CHLT Project
(IST-2001-32745)

Periodic Progress Report

Workpackage 5
Neo-Latin Morphological Analyser

C.N.R.
Istituto di Linguistica Computazionale

Andrea Bozzi
Giuseppe Cappelli
Marco Passarotti
Paolo Ruffolo

December, 2002
Index

1. Summary of key indicators of project progress…………………………...2-13
   1.1 Introduction
      1.1.1 Choosing LEMLAT
      1.1.2 Developing LEMLAT
   1.2 Overall assessment of the main milestones achieved
      1.2.1 Definition of the codes of wordform segments
      1.2.2 SF Coding
      1.2.3 Algorithm for the analysis of a wordform
      1.2.4 Choice and use of a DBMS open source
   1.3 Problems encountered and decisions taken
      1.3.1 New values and new attributes in comparison with EAGLES
      1.3.2 Exceptional wordforms
      1.3.3 Particularities in the application of some SF values
      1.3.4 Addition of the gender code to the LES
   1.4 Correspondance between planned project progress and actual accomplishments

2. Work progress overview…………………………………………………14-15
   2.1 Specific objectives for the reporting period
   2.2 Achievements
      2.2.1 List of Deliverables
      2.2.2 Progress by Workpackage/task
   2.3 Project Reviews

3. Project Management……………………………………………………...16-17
   3.1 Contractual Issues
   3.2 Co-operation within the consortium
   3.3 Participation in workshops and/or conference, publications…
1. Summary of key indicators of project progress

1.1 Introduction
The activities planned in the context of Workpackage 5 have been realized according to the program provided on Technical Annexe. They can be divided in two parts.

1.1.1 Choosing LEMLAT
During the first three months of the project (June – August, 2002), the main engagement has been dedicated to the evaluation of the most suitable strategy, in order to realize a morphological analyser for the lemmatization of Latin texts belonging to a wide diachronic spectrum.

This phase of the project needed a particular care, because, while the experience of lemmatization systems is widespread internationally, the application of linguistic-lexicographic methodologies able to manage textual productions of a wide chronological period is very rare and difficult.

For this reason, it has not been possible to adapt any tool available on the market, or in Research and University Centres, to the analysis of neo-Latin texts: in fact, usually, the adopted linguistic basis does not go beyond the second century after Christ.

The steps of the preliminary phase of the project have been the following:

- Evaluation of neo-Latin graphic-phonetic characteristics and of suitable rules for their automatic analysis;
- Evaluation of the allomorphism phenomena characterizing declensions and conjugations, beginning from the end of the second century after Christ onwards; their appropriate coding for an automatic system;
- Evaluation of the assimilation and dis-similation phenomena characterizing, ever since the classical age, the coupling between prefixes and stems, especially in verbal morphology; their coding for an automatic system.

With reference to the evaluations listed above, some methods to structure a basic lexicon and a set of information (such as prefixes, suffixes, endings) have been produced. These methods seemed to adopt the best by the linguistic items recording way used in the dictionary of the morphological analyser LEMLAT¹.

¹ LEMLAT is the result of the research *Repertorio Lessicale Automatico Latino*. Its authors are A. Bozzi, N. Marinone (linguistic aspect) and G. Cappelli (information aspect). About LEMLAT see:
1.1.2. Developing LEMLAT
In connection with LEMLAT lexical wealth (statistically able to lemmatize about 1.300.000 wordforms from the origins to the fifth/sixth century after Christ), we noticed that the informations referred by the LEMLAT morphological codes must be modified and increased. This enrichment will allow a different kind of lemmatization: in fact, it will favour queries not only through lemma as a key access, but also through the single grammar codes (such as masculine, feminine or neuter nouns, first conjugation verbs, first or second class adjectives and so on). We refer to this kind of lemmatization as analytical, considering the actual LEMLAT lemmatization as synthetical.
In fact, the main aim to be reached by the neo-latin morphological analyser is to allow the widest range of users as possible (from the specialists to the a secondary school student students) to obtain specific answers chosen on the basis of a key access represented even by only one morphological code: for example, to show all the occurrences of the irregular dative wordforms.
The months from september to november 2002 have been dedicated to the first steps to be done, in order to achieve this aim.

1.2 Overall assessment of the main milestones achieved
According to the above considerations, in order to enrich the actual LEMLAT lemmatization, we decided to produce a new coding system of the basical wordform segments (morphemes) recognized by the LEMLAT segmentation module. There are three types of such segments:
1. LES: invariable parts of the inflected wordforms (e.g.\textit{antiqu}-);
2. SM (Segmenti Mediani): paradigmatic suffixes (e.g.\textit{-issim}-);
3. SF (Segmenti Finali): endings (e.g.\textit{orum}).
The new coding of wordform segments has required the steps reported below.

