1 Background issues to the tagging of Urdu

Prior to reporting the various questions investigated in the course of designing the part-of-speech tagging system, there are some background issues that should be touched upon. Firstly, section 1.1 describes the Urdu language itself, this thesis' object of study. Urdu is introduced briefly to provide background for the discussion in later chapters. This was felt necessary because the majority of linguists not specialising in Urdu, particularly in Western Europe, may be assumed to have a passing to no acquaintance with its history, current status and structures. Furthermore there are very few Urdu specialists, and few books on the language (see section 2.3). All in all the language is not widely studied. Therefore this section contains more detail than would be necessary if a more familiar language, such as English or Chinese, were being studied. The reader who is already familiar with Urdu is invited to move over section 1.1.

Secondly, since this study is situated within the corpus linguistic methodology, some preliminary details concerning this methodology are given in section 1.2. Given that a number of detailed overviews of the field have been published in recent years[1], an in-depth survey would be fairly superfluous. However, it is appropriate at this point to give a short overview of what exactly *part-of-speech tagging* is, and how it fits into the greater method of the study of language that is corpus construction and exploitation.

Thirdly, this study is situated within the context of the EMILLE project, a EPSRC-funded project being conducted by researchers at the University of Lancaster and elsewhere. The EMILLE project is discussed in section 1.3.

[1] For example, see Leech (1987, 1991); Church and Mercer (1993); McEnery and Wilson (1996: 1-19).

1.1 The Urdu language

1.1.1 Who speaks Urdu and where is it spoken? [2]

Urdu is principally a language of India and Pakistan, although it is important, particularly as a second language, outside this area as well, for example in Western Europe and America.

Masica (1991: 22) describes Urdu as having no specific territorial base, in the sense that there is no locality or set of localities in the Indian subcontinent that can be pointed at as an Urdu-speaking area. Although the cities of Delhi and Lucknow (India) and Karachi (Pakistan) are all home to many speakers of Urdu, and are listed by Schmidt (1999: xvi) as the loci of development of the standard forms of the language, these cities are also home to many speakers of other languages as well.

Rather than being geographically defined, the Urdu-speaking community is more easily defined in terms of religion. Most speakers of Urdu are Muslims. This is in contrast to speakers of the closely-related language Hindi, who are mostly Hindus. Indeed, it might be argued that were it not for the religious divide, the Hindi/Urdu distinction would not exist (see section 1.1.4 below for more on the distinction between Hindi and Urdu).

Urdu is spoken as a first language by over 60 million people (including 10 million in Pakistan and 48 million in India), or one percent of the population of the

[2] The source of the population figures given in this section is Ethnologue, compiled by SIL International. Available on the World Wide Web: http://www.ethnologue.com

Earth[3]. To put this into context, the most spoken language in the world is Mandarin Chinese, which has around 867 million first-language speakers (14% of the world). This is followed by Hindi (366 million), Spanish (358 million) , and English (341 million) (about 6% each)[4]. Whilst clearly Urdu is not as widely spoken as these languages – although its partial mutual comprehensibility with Hindi means that it is linked to the second largest language in the world – it still possesses a relatively respectable number of speakers. A comparable European language is Italian, with 62 million speakers. In this context, Urdu can be seen to be a significant language in terms of the size of its speaker population.

However, Urdu is demographically significant in another way as well. It is widely used as a second language throughout the Muslim communities of South Asia. As Schmidt (1999: xvi) says, "Urdu is also spoken in Bangladesh, Afghanistan and Nepal, and has become the cultural language and lingua franca of the South Asian Muslim diaspora outside the subcontinent…" Ethnologue lists Urdu as having 104 million speakers (1.7% of the world) when second-language speakers are included. This means that around 44 million speak Urdu as a second language – a high number relative to the number of first language speakers (around 73%). Compare Hindi (only about 33% as many second-language speakers as native speakers) or Mandarin (20%), or even English, nowadays the principal lingua franca of the world (48%). This demonstrates the great importance that Urdu has as a second language.

To give a concrete example, in Pakistan Urdu is spoken as a native language

[3] I assume, for the sake of argument, that the world population is six billion.

[4] Different lists of the world's most spoken languages give different figures depending a) on the year their estimates are based on – population growth being a factor; and b) on whether or not second-language speakers are included in the totals (if they are, English rises drastically, the others somewhat less so). The figures above are those given by Ethnologue; only native speakers are included.

by around 10 million people and is a national language. By contrast Punjabi is spoken by 30 to 45 million and is not, but very many Punjabi speakers use Urdu as a second language. Native speakers of Sindhi, of whom there are around seventeen million in Pakistan, also outnumber native Urdu speakers.

Through emigration, Urdu is now spoken widely across Asia, Europe[5], and North America as well as the Indian subcontinent; there are also speakers in South Africa and elsewhere.

1.1.2 Genetic affiliation and other associations

Urdu is a language of the Indo-European family. The Indo-European family, as its name suggests, also includes most of the languages of Europe and many of the languages of India and the area to the northwest. In terms of the internal classification of this family put forth by Ruhlen (1987: 325), Urdu is classified within the Central group in the Northern India group of the Indic[6] family within the Indo-Iranian branch of Indo-European. Thus, its closest relatives include Hindi, Gujarati, and Punjabi; slightly more distantly related are Sindhi, Marathi, Nepali, and Bengali. The particularly close association of Urdu with Hindi is discussed in more detail in the following section.

