LDVLIB(LEM): A SYSTEM FOR INTERACTIVE LEMMATIZING AND ITS APPLICATION

R. Drewek, M. Erni

Seminar of Romance Languages, University of Zurich, Switzerland

A concrete project like our "Concordanza lemmatizzata delle 'Operette morali' di G. Leopardi" (a lemmatized concordance of an italian text of the 18th century with some archaic phenomena and of about 70'000 tokens and 9'500 types) is a good opportunity to introduce a new software package for linguistic data processing not as mere cumulation of routines or statements but as a comfortable tool just in use.

LDVLIB is no experimental, single language dedicated and fragile collection of algorithms. It tries to provide fast and reliable standard procedures for everyday jobs in linguistic and literary research and sometimes even a bit more. The package consists of 34 programs and 41 modules, mainly written in PL/1. They have been carefully developped in the last seven years and been tested in various research projects since then. The programs can be grouped by purpose:

- text preparation (editing, correcting and printing)
- text corpus handling
- lexical text analysis, lexicostatistics
- statistical string description (length phenomena)
- machine dictionary management
- production of indices, frequency dictionaries and concordances
- lemmatization
- analysis of spoken language texts

- content analysis
- utilities for bibliographies, document preparation, graphics and graphemes

Whereas programs can be used by the non programming researcher communicating with the program by keyword oriented und widely unformatted command language, a set of modules is thought to support the programming linguist in the fields of string manipulation, word and word list manipulation, dictionary handling, VDU fullscreen communications, print plot and other purposes.

All programs which produce numerical output from statistical analysis provide a data interface to input well known statistic software like SPSS or SAS. The text coding rules are oriented on the printed original with a few restrictions which can easily be learned even by non trained personal. The character set is able to receive any roman transliteration of languages using different graphemes, even old Egyptian hieroglyph texts were analyzed by LDVLIB programs.

The complex task of producing a concordance claims a lot of facilities given by LDVLIB programs. The "crucial point" of lemmatization must be discussed to define an appropriate interface in man-machine interaction to obtain reasonable philological results. Our design of an interactive lemmatizer may be useful to show not only man-machine interaction but computational linguist/literary expert interaction as well. And it might reveal the lack of linguistically reliable algorithms for a fully automatic approach to this problem.

LDVLIB(LEM) doesn't lemmatize automatically but it supports lemmatization as follows. It allows to work on single portions of a text and one or more users have access to the on-line machine dictionary at the same time. The user gets presented on the screen:

- in the upper part, from the KWIC-concordance: every token to be lemmatized, with context and references (page, line)
- in the lower part, from the machine dictionary: proposals of lemmatizing relative to the type shown in the upper part.

Interactive lemmatizing consists therefore in recording the (automatically generated) number of the convenient proposal in the line of the token. If there doesn't result any proposal or not a convenient one from the machine dictionary, the user will insert immediately the convenient dictionary entry and record its proposal number in the upper part of the screen. Such a new proposal will be stored in an additional dictionary that is to be transferred periodically into the main dictionary.

The always growing machine dictionary bases on a national language frequence vocabulary of about 25'000 types including about 5'000 lemmata. There has been put a lot of care in the design of the information codes. The machine dictionary entries consist of 4 fields: type (inflected wordform), lemma (deflected keyword), lemma information and type information.

The lemma information includes the following segments:
- word class and additional informations
- additional lemmata (enclitic article, pronouns)
- disambiguation of homography
- cross-reference to the standard lemma (to be generated in the printed output):
  - graphic variant of the lemma (archaic writing)
  - alteration of the lemma (diminutive by suffixation)
- short paraphrase in case of homonymy, where disambiguation is default (in case of polysemy, where disambiguation is optional)

The type information includes the following segments:

- morphological information (gender, number, person, mood, tense, case, gradation)
- morphological variants (archaic inflexion)
- graphic variants (elision, short form)
- special, i.e. idiomatical use
- relation to a distinct vocabulary (e.g. frequence vocabulary)

The users of concordances (lemmatized or not) have different interests. In literary research one may study the single types or even merely the single tokens of a lemma in the order of occurrence in a work. In linguistic research one may be interested in alphabetic order of the types and in subsequent alphabetic order of the right context of the single tokens. These two examples of ordered concordances don't need the type information. But the type information as provided in our machine dictionary will allow to get a more sophisticated internal order of the lemmata: e.g. singular preceeds plural, positive preceeds comparative and superlative, present preceeds past, morphological and graphic variants are distinguished or not, idiomatical uses are ordered separately or not.

The access to a lemmatized concordance will be as to a data base and the linguist interested in certain phenomena may select by options e.g. the substantives and adjectives only or all verbs in passive construction. LDVLIB(LEM) allows always to the user to get full print of the lemmatized concordance or a reduced print of a list of 1 to n lemmata.

It will be shown that support of the philologist's work by a large dictionary is not only useful in concordance making, but as well cumulates a lot of material for subsequent lexicographic work. Looking ahead, two questions must be considered: the integration of a dictionary data base and the productive use of grammatical procedures like ATNs to shift balance between intellectual work and machine support in direction to "a little bit more automatic".