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

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 years[1]. 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 family[2]. 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 modified[3] 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 standard[4]) 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 subtagsets[5] 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.