However, Urdu has also been particularly influenced by two other languages: Persian and Arabic. Persian, an Iranian language, is more distantly related to Urdu

[5] See Baker et al (1999) for an investigation into the nature of the communities speaking Urdu and other South Asian languages in the UK.

[6] The Indic family of languages is also referred to as "Indo-Aryan", presumably to distinguish it from the Dravidian family, which is, like the Indo-Aryan family, located largely in the Indian subcontinent.

than the languages listed above. Arabic is an Afro-Asiatic (specifically Semitic) language and thus unrelated to Urdu. The significance of these languages to Urdu is strongly linked to Islam. As noted, the majority of native speakers of Urdu are Muslims. Arabic is the language of Islamic holy texts, and Islam was brought into India by speakers of Persian.

The influence of these two languages is felt in many areas of the Urdu language. Much vocabulary has been loaned from them; some phrases have been taken intact, thus importing non-native morphology and syntax into the grammar; the script in which Urdu is written in is based on the Persian form of the Arabic alphabet; and even a number of phonemes seem to have been borrowed (see below). Note that the loans were not necessarily native to the language they were loaned from. An example is the derivational morpheme *–cī*, which forms an agentive noun on the basis of a noun. Persian borrowed this morpheme from Turkish before Urdu borrowed it from Persian (Schmidt 1999: 248).

A somewhat less profound, but still notable influence has been exerted upon Urdu by the English language, due to the political domination of the Indian subcontinent by the British Empire from the late eighteenth century until the middle of the twentieth century, and more recently to the use of English as a lingua franca in all parts of the world.

1.1.3 Historical notes

The roots of Urdu lie in the urban dialect spoken in and around Delhi from the twelfth and thirteenth centuries to the present day, known also as Khari Boli (an

Anglicised form of *khāRī bōlī*[7], for which Masica (1991: 28) suggests the translation "standard language"). It shares this origin with Hindi. This language of Delhi was adopted as the language of administration and government firstly by Muslim conquerors such as the Mughals (Masica 1991: 28) and later by the British (Kellogg 1875: xi-xii). Urdu was also by this point being used as a lingua franca throughout North India (Kellogg 1875: xvi) and had an established literature. The Urdu literary tradition is, in fact, now around eight hundred years old (see Bhatia and Koul 2000: 2).

While the standard form of Urdu is considered to be that of Delhi, there are other dialects. The most important is that known as Dakhini, which is centred on Hyderabad in Andhra Pradesh, southern India; however, like standard Urdu, Dakhini has no specific territorial base.

The division of Pakistan from India in 1947 had consequences for Urdu: many Urdu-speaking Muslims migrated from India to Pakistan, considerably increasing the population of Urdu native-speakers in the latter country, and leading to the development of another standard variety, that of Karachi (Schmidt 1999: xvi).

1.1.4 Urdu versus Hindi

Urdu and Hindi are more closely related to each other than either is to any other language. Indeed, their high level of similarity has led some to consider them dialects of the same language (as reported by Bhatia and Koul 2000: ix-x). Masica goes to so far as to suggest that by one definition of a dialect, Urdu and Hindi "are

[7] For details of the transliteration system used throughout, see Appendix 1. Sections of transliterated Urdu text are given in italics when they occur within a paragraph of running English text.

different *literary styles* based on the *same* linguistically defined subdialect" (1991: 27). However, their speakers are not of this opinion, due to what Ethnologue describes as "important sociolinguistic differences" between the groups. For example, they do not share a written standard, as do the dialects of languages such as English or French. But the distinction is not merely social: it is a complicated issue related to register and vocabulary. Depending on the circumstances of use, mutual comprehensibility can range from zero[8] to nearly 100%.

The two languages/dialects both originate from the dialect of the Delhi region (as noted above) and share their phonology, morphology and syntax in all but the smallest details[9]. This means that the colloquial forms of Hindi and Urdu are almost entirely mutually comprehensible. However, more formal and specialised registers – for example the language of literature, religion, philosophy, and law – utilise different vocabulary in Hindi and Urdu. As noted above, Urdu has borrowed a great deal of vocabulary (and its writing system) from Persian and Arabic, and it is particularly in these more elevated registers that this has taken place. On the other hand, Hindi has borrowed its vocabulary in these areas from Sanskrit, the language of the Hindu sacred texts and of classical Indian scholarship (and it uses the Devanāgarī alphabet). At this level, then, mutual comprehension diminishes greatly and may disappear altogether.

There are two other issues that cloud the nature of the relationship between Hindi and Urdu. The first is that the term "Hindi" is frequently used to describe a

[8] Mutual comprehensibility is zero in the written form, since Hindi and Urdu utilise different scripts.

