using XSLT because of the difficulty of accessing nodes across different branches of the tree. Figure 12 shows the correspondence between the nodes from the hierarchy and the position in the table. This problem has been solved for two dimensional tables but remains for higher dimensional tables, where headings are repeated in multiple. This is an acceptable but not optimal result, and possible solutions are being devised. The XSLT transforms are available at <http://www.cs.mu.oz.au/research/lt/-projects/paradigms>, with a prototype implementation available at <http://rimmer.cs.mu.oz.au:3051/paradigms>.

5.4 Software System Architecture

The overall objective of the implementation is to enable the generation of paradigms by querying interlinear text sources, such as those proposed by [4] and [17]. The query

¹The schema is self-referential and is produced using accumulator recursion in the XSLT document.

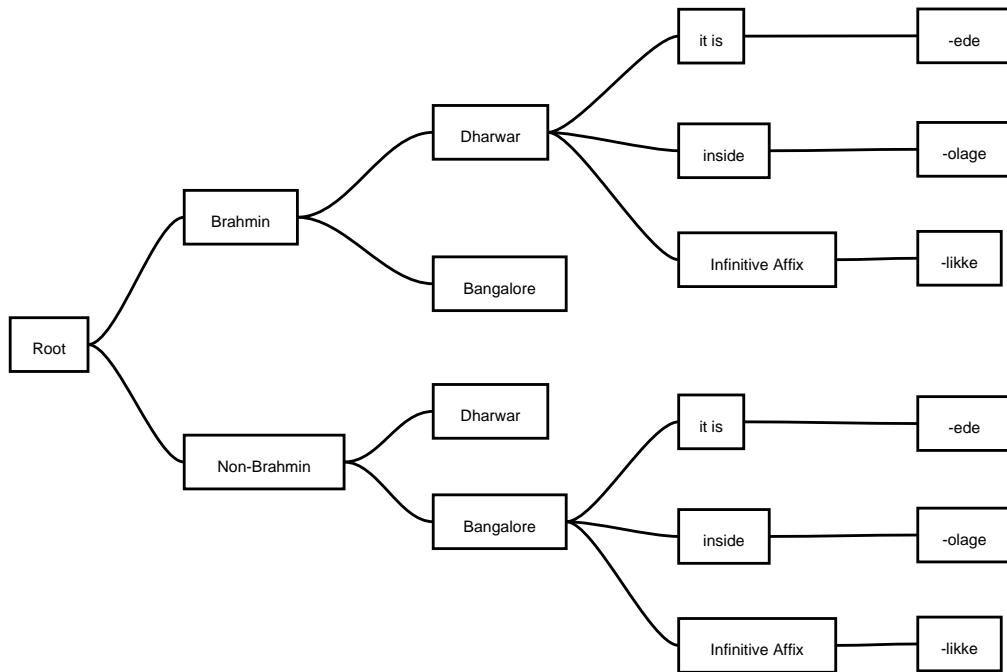
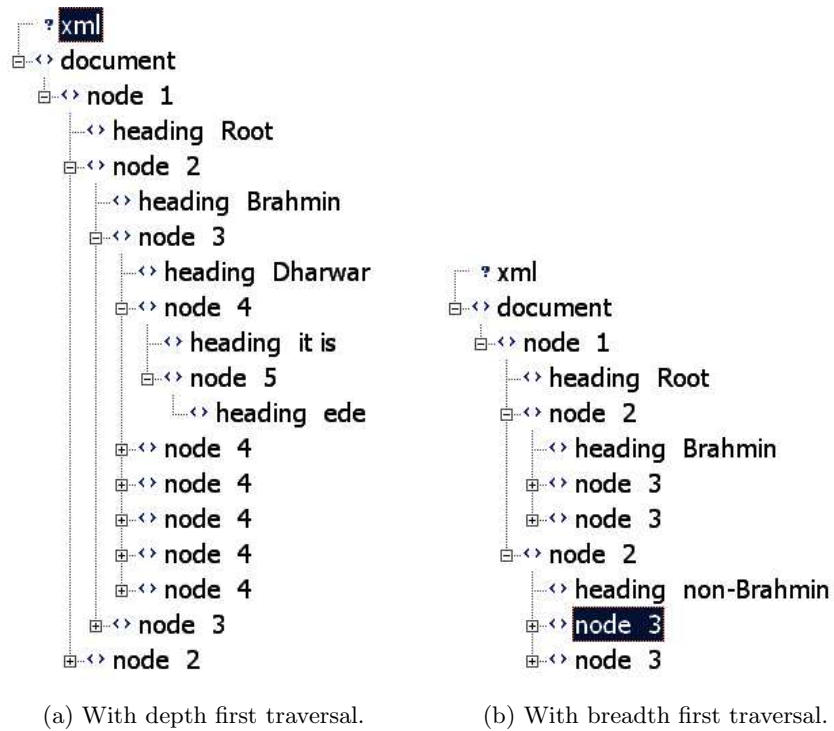


