# 2 Preliminaries to tagset design

Any project to annotate linguistic data presupposes the existence and use of a scheme of annotation. In the case of POS tagging, a POS tagset to categorise and mark up the words of the target text is an absolutely necessary preliminary. Defining such a tagset for Urdu is the primary aim of Chapter 3. My primary aim in this chapter is discuss a number of necessary preliminaries to the definition of the tagset, and to justify my decisions regarding these preliminary matters. A review of previous literature in the field of tagsets is given in 2.1. This includes a review of the EAGLES guidelines (Leech and Wilson 1999), which I use in Chapter 3 to guide the creation of the Urdu tagset. On the basis of the previous work in the field, I define some design principles for the tagset, as laid out in 2.2. A principle claim of this thesis is that these are the most appropriate principles for this type of tagset design, so they are discussed and justified in some depth. A discussion of my choice of a model of the Urdu language for use in creating the tagset is given in 2.3.

## 2.1 Previous work on part-of-speech tagsets[1]

As Voutilainen (1999) points out, "Tagging has been a hot research topic since the early 1980s"; however, research into tagging originated twenty years earlier with such early work as that of Klein and Simmons (1963). Since any attempt to perform tagging, manual or automatic, implies the use of a tagset, it follows that work on tagsets has likewise been ongoing since this early date.

[1] Following Leech (1997b: 20), I define a "tagset" as: a set of word categories to be applied to the word tokens of a text. See also the definition of part-of-speech tagging in section 1.2.

I will therefore present a brief discussion of this early and fundamental work, before going on to look at the development of tagsets during the 1980s. Finally, I will discuss moves during the 1990s towards, on the one hand, greater diversity in tagsets. I will also look at moves towards tagset standardisation through such initiatives as the EAGLES guidelines (see 2.1.3). My aim in this is to discern the design principles operating behind the creation and refinement of tagsets, in order to define a set of design principles for creating an Urdu tagset in the next section (2.2).

Since most work on tagsets has taken place within the context of broader studies (e.g. on corpus construction, parsing, or indeed tagging), I do not attempt to fully summarise any of the cited works, but only comment on them where they touch on tagset design.

### 2.1.1 The earliest work on tagset creation (prior to 1980)

Unsurprisingly, given the prominence of the USA both in linguistics and in computing technologies, the earliest work on tagsets in the 1960s and early 1970s occurred in the US and focussed on the English language. The most important tagsets of this earliest period are those of Klein and Simmons (1963) and Greene and Rubin[2] (1971). As Greene and Rubin note (1971: 2), their work was strongly influenced by Klein and Simmons' earlier study. Other work which touched on tagging was done at this time, for example at the University of Pennsylvania[3].

[2] This tagset was developed for use in tagging the Brown Corpus, and was refined slightly in a later stage of this project (see Francis and Kučera 1982: 3-15).

[3] A historical overview of this project is given by Joshi and Hopely (1997); some details of the parser's workings are given in Harris (1962).

Early work tended to stress the importance of POS tagging in parsing: as Greene and Rubin (1971: 1) point out, "It is generally accepted that, as a prelude to most syntactic analyses of natural language by computer, a text must be annotated with tags indicating parts of speech." Furthermore, they organise the definition and justification of their tagset based on the syntactic functions that a given form may take[4]. Thus, for example, verbal participles are described not with other verbal elements but with nouns, adjectives and determiners under the noun phrase. Likewise, Klein and Simmons' (1963) CGC ("computational grammar coder") software was designed as a component of a parser (itself a component of a system to synthesise human language behaviour). The work directed by Harris (1962) is almost entirely concerned with the computational analysis of syntax. Although some single-character tags are used (e.g. P, D, N, V, A) Harris does not give a defined tagset as such, or consider POS tagging as an operation separate to syntactic analysis.

The tagset of Klein and Simmons contains thirty tags, but their CGC program also outputs information, separate to the main tag, on the number of nouns and verbs; it is also noted if a noun is possessive, so that the actual number of categories distinguished is considerably greater. By contrast, Greene and Rubin's TAGGIT program used 77 tags[5], and information on number is incorporated into the single character string used as a tag.

These two early tagsets display some consistent design features. Both Greene and Rubin and Klein and Simmons incorporate tags for punctuation marks, which are

[4] Rather than organising the definitions by major part-of-speech, as is more common in the later work discussed below.

[5] The later, refined Brown Corpus tagset contained 87 tags (Francis and Kučera 1982); allowing for compound tags, the number of potential analyses for any given orthographic form is 179 (Sampson 1987).

treated as words, a practice which has continued to the present day (as Leech 1997b explains, the treatment of punctuation marks in this manner can be a significant aid in the tagging of other nearby words). In both, a number of highly common auxiliary verbs are tagged differently to main verbs; the CGC tagset goes further and tags the grammatical preposition *of* separately to other prepositions.

Both tagsets also contain some decomposable[6] tags. The pronoun tags in the CGC tagset are decomposable, though others are not. The tags used by Greene and Rubin display more of a tendency to be decomposable. For example, they consistently use the letter Z to indicate third person present tense, the letter D to indicate past tense, the $ symbol to indicate a possessive form. Some tags are entirely decomposable, e.g. the tag WPO, where W = wh-word, P = pronoun and O = objective form. However, unlike some later tagsets, this tagset was not hierarchical[7]: so the special tags for the English auxiliary verbs (DO, HV and BE) and modal auxiliaries (MD) are not presented within the tagset as subcategories of a more general "verb" category. Klein and Simmons' (1963) tagset was not hierarchical either.

Both these early projects also had some means of dealing with ambiguity. The CGC program could produce output such as NOUN/VERB, ambiguity which would

[6] A tag is considered to be "decomposable" if the string that represents that tag contains one or more shorter strings or single characters that are meaningful out of the context of the original tag and may be found elsewhere in the tagset with the same meaning. For example, any noun tag which combines an N for "noun" with other characters to indicate other features of the word is decomposable.

[7] The term "hierarchical", when used of a tagset, means that the categories in that tagset are *structured* relative to one another. Rather than a large number of independent categories, a hierarchical tagset will contain a small number of categories, each of which contains a number of sub-categories, each of which may contain sub-sub-categories, and so on, in a tree-like structure.

be removed at a later stage of parsing the text. Some of the TAGGIT tags were purely for dealing with ambiguous words. For example, the CI tag marks a word which is either a subordinating conjunction *or* a preposition, such as *before*. There are also tags for subordinating conjunction (CS) and preposition (IN). Only CS and IN are needed for an exhaustive classification, but CI is necessary on a pragmatic ground[8]: the limitations of tagging technology.

Ellegård (1978: 96-98) describes another tagset[9] from the early (pre-1980) period, which was used as part of a project to parse texts from the Brown Corpus. It is unlike the tags of Greene and Rubin (1971) and the CLAWS tagsets derived from the Brown tagset, in that the tagset is defined in a decomposed form. There are 25 single-character tags for major word classes, supplemented with further characters indicating inflectional features such as verbal form, noun/pronoun plural and genitive forms, and comparative/superlative forms[10]. However, this tagset cannot be viewed as hierarchical, because the definition of the twenty-five major divisions is entirely flat – there is no indication, for example, of greater linguistic similarity between common nouns (N) and proper nouns (C) than between verbs (V) and nouns.

[8] This tagset contains other tags that deal with ambiguity, e.g. RI (adverb or preposition). A more elegant means of dealing with ambiguity was later adopted in the C5 tagset (see Smith 1997: 147-148), used for tagging the British National Corpus, which contains joint tags for tokens that proved ambiguous between categories (e.g. NN1-VVG). This deals with ambiguity without increasing the number of categories in the tagset proper.

[9] Referred to as the "Gothenburg" tagset, for its author's affiliation, by Sampson (1987).

[10] Sampson (1987) estimates the total number of possible combinations at about 60.

### 2.1.2 Subsequent English tagsets (post-1980)

#### *2.1.2.1 Tagsets used with the CLAWS tagger*

Over the course of the 1980s and 1990s, a sequence of tagsets for English have been devised at Lancaster University for use with the CLAWS tagging software (see Garside 1987). The earliest, CLAWS1[11] tagset was used in the tagging of the LOB Corpus. Since this corpus was designed to parallel the structure of the Brown Corpus, the tags were also parallel, and CLAWS1 or the LOB tagset is very similar to the later version of the Brown tagset (Francis and Kučera 1982). The development of the CLAWS2 tagset was motivated by two requirements: "providing distinct codings for all classes of words having distinct grammatical behaviour", and making the tagset more "systematic in the way that tags are built up from individual characters" (i.e. more decomposable and hierarchical) (Sampson 1987: 167). As a result this tagset contains 166 tags[12]. This is a significant increase on the CLAWS1 tagset (132) and the Brown tagset (87). Furthermore, the tags were redesigned so that, for example, all verbal tags have V as their first character, and as their second character either V again (for a lexical verb) or another character (for a non-lexical verb). This made the tags more hierarchical.

The major subsequent developments in the CLAWS tagset were the C5 and C7 tagsets, developed for the tagging of the BNC and the BNC Sampler (see Leech,

[11] Also known as the LOB tagset. There exist a number of variations of this tagset, described by Sampson (1987).

[12] The CLAWS2 tagset was the basis for the much larger, much finer-grained SUSANNE Wordtag Set (Sampson 1995: 79-149; circa 360 tags).

Garside and Bryant 1994, Leech 1997b, Garside and Smith 1997, Smith 1997). The C7 tagset (146 tags) is the more fine-grained of the two and was used for the 2 million word Sampler; it can be regarded as a further refinement of the CLAWS2 tagset. The C5 tagset is something of a departure from the others, since it has far fewer tags (61) – this was in order to make it useful to the greatest number of end users.