[9] For example, Schmdit (1999: 109) reports that in Hindi, the auxiliary part of the conjunctive participle structure (see section 3.2.2.5) may be omitted in certain circumstances, whereas this is not possible in Urdu.

variety of dialects or local languages spoken throughout the "Hindi area", which consists of the Indian states of Uttar Pradesh, Bihar, Madhya Pradesh, Rajasthan, Haryana and Himachal Pradesh (Masica 1991: 9). These dialects are much less closely related to Urdu than is the official language of the area, standard Hindi, as related by Kellogg:

> … when I first entered India, I was repeatedly assured that the main difference between Hindí and Urdú was one of vocabulary … But the early delusion on this subject was soon dispelled. When we fancied we were speaking something like 'pure Hindí', the villagers stared confounded…
>
> (Kellogg 1875: xii-xiii)

This nineteenth-century experience demonstrates the gap in comprehension between the standard Hindi spoken by Kellogg, and the local languages of the Hindi area spoken by his rural interlocutors – much wider than the gap between Urdu and standard Hindi[10]. Thus, it must be made clear that when I discuss Hindi, I refer to the standard variety, and not to any of the wide range of regional variations found throughout the Hindi area.

The other point of confusion is the existence of the term "Hindustani" to describe the language of which Hindi and Urdu are dialects (if we assume that they are dialects). Although this term is no longer in use, it is still found in some of the literature. One final extra complication is the term "Hindi-Urdu", frequently used by linguists to describe the two languages/dialects together.

Having noted the language/dialect controversy, I do not consider it necessary

[10] Note however that in the century-and-a-quarter since Kellogg's work, standard Hindi has extended in usage considerably through the Hindi area. See also the discussion of Kellogg's terminology in footnote 53 in section 2.3.

to take a stand on one side or another of the debate[11]. Apart from anything else, it is a question which is impossible to resolve without a precise definition of the terms "language" and "dialect". It is not particularly important, for purposes of part-of-speech tagging, whether Hindi and Urdu are two languages or two forms of one language. Nor indeed is it relevant exactly how we choose to define "a language".

For the sake of clarity, I will use the terms discussed above as follows[12]:

- Urdu = the official language of Pakistan, originating from the dialect of Delhi, containing much Persian and Arabic vocabulary.
- Hindi = the most commonly spoken language of India, originating from the dialect of Delhi, containing much Sanskrit vocabulary.
- Hindustani (a term which, due to its current unpopularity, I will avoid wherever possible) = the colloquial form of Hindi and Urdu. This consists of the area of overlap between Hindi and Urdu – i.e. the phonology and grammar and a certain amount of core vocabulary.
- Hindi-Urdu = the whole of all registers of both Hindi and Urdu.

This usage can be represented diagrammatically as follows:

[11] That said, it would seem that at least for the moment, political and cultural considerations have carried the day, and Urdu and Hindi will be treated for the foreseeable future as separate languages.

[12] Historically, the meaning of these terms has varied greatly as different powers have controlled the Indian subcontinent. For example, during the period of British rule, "Hindustani" (or a variant of that word) was used by many European writers to refer to what would today probably be considered to be Urdu (e.g. Kellogg 1875: footnote on page 2). However, in this thesis I will keep to the modern meanings outlined above. Some other terms, such as "Moors" (a pidgin of English and Hindi-Urdu, sometimes used in the nineteenth century to refer to Hindi or Urdu proper), will be avoided altogether.