Figure 7: A partial decision tree for the Kanarese paradigm.



(a) With depth first traversal.

(b) With breadth first traversal.

Figure 8: A hierarchical presentation of the Kanarese paradigm.

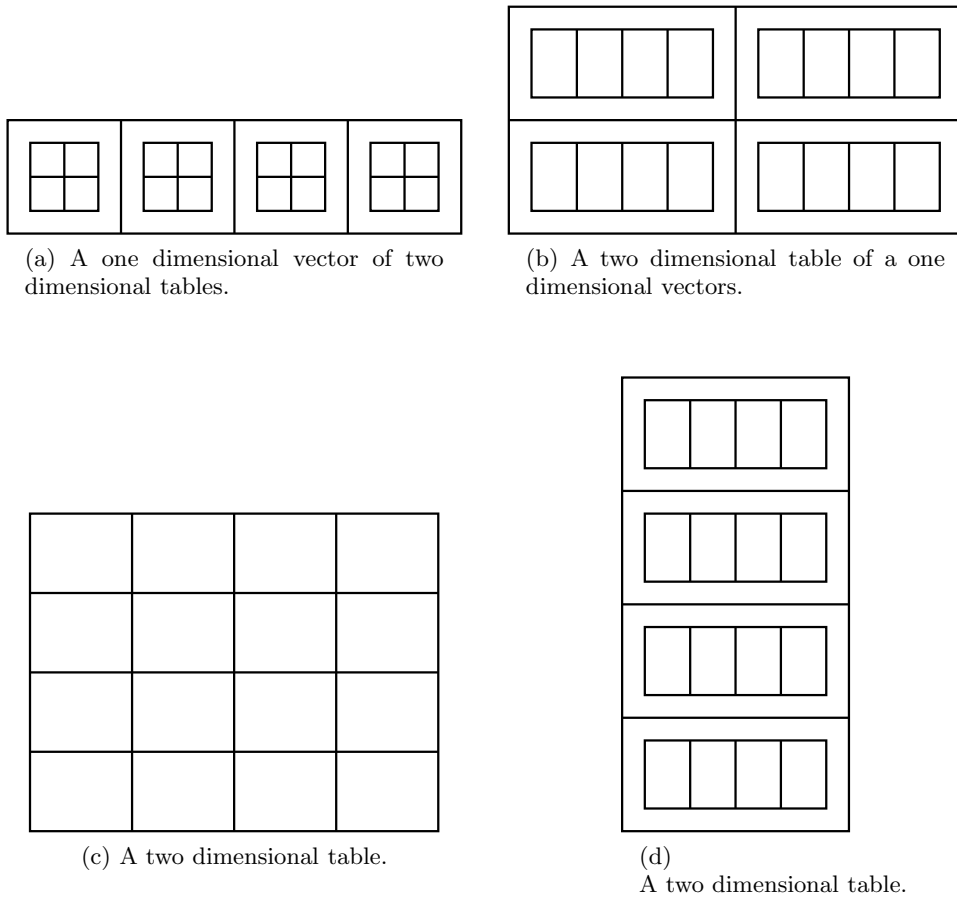


Figure 9: A sequence of different tabular representations.