Cloeren (1999: 50) characterises the C5 tagset as "flat", i.e. non-hierarchical. In fact, although none of the CLAWS tagsets are laid out in the hierarchical fashion described by Cloeren, the C7 tagset *is* hierarchical in conceptual terms (see Leech 1997b: 27-28). The same is true for most of the C5 tagset (there are exceptions – for example a cardinal number is CRD, an ordinal ORD: this cannot be reconciled with a left-to-right tag hierarchy). Furthermore, both C5 and C7 are largely decomposable – the C7, again, to a greater extent. For example, the tag PPHO2 is analysable as P = pronoun, P = personal, H = third person, O = objective case, 2 = plural.

#### *2.1.2.2 Other English tagsets*

##### *2.1.2.2.1 The TOSCA scheme*

The TOSCA analysis scheme (described by van Halteren and Oostdijk 1993) includes a POS tagset. This tagset differs considerably from the CLAWS tagsets, firstly in its form: it is made up of only 32 word class tags. However most word classes allow subclassification to be annotated in a feature list following the tag, meaning that the actual number of combinations is much higher. The TOSCA tagset is also notable in that it makes many more distinctions relating to the syntactic function of the word than the CLAWS tagsets: for example there are three major word class

tags for the word “it” depending on whether it is a pronoun, a formal “it”, a cleft “it”, or a provisional “it” (van Halteren and Oostdijk 1993: 160). Another example of such a difference is that the feature list for verbs can include information on transitivity of the verb, which is clearly a syntactic feature rather than a morphosyntactic one. These differences are probably attributable to the primary purpose of the corpus tagged with the TOSCA scheme, which was to study syntactic variation.

#### *2.1.2.2.2 The ICE tagset*

An important development from the TOSCA tagset is the ICE[13] tagset, described by Greenbaum and Yibin (1996). It distinguishes 19 word classes (a substantial reduction) but, like the TOSCA tagset, gives most words a feature list as well as a major word class tag. This means that the tagset contains, in effect, around 260 tags. This tagset as well contains significant differences of classification from the CLAWS tagsets: for example, the verb “be” is tagged as both an auxiliary and a verb depending on its function (AUX and V being different major categories in this system of description).

#### *2.1.2.2.3 The Penn tagset*

The tagset used in Penn Treebank (described by Marcus et al. 1993: 314-318) is also based on the Brown Corpus tagset. However, it has been modified in the direction of simplification, rather than complexity, as is the case with the CLAWS tagsets. Thus there are significantly fewer tags (36). It makes fewer of what Marcus et

[13] International Corpus of English: see Greenbaum (1996).

al. describe as "lexically recoverable distinctions", i.e. the distinction between main verbs and the verbs *be*, *do* and *have* is not preserved in this tagset since that distinction is made by the forms of the words themselves. Also, information that could be recovered from the parsing information in the Penn Treebank has been excluded from the tagset. Marcus et al. also suggest the risk of inconsistency in tagging (for example, in the Brown Corpus, between deictic adverbs and nominal adverbs) as a reason for simplification: "It is clear that reducing the size of the tagset reduces the chances of such tagging inconsistencies."

#### *2.1.2.2.4 The Lund tagset*

The tagset designed for the annotation of the London-Lund Corpus of Spoken English, described by Svartvik (1990), represents a tagset significantly different from the Brown Corpus/CLAWS tagset tradition. It is more fine-grained, consisting of just over 200 tags. Since it was designed for spoken texts, it includes tags for a variety of "discourse item"-type adverbs not usually distinguished in the tagging of written texts, as well as tags for other features of speech such as swearing . Similarly it lacks punctuation tags. Moreover, this tagset is entirely (and strictly) hierarchical and decomposable into single characters (or 2-3 character strings) that indicate given features[14].

[14] An earlier version of this tagset (Svartvik 1990: 96-98) was not as fine-grained as the version described, but was equally decomposable and hierarchical.

*2.1.2.2.5 The EngCG tagset*

The tagset used by the EngCG tagger (which is a component of the EngCG parser: see Karlsson et al. 1995) is somewhat different from the tagsets reviewed above. It is described by Heikkilä (1995: 111) as a "feature system" of "139 morphological or morphosyntactic features" rather than as a tagset *per se*. Heikkilä does not say how a feature system is different to a tagset. But it may be because in the tagging process, each word is normally given a single tag[15], but a single word may receive more than one of Heikkilä's features. 16 of the features (which are non-decomposable) are major part-of-speech features. As well as the more usual major parts-of-speech, these include classes such as "-*ing* forms" and "-*en* forms", and separate classes for co-ordinating and subordinating conjunctions.

The remaining features indicate inflectional and auxiliary[16] features, such as case, number, person, derivational features, and (for verbs) features for tense, transitivity and type of complement. A list is given by Heikkilä (1995: 115-131). Heikkilä also makes it clear that the syntax by which the features are annotated onto a word differs depending on the major part-of-speech given to the word. Thus, the full set of features given to any given word could be seen as equivalent to a single tag in a tagset using unitary tags. In that case, the number of possible "tags" would be huge indeed. Furthermore, rather than being made up of non-decomposable features, this tagset would instead embody the ultimate in decomposability, where decomposable elements are realised as separate character strings.

[15] The exception would be in cases where an automated tagger is programmed to leave a certain amount of ambiguity in the tagged output.

[16] That is "auxiliary" in the general layman's sense, not in the sense of "auxiliary verb".

### 2.1.3 A standard for part-of-speech annotation: The EAGLES guidelines[17]

When POS tagging came to be applied to languages other than English, the need for the creation of standards became clear. The most important recent standard on part-of-speech tagsets is the EAGLES guidelines (Leech and Wilson 1999). The aim of these guidelines is standardisation of tagsets used in different projects and/or for different languages. This is, as Leech and Wilson outline, is desirable for the following reasons:

> "In the interests of interchangeability and re-usability of annotated corpora, it is important to avoid a 'free-for-all', or a 'reinvention of the wheel' every time a new project begins… At the cross-linguistic level, annotations used for one language should as far as possible be compatible with annotations used for another. Compatibility here means that where there are descriptive categories in common between different languages, these should be recognised in the annotation scheme and recoverable from the annotations applied to texts in different languages."
>
> (Leech and Wilson 1999: 55-56.)

To this end, the EAGLES guidelines outline a set of features that tagsets should/may include. Simultaneously, a scheme of encoding all these features into an "intermediate tagset" is given. The choice of how the features are encoded within the scope of a particular corpus or research project is left to the user. The purpose of the

[17] The EAGLES guidelines actually make recommendations for standards for a range of language engineering resources. However, since the other parts of the project are not relevant for current purposes, I shall use the term "EAGLES guidelines/recommendations" to refer solely to the guidelines on morphosyntactic annotation of texts.

intermediate encoding is to allow "translation" between any two tagsets created in compliance with the EAGLES guidelines, thus ensuring their compatibility (in the sense given by Leech and Wilson).

The EAGLES guidelines for morphosyntactic annotation are structured on three levels: 1) what is considered obligatory; 2) what is recommended; 3) optional extensions for properties that are language-specific or marginal to *morphosyntactic*[18] annotation. At each level, tags are defined as morphosyntactic attribute-value pairs (e.g. *Gender* is an attribute that can have the values *Masculine*, *Feminine* or *Neuter* in the EAGLES recommendations). These attribute-value pairs may be structured as a hierarchy but need not be. In the intermediate tagset, they *are* so structured, inasmuch as some attributes are only applicable if another attribute has taken a particular value. For example the attribute *Pron.-Type* in the *Pronouns and Determiners* class is not applicable if the higher-level attribute *Category* takes the value *Determiner*.

The property suggested by the EAGLES guidelines as obligatory to any part-of-speech tagset is that of "major word categories", of which thirteen are proposed: noun, verb, adjective, pronoun/determiner, article, adverb, adposition, conjunction, numeral, interjection, unassigned/unique, residual, and punctuation. The recommended and optional attributes are then organised by these major word categories, and do not necessarily correspond across word classes. For example, the attribute numbered (i) (i.e. the first recommended attribute after the obligatory attribute of *Major Word Category*) is *Type* (*Common*/*Proper*) for nouns but *Person* (*First*/*Second*/*Third*) for verbs and *Degree* (*Positive*/*Comparative*/*Superlative*) for adverbs.

The recommended attributes also include number, gender, case, finiteness,

[18] That is, morphosyntactic as opposed to syntactic, semantic, or other annotation.

tense, voice, and other miscellaneous subcategorisation features. The optional part of the recommendations consists of similar attributes of lesser applicability, and some additional values – mainly specific to one language or a small group of languages – for the recommended attributes.

The EAGLES guidelines provide a flexible framework that encompasses all the basic things which one would want to mark up, without restricting the freedom of the tagset designer. Many of the tagsets discussed above include features which could not easily fit into the EAGLES guidelines (for example, the information on derivation included in the EngCG tagset, or the discourse features annotated by the Lund tagset; see 2.1.2.2.4, 2.1.2.2.5 above). However, these are exactly the features which are impossible to "translate" to another tagset – since no other tagset includes them – so to cover them in the EAGLES intermediate tagset would be superfluous. The value of this framework is that it promotes consistency and reusability of linguistic resources for different languages and discourages "wheel reinvention".

The main drawback to the EAGLES guidelines, however, is that they cover only a tiny fraction of the world's languages. As a project of the European Union, it covers only English, Dutch, German, Danish, French, Spanish, Portuguese, Italian and Greek: nine languages in all, which are moreover typologically similar, geographically confined to Western Europe, and closely related. As Leech and Wilson point out, "It remains to be seen how far these guidelines can be extended, without substantial revision, to other languages" (1999: 58). Therefore, the use of the EAGLES guidelines may or may not prove to be appropriate in the construction of an Urdu POS tagset.