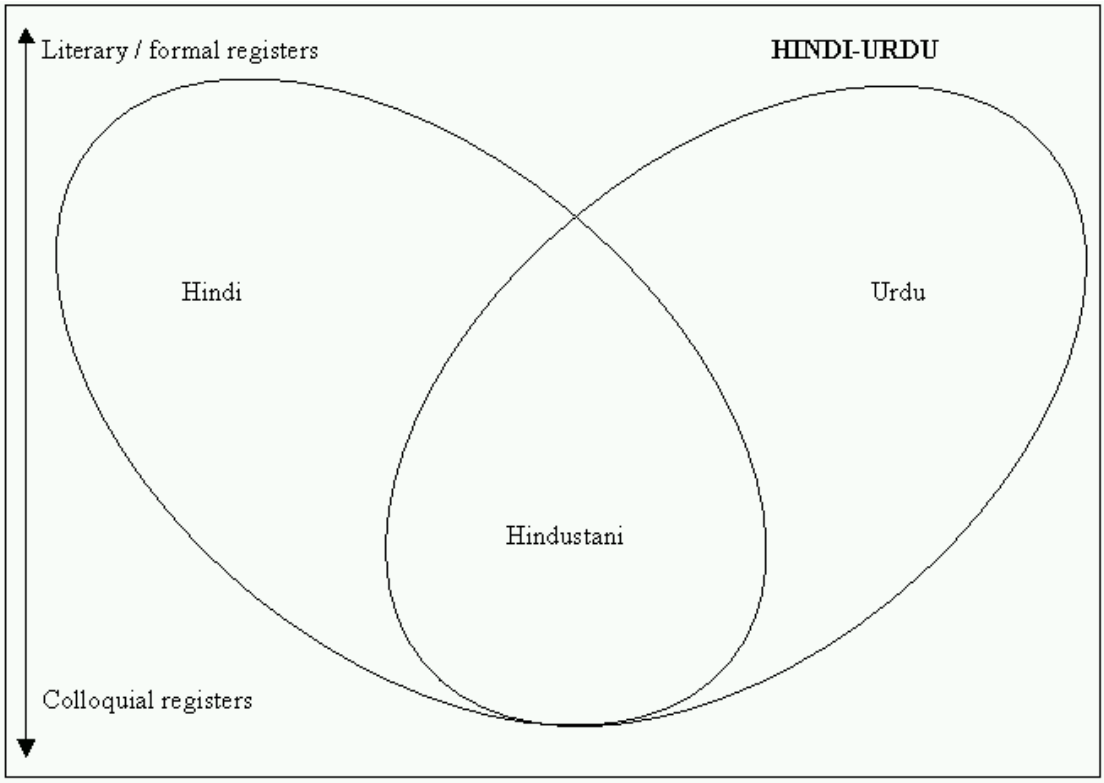


Fig. 1.1 Hindi, Urdu and Hindustani

1.1.5 A brief overview of the structure of Urdu

To provide some background to the more detailed discussions of Urdu morphosyntax in later chapters of this thesis, there follows a necessarily brief summary of some of the main structural features of the Urdu language. In this summary I have deliberately avoided pre-empting any subsequent discussion of controversial issues. In the sections on morphology and syntax, there are, however, a number of references forward to my discussion of such issues. So far as is possible, I have restricted myself to facts that are commonly known and generally accepted. I have also focussed on the unexpected as opposed to the commonplace. For example, in my discussion of phonology, I do mention the occurrence in Urdu of retroflex consonants, but do not discuss the nasals, since the nasals in Urdu follow a pattern

found in many languages.

My sources for this information, where not otherwise noted, are Schmidt (1999), Kachru (1990) and Bhatia and Koul (2000).

1.1.5.1 Phonology[13]

The following short account of Urdu phonology is drawn primarily from Kachru (1990), who gives the Hindi-Urdu phoneme inventory as tabulated below[14]. I have replaced Kachru's symbols with IPA notation. The phonemes in brackets occur only in what Kachru refers to as "highly Sanskritised or highly Persianised varieties"; however, since Urdu contains substantially more Persian vocabulary than Hindi, it will be convenient to regard them as full phonemes of Urdu.

Table 1.1: Vowel Phonemes (after Kachru 1990: 55)

	Front	**Centre**	**Back**
High	i		u
	ɪ		ʊ
Mid high	e	ə	o
Mid low	ɛ		ɔ
Low		a	

[13] For more information relevant to the phonology and writing system of Urdu, see Appendix 1, which details the system of transcription that I have used throughout this thesis.

[14] I was unable to find an author dealing solely with the phonology of Urdu, and was thus forced to rely on Kachru's composite Hindi-Urdu overview.

Table 1.2: Consonant Phonemes (after Kachru 1990: 55)

			Labial	**Dental**	**Retro-flex**	**Alveo-Palatal**	**Velar**	**Back Velar**
Stop	v.less	Unasp.	p	t	ʈ	c	k	[q]
		Asp.	p^h	t^h	$ʈ^h$	c^h	k^h	
	voiced	Unasp.	b	d	ɖ	ɟ	g	
		Asp.	b^h	d^h	$ɖ^h$	$ɟ^h$	g^h	
Nasal			m	n	[ɳ]	[ɲ]	[ŋ]	
Flap	voiced	Unasp.			ɽ	r		
		Asp.			$ɽ^h$			
Lateral						l		
Fricative	v.less		[f]	s	[ʂ]	[ʃ]	[x]	
	voiced			[z]		[ʒ]	[ɣ]	
Semi-vowels			w[v]			y		

Urdu is notable for its wide range of plosive consonants. At each of five places of articulation (labial, dental, retroflex, palatal, and velar) there are voiceless aspirated, voiceless aspirated, voiced unaspirated, and voiced aspirated plosives. This feature is shared with many Indo-Aryan languages (Masica 1991: 94, 101). Also notable is that the majority of the fricatives (other than / s /) are borrowed from Persian or Arabic.

Like English, Urdu possesses two groups of vowels that are distinct in terms of both length and quality. The tense vowels are also long.

1.1.5.2 Writing system[15]

1.1.5.2.1 The Perso-Arabic script

Urdu is written in the Perso-Arabic script. This first developed for writing Arabic[16], having its roots in an ancestor script used to write some earlier Semitic language (Sampson 1985: 77-82). Many other writing systems of the Middle East and Mediterranean share this source[17].

The most obvious visual qualities of the Perso-Arabic script are that it is written horizontally from right to left, and that it is always cursive, even in its printed form. As a consequence characters are realised differently depending on their position in the word (initial, medial, or final), and the majority of characters are joined to the characters on either side in writing and in print. There are also some ligatures – special forms which occur when particular characters are together.

---

[15] The writing system of a language is usually not considered a part of its basic structures, and many works on Urdu grammar (e.g. Schmidt 1999) overlook it altogether. However, since this thesis deals solely with the written form of Urdu (when speech is considered, it is in a transcribed – i.e. written – form) it seemed appropriate to summarise the writing system at this point.

