THE LEXICAL ANALYSIS OF FRENCH
Max Silberztein
Laboratoire d'Automatique Documentaire et de Linguistique1, Centre d'Etudes et de Recherches en Informatique Linguistique2

0. Introduction

Computers are increasingly used for text processing. Therefore, large amount of documents are stored at one point or another on magnetic supports, and thus can be used for computer or linguistic processing. Such available texts constitute basic linguistic material, hence, it is normal that linguists should try to use computer-based methods to analyze them.

Linguistic analysis at any level (whether syntactic or semantic) should not be started without a complete list of the words of the text with minimal grammatical information. Hence, compiling such an elementary lexicon is a necessary step in any linguistic analysis. The process of recognizing the words of a text is called lexical analysis. It requires purely computer-based calculations as well as consulting linguistic data bases. Building a lexical parser that recognizes the words of a text (without any approximation) is indispensable. For instance, it is not possible to analyze the sentence:

\[ \text{Il table sur les pieds noirs pour être élu} \]

\((He \text{ banks on the Algerian-born Frenchmen to be elected})\)

if one doesn't know that the token \text{table (banks)} can be a conjugated form of the verb \text{tabler (to bank)}, and \text{pieds noirs (Algerian-born Frenchmen)} can be a compound noun.

Building a lexical parser requires that some linguistic problems be formulated within a formal framework. The lexical analysis can be seen as a projection of the tokens of a text into the system of dictionaries. A token is recognized when it is associated with the linguistic informations stored in the dictionaries of the system. On the one hand, an alphabetic representation must be defined for the dictionary. On the other hand, texts are available in a wide variety of typographic forms. One of the problems we address is the matching of the two representations of linguistic elements. We decompose this question in a large number of simple cases, each well-known. But the difficulty lies in the formalization of not always well-defined habits of writing, and in the interaction of all these elementary phenomena.

DELA is LADL's system of dictionaries. It contains the DELAS, the DELAF, the DELAP and the DELAC. DELAS is LADL's Electronic Dictionary for Simple words, which was constructed by Blandine Courtois (1987). It now contains 70,000 simple entries, with which are associated grammatical information (verb, noun, adjective, etc.) and morphological information indicating how verbs are conjugated and how nouns and adjectives are inflected. By using this morphological information, Blandine Courtois generated automatically the DELAF (LADL's Electronic Dictionary for inflected

1. Unit 819 of the CNRS. This work has been partly financed by the Programme de Recherches Coordonnées «Informatique Linguistique» of the Ministry of Research and Technology.
2. FIRTECH Industries de la langue française.
Forms) which contains 550,000 word forms, like *mangerions* (*we would eat*), *chaises* (*chairs*). The DELAP (LADL's Electronic Dictionary for Phonemics) has been constructed by Eric Laporte; it has the same entries as the DELAS, and associates with each entry a phonemic representation. The DELAC (LADL's Electronic Dictionary for Compound words) contains 60,000 compound words, like *pomme de terre, pied noir*. DELA has special provisions for handling proper names.

Formal definitions have to be given for the fundamental linguistic objects in order to recognize them:

Basic objects

1. Letters
2. Delimiters
   2.1. Orthographic signs
   2.2. Punctuation marks
   2.3. Digits and symbols

Strings

1. Words
   1.1. Simple words
   1.2. Affixes
   1.3. Abbreviations
   1.4. Compound words

2. Numbers
3. Punctuation
4. Sentences

If the text is available only on paper, the 'basic' automatic recognition of the characters leads to unresolved problems; fortunately, we study texts available on computers, in which characters are already identified.

We describe in this paper some contextual constraints in French which are used for the automatic recognition of simple and compound words. These constraints can be represented in automata form, using a language similar to the LEX language ([M.E. Lesk and E. Schmidt, 1978]).

### 1. Basic objects

Computers process alphabetical, numerical, punctuation and semi-graphical characters. These characters are encoded in a normalized digital form (the most usual are ASCII and EBCDIC).

#### 1.1. The letters

We first describe the alphabet used in the DELA system. We then describe the alphabet used in current French texts, and the problems which arise during the projection from the text into the DELA system.

#### 1.1.1. The DELA alphabet

The DELA alphabet contains 45 letters:

\[
a \ddot{a} \ddot{a} \ddot{a} \ddot{a} \ddot{a} b c \ddot{c} d \ddot{e} \ddot{e} \ddot{e} \ddot{e} \ddot{e} f g h i j k l m n \ddot{n} \ddot{o} \ddot{o} \ddot{o} \ddot{o} p q r s t u \ddot{u} \ddot{u} \ddot{u} \ddot{v} w x y z
\]