### 2.1.4 Some recent tagsets based on the EAGLES guidelines

Since the release of the EAGLES guidelines on part of speech tagsets, they have been applied to many different languages. Leech and Wilson report (1999: 57) that around 20 users in Europe had adopted the EAGLES guidelines for one use or another.[19] While it is not possible to review all of these in depth, the use that three of these projects have made of the EAGLES morphosyntactic framework is discussed below. Each employs the framework for a different purpose. The MULTEXT project uses it to ensure comparability between newly-created resources, including "lexical specifications", for several different languages. The GRACE project, on the other hand, is concerned with multiple resources for a single language, and the CRATER project uses the framework to ensure cross-linguistic comparability between previously existing resources.

#### *2.1.4.1 The MULTEXT[20] project*

The MULTEXT project's aim is to develop "tools, corpora, and linguistic

[19] From the editor's introduction to the EAGLES website: "the EAGLES morphosyntax proposals are already being applied -- and consequently tested and evaluated -- in a number of national and European projects, such as LRE DELIS, RENOS, CRATER, MECOLB, MULTEXT, COPERNICUS MULTEXT-East and TELRI, MLAP-PAROLE, ESPRIT-ELSNET, French GRACE, German Textcorpora und Erschliessungswerkzeuge, LE-SPARKLE, ELRA, EUROWORDNET and PAROLE." (Calzolari, McNaught and Zampolli 1996).

[20] MULTEXT: Multilingual Text Tools and Corpora. An broad overview is given by Ide and Véronis (1994).

resources for a wide variety of languages"[21], starting initially with English, French, German, Italian and Spanish, but with an extension into many other languages. For example the MULTEXT-EAST[22] project extends the work done in MULTEXT to six languages of Central and Eastern Europe (including some non-Indo-European languages). The crucial innovation of MULTEXT was to introduce a distinction between an ideal tagset (in MULTEXT terminology not a "tagset" as such but rather as a set of "lexical specifications") and tagsets for actual use in the annotation of corpora.

The lexical specifications (described by Calzolari and Monachini 1996) represent a fairly direct implementation of the obligatory and recommended parts of the EAGLES guidelines. There are minor exceptions: some features are present in the MULTEXT lexical specifications that are not in the EAGLES, e.g. a feature "type" for adjectives (qualificative, ordinal, cardinal, indefinite, possessive). Similarly some of the attributes have been reordered (e.g. the features marked on pronouns), and the notation differs from that of EAGLES in its details (e.g. in the use of alphabetic characters rather than numerals).

The lexical specifications are intended to be independent of the language they describe (so the same system is used for English, French, German, etc.) and also of factors such as tagger limitations or the demands made on annotation schemes by the goals of particular projects. As Véronis and Khouri (1995) explain, tagsets used in

[21] Quoted from the front page of the project's website: www.lpl.univ-aix.fr/projects/multext/ , which also gives as examples of languages covered "Bambara, Bulgarian, Catalan, Czech, Dutch, English, Estonian, French, German, Hungarian, Italian, Kikongo, Occitan, Romanian, Slovenian, Spanish, Swedish and Swahili".

[22] Multilingual Text Tools and Corpora for Central and Eastern European Languages Project: a report from this project is available on the internet at www.lpl.univ-aix.fr/projects/multext-east/

corpora vary both between languages and within languages to such a great extent as to render them incompatible. This results from differences in what is marked and not marked by the tagsets, theoretical divergences, demands of the tagging process, and differences over the extent of the categories marked by the tags. Divergence between languages only exacerbates the problem.

Véronis and Khouri (1995) suggest dealing with incompatibility of tagsets – with each other and with the MULTEXT lexical specifications – by means of a two-level model. They suggest that the lexical specifications be used in lexicons, but that the actual tagsets used in corpora should be language- and purpose-specific. However, the two should be related in that any tagset should be an underspecification of the notation used in the lexical specifications. Thus, one tag may map onto one or many lexical specifications, but each lexical specification should be represented by only one tag.

As an example of how this works, Véronis and Khouri give the set of French verbs *viens*, *venais*, *viendrai*, *vienne*, *vinsse*, *viendrais*, and *vins* (all forms of the verb *venir*). In a particular French tagset, all these forms could be tagged VM1S (first person singular main verb), whereas in the lexical specifications, each of these would receive a different tag (since the lexical specifications cannot ignore mood and tense, as the VM1S category does). This allows tagsets of varied size and granularity[23], making varied decisions on which categories to mark, to be mapped onto the same basic morphosyntactic categories – essentially those of the EAGLES guidelines.

Thus, tagsets can be devised which are compatible not only with other tagsets for the same language, but with tagsets for other languages covered by MULTEXT/EAGLES. This demonstrates a principal advantage of adhering to an

[23] The granularity of the tagset may be a factor in the success of a tagging procedure (Smith 1997).

established standard.

#### *2.1.4.2 The GRACE project*

The aim of the GRACE[24] project is the evaluation of taggers for one language, French. In the course of this, it utilises a set of tags based on the work of MULTEXT and EAGLES (see Rajman et al. 1997). The way in which GRACE builds on MULTEXT and EAGLES is exemplary of what may be done within the stricture of the standards they propose.

The modifications to the EAGLES recommendations made in MULTEXT are retained in GRACE, and other minor changes are added. Two examples: there is no separate major category for numerals, which are instead coded as subcategories of other word classes; and the value neuter for the gender attribute has been removed, since it is not needed for French. Thus, a tagset for French (312 tags) was created for use in the evaluation project, without using the underspecification strategy described by Véronis and Khouri (1995). Because of its compliance with the established standards, and the high degree of decomposability that the said standards impose, this tagset is moreover very easily comparable to tagsets for other languages.

#### *2.1.4.3 The CRATER project*

The CRATER project (summarised by McEnery et al. 1997) was concerned with the creation of an aligned, parallel, tagged corpus in three languages (English,

[24] GRACE: Grammars and Resources for Analysers of Corpora and their Evaluation: details of the project are available on the internet at www.limsi.fr/TLP/grace/www/grace.html

Spanish, and French). For English and French, separate pre-existing taggers and tagsets were used; for Spanish a tagger was retargeted and a new tagset devised.

This new tagset, unlike the English and French tagsets, is fully conformant to the EAGLES guidelines. However, mappings from the non-EAGLES English and French tagsets to EAGLES-conformant representations were also devised. These mappings, given by McEnery (1996 – English) and CRATER (1996 – French), are, as can be seen, simple and unproblematic, despite very great differences in form and organisation of categories between the two source tagsets.

This demonstrates the value of a standard in unifying diverse annotation schemes. Furthermore, as McEnery et al. point out, "The use of EAGLES-conformant annotation should increase the utility of the CRATER deliverables to the European language engineering community" (1997: 224).

### 2.1.5 Some recent tagsets for languages not covered by the EAGLES guidelines

As Calzolari and Monachini (1996) point out, in the linguistics of Europe and North America, "Classification of lexical items relies on the old tradition of Greek and Latin grammar." Thus, this classification becomes more difficult when dealing with languages that are unrelated to, and mostly uninfluenced by, Latin and Greek. This includes most of the languages of the world, and at least two languages of considerable international significance, namely Chinese (Mandarin Chinese being, as noted in section 1.1.1, the first language of 14% of the human species) and Arabic

(forms of which are the native language of around 280 million people[25]). As I shall demonstrate below, tagsets for these languages are perforce rather different from the English and European tagsets reviewed above, and could not be designed in compliance with the EAGLES standards.

#### *2.1.5.1 Tagset design for Arabic*

The most significant Arabic tagset is that of Khoja et al. (2001). This tagset is based on the traditional description of Arabic grammarians. Words are divided into three classes (nouns, verbs and particles). What in European tagsets, including most of those discussed above, are the other major parts of speech (e.g. adverb, preposition) are dealt with as subcategories of these main three divisions. Thus the categorisation system is strictly hierarchical. The tagset given by Khoja et al. based on this categorisation is also fully decomposable.

This system is fairly incompatible with an European-style tagset, such as one based on the EAGLES tagset. However, as Khoja et al. demonstrate, it is extremely appropriate for Arabic, because of the way in which subcategories inherit properties of their parent classes. For example, adjectives inherit the marking of definiteness and indefiniteness that is a characteristic of the noun class of which adjectives are a subcategory.

Furthermore, taking a traditional Arabic approach facilitates the coding of

[25] The source for Arabic data is http://www.unhchr.ch/udhr/lang/arz.htm, website of the UN High Commission for Human Rights. Ethnologue, the source of my other speaker population figures, does not give a figure for the total number of Arabic speakers, instead listing each form of colloquial Arabic as a separate language.

features, such as jussive mood in verbs or dual number in nouns, that are fully relevant in Arabic but absent in EAGLES (Leech and Wilson 1999: 62, 63) and the tagsets it was designed to standardise. Khoja et al. also make the point that any attempt to use a European-style tagset for Arabic would go against the way that native speakers perceive the structure of Arabic[26].

This demonstrates that the value of the EAGLES guidelines diminishes once one moves beyond the bounds of the languages of the European Union for which the standard was developed. This value declines further when one considers language which are not closely related to those EU languages. Similarly, the value of the precedent set by European tagsets declines[27] (see my discussion of various tagsets' precedent above: 2.1.1 to 2.1.4.3).

#### *2.1.5.2 Tagset design for Chinese*

Two recent tagsets for Chinese are the CKIP tagset, developed in Taiwan, and the Jasmine tagset, developed in Hong Kong (both described by Piao 2000: 53-64). Piao also defines a tagset of his own, the CEPC tagset. Each of the three contains a number of features incompatible with EAGLES guidelines – although not the same features in each tagset. These include the existence of "aspect markers", "structural

[26] This could theoretically be true of imposing Latin or Greek grammatical categories onto non-Latin/Greek European languages such as English or French. However, while speakers of English or French *might* not agree with a model based on Latin and Greek, we have no way to ascertain this, since there is no alternative model for the categories of English or French. In the case of Arabic, such a model does exist, and accords better with native speaker perceptions of the language.