[16] The usage of the script for writing Arabic differs in a number of significant details from the way it is used to write Urdu. When discussing the script, unless otherwise specified, I refer to the Urdu usage. A brief overview of Arabic usage is given by Kaye (1990: 179-180).

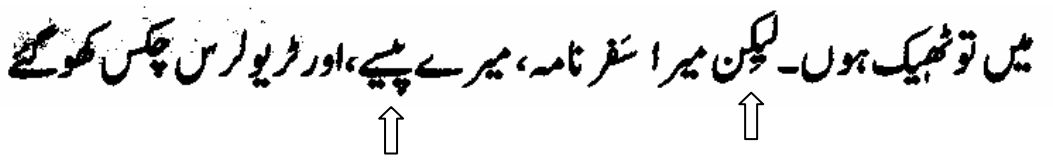
[17] The Hebrew, Greek, and (ultimately) Roman and Cyrillic alphabets are also derived from the Semitic, as described by Sampson (1985: 77-110). Some have also suggested that the Devanāgarī script used to write Sanskrit and Hindi, and related Indian scripts, is descended ultimately from an alphabet of the Semitic family (e.g. Hetzron 1990: 163).

1.1.5.2.2 nasx *versus* nasta'liq

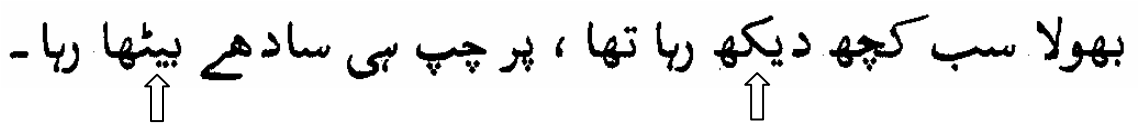
There are many forms of Perso-Arabic printing and calligraphy. For Urdu, the most important distinction is between the forms of writing called *nasx* and *nasta'liq*. As Bhatia and Koul explain:

> The first style [*nasx*] is employed for the Quran and all Arabic publications are printed in this style. This style is also produced by Urdu typewriters. The second style [*nasta'liq*] is beautifully handwritten by professional scribes and then lithographed. It is most commonly used in Urdu publications.
>
> (Bhatia and Koul 2000: 8)

More specifically, nasx is written from right to left horizontally. This is also true of the text in general in nasta'liq, but individual words tend to move diagonally, from upper right to lower left. Also in nasta'liq, characters tend to be blended together to a greater degree: the lines of the letters flow together and are frequently made very short, leaving only superscript and subscript dots to show what the characters were. Letters also sometimes occur on top of one another. Examples of both styles follow, enlarged in both cases; the arrows indicate words that are particularly good examples of the horizontal / diagonal distinction.



(nasta'liq; from Bhatia and Koul 2000: 247)



(nasx; from Schmidt 1999: 132)

Urdu grammars and dictionaries have been printed in both styles. Schmidt (2001) uses nasx, whereas Haq (2001) and Bhatia and Koul (2000) use nasta'liq. Aesthetic judgements aside, nasta'liq is harder to read for the beginner, and harder for a computer to produce: for this reason all Urdu text in this thesis is given in the nasx style. However, it should be noted that the association of nasx with Arabic is such that one Urdu speaker working on the EMILLE project commented that nasx was "Arabic writing" and not "Urdu writing"[18].

1.1.5.2.3 Urdu modifications to the Arabic alphabet

In Arabic, short vowels are not customarily written[19]. Long vowels and diphthongs are indicated by means of some associated consonant symbol. Urdu has considerably more vowels than Arabic, possessing ten contrasting oral vowels as opposed to six[20]. Urdu also has nasal vowels which may contrast with the corresponding oral vowels – this distinction is important, for example, in distinguishing between some case endings. However, Urdu maintains the Arabic practice of mostly omitting vowel diacritics. This means that there is a great deal of ambiguity for long vowels, since the Arabic consonant characters are pressed into

[18] R. Iqbal, personal communication.

[19] In books for early literate children, and in the Qur'an, superscript vowel diacritics are used; otherwise they are generally omitted.

[20] The figure of six vowels for Arabic does not count diphthongs, but does count both long and short vowels. Kaye (1990: 175) suggests that the long vowels are geminated forms of the short vowels, so an equally valid estimate would be three. Note that the various "dialects" of Arabic may differ considerably from Classical or Modern Standard Arabic with regard to vowel distinctions.

service for several purposes. For example, the medial form of the letter *ye* can stand for any of the vowels *ī*, *ē*, *ai*[21], or the consonant *y*. The letter *vāo* represents the vowels *ō*, *ū*, *au* and the consonant *v*. Thus, for these characters, there is a one-to-many mapping to phonemes.

Urdu's consonant inventory is somewhat different to that of Arabic. Distinctions that existed in Arabic are neutralised in Urdu loans from Arabic and thus in the alphabet. The symbols that in Arabic representing dental / t / and emphatic / t / are pronounced identically in Urdu[22]. The symbols representing / θ / , dental / s / and emphatic / s / are all *s* in Urdu. Dental and emphatic / ð /, emphatic / d / and / z / are all pronounced *z*. The voiceless pharyngeal fricative is pronounced as *h* in Urdu, as is the contrasting Arabic / h /. Thus, for these characters, there is a many-to-one mapping to phonemes.

