

The computational analysis of morphosyntactic categories in Urdu

Andrew Hardie

August 2003

(revised version February 2004)

Thesis submitted for the degree of PhD

Department of Linguistics and Modern English Language

Lancaster University

Abstract

Urdu is a language of the Indo-Aryan family, widely spoken in India and Pakistan, and an important minority language in Europe, North America, and elsewhere. This thesis describes the development of a computer-based system for part-of-speech tagging of Urdu texts, consisting of a tagset, a set of tagging guidelines for manual tagging or post-editing, and the tagger itself.

The tagset is defined in accordance with a set of design principles, derived from a survey of good practice in the field of tagset design, including compliance with the EAGLES guidelines on morphosyntactic annotation. These are shown to be extensible to languages, such as Urdu, that are closely related to those languages for which the guidelines were originally devised. The description of Urdu grammar given by Schmidt (1999) is used as a model of the language for the purpose of tagset design.

Manual tagging is undertaken using this tagset, by which process a set of tagging guidelines are created, and a set of manually tagged texts to serve as training data is obtained.

A rule-based methodology is used here to perform tagging in Urdu. The justification for this choice is discussed. A suite of programs which function together within the *Unitag* architecture are described. This system (as well as a tokeniser) includes an analyser (*Urdutag*) based on lexical look-up and word-form analysis, and a disambiguator (*Unirule*) which removes contextually inappropriate tags using a set of 274 rules. While the system's final performance is not particularly impressive, this is largely due to a paucity of training data leading to a small lexicon, rather than any substantial flaw in the system.

Contents

Introduction	13
Chapter 1 – Background issues to the tagging of Urdu	19
1.1 The Urdu language	20
1.1.1 Who speaks Urdu and where is it spoken?	20
1.1.2 Genetic affiliation and other associations	22
1.1.3 Historical notes	23
1.1.4 Urdu versus Hindi	24
1.1.5 A brief overview of the structure of Urdu	28
1.1.5.1 Phonology	29
1.1.5.2 Writing system	31
1.1.5.2.1 The Perso-Arabic script	31
1.1.5.2.2 <i>nasx</i> versus <i>nasta'liq</i>	32
1.1.5.2.3 Urdu modifications to the Arabic alphabet	33
1.1.5.3 Morphology	35
1.1.5.4 Syntax	36
1.2 Part-of-speech tagging in corpus linguistics	38
1.3 The EMILLE project	41
1.4 Concluding remarks	43
Chapter 2 – Preliminaries to tagset design	45
2.1 Previous work on part-of-speech tagsets	45
2.1.1 The earliest work on tagset creation (prior to 1980)	46

2.1.2 Subsequent English tagsets (post-1980)	50
2.1.2.1 Tagsets used with the CLAWS tagger	50
2.1.2.2 Other English tagsets	51
2.1.2.2.1 The TOSCA scheme	51
2.1.2.2.2 The ICE tagset	52
2.1.2.2.3 The Penn tagset	52
2.1.2.2.4 The Lund tagset	53
2.1.2.2.5 The EngCG tagset	54
2.1.3 A standard for part-of-speech annotation: The EAGLES guidelines	55
2.1.4 Some recent tagsets based on the EAGLES guidelines	58
2.1.4.1 The MULTTEXT project	58
2.1.4.2 The GRACE project	61
2.1.4.3 The CRATER project	61
2.1.5 Some recent tagsets for languages not covered by the EAGLES guidelines	62
2.1.5.1 Tagset design for Arabic	63
2.1.5.2 Tagset design for Chinese	64
2.1.5.3 Tagset design for Korean	65
2.1.5.4 Tagset design in the Paninian tradition	66
2.2 Design principles for an Urdu tagset	67
2.2.1 Standards	68
2.2.1.1 The general advantages of compliance with standards	68
2.2.1.2 The particular advantages of compliance with standards when working with Urdu	69
2.2.1.3 Adherence to the EAGLES guidelines	70