[27] See my discussion of such tagsets (sections 2.1.2 to 2.1.4), and of the design features which can be derived from their precedent (section 2.2).

markers" and "classifiers" as major word classes on a level with "noun" and "verb"; the part-of-speech "locative noun" as a sub-class of "noun"; subclassification of verbs according to their transitivity and stative/non-stative status; a major distinction between "content" and "functional" words (i.e. lexical and non-lexical categories) which splits apart verbs from auxiliaries of mood – difficult to reconcile with the EAGLES guidelines, which do not specify any hierarchy above the major word classes and class auxiliaries as a type of verb; and so on.

As with Arabic, we see that the value of the EAGLES guidelines is less with regard to languages other than those they were designed to cater for.

#### *2.1.5.3 Tagset design for Korean*

An example of a tagset for the Korean language is that described by Chae and Choi (2000). As with Khoja et al.'s (2001) tagset for Arabic, the structure of the tagset suggests a language that would not be compatible with the EAGLES guidelines. While most of the familiar POS categories are present, their grouping is not that found in European languages. For example, adjectives are classed with verbs in an overarching category "predicates", and there is a category "modifiers" which includes adverbs and adnouns. There are also separate major categories for particles, endings and suffixes. Thus, again, it can be seen that it would most likely be inappropriate to attempt an EAGLES-based tagset for this language.

However, despite the fact the categories described are emphatically not those of the EAGLES guidelines, the tagset itself has much in common with both EAGLES and other tagsets – such as being strictly hierarchical (on three levels) and fully decomposable.

#### *2.1.5.4 Tagset design in the Paninian tradition*

The languages of South Asia have their own grammatical tradition in the heritage of ancient grammarians such as Panini, who lived sometime in the middle of the first millennium BC. Panini's major work was a detailed description of the grammar of Sanskrit known as the Astādhyayī (Cardona 1976: 139). Panini did not work in isolation: his work was rooted in that of earlier grammarians and was followed by many centuries of continuous grammatical writings (Misra 1966: 14-17, 24-28). This tradition is entirely separate from the European Latin/Greek-based tradition discussed above, and indeed continues to the present day.

Recently, some work has been done on natural language processing in the Paninian tradition, for example Bharati et al. (2000), who describe some tags applied to text in Hindi from a tagset rooted in the analysis of Panini. There are several differences between this analysis and the European-tradition POS tagging. For instance, the distinction between POS tagging, parsing, and tagging of grammatical relations and semantic roles does not seem to be as clear cut in the analytic scheme of Bharati et al. So the tags *k1* and *k2*, representing *karta karaka* and *karma karaka* respectively, are in some ways like noun POS tags – but they also give information on semantic roles[28].

Note that for Hindi, Urdu and similar languages it is not impossible, as it seemed to be for Arabic, Chinese, and Korean, to design tagsets based on the

[28] Misra (1966: 31) glosses these terms as "agent" and "goal" respectively, which corresponds to Bharati et al.'s usage of the k1 and k2 tags. It should be noted that these semantic/syntactic tags are not intended either to be used in part-of-speech tagging *per se* or to be applied by computers.

European grammatical tradition. Indeed, I do so in the next chapter (for reasons explained in 2.3 below). One would expect this to be the case: the Indo-Aryan languages are genetically related to the languages of Europe and share many features with them. Sanskrit is in fact not incredibly dissimilar to Greek and Latin. However, for Bharati et al., the advantage of instead using a Paninian mode of analysis is twofold. It allows the process of tagging to draw on the expertise of analysts well versed in the Paninian tradition, and it is better equipped to deal with the structures of Indian languages due to being based on Sanskrit.

Here we see that beyond the original domain of the EAGLES guidelines, even where it is possible to apply them, it may prove advisable to use another approach due to some particular local circumstance (in this case, the influence of grammatical tradition).

## 2.2 Design principles for an Urdu tagset

There has been less written on the basic design principles of tagsets than on the actual creation of tagsets; nevertheless, there is a body of previous work in this area. Notable examples are Leech (1997b) and Cloeren (1999). In the section that follows I will outline the design features to which I will adhere in devising a tagset for Urdu. Since I am claiming that these represent the optimal design principles for a tagset of this nature (see Introduction), I describe the motivation and authority for including each in some detail.

### 2.2.1 Standards

I have made compliance with existing standards a design feature of the tagset for two reasons: firstly, such compliance is in general a desirable feature, and secondly, in the case of Urdu it is even more necessary because of the strong likelihood that an Urdu corpus would need to be used in multilingual studies. In this section I provide evidence for these claims, before discussing the choice of the EAGLES guidelines as the standard which has been complied with.

#### *2.2.1.1 The general advantages of compliance with standards*

Compliance with existing standards is desirable in the creation of any resource, tool or scheme for annotating or exploiting corpora. It increases the comparability of annotations on a corpus with those from other corpora, in the same language or across languages. It also increases the comparability of results derived from these corpora (Kahrel et al. 1997: 232). It helps make the corpus compatible with other resources and tools, such as parsing technology and computer-based lexicons and grammars (Leech and Wilson 1999: 55). It decreases the effort the end user must make to become familiar with a tool or annotation scheme. It also makes the set of potential end users as wide as possible by making the resource reusable, thus saving time and money (Kahrel et al. 1997: 232)[29].

[29] It is for these same reasons that it is worth deriving design principles from previous work in the field at all, rather than inventing them *a priori*.

#### *2.2.1.2 The particular advantages of compliance with standards when working with Urdu*

It is particularly likely that tagged Urdu texts will be used for multilingual studies and applications. As outlined in section 1.3, the MILLE survey (McEnery et al. 2000) found that among researchers in the field of human language technology, there was a significant demand for bilingual and multilingual data – this being the same group that will probably constitute the end users of any tagged Urdu corpus. Furthermore, the EMILLE project itself – of which the POS tagging system described in this thesis forms a part – contains a significant multilingual element, in that each of the corpora created by the project contains a significant proportion of parallel data.

Even had this demand not been apparent in the MILLE survey, it would still be the case that bi- and multilingualism is a more important issue for speakers of Urdu than for many European languages. This is obviously the case for Urdu-speaking communities outside the Indian subcontinent, e.g. in the UK or South Africa, whose members must be at least bilingual if they wish to communicate with the larger community. But even within Pakistan, more people speak Urdu as a second language than as a first, and in India, Urdu is everywhere spoken alongside other languages (see section 1.1.1 for details).

Thus it can be seen that use in multilingual studies and applications means that the Urdu corpus will need to be used alongside other corpora, most likely POS tagged by another system, in other languages. Making the tagset cross-linguistically comparable to facilitate cross-linguistic analyses and exploitations by the end user must therefore be a key aim. The best way to achieve this is through compliance with existing standards.

#### *2.2.1.3 Adherence to the EAGLES guidelines*

For the reasons outlined in 2.2.1.1 and 2.2.1.2 above, the Urdu tagset will adhere to an established standard. The standard in question will be the EAGLES guidelines. This is not a straightforward decision, since Urdu is not one of the languages that these guidelines were designed to cover. As has been demonstrated above (2.1.5), tagsets for languages outside this domain are often of a nature that renders them incompatible with the EAGLES guidelines (e.g. Khoja et al. 2001, Piao 2000). The practicality of attempting to create an Urdu tagset within the EAGLES framework is thus questionable.

However, the case of Urdu is somewhat different to that of Arabic or Chinese. Although Urdu is not a European language, it is still part of the Indo-European family (see 1.1.2, and also 2.1.5.4), as are the languages covered by EAGLES. Thus we can expect guidelines drawn up for the languages of Europe to remain largely relevant. However, it will be borne in mind that the EAGLES guidelines are here being extended to a domain that they were not expressly intended to cover. Aspects of Urdu morphosyntax that cannot be mapped in a straightforward manner to the EAGLES intermediate tagset are to be expected: every language has its particular idiosyncrasies. The use of the EAGLES guidelines in creating a tagset for Urdu will allow the claim stated in the Introduction, that this standard is useful beyond its original domain and can validly be applied to Urdu, to be evaluated.

### 2.2.2 Information to include

The most absolutely fundamental question of tagset design is "What should the tags tell the user?" or, to put it another way, "What information should be included?" According to Cloeren (1999: 38), "all tagsets account for major wordclass information". What then are these major word classes or parts of speech? A good working definition of a "part of speech" is that of Greene and Rubin (1971: 3): "categories [that] group lexical items which perform similar grammatical functions"; see also my own definition of POS tagging in section 1.2. There has developed an established consensus on what these categories, Cloeren's "major word classes", should be. Cloeren lists adjective, adposition, adverb, article, conjunction, interjection, noun, numeral, pronoun/determiner, and verb, with punctuation often counted as a major class, and two classes, unique and residual, accounting for one-member classes and items which do not fit elsewhere in the scheme of analysis. These are, indeed, the top-level obligatory categories given in the EAGLES guidelines. To include these categories must clearly be a design principle, to maintain compliance with this international standard as well as with existing practice.