Similarly, two Arabic consonants are sufficiently alien to Urdu phonology that they are not pronounced at all. The voiced pharyngeal fricative is not pronounced, or is pronounced as a vowel or glottal stop, and the glottal stop symbol[23] is used in writing to indicate a vowel cluster.

For the consonant distinctions which do not exist in Arabic, but are found in either Persian or Urdu, additional characters have been added. Those consonants shared by Persian[24] and Urdu include the symbols for *p*, *c*, *ź*, and *g*. Urdu-specific symbols include the systematically-created characters for the retroflex consonants *R*,

[21] Bhatia and Koul (2000) report that in some dialects of Urdu, the vowel I transcribe as *ai* is a diphthong rather than a long vowel; however, in the standard language of Delhi, it is a pure vowel.

[22] For details of Arabic phonology, see Kaye (1990).

[23] I refer here to the superscript *hamza*. The character *alif*, on which *hamza* is placed, indicates a vowel in Urdu.

[24] For the Persian alphabet and phonology, see Payne (1990).

T, and *D*.

Three characters have been split. The first of these is the Arabic symbol for / y /. This has been split into two characters, each of which represents two phonemes. Though both sides of the split in Urdu have the same initial and medial forms, their independent and final forms are somewhat different. The second is the Arabic / h /, which has been split into an *h* and a symbol marking aspirated consonants (also transcribed as *h*; see Appendix 1). Thirdly, a special form of the symbol for / n / is used to represent finally-occurring nasal vowels.

The split characters have shapes that are necessarily somewhat modified from those of Arabic; this is particularly true for the *h* and aspiration symbols. However, there are other characters whose Arabic shapes are modified in Urdu, for example *k*.

1.1.5.3 Morphology

Urdu displays a wide variety of derivational phenomena. Compounding is common, as is reduplication and echo-compounding (the compounding of a word with a reduplicated form of itself that has a different initial consonant; see Kachru 1990: 62). There are also many derivational affixes. Some have been borrowed from Persian and Arabic (Schmidt 1999: 246-271), for example *ɣair–*, "un–", or *–dār*, an adjective-forming suffix. Others are ancestral in Indo-Aryan, for example the *–ā* and *–vā* suffixes which create transitive and causative verb roots from intransitive verb roots. This process is also accomplished by means of ablaut of the verb root's vowel: see Kachru (1990: 57-58, 63) and Schmidt (1999: 157-175).

Urdu inflection is based on suffixation; the suffixes are fusional – consisting overwhelmingly of a single syllable, or even a single vowel, that may mark multiple

features (see also 3.1.5). Gender, number and case[25] are marked on nouns and adjectives (see also 3.1, 3.3). Verbs are marked for agreement in gender and number or person and number (see also 3.2). Verbs are not marked for tense, with the exception of the irregular auxiliary *hōnā*. Instead auxiliaries are used in combination with a non-finite form (the root, the infinitive or one of the participles) or the subjunctive[26].

In many cases, identical forms occur within or between inflectional paradigms. For example, the suffix *–ē* indicates masculine oblique case or plural number on adjectives, but also indicates the subjunctive form of verbs. See also section 3.1.5.

1.1.5.4 Syntax

The basic word order of the Urdu clause is generally given as subject – object – verb (SOV), an extremely common word order in the world's languages (Whaley 1997: 80-83). However, variation in this word order is common, particularly the reordering of nominal constituents "for thematic purposes" (Kachru 1990: 67-69). It should be noted, however, that Butt (1995: 21) among others has argued that Urdu is non-configurational, that is, that the ordering of elements of the sentence is not restricted.

While Urdu does possess subordinating conjunctions and relative pronouns, the relations that English are often expressed by subordination can also be expressed by the means of the verbal participles. These have a very wide range of uses (see

[25] Note that while opinion differs over what exactly constitutes the cases of Urdu (see 3.1.3), there is general agreement that the language does have cases.

[26] See also the discussion of compound verbs in section 1.1.5.4 below and in section 3.2.2.5.

Schmidt 1999: 108-111, 118-132), not only within the verb phrase, but also adverbially and adjectivally. With regard to relative clauses, Kachru (1990: 71) reports that they can follow their head noun, but the more usual situation is for them to precede or follow the clause in which their head noun is situated.

Urdu is notable for its use of verb phrases containing more than one verb (called "compound verbs" by Schmidt 1999; see Butt 1995 for an in-depth discussion of such structures). These consist of a lexical verb plus some sort of auxiliary or semi-auxiliary[27]. Depending on how much semantic content the auxiliary has, it may merely indicate tense, or it may add some shade of meaning to the lexical verb. Within the verb phrase, the auxiliary verb follows the lexical verb. In some cases (such as the future marker *gā* / *gē* / *gī*) this has led to uncertainty as to whether the morpheme is an independent auxiliary verb or actually a tense-marking suffix. For example, Kachru (1990), Schmidt (1999) and Bhatia and Koul (2000) all describe the future marker as a "suffix", but differ on whether this means it should be written as a single word with the verb or not.