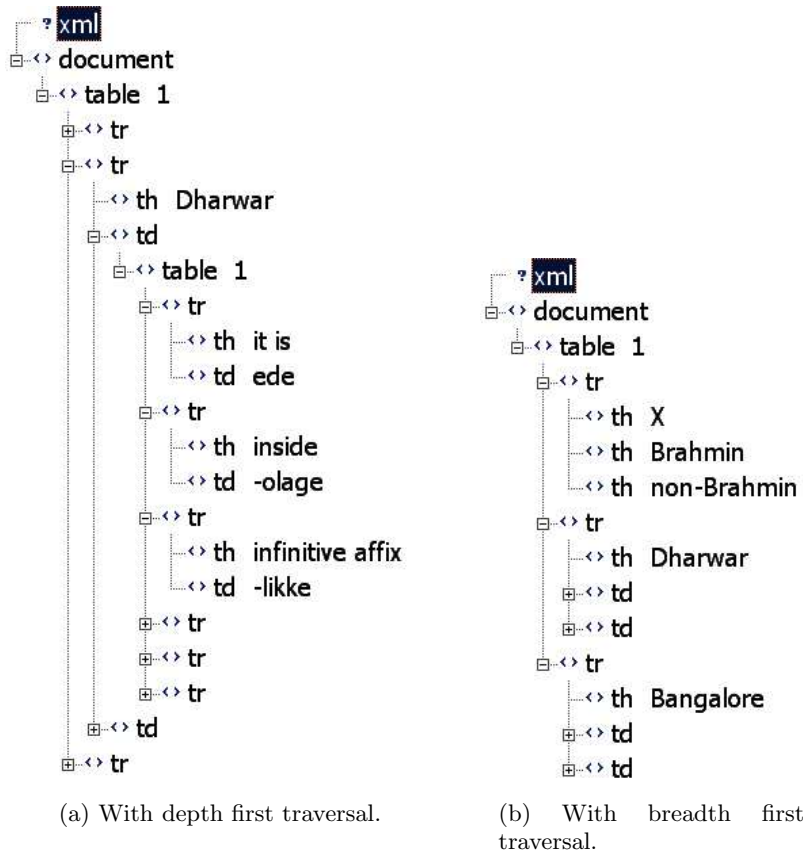


Figure 10: Different views of the XHTML representation of the Kanarese paradigm.

X	Brahmin		non-Brahmin	
Dharwar	it is	ede	it is	ayti
	inside	-olage	inside	-aga
	infinitive affix	-likke	infinitive affix	-ak
	participle affix	-o	participle affix	-a
	sit	kut-	sit	kunt-
	reflexive	ko	reflexive	kont-
	Bangalore	it is	ide	it is
inside		-alli	inside	-aga
infinitive affix		-ok	infinitive affix	-ak
participle affix		-o	participle affix	-a
sit		kut-	sit	kunt-
reflexive		ko	reflexive	kont-

(a)

X	Brahmin		non-Brahmin	
it is	Dharwar	Bangalore	Dharwar	Bangalore
	ede	ide	ayti	ayti
inside	Dharwar	Bangalore	Dharwar	Bangalore
	-olage	-alli	-aga	-aga
infinitive affix	Dharwar	Bangalore	Dharwar	Bangalore
	-likke	-ok	-ak	-ak
participle affix	Dharwar	Bangalore	Dharwar	Bangalore
	-o	-o	-a	-a
sit	Dharwar	Bangalore	Dharwar	Bangalore
	kut-	kut-	kunt-	kunt-
reflexive	Dharwar	Bangalore	Dharwar	Bangalore
	ko	ko	kont-	kont-

(b)

Figure 11: The Kanarese paradigm viewed through a web browser.

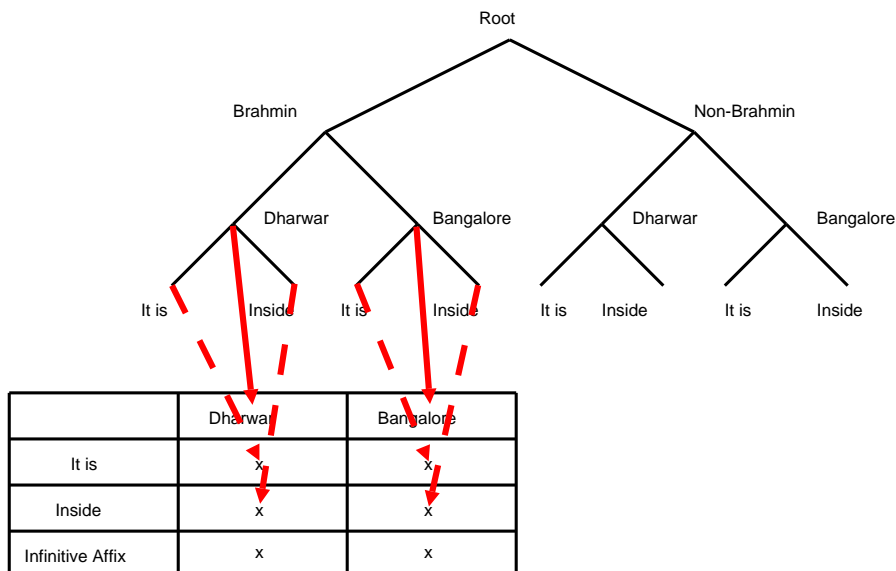


Figure 12: One mapping of high level headings from hierarchy to table.