There are also distinctions which Cloeren refers to as "subclassifications of the major wordclasses" (1999:39). These include such contrasts as common versus proper (nouns), main versus auxiliary (verbs), degree versus general (adverbs), prepositions versus postpositions, and so on. There is also morphological information, noting such features as number, person, gender and case. In line with the principle of compliance with standards, it will be a design principle to include so many of these as are licensed by the EAGLES guidelines, or which seem to be particularly relevant to the matter at hand (i.e. Urdu morphosyntax). There will not be any particular distinction made

between the three classes of information: major word class, subcategory and morphological data will all be given in a single tag. The reason for this is that there may not be full agreement among all researchers as to which linguistic feature belongs where in this scheme. For example, the status of the main verb versus auxiliary verb distinction is seen as a difference of subcategorisation in the EAGLES guidelines, but some tagsets (for example, the ICE tagset: Greenbaum and Yibin 1996) consider this to be a difference of major word class. Furthermore, the syntax of the decomposable (see the following section 2.2.3) unitary tags will allow separation of the encoded information into major word class, subcategory and morphological data by any end user who should wish to use this three-way classification. Therefore, , to build this distinction into the design principles of the tagset would be superfluous.

Given that the information discussed above will be included in the Urdu tagset, the next question should be: "What should *not* be included?"

- No derivational information will be included in the tagset.

It is conceivable that such information might be relevant to morphosyntax, on the grounds that the derivation of a word may affect its inflection (i.e. by determining what paradigm it belongs to). However, to do so would take the tagset considerably beyond what is recommended by the EAGLES guidelines. Very few of the tagsets reviewed above include derivational information (the EngCG tagset is an example of one that does: see section 2.1.2.2.5). To maintain comparability with earlier tagsets, no derivational information will be included.

- No etymological information will be included in the tagset.

The same considerations apply here as to derivational information. It could be useful, as the inflection of Arabic and Persian loans can differ from that of native vocabulary. However, it is marginal to morphosyntax *per se*, and would take the tagset beyond the bounds of the EAGLES guidelines.

- No syntactic information will be included in the tagset.

No information on syntactic roles (subject, object and so on) will be included; nor will information on transitivity, the kinds of complements demanded by verbs, etc. Although some tagsets – particularly those linked to parsing projects – have included them, such details would be marginal to morphosyntax and would reduce compatibility with the EAGLES guidelines.

- No semantic information will be included in the tagset.

Some tagsets have included semantic information in their morphosyntactic annotation. The original Brown Corpus tagset includes at least one wholly semantic tag, for example: JJS, semantically superlative adjective, a category distinct from JJ only in terms of its meaning. The C7 tagset also includes some semantic features, for example, the names of places as opposed to other proper nouns. However, semantics is a separate field to morphosyntax, and although it can be marked on a corpus text it is separate from part-of-speech tagging. Therefore no semantic information will be included in the tagset.

The inclusion of a proper/common distinction for nouns could be seen as an exception to this rule. In Urdu, common and proper nouns are much more alike than in English or other Western European language, since there is no uppercase to mark a proper noun with, and there are no articles to mark common nouns. Thus the distinction is much closer to a semantic distinction than in English, where "a man" is clearly differentiated from "Mr Smith". However it will be retained for the Urdu tagset, to maintain compatibility with the EAGLES guidelines – the common/proper distinction will be easily enough to suppress if it proves undesirable for a particular purpose.

- No discourse information will be included in the tagset.

Information on discourse would be even more marginal to morphosyntax than semantic information.

### 2.2.3 Hierarchy and decomposability

In the discussion of previous tagsets above, I have shown that tagsets have become increasingly *hierarchical* and *decomposable* over the years. Indeed, these seem intuitively to be useful features for a tagset. For the human user[30], it is easier to memorise a small number of decomposable elements than a large number of tags. Also, as Leech (1997b: 26) points out, the use of decomposable tags allows searches

[30] Even though the overall aim of this thesis is to create an automated POS tagger, the tagset will need to be used extensively by humans, e.g. in the process of creating training data or in the post-editing stage of mark-up – and of course the end output must ultimately be interpreted by a human analyst.

of a text or corpus to be carried out with an underspecified search string (e.g. N* for all nouns, NN* for all common nouns, and so on, where * is a wildcard character). Interpretation of the tag is easier, furthermore, if the decomposable elements are arranged hierarchically[31]. Cloeren (1999: 39-40) suggests that of the three types of information in a tagset, major word classes should be highest in the hierarchy, followed by subclassifications, and lastly morphological features. This is indeed a common approach in hierarchical tagsets – for example, Khoja et al. (2001) and Chae and Choi (2000) use exactly such an approach. Therefore, I will adhere to this practice: the Urdu tagset will be fully decomposable and hierarchical.

Support for decomposability as a desirable feature of a tagset is not universal. In his description of the SUSANNE tagset, Sampson (1995: 79-82) argues that "unitary [i.e. non-decomposable] wordtags are preferable to sets of features" for two reasons: the use of feature bundles focuses attention upon grammatical features (such as number) which are found for many word classes, and also equates "similar features that occur in different word-types". For example it suggests that plurality is the same phenomenon in any part of speech, whereas the comparability of pronoun plurality and verbal plurality is not certain.

In light of this objection, it should be emphasised that the decomposable elements of the tagset will indicate features in a hierarchy, not a matrix. This distinction is crucial in that it ensures that the significance of any feature is context-dependent in the tag. For example, if 1 = singular number and 2 = plural number, we might propose the following very simple tags for singular and plural nouns and verbs: N1, N2, V1, V2. Since the number feature is lower in the hierarchy than the noun-verb distinction, the 1~2 distinction need not be the same linguistic phenomenon in

[31] What I describe as a "hierarchy" is described by Leech (1997b) as a "logical tagset".

the N* context as it is in the V* context (whereas the N~V distinction is the same regardless of the number of the items being distinguished). In a feature matrix, by contrast, we might expect all features to be independent. Employing this principle reconciles Sampson's second objection[32] to feature-bundle tags with the desirability of a decomposable, hierarchical tagset.

### 2.2.4 Theoretical Neutrality

As Leech points out, no annotation scheme can claim to be fully definitive or to provide an absolute gold standard of annotation. For this reason,

> to avoid misunderstandings and misapplication, it is a good idea for annotation schemes to be based as far as possible on consensual or theory-neutral analyses of the data … While annotators are bound to face some theoretically sensitive decisions, their goal should be to adopt annotations which are as widely accepted and understood as can be managed.
>
> (Leech 1997a: 6-7) [33]

This is desirable to ensure that both the tagset itself and text tagged with it are reusable to as wide a group of end users as possible, not limiting their utility to "those who have adopted a particular theoretical framework" (Leech and Wilson 1999: 55-58).

---

[32] Sampson's first objection – that features applicable to only one or two classes of words are more likely to be ignored if attention is focussed on features that can be marked in every word class – is also less of a problem with a hierarchical tagset, since it is quite possible for a feature to appear only on one sub-branch of the hierarchy, should that be deemed desirable. Indeed, I have done something of this sort with regard to Urdu's symmetrical y-v-k-j word sets (see sections 3.4.2, 3.6.2).

[33] A similar point is made by Véronis and Khouri (1995).

Adherence to the EAGLES standard should go some way towards ensuring that the tagset is theory-neutral. However, theoretical neutrality will also be a design principle of the tagset in its own right[34].

### 2.2.5 Granularity of the tagset

The size of the tagset and its granularity are inextricably linked: the more fine-grained the analysis, the greater the number of tags must be. As Véronis and Khouri (1995) point out, the size of the tagset can affect the performance of taggers – particularly probabilistic taggers which require training data. While it was originally assumed that the lower the granularity of the tagset, the greater the accuracy of the tagger, it has been shown by researchers on the CRATER project that a tagset of higher granularity may actually give better results (McEnery et al. 1997)[35]. However, since the tagset here is defined prior to any work on the actual tagger, it cannot be known how fine-grained a tagset will prove optimal for tagging purposes. Similarly, Leech (1997b: 24-25) points out that linguistically desirable distinctions in a tagset may not be feasible to implement in an automatic tagging system – again, it cannot be known in advance which distinctions will prove unfeasible.

Therefore, the first step must be to make a linguistically ideal tagset – the tagset which we would like to apply to our texts in a perfect world, even though in practice we may not be able to. This ideal tagset will be the largest conceivable within

[34] The principle of theoretical neutrality is particularly applicable to the issue of Urdu's status as an ergative language (see section 1.1.5.4).

[35] Smith (1997: 140-141) describes how this phenomenon was exploited with regard to the annotation of the BNC.

the parameters laid out by these design principles, on the basis that it is always easier to remove distinctions than add them should revisions be needed later. The key design principle here is thus that of maximum granularity.

On the basis of that ideal tagset, more restricted tagsets (henceforth referred to as "subtagsets") for actual use in tagging texts may be defined subsequently. Note that it would not be theoretically impossible to tag using the full tagset[36]; it is merely anticipated that some subtagset may be preferable for at least some purposes. These subtagsets should be created by neutralising distinctions made in the main tagset and thus merging entire categories. This is the approach used by Véronis and Khouri (1995) for defining the relationship between the MULTEXT lexical specifications and corpus tagsets. Leech points out (1997b: 25) that "there normally has to be a trade-off between what is linguistically most desirable and computationally feasible". The tagset-subtagset approach means that this trade-off can be postponed as long as possible, and thus constitute a less basic factor of the tagset.

The relationship between the tagset and its subtagsets is not the same as the relationship between the EAGLES guidelines and the tagset. The EAGLES guidelines inform and guide the development of the tagset by presenting a large range of possible annotations that might fit any of a wide range of languages, which the tagset restricts and adapts to one particular language (in this case Urdu). By contrast, the relationship between the tagset and subtagsets is very much more rigid: the subtagsets must consist of a subset of the main tagset. However, because the tagset will be created in accordance with the EAGLES guidelines, this means that any subtagset defined from it will also be EAGLES-compliant, and would be so even considered in isolation from

[36] Contrast the EAGLES intermediate tagset: it would *not* be practical to tag text directly using EAGLES intermediate tags. This is not what they were intended for.

the tagset.