1.2.1 Definition of the codes of wordform segments

The codes have been chosen according to the coding conventions developed by EAGLES (Expert Advisory Group on Language Engineering Standards), with the following advantages:

- EAGLES is a coding standard largely accepted and tested on many European languages;²
- EAGLES coding is so flexible that allows the lexicographer a wide spectrum of personalization. This is particularly important for us: in fact, since the application of EAGLES standards on LEMLAT is the first one on a dead language, it implies some adaptations of it. Besides the use of codes already defined in previous applications, EAGLES flexibility has allowed us to invent new codes for the particular coding necessities required by latin language.

1.2.2 SF Coding

The use of EAGLES recommendations for LEMLAT development differs from the previous applications of this standard for two reasons at least: the first one is that LEMLAT operates on a dead language; the second one is that the codes are applied not to words, but to “pieces” of words (SF).

Therefore, since the first and second position codes recommended by EAGLES are the ones referring to PoS (first position) and Type (second position)³, they cannot be applied to SF, because these informations are transmitted not by SF, but by LES. SF, in fact, is not a PoS on its own, but it binds itself to a PoS.

For example, in the wordform *pulchro*, the LES *pulchr* is the bearer of the informations related to the PoS (adjective) and to the Type (qualifying), whereas the SF (-o) bears the number (singular), the genders (masculine and neuter) and the cases (dative and ablative).

The allocation of the codes of PoS and Type to the analysed wordforms occurs through a table allowing the automatic conversion between LEMLAT codes of the LES paradigmatic category and the PoS and Type EAGLES codes. For example, the LES *ros* has the paradigm n1 (first declension nouns): this code corresponds to PoS N (noun) and Type c (common).

Each code represents a value of an attribute; each code has a specific position in the coding system.

We report below a table resuming the codes used for the development of LEMLAT system.⁴

---

³ In EAGLES, for Type we mean the possible types of each PoS: for example, the nouns can be proper or common Type, the adjectives qualifying, ordinal, cardinal, indefinite and possessive Type and so on.
⁴ In this table. P: Position; ATT: Attribute; VAL: Value; C: Code.
<table>
<thead>
<tr>
<th></th>
<th>ATT</th>
<th>VAL</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Flexional category</td>
<td>I decl</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>II decl</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td></td>
<td>III decl</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IV decl</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td></td>
<td>V decl</td>
<td>E</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I conjug</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>II conjug</td>
<td>G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>III conjug</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IV conjug</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Conjug e/i</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Irregular conjug</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>None</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>Mood</td>
<td>Active indicative</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pass/Dep indicative</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Active subjunctive</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pass/Dep subjunctive</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Active imperative</td>
<td>e</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pass/Dep imperative</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Active infinitive</td>
<td>g</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pass/Dep infinitive</td>
<td>h</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Active participle</td>
<td>j</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pass/Dep participle</td>
<td>k</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Active gerund</td>
<td>m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pass/Dep gerund</td>
<td>n</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Active supine</td>
<td>p</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pass/Dep supine</td>
<td>q</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pass/Dep gerundive</td>
<td>r</td>
</tr>
<tr>
<td></td>
<td></td>
<td>None</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>Tense</td>
<td>Present</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Imperfect</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Future</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Perfect</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Plusperfect</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Future perfect</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>None</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>Case</td>
<td>Nominative</td>
<td>n</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Genitive</td>
<td>g</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dative</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accusative</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vocative</td>
<td>v</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ablative</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Adverbial</td>
<td>r</td>
</tr>
<tr>
<td></td>
<td></td>
<td>None</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>Gender</td>
<td>Maschuline</td>
<td>m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Feminine</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Neuter</td>
<td>n</td>
</tr>
<tr>
<td></td>
<td></td>
<td>None</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>Number</td>
<td>Singular</td>
<td>s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Plural</td>
<td>p</td>
</tr>
<tr>
<td></td>
<td></td>
<td>None</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>Person</td>
<td>First</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Second</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Third</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>None</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>Degree</td>
<td>Positive</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparative</td>
<td>2</td>
</tr>
</tbody>
</table>
We report below the first part of the archive where the SF coding has been done (the so-called SF archive).