would generate the XML representation of the paradigm, and subsequently transform the paradigm according to display preferences.

Figure 13 shows the system with a model-view-controller (MVC) architecture. The ‘model’ component is comprised of the XML representation of the paradigm and the transformation which creates a hierarchical XML file. The transform that produces an XHTML document for the web browser is the ‘view’ component. Other components that are possible are indicated in Figure 13. The ‘controller’ component is an XSLT style sheet containing the transformation logic, which is responsible for parameterising user input in order to generate the appropriate visualisation.

6. FUTURE WORK

This section describes a number of areas for future work on the paradigm model as described in this paper. We believe that a linguistically-grounded and computationally implemented model will enable linguists to manipulate and manage linguistic data in new ways. There are several areas in which further work is required: extending the model and implementation to support the full range of observations made in the survey; creating methods for constructing paradigms; enabling queries of the model; and developing a systematic model for user interactions.

The survey identified many features of paradigm visualisation that are not yet supported by our implementation. For instance, the German paradigm in Figure 1 collapses gender distinctions in the plural form. To get this display using the current implementation we would have to collapse gender and number into a single domain having four values: masc-sg, fem-sg, neut-sg, pl. The existence of a mapping for linguistic paradigms to a relational model opens up the possibility of using integrity constraints to capture such patterns in the data. For instance, we can require that the distinction between masculine and feminine is not made in the plural as follows:

$$\begin{aligned}
 \{t \mid & t \in \text{GermanParadigm} \\
 \wedge & t[\text{number}] = \text{'plural'} \\
 \wedge & t[\text{masc}] \neq t[\text{fem}]\} \\
 = & \emptyset
 \end{aligned}$$

The implementation would also need to be extended to exploit such constraints.

There are two methods for creating paradigms, through transformations from existing encoded data or direct data entry. Transformations are ideal in the case of fieldwork where data is prepared as an interlinear text or lexicon. The basis of the transform is either annotation (e.g. select all masculine pronouns and compare to feminine pronouns) or formula (e.g. select all words beginning ‘un-’). There is an obvious preference toward data entry when features are only evident from manually selected data. The first scenario brings challenges in the areas of linguistic analysis, information retrieval and artificial intelligence. The second relates to user models and intelligent interfaces. These are not contradictory goals, however, yet integrating both of these objectives into a single system remains an goal for future work.

In one sense the division between creating, viewing and editing data is arbitrary. Building a system that supports seamless integration with querying is more challenging. Immediate plans include investigating different data views such as data slices. Research will include investigating the application of XML query technologies to paradigm data.

There are a range of low level implementation issues which are not addressed at the time of writing, including the removal of redundant column and row headings, handling and displaying incomplete data, sorting, displaying partial information, displaying information for printed output and including more complex data. The first issue is the most pressing and the authors have every confidence of developing a model which better handles high-level headings. Not only must the model handle incomplete data it must also provide a mechanism for indicating the reasons for incomplete data.