Urdu, like Hindi and other closely related languages, is sometimes described as having ergative marking of grammatical relations[28] in constructions built on the

[27] While the auxiliary with the widest range of uses, *hōnā*, is nearly always described as such, a variety of terminology is used with regard to the auxiliaries in the verb phrases that Butt (1995) calls "aspectual complex predicates" and Kachru (1990) and Schmidt (1999) describe as "compound verbs". These auxiliaries are called "vector verbs" (Schmidt 1999) or "light verbs" (Butt 1995, who does not view these verbs as auxiliaries), and have also been known as "intensifying verbs", "compound auxiliaries" and "explicator verbs" (see Schmidt 1999: 143).

[28] Languages with ergative case marking (or, more accurately, ergative-absolutive marking) use the same morphological marking for the subject of an intransitive verb as for the direct object of a transitive verb. This is in contrast with nominative-accusative languages, which use the same marking

perfective participle[29] (e.g. Dixon 1994: 190; Masica 1991: 341-346). However, this is not an uncontroversial point of view. It has been contested, for example, by Butt, who views the so-called ergative marker (which I will refer to neutrally as the postposition *nē*) not as indicating a grammatical relation but as having "been invested with semantic content… as a marker of agentivity or volitionality" (Butt 1995: 14). Therefore, I do not intend to take a position on this matter, but note only that the tagging scheme developed in Chapter 3 is equally compatible with either of these two theoretical standpoints (see also my discussions of theoretical neutrality (2.2.4), case (3.1.3), verb agreement (3.2), and adpositions (3.7)).

A few other minor details may be noted. Urdu has postpositions rather than prepositions, and uses many phrasal postpositions (also known as complex postpositions). In the noun phrase, demonstratives, postposition phrases and adjectives precede their head nouns.

1.2 Part-of-speech tagging in corpus linguistics

As has already been mentioned, a great deal has been written on the development and current state-of-the-art of the corpus linguistic methodology. It is sufficient here to point out the important part that part-of-speech tagging has played in the development of corpus linguistics.

There are two defining characteristics of the collections of text known as

---

for the intransitive and transitive subjects and a different marking for the direct object. Since Urdu displays both patterns, it is sometimes described as having "split ergativity". For further details of ergativity see Dixon (1979).

[29] See section 3.2.1.3 for further details of the Urdu perfective participle.

corpora, characteristics which have been found in every modern corpus. Firstly, they are typically large – one million words, in the case of the early Brown Corpus[30], and increasing in size ever since[31]. Secondly, they are machine readable – i.e. the text is stored in digital form on computer, rather than solely on paper. These characteristics are key to the value of corpora in investigating language. A computer can search through, and compile data on the basis of, a vast body of machine-readable text comparatively quickly – and in doing so can accurately sift through much more data than the unaided human analyst could handle in a lifetime. A variety of analyses and statistics can be obtained in this manner – word frequency counts, concordances, finding collocations and keyword analysis being among the most common such operations.

However, the exploitation of corpora as a resource for linguistic investigation is increased considerably when corpus texts are marked up or annotated with additional information. There are various kinds of corpus annotation. Since a corpus text typically consists of nothing but running ASCII-encoded text, one basic type of encoding marks out features of the text other than the actual words, such as chapter divisions, distinctions between headings and body text, and so on[32]. Other types of

[30] For more details on the Brown Corpus, see Francis and Kučera (1982).

[31] This increase in size is largely due to computer memory becoming cheaper and data collection easier. For example, the British National Corpus, constructed in the early to mid 1990s, contains 100 million words. However, it should be remembered that "large" is always a relative term. There are research questions for which a corpus of a few thousand words could be adequate.

[32] This type of annotation is also referred to as "encoding". A range of different encodings have been used over the years; the most common of recent times has been SGML (Standard Generalised Markup Language), although the SGML subset known as XML (Extensible Markup Language) is growing in

annotation mark onto the text some aspect of linguistic analysis which it is anticipated will aid the user of the corpus in their research goals. Such forms of annotation may be performed manually, or may be automated and done by computer. Typical examples are semantic tagging, parsing or other marking up of syntactic constituents, and tagging of discourse features[33]. However, perhaps the most common form of corpus annotation is part-of-speech tagging[34].

Part-of-speech tagging may be defined as the process of assigning to each word in a running text a label which indicates the status of that word within some system of categorising the words of that language according to their morphological and/or syntactic properties (frequently known as a "tagset")[35]. These "categories" are often similar to, or subdivisions of, the eight parts of speech recognised by grammarians in the Latin/Greek tradition (see Voutilainen 1999a: 3-4). This is the source of the name "part-of-speech tagging"[36].

There are a wide variety of applications, both potential and actual, of part-of-speech tagging software and tagged text. These include information retrieval, word processor spelling- and grammar-checking, speech processing, handwriting

---

popularity. Standards have been developed for corpus encoding, the most noteworthy being the Text Encoding Initiative (TEI: see http://www.hcu.ox.ac.uk/TEI/Guidelines/index.htm for details).

[33] There are of course many "non-typical" annotation systems as well, as well as "one-off" systems devised for a specific purpose. Examples include the system used by Short, Semino and Culpeper (1996) to mark up features relating to the presentation of speech in texts, or that used by McEnery, Baker and Hardie (2000) to analyse swearing and abuse.