The principle of maximum granularity has implications for whether the tagset will tag words by function or by form. An example of the form/function distinction can be seen in the way that tagsets for English deal with the base form of the verb (i.e. the uninflected form, such as *do*, *walk*, *die*, etc.). This form can represent the non-third-person-singular present indicative, the present subjunctive, the infinitive, or the imperative. Some tagsets give most or all of these uses the same tag (e.g. the C7 tagset only distinguishes VB0 (base form) and VBI (infinitive)) – this can be described as tagging by form. Other tagsets, e.g. the EngCG tagset, provide some means for distinguishing different uses of the same form – the EngCG tagset can distinguish indicative, subjunctive and imperative, for example. This can be described as tagging by function[37]. The high granularity of the tagset will mean that the tags mark function rather than form, although subtagsets which collapse a number of distinctions may come much closer to tagging by form.

### 2.2.6 Dealing with tokenisation problems and word-token mismatch

#### *2.2.6.1 The difficulties of tokenisation in Urdu*

Before considering word-token mismatch, it is necessary to digress somewhat and discuss how tokenisation is to be carried out, in order that it is clear what exactly the tokens are that the tagger, computer or human, will be faced with. Dividing a text

[37] This form/function dichotomy has been known since the earliest days of tagging: Greene and Rubin (1971) point out that their approach to ambiguity (see 2.2.7) means that in some cases, form is tagged in preference to function.

into tokens is in itself not a trivial task in Urdu. Although the text makes clear word breaks by means of spaces or the use of final forms of letters, not all the orthographic word breaks are necessarily "actual" word breaks. This becomes clear when one considers that many writers, when making semi-phonetic transcriptions of Urdu, omit some word breaks.

For example, the transcriptions of future-tense verbs given by Bhatia and Koul (2000: 97-98) give these verbs as single words, and go far as to state that "it [the future tense] is only one unit in Urdu". However, in their Perso-Arabic examples, the future marker *gā / gē / gī* is invariably written as a separate word (and in my own transcriptions it is written as such). In all such cases, tokenisation in this thesis will be done on the assumption that such word breaks in the orthography represent genuine token breaks in Urdu. This is for three reasons. Firstly, it makes the tokenisation process much easier to automate. Secondly, it simplifies the system of categories in the tagset considerably, since the set of forms that are considered inflectional affixes is reduced (see also section 3.2.2.1 for more on the future auxiliary *gā / gē / gī*). Finally, it is hoped that by sticking closely to the orthography, the tokenisation may produce a result that is close to how native speakers perceive their language.

There is another, more problematic type of "unreal" orthographic word breaks. For example, the word that Schmidt (1999: 249) transcribes as "*zimmēdār*", which means "responsible", clearly contains a word break in its Perso-Arabic spelling (before the *dār*[38]). However the same derivational morpheme *dār*, which forms adjectives from nouns, is written without a word break in other words such as *samajhdār*, "sensible". In line with the principle laid out above, the word break in *zimmah dār* will be considered genuine: *zimmah dār* is then two tokens, and

[38] For this reason, I transcribe the same word as *zimmah dār*.

*samajhdār* is one. This creates something of a problem for tagging, discussed in section 3.12.2.

#### *2.2.6.2 Word-token mismatch*

A problem is poised by the issue of multi-unit tokens and multi-token units (see Cloeren 1999: 44-46), i.e. contractions and idioms[39]. A design principle of the tagset will be that it should provide some means of dealing with these.

##### *2.2.6.2.1 Contractions*

Contractions (two or more words realised as a single token, such as French "au" or English "won't"; described as "mergers" by Leech 1997b) would ideally be given two tags. However, Cloeren suggests two methods of doing this: to give both tags to the entire token (i.e. *won't = [modal verb + negative marker]*), or to split the word up and give each part its own tag (i.e. *wo = [modal verb] n't = [negative marker]*).

The result of the former approach is that "further processing becomes more complicated" (Cloeren 1999: 44). The latter approach would ultimately result in attempting to split such forms as French *au* into units taggable as preposition and article. This would be linguistically very dubious, since this digraph represents a single vowel. Nevertheless, it is the latter approach which will be followed, since it is of paramount importance to keep the tagging system as simple as possible, to allow

[39] I use the word "idiom" to refer to a lexical item that has the form of a syntactic phrase rather than a word. This concept is explained by Katamba (1993: 296-299).

for the widest possible use. It also allows words to be tagged in the same way regardless of whether or not they are clitics / have clitics hanging onto them. It is anticipated that forms as thoroughly fused together as *au* will prove rare[40].

Thus it is possible to state as a design principle that every token in the text shall receive exactly one tag. Clitics will be tagged separately from the word to which they are attached[41]; this does not go against the stated principle of preserving all orthographic word breaks since it involves adding additional word breaks rather than suppressing existing ones.

#### *2.2.6.2.2 Idioms*

For idioms such as French "pomme de terre" or English "know how" (in the sense of "expertise"; examples from Cleoren 1999: 44) the problem is less acute than with contractions, since it is always possible to tag the idiom as if it were a phrase. Cloeren suggests a more sophisticated strategy: giving an interior analysis (e.g. for "pomme de terre": [noun] [preposition] [noun] ) plus an exterior analysis ( [noun31]

[40] That is, rare in the written form of language, which is the form we are concerned to tag (spoken texts being tagged in orthographically transcribed – i.e. written – form). One might anticipate that completely fused words would be more common in the spoken form. However, special provision can be made for such awkward cases as French *au*: for instance the EAGLES guidelines allow a feature "Fused preposition/article" to be annotated (Leech and Wilson 1999: 67-68).

[41] Note that this has implications for tokenisation: either the tokeniser must be able to separate clitics from the words to which they attach, or else the tagger (human or automatic) will need to have some means of correcting the tokenisation.

[noun32] [noun33] )[42]. However, to do this would mean giving more than one tag to the tokens within the multiword, in contravention of the principle stated above. To give just the exterior analysis would give less than one tag to each word (even though the tag may be encoded onto each of the words, it is still a single tag applied to more than one word). This is equally a breach of the stated principle.

Thus, no multiword or idiom tags will be defined in the ideal tagset (although it may later prove desirable to define some for the subtagset used for actual tagging practice). This is for three reasons. Firstly, their use would introduce needless theoretical controversy into the tagset. There will probably never be complete agreement on where to draw the line between idioms and non-idiomatic frequently-occurring phrases. Secondly, even a native speaker of Urdu would not be able to sit down and write a list of all the language's multiword units: such units would, more likely, be encountered in the process of work on tagged texts – a stage which cannot precede the definition of a tagset! Finally, idiom tagging ought not to require any new tags – merely some means of extending the same tags over more than one token – and thus should not affect the initial composition of the tagset. However, care will be taken that nothing is included in the tagset which prohibits the introduction of some form of idiom tagging in a subtagset at a later stage.

### 2.2.7 Dealing with ambiguity

One form of ambiguity occurs when one orthographic form realises several grammatical forms and functions. Leech and Smith (1999: 26) give a prime example

[42] Although as Cloeren points out, such a strategy runs into difficulties dealing with discontinuous idioms such as "*provided* after all *that*".

of ambiguity with regard to the English word "cut". This can be a noun, an adjective, or any one of six different forms of a verb: context usually makes it clear to the human which it is. For Leech and Smith, "Where wordclass tagging is a preliminary to other levels of annotation, its primary use is to resolve the homograph ambiguities". Another type of ambiguity arises when a category has an unclear boundary, so even a human being is hard pressed to decide on an appropriate tag.

As noted above, most tagsets provide some means of dealing with homograph ambiguity, for example the C5 tagset's portmanteau tags (e.g. NN2-VVZ for "plural noun or third person singular present indicative verb"), or the CI tag (conjunction or preposition) in the Brown tagset. Both these approaches avoid making a commitment in cases of homograph ambiguity, but do mark off the ambiguous cases from the non-ambiguous. However, such solutions contravene, implicitly or explicitly, the principle that it should be possible to give every token exactly one tag. They would also contravene the principle of tagging by function over form. The very definition of a homograph is that it has one form but more than one function: therefore in line with the "tag for function" principle, every separate function of a homograph should be tagged differently. Portmanteau tags result in the different functions of the form being tagged alike.

Therefore, no such tags will be included in the tagset. If they should prove necessary in the process of automatically tagging texts, they may be included in future subtagsets but they have no place in the somewhat abstract ideal tagset. This fits with tagset-subtagset approach being taken in line with the two-level model described by Véronis and Khouri (1995) (see section 2.1.4.1 above).

The other problem, of categories with unclear boundaries, can be ameliorated with extensive and clear tagging guidelines to ensure that all human users are

operating with the same category definitions. Thus, the tags will always be used alike. Creating clear boundaries between the categories may require arbitrary decisions, which is regrettable, but these decisions will be consistent.

### 2.2.8 Summary

The following is a summary list of the design features outlined above:

- The tagset will adhere to the EAGLES guidelines (Leech and Wilson 1999);
- The tagset will include information on major word classes, subclassifications, and morphology;
- The tagset will not include any derivational, etymological, syntactic, semantic or discourse information;
- The tagset will be fully decomposable and hierarchical;
- The tagset will be theoretically neutral;
- The tagset will be as fine-grained as possible (thus allowing for the greatest possible freedom in constructing subtagsets for actual application);
- The tagset will tag by function rather than by form;
- Every word token (with tokenisation determined principally by orthography) will receive exactly one tag, with clitics tagged separately from the word they are attached to;
- The tagset will contain no idiom tags;
- The tagset will contain no portmanteau tags (or other tags whose sole purpose is to deal with ambiguity).

In the preceding detailed discussion of these points, I have provided sufficient justification for my claim, that the above features represent an optimal set of design features for an Urdu tagset, to be taken as substantiated.