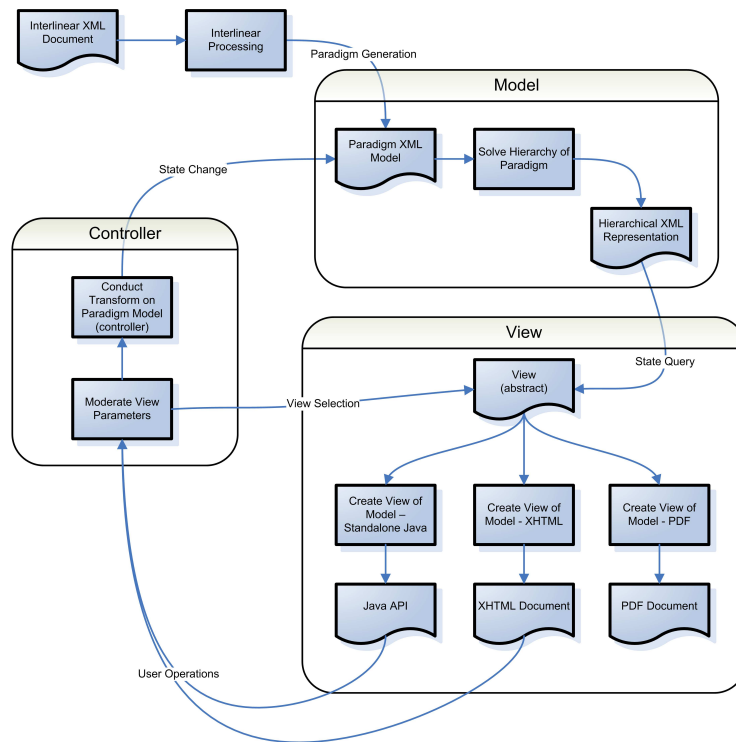


Figure 13: The architecture of the system for manipulating paradigms.

7. CONCLUSION

We have shown that a general model for linguistic paradigms can be conceived, and handle a wide range of conventional representations. An XML expression of the encoding model, together with relevant XSL machinery, provides a powerful tool for the exploration of flexible paradigmatic data in a fashion difficult to achieve in other contexts.

8. REFERENCES

- [1] T. I. P. Association. The international phonetic alphabet (revised to 1993), 1993.
- [2] P. Austin. *A grammar of Diyari, South Australia*. Cambridge, UK: Cambridge University Press, 1981.
- [3] S. Bird. Multidimensional exploration of online linguistic field data. In P. Tamanji, M. Hirotani, and N. Hall, editors, *Proceedings of the 29th Annual Meeting of the Northeast Linguistics Society*, pages 33–47. GLSA, University of Massachusetts at Amherst, 1999. Also appeared in *Notes on Linguistics (SIL)*, Vol. 2, pp. 125–144.
- [4] C. Bow, B. Hughes, and S. Bird. Towards a general model of interlinear text. In *Proceedings of EMELD Workshop 2003: Digitizing & Annotating Texts & Field Recordings*. Electronic Metastructure for Endangered Language Data (EMELD) Project, 2003.
- [5] W. W. W. Consortium. *Document Object Model (DOM) Level 1 Specification Version 1.0*. W3C, 1998. <http://www.w3.org/TR/1998/REC-DOM-Level-1-19981001/>.
- [6] W. W. W. Consortium. *XSL Transformations (XSLT) Version 1.0*. W3C, 2001. <http://www.w3.org/TR/XSL/>.
- [7] W. W. W. Consortium. *XHTMLTM 1.0 The Extensible HyperText Markup Language (Second Edition)*. W3C, 2002. <http://www.w3.org/TR/XHTML1/>.
- [8] W. W. W. Consortium. *Extensible Markup Language (XML) 1.0 (Third Edition)*. W3C, 2004. <http://www.w3.org/TR/2004/REC-xml-20040204/>.
- [9] T. Crowley. *An Introduction to Historical Linguistics, second edition*. Auckland, NZ: Oxford University Press, 1992.
- [10] T. Crowley, J. Lynch, J. Siegel, and J. Piau. *The Design of Language: An Introduction to Descriptive Linguistics*. Auckland, NZ: Addison Wesley Longman New Zealand Ltd, 1995.
- [11] P. Daniels. Writing systems. In M. Aronoff and J. Rees-Miller, editors, *The Handbook of Linguistics*, pages 43–80. Oxford, UK: Blackwell Publishers, 2001.
- [12] S. Farrar and D. T. Langendoen. A linguistic ontology for the semantic web. *GLOT International* 7(3), 7(3):97–100, 2003.
- [13] E. Finegan. *Language: its structure and use*. Fort Worth: Harcourt Brace, 1999.
- [14] M. Gyssens, L. V. S. Lakshmanan, and I. N. Subramanian. Tables as a paradigm for querying and restructuring. In *Proceedings of the 15th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. Montreal, Quebec, Canada.*, pages 93–103, 1996.
- [15] J. Haiman. Hua (papuan). In A. Spencer and A. M. Zwicky, editors, *The Handbook of Morphology*, pages 539–562. Oxford, UK: Blackwell Publishers, 1998.
- [16] R. J. Hayward. Qafar (east cushitic). In A. Spencer and A. M. Zwicky, editors, *The Handbook of Morphology*, pages 624–647. Oxford, UK: Blackwell Publishers, 1998.

