Inflectional Morphology, Reverse Similarity and Data Mining – Finding and Applying Compact and Transparent Descriptions of Verb Systems of Natural Languages

Alfred Holl
Fakultät Informatik
Georg-Simon-Ohm-Hochschule Nürnberg
Postfach 210320, 90121 Nürnberg
Germany
Alfred.Holl@ohm-hochschule.de

ABSTRACT: *Under the term "data mining", the field of computer science includes many different techniques for data analysis, among them methods of cluster analysis. In the approach presented, a special method is designed for the analysis of inflectional systems. The algorithm is independent of individual natural languages and parts of speech. It finds two types of clusters: morphologically homogeneous ones, which contain reversely similar (which possess the same trailing letters), morphologically analogous lexemes of the examined language-part-of-speech combination, and morphologically inhomogeneous ones, in which the largest part of the lexemes is morphologically homogeneous. The resulting registers are compact and transparent as well as easily extensible and correctible. In condensed form, they provide linguistically and didactically usable, structural results on inflectional systems. For instance, it is possible to assign arbitrary lexemes to clusters with a detailed explanation based upon the structure of an inflectional system. The approach is most often applied to verb systems in inflecting and agglutinating languages.*

Key words: Analogy, Aphasia Rehabilitation, Cluster, Conjugation, Data Mining, Inflection, Inflectional Class, Inflection Form, Inflection Type, Inflectional Feature, Inflectional System, Inflectional Morphology, Language Learning, Language Teaching, Morphological Analogy, Morphology, Pattern Verb, Principal Part, Reverse Order, Verb

Received: 9 January 2011, Revised 18 February 2011, Accepted 2 March 2011

1. Introduction and Overview

This research approach combines data mining and structural linguistic analysis of inflectional systems.

Linguistic data analysis is nothing new. When one establishes morphological, syntactical or phonological rules, data analysis is necessary to analyze linguistic material (texts, grammar books, dictionaries and especially inflectional systems). Every existing list of irregular verbs, for example, is the result of a manual data analysis.

My approach, however, attempts to put the analysis of inflectional systems on a consistent formal and automated platform using methods of computer science. For that purpose, I recur to my analytic algorithm, independent of individual languages and parts of speech, first published in Holl / Behrschmidt / Kühn 2004, II.55-II.73, improved in Holl / Suljiæ2010, 54-57, a previous version in Holl 1988, 183-184.

The algorithm was successfully used for the examination of a couple of language-part-of-speech combinations in Indo-European languages: the noun system of Swedish (2007) and the verb systems of Latin (1988), Catalonian (1988), Portuguese (1988), Rumanian (1988), Italian (1988, 2002), Spanish (1988, 2002), French (1988, 2002, 2003), German (2002, 2004), Russian (2004), Greek (2006), Ancient Greek (2006), English (2002, 2007), Swedish (2002, 2007) and Croatian (2010). I focus mostly on verb systems, as they are a lot easier to describe than noun systems.

The basic assumptions and strategies of my approach remained the same since the 1980s although I had to adjust some parameters during the course of time and learnt some new aspects with each language-part-of-speech combination I examined. So when looking in one of my earlier publications in this field, you will find slight differences compared with the more mature and more elaborate version of my approach presented here. This is especially due to the recent exact investigation of the requirements to automatically assign arbitrary lexemes to clusters in Holl / Zimnik 2009.

This paper is intended to be easily readable for all kinds of researchers in the area of computer linguistics. Therefore, I start with some introductory definitions regarding inflection (Section 2). My approach is motivated by the discussion of analogical reasoning strategies based upon the reverse similarity of lexemes (Section 3). The goal is to find linguistically interesting clusters which contain reversely similar lexemes. Possible types of clusters are presented in Section 4. In order to apply my approach, I have to take some principal decisions independent of the language-part-of-speech combination examined (Section 5). In the analytic part, the divisive cluster algorithm, which automatically structures pre-processed linguistic material, is presented. The result of the algorithm is manually post-processed and improved to generate a register which contains different types of clusters (Section 6). In the synthetic part, a search algorithm is designed which uses the register as a collection of rules in order to automatically detect which cluster some arbitrary lexeme belongs to (Section 7). The paper terminates with some perspectives on future research work (Section 8).

In this paper, I will omit some details of my approach and concentrate on its main ideas and procedures. Examples from the English verb system will be presented in order to ensure broader understanding, although modern English has only a reduced inflectional system.

2. Basic Terminology

Linguists use the term lexeme as label for a “word” in a lexicographic sense. A **lexeme** is an abstract basic unit of lexicography which can occur in different inflection forms. In a dictionary entry, a lexeme is represented by the one of its inflection forms, which is considered as most important, namely its **lexical base** (lemma). The lexical base of a lexeme is used as the “name” with which a lexeme is referred to. In the case of a verb, its lexical base is – depending on the language – its present infinitive of active voice, briefly infinitive, or its 1st person singular of present tense. In addition, I will use the term **lexeme ending** as an abbreviation of the term **ending of the lexical base of the lexeme** by identifying a lexeme by its lexical base.

Applying the above definition, **inflection** means the rule-based modifications of lexemes in order to mark certain inflectional categories. Inflection is an umbrella term comprising declension and conjugation. **Declension** comprises the modifications of nouns, pronouns and adjectives to indicate case, number and gender; **conjugation** those of verbs to indicate tense, mode, voice, person and number. **Inflectional morphology** (in this paper briefly morphology) is the linguistic discipline which deals with inflection.

The phenomenon of inflection can be found in inflecting and agglutinating languages. In inflecting languages, one can split a word form into a stem which carries the semantic meaning (and sometimes also a part of the inflectional meaning) and an ending which carries the largest part of the inflectional meaning. In agglutinating languages, the splitting of meanings is a lot clearer, as uniting several components of the meaning on one morpheme rarely occurs. The border between inflecting and agglutinating languages can often not be drawn exactly.

Morphological analogy of two or more lexemes means that they possess the same morphological properties, that is, the same inflectional features. In other words, their inflection forms possess the same structure and the same components. When an inflection form of a lexeme is known, the corresponding one of an analogous lexeme can be derived using analogical reasoning, e.g. the inflection forms of the verb *cling* can be derived from those of the verb *fling*.

The set of the inflectional features of a lexeme is called its **inflection type**. The ordered set of the inflection forms of a lexeme is its **averbo**.