[34] See Leech (1997a) for more on the history of corpus annotation and part-of-speech tagging.

[35] Similar definitions are given by Leech (1987: 8) and van Halteren (1999: xiii).

[36] The process of "part-of-speech (POS) tagging" is also frequently referred to as "morphosyntactic annotation" (e.g. by Leech and Wilson 1999), or "(syntactic) wordclass tagging" (e.g. by authors in van Halteren 1999). I use these terms interchangeably.

recognition, machine translation, production of corpus-based dictionaries and grammars, and applications in the teaching of foreign languages and knowledge of grammar (see Leech and Smith 1999 for more details of these applications; see also Brill 1995: 552). It is thus clear that POS tagging is a key component of human language technology and corpus research.

Contemporaneous with the ongoing development of sophisticated forms of annotation has been the expansion of corpus linguistics into languages other than English. As is discussed in greater detail below, corpora are now available or in preparation for practically all the major languages of the world[37]. Clearly, part-of-speech tagging would be as valuable applied to these corpora in "new" languages, as it has proven in "old" languages such as English.

The aim of this thesis is therefore to combine a new direction in corpus linguistics with a very old one: to apply the tried and tested method of part-of-speech tagging to a corpus developed for a language that hitherto has not been studied via the corpus linguistic methodology – Urdu.

1.3 The EMILLE project[38]

This project is situated within the context of the EMILLE (Enabling Minority Language Engineering) project, being conducted by researchers at the University of

[37] Here it will perhaps be convenient to consider a major language as one in which a sufficiently large number of texts are produced and available in electronic form to make the collection of texts for a corpus practical.

[38] Details of the EMILLE project are available on the World Wide Web at http://www.emille.lancs.ac.uk/ .

Lancaster, the University of Sheffield and elsewhere. The starting point for EMILLE was the results of the earlier MILLE project. This investigated the availability of and demand for human language technology resources for a range of UK non-indigenous minority languages. On the basis of these results, the EMILLE project is intended to supply the major non-existent language resources for which a demand was identified. The results of the MILLE survey are described by McEnery et al. (2000); it is sufficient here to note that the main languages for which there was a demand in the human language technology community but no human language technology resources available or under development were the languages of South Asia – including Urdu.

Therefore, the aim of the EMILLE project is to develop corpus resources including nine million word corpora for a number of South Asian languages, including Hindi, Urdu, Bengali, Gujarati, Singhalese, Tamil, Punjabi, and Urdu. The most significant of these resources will be corpora of nine million words per language, incorporating written, spoken and parallel data.

It should be clear from what has been said above that for these corpora to be annotated with part-of-speech tags would be highly advantageous. However, tagging technology can take significantly longer to design and implement than any actual corpus. For this reason it was only considered practicable to attempt to tag one of the languages in question.

Urdu was selected as a suitable language for which to develop a part-of-speech tagging system[39] for three reasons. Firstly, MILLE found that the demand for Urdu resources was relatively high (McEnery et al. 2000: 803). Of all the minority languages on the MILLE questionnaire, Urdu was the seventh most in demand; of

[39] I refer here to an *automated* tagging system. Although it is theoretically possible for POS tagging to be done entirely by hand, this is not practical for corpora of the size being built by the EMILLE project.

those for which no reasonable amount of corpus data was found to be available, Urdu was the third most in demand. Thus it is reasonable that Urdu should be the language to be tagged.

Secondly, the demographic significance of Urdu, as detailed above, makes it a suitable candidate for the development of tagging technology. By these first two criteria, however, there would be better candidates – for example, Hindi. However, Urdu possesses a number of features that complicate part-of-speech tagging in that language. Some are grammatical features arising from the strong influence on its structure exerted by Persian and Arabic (see for example the discussion of the use of the Arabic definite article in Urdu in section 3.5). However, the main complicating factor is the Perso-Arabic script, which has a different directionality to the Latin text most frequently used for purposes of corpus annotation, which creates a number of interesting problems to be discussed later in this study. Thus, Urdu was selected as an appropriate language to be tagged because tagging Urdu seemed likely to be more difficult than tagging any other language in the EMILLE project.

One final reason for focussing on Urdu is that the MILLE project also showed that bilingual and multilingual data was most in demand (see McEnery et al. 2000: 803). As discussed above, Urdu is highly significant as a lingua franca and in bilingual contexts; it is therefore an ideal subject language for an experiment in part-of-speech tagging to be conducted.

1.4 Concluding remarks

In this initial chapter I hope to have demonstrated why Urdu is interesting – both linguistically in terms of its structures, and sociolinguistically in terms of its

history and speaker community. I have also detailed briefly the EMILLE project to clarify the context in which this thesis has been written.

Having covered these preliminary issues, the next task is to embark on the development of the part-of-speech tagging system for Urdu. The necessary first component to this system is the tagset, which is described in Chapter 3. However, before the tagset itself can be composed, it will be beneficial to examine tagset construction – both its history and the current best working practice in corpus linguistics. This will enable the formulation of some design features for the Urdu tagset, and it is this which is the topic of the following chapter.