### 2.2.9 The superficial features of tagset design

#### *2.2.9.1 Principles of the tagset's appearance*

The essence of a tagset composed according to the design features discussed above is a set of categories, into which any token in the language or variety to be tagged is theoretically classifiable, though in practice there will probably be numerous tokens that are ambiguous between categories even to a skilled human analyst. Independent to a degree from the categories is the form by which they are encoded – that is, the actual character strings or tags that are inserted next to the tokens in the tagged text[43]. The form of these tags is the topic of this section.

The strings *could* be entirely arbitrary – e.g. the category "singular common noun" could theoretically be noted by the string "@RE$8%" – but in reality it is preferable for the shape of the tag to reflect its meaning. As Cloeren (1999: 49) points out: "For reasons of readability there is a preference for *mnemonic* tags… Full-length names may be clearer individually, but make the annotated text virtually unreadable." For this reason, almost all tagsets[44] have tags that are effectively abbreviations of the

[43] This distinction is made by Leech (1997b), although he uses the term "tag" to refer to the category and the term "label" to refer to the string encoding the category.

[44] But not all: as Cloeren (1999) points out, numerical codes can also be used, as for example in the EAGLES intermediate tagset. However, the use of such non-mnemonic codes is more suited to

linguistic terms that describe their category (e.g. NN1 in the C7 tagset), or that reflect some phonetic or orthographic feature of the category annotated (e.g. VVZ for the verbal form in *–s*, VVD for the form in *–ed*). This is a practice that I shall follow. Furthermore, those tagsets that are hierarchically structured (e.g. parts of C7, the ICE tagset) reflect this in their visual realisation; I shall do likewise.

There are a number of other features that have become commonplace in the presentation of tagsets (although none is universal). For example, in the Brown tagset/CLAWS tagset tradition, all tags consist of a single string unbroken by white space or punctuation characters. This is by no means a necessary feature of a tagset – and many more recent tagsets do not follow it, especially those that use explicit feature hierarchies (e.g. the EngCG tagset, discussed above in section 2.1.2.2.5). Cloeren (1999:50) suggests that hierarchical tagsets should use delimiters to mark off the different decomposable elements of a tag from one another. However, this makes the tag less concise, which is undesirable (Leech 1997b: 25). Delimiters will also increase the ultimate size of the tagged text file considerably – where sixteen-bit Unicode files are used, as will be the case for the Urdu texts that will ultimately be dealt with, this is a non-trivial consideration. The use of delimiters can however be avoided – and the length of tags minimised – if every decomposable element is precisely one character long, and every character in every tag represents a decomposable element. The boundary between characters then serves as a kind of zero-delimiter.

There is also a notable tendency for tags to consist of uppercase characters

---

automatic mapping (the *raison d'être* of the EAGLES intermediate tagset) than to use by human analysts.

only[45], or of uppercase characters followed by lowercase characters (the assorted CLAWS tagsets, the Penn tagset, and the EngCG tagsets exemplify the former, the MULTEXT tagset the latter). As Calzolari and Monachini (1996) point out, this is useful to preserve the distinction between the tags and the actual words of the text. For this reason, and also because this superficial feature is well-established, I will comply with this. Thus, in summary, the forms of the tags in the Urdu tagset will obey the following rules:

- All tags (and elements of tags) will have, so far as possible, mnemonic value;
- All tags will consist of a single unbroken string of characters;
- Only uppercase letters and the numeric symbols 0 to 9 will be used[46];
- Sequences of numerals will not be used (to improve readability);
- Each character in a tag string will represent exactly one decomposable element, where a decomposable element shows what value the tag indicates for a given feature[47];

[45] One suspects that the original motivation for the use of uppercase letters was the inability of early computers to produce lowercase characters: e.g. the tags used by Klein and Simmons (1963) for "noun" and "verb" are NOUN and VERB respectively.

[46] There is a single exception to this: where a punctuation mark is tagged as itself. In this case, the tag will be no more than a single character in length.

[47] There is also an exception here: where a single-member group of tags (for example, the definite article tag, or the tag for a question marker) contains a single decomposable element, not found elsewhere in the tagset. Sometimes such cases are better represented as a two-character string which makes up the entirety of the tag, in the interests of mnemonic ease (and of not running out of available letters). So: AL (definite article), QQ (question marker). Tags of less than two characters have not been used (except for punctuation tags) as they would be hard to spot by eye in the midst of the text.

- The sequence of characters from left to right[48] will represent a hierarchy of features ordered from the most general to the most specific (where there is no clear difference in specificity between two features, the choice will be made arbitrarily).

By following these principles, it is my hope that the superficial appearance of the tagset will reflect its hierarchical and decomposable nature while maintaining as far as possible the reader-friendliness of mnemonic tags.

#### *2.2.9.2 The Perso-Arabic tagset*

When a Western model of language, such as that used below (see 2.3), is applied to a language from outside Western Europe, it is very easy to fall into modes of analysis that imply the languages of Europe to be the norm and everything else a deviation from it, or else capable of being shoehorned into a European analysis. An example of such a mode of analysis would be a tagset for Urdu based on the Roman alphabet and the English language (e.g. N for noun, V for verb). The symbols of the tags would be alien – although probably not unfamiliar – to the Urdu-speaking reader, as would the terminology.

For the tags and the tokens they are to classify to be drawn not only from different languages but also from different alphabets seems on the face of it a

[48] By "from left to right" I mean "from the end of the string nearest the start of the disk file to the end of the string nearest the termination of the disk file". In the Perso-Arabic form of the tagset, which – like the Perso-Arabic script in general – is written right-to-left, the most general element is thus displayed on the right.

somewhat perverse approach. However, it is undeniable that the vast majority of literature in linguistics is published in English, even when English is neither the object of study nor the first language of the author. Thus it would be inappropriate to make sole use of Urdu-based abbreviations and codes, since it would hinder comparability with other tagsets and put unnecessary difficulties in the way of many readers. As noted in Chapter 1, very few linguists in Europe and North America have even a passing knowledge of Urdu. Fortunately, since the underlying categories can be linked to any set of character strings, it is possible to have two tagsets. The first will use Roman characters and English abbreviations[49], the second will use Perso-Arabic characters and abbreviations of Urdu words. Mapping between these two tagsets can be made entirely automatic. The same surface design principles will be applied to the Perso-Arabic tagset as to the English tagset.

The Perso-Arabic tagset is given parallel to the Roman tagset in the discussion of the grammatical categories in the next chapter, and a brief explanation of their composition is given in Appendix 3.

#### *2.2.9.3 Other potential encodings*

Many tagsets indicate their hierarchical nature more overtly than the design principles above allow for. This does not only include XML or SGML attribute-value encodings: for example, Véronis and Khouri (1995) give N[type=common gender=masculine number=singular] as an expanded form of the MULTEXT tag

[49] The Roman form of the tagset will be referred to exclusively in my commentary.

Ncms-[50]. Cloeren (1999: 50) suggests yet another encoding, for example PRN(pos,pl,1) for a first person plural possessive pronoun.

There is no reason why the categories in the Urdu tagset created below could not be represented in this manner, although (for reasons given above) it is not the primary method of my choice. Thus, for instance, the tag NNMM1N could easily be written as N[type=common marked=yes gender=masculine number=singular case=nominative], or as N(com,mkd,masc,sng,nom), or as any other decomposable notation one might conceive of. Automated mapping between two tagsets which encode exactly the same distinctions is a computationally trivial task. All that would be necessary would be to run a search-and-replace program based on completely unique strings.

## 2.3 The choice of a model of the grammar of Urdu

To create the categories of the tagset, it is necessary to have a model of the language to categorise. An ideal approach would be to derive this model from empirical data – however, this cannot be done prior to the creation of a tagset. A native speaker of a language could use their own intuitions about the language as a model, but as I am not a native speaker of Urdu, this is not an option. It would be theoretically possible to use another person who is a native speaker as the model. However, to extract the large amount of grammatical information needed to define the tagset would be a long, laborious and error-prone process, and not practicable in the

[50] The character "-" does not have an analogue in the expanded form, since the attribute that it relates to has the value *n/a*, which in the expanded form is indicated by its absence.

scope of this thesis[51].

The only remaining option is to make use of a published description of Urdu grammar as a model of the language. This may, in fact, be preferable, because it means that the terminology used will be compatible with previous work on Urdu, which will make the tagging system more immediately accessible to linguists who have worked with Urdu in the past. This will in turn allow them to make use of tagged texts and corpora in improving their grammars, meaning that the next generation of published descriptions may be informed by the type of large-scale empirical analysis which is now, as I have explained above, impracticable.

The selection of a published description to serve as a model is also no simple matter. The first choice that presents itself is between a grammar rooted in the European tradition of linguistics and a grammar based on the Paninian tradition (see section 2.1.5.4 above). However, this turns out to be a non-choice. The Paninian tradition, rooted as it is in Sanskrit, was not the major influence on grammatical research in Hindi and Urdu (Bhatia 1987: 11). Rather, the European tradition was a much greater influence: as Bhatia (1987: 15) concludes, "in no serious sense is the grammatical tradition of the Hindi language a representation of the tradition handed over to it by Sanskrit grammar".

It is therefore unsurprising that, so far as I am able to ascertain, no full Paninian descriptive grammar of Urdu has been published in Europe. Bhatia (1987) reports that such a grammar was written for Hindi in 1855 by Pandit Shrīlāl. But the unavailability of this and similar grammars to the European academic make using them as a model problematic. One further reason to avoid an analysis rooted in Panini

[51] While native speaker input cannot be the sole source of the model of the language, it can be of use in clarifying points of uncertainty in the model actually employed (see below and also Chapter 4).

is that without first acquiring a knowledge of Sanskrit, it might well be difficult to comprehend Paninian descriptions; this would impair the creation of the tagset.