2.2.2 Information to include	71
2.2.3 Hierarchy and decomposability	74
2.2.4 Theoretical neutrality	76
2.2.5 Granularity of the tagset	77
2.2.6 Dealing with tokenisation problems and word-token mismatch	79
2.2.6.1 The difficulties of tokenisation in Urdu	79
2.2.6.2 Word-token mismatch	81
2.2.6.2.1 Contractions	81
2.2.6.2.2 Idioms	82
2.2.7 Dealing with ambiguity	83
2.2.8 Summary	85
2.2.9 The superficial features of tagset design	86
2.2.9.1 Principles of the tagset's appearance	86
2.2.9.2 The Perso-Arabic tagset	89
2.2.9.3 Other potential encodings	90
2.3 The choice of a model of the grammar of Urdu	91
2.4 Concluding remarks	98
Chapter 3 – Specification of the tagset	100
3.1 Nouns	101
3.1.1 Gender	102
3.1.2 Number	103
3.1.3 Case	103
3.1.4 EAGLES attributes for nouns not used in this tagset	105
3.1.5 The problem of ambiguous suffixes	106

3.1.6 The tags for nouns	109
3.2 Verbs	113
3.2.1 Lexical verbs	119
3.2.1.1 The root	120
3.2.1.2 The infinitive	120
3.2.1.3 The participles	122
3.2.1.4 The subjunctive	125
3.2.1.5 The imperative	127
3.2.2 Auxiliary verbs	127
3.2.2.1 <i>gā</i>	128
3.2.2.2 <i>rahā</i>	129
3.2.2.3 <i>cāhiē</i>	130
3.2.2.4 <i>hōnā</i>	131
3.2.2.5 Modal and vector verbs	137
3.3 Adjectives	143
3.4 Pronouns and determiners	148
3.4.1 First and second person personal pronouns	151
3.4.1.1 The non-existence of third person personal pronouns	151
3.4.1.2 The problematic honorific pronoun <i>āp</i>	152
3.4.1.3 The tagging of first and second person personal pronouns	155
3.4.2 Third person pronouns/demonstratives, interrogative and relative pronouns and determiners	160
3.4.3 Reflexive pronouns	168
3.4.4 Other pronouns and determiners	170
3.5 Articles	171

3.6 Adverbs	172
3.6.1 Lexical adverbs	172
3.6.2 Non-lexical adverbs	174
3.7 Adpositions	176
3.8 Conjunctions	179
3.9 Numerals	180
3.10 Interjections	183
3.11 Punctuation	184
3.12 Unique/unassigned (including particles, clitics and tags)	187
3.12.1 Tags for the unique categories	187
3.12.2 The <i>zimmah dār</i> problem	192
3.13 Residual	195
3.14 The tagset defined as a hierarchy	196
3.15 The extensibility of the EAGLES guidelines	201
3.16 Concluding remarks	202
 Chapter 4 – Manual tagging of Urdu texts	 204
4.1 Why undertake manual tagging?	204
4.2 Modifying the tagset	207
4.2.1 Changes to the tagset on the basis of manual tagging	208
4.2.1.1 Deletion of the tags for marked predicate-only adjectives	208
4.2.1.2 Deletion of the tag for the inclusive emphatic particle	209
4.2.1.3 Addition of tags for forms of fused adverbs plus <i>hī</i>	209
4.2.1.4 Addition of a tag for “verbal postposition”	210
4.2.1.5 Addition of a further punctuation tag	211

4.2.1.6 Tabulated definition of the new tags	212
4.2.1.7 Evaluating Schmidt's model in practical applications	213
4.2.2 Collapsing the tagset	213
4.2.2.1 Proper nouns versus common nouns	213
4.2.2.2 Oblique case versus vocative case nouns	214
4.2.2.3 Predicate-only adjectives versus general adjectives	215
4.2.2.4 Removing distinctions in the subtagsets	215
4.3 Categorisation difficulties in manual tagging	217
4.4 The tagging guidelines	219
4.5 Creating the manually tagged dataset	223
4.6 Concluding remarks	225
 Chapter 5 – A review of part-of-speech tagging technology	 227
5.1 A proposed typology of disambiguation methodologies	228
5.2 Rule-based approaches to disambiguation	232
5.2.1 Early work with rule-based disambiguation approaches	234
5.2.2 Work in the Constraint Grammar framework	237
5.3 Probabilistic approaches to disambiguation	243
5.3.1 Early work with probabilistic approaches	245
5.3.2 Later work on probabilistic taggers using Markov models	247
5.3.2.1 Markov models: an overview	248
5.3.2.2 Selecting an appropriate tag: additive probabilities in CHAINPROBS	251
5.3.2.3 Selecting an appropriate tag: the Viterbi algorithm in VOLSUNGA	254