<table>
<thead>
<tr>
<th>SF</th>
<th>CodLes</th>
<th>SF Codes</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>n1</td>
<td>A--bfs--</td>
<td>ros-a</td>
</tr>
<tr>
<td>a</td>
<td>n1</td>
<td>A--bms--</td>
<td>pirat-a</td>
</tr>
<tr>
<td>a</td>
<td>n1</td>
<td>A--nfs--</td>
<td>ros-a</td>
</tr>
<tr>
<td>a</td>
<td>n1</td>
<td>A--nms--</td>
<td>pirat-a</td>
</tr>
<tr>
<td>a</td>
<td>n1</td>
<td>A--vfs--</td>
<td>ros-a</td>
</tr>
<tr>
<td>a</td>
<td>n1</td>
<td>A--vms--</td>
<td>pirat-a</td>
</tr>
<tr>
<td>a</td>
<td>n1e</td>
<td>A--bfs--</td>
<td>plastic-a</td>
</tr>
<tr>
<td>a</td>
<td>n1e</td>
<td>A--bms--</td>
<td>poet-a</td>
</tr>
<tr>
<td>a</td>
<td>n1e</td>
<td>A--nfs--</td>
<td>plastic-a</td>
</tr>
<tr>
<td>a</td>
<td>n1e</td>
<td>A--nms--</td>
<td>poet-a</td>
</tr>
<tr>
<td>a</td>
<td>n1e</td>
<td>A--vfs--</td>
<td>plastic-a</td>
</tr>
<tr>
<td>a</td>
<td>n1e</td>
<td>A--vms--</td>
<td>poet-a</td>
</tr>
<tr>
<td>abus</td>
<td>n1e</td>
<td>A--bfp--</td>
<td>de-abus</td>
</tr>
<tr>
<td>abus</td>
<td>n1e</td>
<td>A--dfp--</td>
<td>de-abus</td>
</tr>
<tr>
<td>ae</td>
<td>n1</td>
<td>A--dfs--</td>
<td>ros-ae</td>
</tr>
<tr>
<td>ae</td>
<td>n1</td>
<td>A--dms--</td>
<td>pirat-ae</td>
</tr>
<tr>
<td>ae</td>
<td>n1</td>
<td>A--gfs--</td>
<td>ros-ae</td>
</tr>
<tr>
<td>ae</td>
<td>n1</td>
<td>A--gms--</td>
<td>pirat-ae</td>
</tr>
<tr>
<td>ae</td>
<td>n1</td>
<td>A--nfp--</td>
<td>ros-ae</td>
</tr>
<tr>
<td>ae</td>
<td>n1</td>
<td>A--nmp--</td>
<td>pirat-ae</td>
</tr>
<tr>
<td>ae</td>
<td>n1</td>
<td>A--vfp--</td>
<td>ros-ae</td>
</tr>
<tr>
<td>ae</td>
<td>n1</td>
<td>A--vmp--</td>
<td>pirat-ae</td>
</tr>
<tr>
<td>ae</td>
<td>n1e</td>
<td>A--dfs--</td>
<td>plastic-ae</td>
</tr>
<tr>
<td>ae</td>
<td>n1e</td>
<td>A--dms--</td>
<td>poet-ae</td>
</tr>
<tr>
<td>ae</td>
<td>n1e</td>
<td>A--gfs--</td>
<td>plastic-ae</td>
</tr>
<tr>
<td>ae</td>
<td>n1e</td>
<td>A--gms--</td>
<td>poet-ae</td>
</tr>
<tr>
<td>ae</td>
<td>n1e</td>
<td>A--nfp--</td>
<td>plastic-ae</td>
</tr>
<tr>
<td>ae</td>
<td>n1e</td>
<td>A--nmp--</td>
<td>poet-ae</td>
</tr>
<tr>
<td>ae</td>
<td>n1e</td>
<td>A--vfp--</td>
<td>plastic-ae</td>
</tr>
<tr>
<td>ae</td>
<td>n1e</td>
<td>A--vmp--</td>
<td>poet-ae</td>
</tr>
</tbody>
</table>

The part reported above refers to the SF a, abus and ae, belonging to the paradigms of first declension\(^5\). In it, there are four fields:
1. SF;
2. LEMLAT code (CodLes) of the morphological paradigm where SF appears;
3. codes of the informations relative to SF;
4. one example for each SF.