Thus, we are left with works on Urdu in the European grammatical tradition. The field of descriptive grammars of Urdu in English is quite narrow[52], although rather more has been written on the subject of Hindi. As a result there is no accepted standard grammar of Urdu. In any case, most of the work in the field falls short of what one would expect from a standard grammar in at least one of a number of ways which will now be discussed.

A considerable amount of descriptive work on Urdu was done in the days of British rule in India, for example Platts (1884), Kellogg (1875). However, although detailed, such work is now out of date in regard to the language itself. For example, even Kellogg's usage of the terms "Hindí" and "Urdú" is not fully consistent with modern usage[53], leading him to report greater differences between the two than exists between the modern standard forms. His description of pronouns (Kellogg 1875: 168-174) is another example of a feature which conflicts with later descriptions. To base the tagset on an outdated description of the language would unnecessarily complicate

[52] Even more scarce in the UK are good, recent Urdu-English dictionaries. Those used in this thesis are Oriental Book Society (date unknown) and Haq (2001). Both have flaws. For example, the first is rather narrow in its lexical coverage, and the latter suffers from blurred printing that renders some pages illegible.

[53] For Kellogg, "Urdú" is a Persianised form of Hindí, where "Hindí" refers to a wide collection of dialects that Masica (1991: 9) describes as the "regional languages of the Hindi area". Kellogg specifically rejects a definition of "Hindí" which refers only to the Sanskritised standard language of that area – which is the most usual meaning of the word "Hindi" today (see section 1.1.4). One reason for this, as Masica points out, is that in Kellogg's day this standard form was less established as a spoken language.

usage by those versed in modern linguistics. I have thus eliminated anything written before 1950 as a model of Urdu for the tagset.

The remainder of the work on Urdu tends to fall into two camps. The first group are works in theoretical linguistics or typology or language surveys that touch on Urdu, or, more typically, "Hindi-Urdu", suggesting that many writers are not concerned with the differences between the two (see section 1.1.4). The second are works with a pedagogical bias, i.e. their purpose is to facilitate the teaching of Urdu as a foreign language. They are thus aimed squarely at the non-linguist.

It would therefore seem that the first group are probably the better choice, as they provide sufficient linguistic detail for devising the tagset. However, in actual fact this is not the case. Works in theoretical linguistics which concentrate on Hindi-Urdu tend to focus on one aspect of the language to the exclusion of the rest. Thus they do not provide a complete model of the language.

A typical example of this tendency is Butt (1995). Butt's goal is to analyse the "complex predicates" of Urdu, i.e. verb phrases consisting of more than one verb, within the framework of lexical functional grammar. She also touches briefly on the issue of ergativity in Urdu (see 1.1.5.4). She does not however cover the grammar as a whole – for example, she makes little mention of any of Urdu's inflections, it simply being irrelevant to her topic. Most of the papers listed in Masica's comprehensive bibliography (1991: 493-497, 510) are also of this kind.

Similarly, language surveys, while they cover the whole language, are not sufficiently detailed to constitute adequate models. For example, Masica (1991) surveys all the languages and many dialects of Indo-Aryan. This perspective leads to an emphasis on the similarities and differences between related languages, rather than on how any one of those languages is structured. Kachru's (1990) survey is only of

Hindi-Urdu, but is too brief to provide a full model (for example, the entire inflectional morphology of nouns, adjectives and verbs are covered in just six pages; the entire phonology in three). A problem with using any language survey as a model is that the differences between Hindi and Urdu tend to be downplayed, as their similarities tend to become more significant. This tendency to refer to Hindi-Urdu as a single language is evident in most of Masica's and Kachru's summaries.

This leaves pedagogical works, "teach yourself" books and so on. These include[54] Bhatia and Koul (2000), Barz (1977), and Bailey et al. (1956). These have a number of flaws in common. Firstly, since they are all aimed at beginners, they cover what it is anticipated a learner would need to know first, and are thus too partial to be adequate models[55]. For example, Bhatia and Koul (2000) include no discussion at all of relative pronouns/relative clauses. Furthermore, in their discussion of noun gender, they provide "rules of thumb" for guessing the gender of a noun that fall far short of a complete description of various nominal endings and the genders associated with them (compare, for example, Bhatia and Koul 2000: 313-314 to Schmidt 1999: 1-5).

Secondly, because they are aimed at laymen, pedagogical works do not use linguistic terminology fully. It can thus be hard to extract hard linguistic facts from them. For example, Bailey et al. declare (1956: 18) that "Adjectives are often used as adverbs; when so used they agree with their nouns or pronouns like adjectives." For a linguist this begs several questions. How is it decided which noun or pronoun an

[54] Again, I discuss here only books which provide instruction on Urdu, Hindi-Urdu, Hindi *and* Urdu or Hindustani. There are others which restrict themselves to Hindi (e.g. McGregor 1972) and thus are immediately excluded from consideration as a model for Urdu.

[55] This should not be interpreted as a criticism of these works. I am well aware that to use a pedagogical manual as a model for tagset definition would be putting it to a purpose it was never intended to fulfil. If pedagogical manuals are not up to the task, this does not reflect badly on them!

adverb is to agree with? If an adjective used as an adverb has adjectival agreement, on what basis is it deemed that they are adverbs and not adjectives in the first place? However Bailey et al. do not elaborate.

The closest work to a complete reference grammar is Schmidt (1999), but this too has a fairly pedagogical approach, as the author makes clear in her introduction (1999: xvi). Some features are absent which we would expect in a reference grammar. For example, there is no morph-by-morph (or even word-by-word) gloss of the examples, only full-phrase translations/equivalents; phonology is excluded; the paradigms of irregular forms are not exhaustive; and discussion is given to matters such as formal and pious idioms, which would probably not be found in a standard reference grammar. However, Schmidt's grammar is sufficiently descriptive (rather than pedagogical) in nature, and contains few enough holes, that it is appropriate for use as a model. It is also very recent, which would allow the tagset to be based on the most up-to-date study of Urdu available.

A final option might be to attempt a synthesis of the various sources on Urdu grammar outlined so far. However, this is a highly problematic approach. Firstly, to attempt to synthesise such widely varying materials would be a major undertaking in itself, let alone to do so (as would be the case here) as the first step in a project directed at the ultimate goal of automated part-of-speech tagging. As such it is a task beyond the scope of this thesis. Secondly, even if a synthesis were practical, there are great difficulties for the non-speaker to resolve. If one author reports a phenomenon which another author does not, we might assume that the second author has omitted to mention that detail, and include it in the synthesis. However, an assumption which is *prima facie* equally valid is that the difference might arise from a difference in the version of the language being described (in terms of time or in terms of dialect). If

that is the case, should the detail be included in the model or not? This is a far from hypothetical scenario. For instance, Platts (1884) describes distinct plural forms for the pronouns *yah* and *vah* that are not mentioned by Schmidt (1999), who reports that these pronouns are both singular and plural. In the absence of other evidence, one can only speculate on the reason for this. Possibly the plural forms have fallen out of use; possibly they were only ever in restricted use, and Platts is being more exhaustive in his description. Be that as it may, it is almost impossible to imagine how one could synthesise these two reports; this question can only be decided by reference to a native speaker or a corpus, which lies outside the scope of a literature synthesis.

For these reasons, Schmidt (1999) alone is used as the model of Urdu grammar for the definition of the tagset. This has necessitated taking Schmidt at her word and assuming that the model of Urdu she presents is identical to actual Urdu. This is almost certainly not the case – a model is by definition not identical to the thing it models. But it is to be hoped that other than the necessary lesser degree of detail that is characteristic of a model, the discrepancies between the model and the reality are minor.

At some points Schmidt's account of the grammar of Urdu proved inadequate. Such examples are highlighted in the definition of the tagset in the following chapter. They include instances where Schmidt is silent or vague on some point of importance for the creation of the tagset's categories. There are two strategies for dealing with such points of uncertainty. The first is to refer to older grammars, to pedagogical manuals, or to any linguistic studies which happen to deal with the relevant part of the grammar. This has been done where necessary, *ad hoc*, as it were, and references to such works will be found throughout the following chapter. When this is done, priority is always given to authors who deal exclusively with Urdu, although writers

on Hindi or Hindi-Urdu have also been consulted. Of course, first priority has been given to Schmidt's description at all points, to avoid falling into the difficulties associated with a literature synthesis as described above.

The other potential strategy is to take recourse to a native speaker's intuitions. This strategy have also been employed, with regard to points that are mentioned in the following chapter.

It is necessary for current purposes to assume that Schmidt's model is both accurate and suitable for the purpose it is here applied to. However, this assumption will not be left permanently untested. The test of the suitability of Schmidt's grammar as a model for tagging will be how well a tagset defined on the basis of it performs when applied to natural samples of Urdu text. This process is described in Chapter 4. It should be understood that the definition of the tagset is open to revision if flaws are made evident through the process described above, or when the tagset or any subtagset derived from it are applied to text.

## 2.4 Concluding remarks

In this chapter, I have fulfilled three aims. Firstly, I have given a summary of the history and "state of the art" of tagset design, in particular discussing the EAGLES guidelines (2.1.3) as a major recent multilingual standard for part-of-speech tagsets. For reasons explained in 2.2.1, this standard is used as the basis for the design of the Urdu tagset. I have justified my claim to have devised a set of design principles (2.2) which will, in addition to the EAGLES guidelines, guide the creation of the tagset. Having constructed a framework for tagset design, I justified the choice of Schmidt (1999) as a model of the language for creating the system of morphosyntactic

categories underlying the tagset. With these necessary preliminaries in place, it is now possible to define a tagset for Urdu texts and corpora – itself a necessary preliminary to the ultimate aim of achieving accurate automated part-of-speech tagging in Urdu. The definition of this tagset is the subject of the next chapter.