5.3.2.4 The use of lexical probabilities in Markov models	256
5.3.2.5 Acquiring Markov model parameters	258
5.3.2.6 Some variations within Markov model disambiguation	260
5.3.2.7 The performance of Markov model taggers	263
5.4 Approaches utilising corpus-derived rules	263
5.4.1 The parser-based approach	264
5.4.2 The transformation-based approach	267
5.4.2.1 Tagging by transformation-based error-driven learning	267
5.4.2.2 The advantages of the transformation-based approach	268
5.4.2.3 A summary of Brill's algorithm	269
5.4.2.4 The form of transformations	272
5.4.2.5 Extensions to Brill's basic method	274
5.4.2.5.1 N-best tagging in the transformation-based approach	275
5.4.2.5.2 Unsupervised training of a transformation-based model	275
5.5 General machine learning approaches to disambiguation	278
5.5.1 Overview	279
5.5.2 The application of neural networks in disambiguation	280
5.6 Combining and comparing disambiguation methods	284
5.6.1 The difficulty of comparing different taggers	285
5.6.2 Enabling a meaningful comparison of disambiguation methodologies	288
5.6.3 Combining different tagging methodologies: hybrid taggers	290
5.7 Selecting an approach for disambiguation in Urdu texts	293
5.7.1 General factors	294
5.7.2 Factors specific to this application in the tagging of Urdu	296

5.7.2.1 The Urdu language	296
5.7.2.2 The nature of the tagset	297
5.7.2.3 Practical restrictions	298
5.8 Concluding remarks	300
Chapter 6 – Implementing a tagger for Urdu	302
6.1 Measuring performance in a tagger experiment	303
6.2 A description of the tagger system	306
6.2.1 General system philosophy and architecture	306
6.2.1.1 Classification of disambiguation systems within Unitag	308
6.2.1.2 The structure of Unitag	310
6.2.1.3 The Unitag file format	314
6.2.1.4 Defining an instantiation of Unitag	319
6.2.2 The design of the tokeniser program	321
6.2.3 An analyser program for Urdu	322
6.2.3.1 Lexical lookup in Urdutag	322
6.2.3.2 Character type analysis in Urdutag	323
6.2.3.3 Morphological analysis in Urdutag	324
6.2.4 The design of the disambiguator program	337
6.2.4.1 The formalisms of rules in Constraint Grammar and Brill's tagger	337
6.2.4.2 The Unirule formalism	340
6.2.4.2.1 Actions	341
6.2.4.2.2 Conditions	343
6.2.4.2.3 Some example rules	346

6.2.4.3 How Unirule works	348
6.3 Creating and optimising the lexicon	349
6.3.1 The Unilex software	349
6.3.2 Manual lexicons	351
6.3.3 Optimising the lexicon	353
6.3.3.1 Variables in lexicon creation	353
6.3.3.2 Deducing the optimal threshold	356
6.3.3.3 Enriching a lexicon	359
6.3.3.4 Different combinations of lexicons	360
6.3.3.5 Summary: the optimal lexicon	361
6.4 Developing a rule list for Urdu	362
6.4.1 How the rule list was developed	362
6.4.2 The nature of the rules	363
6.4.3 The remaining ambiguity	366
6.4.4 What can be learnt from the rules?	367
6.4.5 The order of the rules	368
6.4.6 The number of cycles of disambiguation	369
6.5 Concluding remarks	371
Chapter 7 – Conclusion	372
7.1 Results of this study	372
7.1.1 Resources developed	372
7.1.1.1 The tagset	372
7.1.1.1.1 Possible improvements to the tagset	373
7.1.1.2 The tagging system	374