- [17] B. Hughes, S. Bird, and C. Bow. Encoding and presenting interlinear text using xml technologies. In *Proceedings of the Australasian Language Technology Workshop 2003*. Australasian Language Technology Association / University of Melbourne, 2003.
- [18] J. Lynch. *Pacific Languages: an Introduction*. Honolulu: University of Hawai'i Press, 1998.
- [19] J. Simpson. Warumungu (australian - pama-nyungan). In A. Spencer and A. M. Zwicky, editors, *The Handbook of Morphology*, pages 707–736. Oxford, UK: Blackwell Publishers, 1998.
- [20] A. Spencer. Morphophonological operations. In A. Spencer and A. M. Zwicky, editors, *The Handbook of Morphology*, pages 123–143. Oxford, UK: Blackwell Publishers, 1998.
- [21] P. Trudgill. *Sociolinguistics: An Introduction to Language and Society*. London: Penguin Books, 1974.
- [22] D. Tufis and A.-M. Barbu. *A Reversible and Reusable Morpho-Lexical Description of Romanian*. Editura Academiei, 1997.
- [23] W. Yu, Z. M. Ozsoyoglu, and G. Ozsoyoglu. Xml restructuring and integration for tabular data. In *Proceedings of the 14th International Conference on Databases and Expert Systems. Prague, Czech Republic*, 2003.

9. ACKNOWLEDGEMENTS

The research in this paper has been supported by the National Science Foundation through Grant Number 0317826 (Querying Linguistic Databases).

APPENDIX

A. KANARESE PARADIGM IN XML

```
<?xml version="1.0"?>
<!DOCTYPE document SYSTEM "para1.dtd">

<document>

<attributes>
  <name name="caste" to="portrait">
    <value value="Brahmin"/>
    <value value="non-Brahmin"/>
  </name>
  <name name="town" to="portrait">
    <value value="Dharwar"/>
    <value value="Bangalore"/>
  </name>
  <name name="morpheme" to="landscape">
    <value value="it is"/>
    .
    .
    <value value="reflexive"/>
  </name>
  <name name="content" to="portrait">
    <value value="-a"/>
    .
    .
    <value value="kut-"/>
  </name>
</attributes>

<paradigm>
  <form>
    <attribute name="caste" value="Brahmin"/>
    <attribute name="town" value="Dharwar"/>
    <attribute name="morpheme" value="it is"/>
    <attribute name="content" value="ede"/>
  </form>
  <form>
    <attribute name="caste" value="Brahmin"/>
    <attribute name="town" value="Dharwar"/>
    <attribute name="morpheme" value="inside"/>
    <attribute name="content" value="-olage"/>
  </form>
  <form>
    <attribute name="caste" value="Brahmin"/>
    <attribute name="town" value="Dharwar"/>
    <attribute name="morpheme"
      value="infinitive affix"/>
    <attribute name="content" value="-likke"/>
  </form>
  <form>
    <attribute name="caste" value="Brahmin"/>
    <attribute name="town" value="Dharwar"/>
    <attribute name="morpheme"
      value="participle affix"/>
    <attribute name="content" value="-o"/>
  </form>
  .
  .
  <form>
    <attribute name="caste" value="non-Brahmin"/>
    <attribute name="town" value="Bangalore"/>
    <attribute name="morpheme" value="reflexive"/>
```

```
    <attribute name="content" value="kont-"/>
  </form>
</paradigm>
</document>
```

B. KANARESE PARADIGM DTD

```
<!ELEMENT attribute EMPTY>
<!ATTLIST attribute
name (caste|content|morpheme|town) #REQUIRED
value CDATA #REQUIRED
>
<!ELEMENT attributes (name+)>
<!ELEMENT document (attributes,paradigm)>
<!ELEMENT form (attribute+)>
<!ELEMENT name (value+)>
<!ATTLIST name
name NMTOKEN #REQUIRED
to NMTOKEN #REQUIRED
>
<!ELEMENT paradigm (form+)>
<!ELEMENT value EMPTY>
<!ATTLIST value value CDATA #REQUIRED>
}
```