7.1.1.2.1 Reasons for relatively poor performance in the tagging system	375
7.1.1.2.2 Possible improvements to the tagging system	378
7.1.2 The duration of the project	379
7.1.3 Discoveries concerning the structure of Urdu	381
7.2 Possible future research	382
7.3 An overall summary of the thesis	383
7.4 Concluding remarks	385
Appendix 1 – Description of the transcription scheme	387
Appendix 2 – Glossing system	398
Appendix 3 – Creating the Perso-Arabic tagset	401
Appendix 4 – The tagging guidelines	408
References	461

Introduction

In this thesis, I bring an established procedure of corpus linguistics, part-of-speech tagging, together with a language, Urdu, to which it has not so far been applied.

Part-of-speech tagging of texts and corpora has been of interest to computational and corpus linguistics for over thirty years¹. Its wide range of uses – both as a basis for additional corpus annotation, and in its own right – make it central to much research in the field of corpus linguistics. Thus the corpus linguistic methodology cannot be applied to its full extent to a language for which no part-of-speech tagging has been accomplished. For this reason, it is always a worthwhile goal to extend part-of-speech tagging technology and practices to such a language. As every language is to some degree unique, every language will present its own particular problems in developing part-of-speech tagging technology, making the extension of this tagging technology a perennially novel and interesting task.

However, there are additional reasons, particular to Urdu, that make the development of tagging technology a topic of even greater interest. Firstly, it is not merely a language for which no part-of-speech tagging has been done. It is a member of an entire family of languages for which no part-of-speech tagging has been done, the Indo-Aryan family². As such, it may be hoped that the experience of creating a part-of-speech tagging system for Urdu may prove of benefit to later attempts to do the same for other Indo-Aryan languages. Secondly, the nature of Urdu as an Indo-Aryan language influenced very strongly by Persian and Arabic, being written in a

¹ See 2.1.1.

² See 1.1.2 and McEnery et al. (1997).

slightly modified³ form of the Arabic alphabet, means that it presents a number of interesting and possibly unique problems. For example, how does one tag words loaned from Arabic, which is structurally quite different to Urdu? How does one tag those Persian affixes which are written with a word break between them and the bases they are attached to? The opportunity to confront and solve such problems as these is a significant part of the interest of this thesis.

The main aim of this thesis is to achieve functional automated part-of-speech tagging in Urdu. However, in the process of fulfilling this aim, I will deal with a number of subsidiary aims, and examine and establish several claims. The most important aims of the thesis, contributory to the main aim stated, are as follows:

- 1) to develop a tagset for Urdu;
- 2) to develop a set of tagging guidelines;
- 3) to create an actual tagger program or suite of programs.

The tagset and tagging guidelines, although they may constitute a part of the tagging system in their own right, are also necessary prerequisites to automated tagging software. The tagset is necessary because it is impossible to mark up morphosyntactic categories without the existence of an annotation scheme. Because many tagging technologies require training data, hand-tagged text may be a prerequisite to an automated tagger, and tagging guidelines are required to create this.

In the fulfilling these three principal aims, I will make and justify the

³ The modifications that separate the Urdu alphabet from the Arabic are very slight in comparison to, for example, the modifications that distinguish the Cyrillic alphabet from the Greek alphabet. The Urdu and Arabic forms are very clearly aspects of the same writing system.

following claims about the methods I use towards these aims:

- 1) that the design principles for the tagset that I devise in Chapter 2 are appropriate for the task;
- 2) that the EAGLES guidelines on morphosyntactic annotation (a major international standard⁴) are extensible to Urdu and thus a suitable framework for the definition of a tagset;
- 3) that the methodology of morphosyntactic tagging based on rules devised by a linguist is the best one to utilise when approaching the tagging of Urdu.

While Chapter 1 is a general introduction, Chapters 2 and 3 address the first aim, of creating a tagset, and the first two claims. Chapter 4 deals with the second aim, creating tagging guidelines, in the context of a phase of manual tagging. Chapter 5 justifies the third of the claims, and Chapter 6 is concerned with the aim of creating the actual tagger.

Chapter 1 provides a short discussion of some introductory matters with regard to Urdu, a language of which I am not a native speaker. In it I claim that Urdu is a language of demographic and social significance. Therefore, it is a suitable language for which to develop part-of-speech tagging. I also aim in this chapter to provide sufficient background to the Urdu language to make comprehensible to all my discussion in subsequent chapters of matters concerning its structure. This chapter also discusses the role of part-of-speech tagging in corpus linguistics in general and the EMILLE project, which this study is a part of, in particular, with the aim of contextualising the work done in this thesis.

⁴ See also 2.1.3.

Chapter 2 moves onto the first main aim, of creating a tagset. Before a tagset can be created, there are certain necessary preliminaries, which are dealt with in this chapter. The first (section 2.1) is a review of previous work in the field of tagset creation, in which I aim to show that there has developed a general consensus on at least some of the design principles of a good tagset. This is the basis for the design principles underlying my tagset, which are outlined and justified in section 2.2. This part of the thesis will provide evidence for my claim that these are appropriate design principles for the task in hand. One particularly important design principle (2.2.1) is that of compliance with existing standards, and my decision to create the tagset in line with the EAGLES guidelines is discussed and justified. The third necessary preliminary is a model of the language to be used as a basis for the categories in the tagset. In 2.3 I justify my decision to use the grammar of Schmidt (1999) as a model, and discuss some connected issues.

With these preliminaries dealt with, Chapter 3 accomplishes the first main aim of the thesis by actually defining the U0 tagset, by means of going through the EAGLES guidelines category by category. At all stages the design principles discussed in Chapter 2 are employed. As a result of doing this, I am able to confirm my claim that the EAGLES guidelines are extensible to the Urdu language.

Chapter 4 ends my work on tagsets by describing the process in which the U0 tagset was first put into practice in a phase of manual tagging by a native-speaker informant. This allows me to assess whether or not the tagset is adequate to describe all the categories of Urdu. I also aim here to put to the test the utility of Schmidt's grammar in practical applications, in order to give some empirical support to the decision made in chapter 2 to use this description of Urdu as my model. As a result of this assessment, certain changes to the tagset are outlined and justified. Furthermore,

two subtagsets⁵ for use in the tagging of texts are defined. Also in this chapter, I substantiate the need for a set of tagging guidelines for the Urdu tagset, and describe their creation.

Chapter 5 aims to provide support for my claim that the use of disambiguation rules written by a linguist is the best possible approach to the tagging of Urdu. To this end, I review literature in the field of part-of-speech tagging technology. I look in depth at a number of different tagging methodologies, including tagging based on rules written by a linguist, probabilistic tagging using Markov models, and tagging based on rules learned automatically from a tagged corpus. I do this in order to be able to justify my choice of the first of these approaches. This choice is made in the light of a number of factors which are also discussed in this chapter. I also look at the process of comparing different taggers, making the claim that such comparison is highly problematic and thus cannot be of assistance in selecting an approach to automated tagging.

Chapter 6 describes the process by which the main aim of this thesis, a functioning tagging system for Urdu, was achieved. I discuss how the performance of the tagger is to be measured, and then go on to outline the various component programs written for the tagging system. These include *Unitag* (an overall architecture and file format), *Verticalise* (a tokeniser), *Urdutag* (an analysis program which performs lexical lookup and morphological analysis), and *Unirule* (a tag disambiguation program which applies rules written in the Unirule format by the user); also discussed is *Unilex*, which creates and manages lexicons. The procedure by which an optimal lexicon was created is described, as is the process of creating a list of disambiguation rules (ultimately consisting of 274 rules).

⁵ See 2.2.5.

Chapter 7 is my conclusion and looks back across the preceding seven chapters, considering the results of the study and possible avenues of further research.