In particular, we report the semantics of the six values relative to the SF a in the n1 paradigm:

---

\(^5\) The LEMLAT morphological paradigms of first declension nouns are two: \textit{n1} (first regular declension nouns) and \textit{n1e} (first irregular declension nouns).
1. A--bfs--: first declension (A), ablative (b), feminine (f), singular (s);
2. A--bms--: first declension (A), ablative (b), masculine (m), singular (s);
3. A--nfs--: first declension (A), nominative (n), feminine (f), singular (s);
4. A--nms--: first declension (A), nominative (n), masculine (m), singular (s);
5. A--vfs--: first declension (A), vocative (v), feminine (f), singular (s);
6. A--vms--: first declension (A), vocative (v), masculine (m), singular (s).

1.2.3 Algorithm for the analysis of a wordform

To associate the new morphological codes to a wordform analysed by LEMLAT (e.g. rosas), the new LEMLAT module (from now: CHLT LEMLAT) executes the following steps:

1. the segmentation system identifies, among the morphemes of the wordform, the LES and the SF.
   \[ \text{Ø ros as} \]
   \[ \text{Ø as} \]

2. LEMLAT code of the LES paradigmatic category is automatically converted into EAGLES codes of PoS and Type through the conversion table mentioned above.
   \[ \text{Ø ros has the paradigm n1, that is converted to PoS N (noun) and Type c (common)} \]

3. in the first field of the SF archive, the SF homograph to the one identified in the input wordform (in this case, as)\(^6\) are selected.

   \[
   \begin{array}{llll}
   \text{as} & \text{n1} & \text{A--afp--} \\
   \text{as} & \text{n1} & \text{A--amp--} \\
   \text{as} & \text{n1e} & \text{A--afp--} \\
   \text{as} & \text{n1e} & \text{A--amp--} \\
   \text{as} & \text{n1e} & \text{A--nfs--} \\
   \text{as} & \text{n1e} & \text{A--nms--} \\
   \text{as} & \text{n3e} & \text{C--amp--} \\
   \text{as} & \text{n3e} & \text{C--afp--} \\
   \text{as} & \text{n6} & \text{---afp-1}
   \end{array}
   \]

4. Among the selected homograph SF, that one (or those ones) having, in the second field, a paradigm code compatible with the LES code is chosen.
   In the example, the SF as n1 is chosen, because the LES ros is n1.

   \[
   \begin{array}{llll}
   \text{as} & \text{n1} & \text{A--afp--}
   \end{array}
   \]

---

\(^6\) Since, presently, the coding is restricted to the SF of nominal and adjectival inflexions, we report here only the values of the SF until now coded.
5. CHLT LEMLAT reads the third field of SF archive and chooses the sequence of codes that, in fifth position, has a gender code compatible with the LES one⁷.
   Ŷ The sequence A--afp-- is chosen, because the LES ros is feminine (f)
6. The PoS and Type codes (point 2.) are added to the SF ones (point 5.).
   Ŷ Nc+A--afp--
7. The codes are automatically translated in a form easily interpreted by the user⁸.
   Ŷ NcA--afp-- is translated as “Noun (N), common (c), first declension (A), accusative (a), feminine (f), plural (p)”
8. In output, the analysed wordform is repeated and coupled with its lemma and with the new morphological informations.
   Ŷ Wordform: rosas
   Ŷ Lemma: rosa
   Ŷ Morphological informations: Noun, common, first declension, accusative, feminine, plural

SF are being coded: at the moment, the part related to the nominal and adjectival inflexions has been performed.

1.2.4 Choice and use of a DBMS open source
Some DBMS (DataBase Management System) have been examined to choose the most suitable one to be used in CHLT LEMLAT.
The main specific requirements have been the following:
   Ŷ availability of the system in open source version, to fulfill the requirements of the overall system;
   Ŷ good diffusion of the product, so to have an indirect test of its efficiency and its reliability;
   Ŷ user friendly interface with C functions, to make the LEMLAT updating easy and fast;
   Ŷ availability of very precise reference guides.

MySQL system has completely fulfilled the requirements above-mentioned.
We have planned and implemented a MySQL database containing the LES informations requested by LEMLAT modules and we have written some C functions to allow an easy use of it.

---
⁷ On this subject, see also 1.3.4.
⁸ This conversion comes off through a conversion table between the codes and the description of the function represented by each of them.
We have implemented and tested LEMLAT procedures for the consultation of LES archive. Moreover, we have implemented some applications to operate some specific handlings of the informations in the database.

1.3 Problems encountered and decisions taken

1.3.1 New values and new attributes in comparison with EAGLES

The definition of the codes has been operated with the greatest respect of EAGLES recommendations, but the application to latin language requested some modifications and additions in comparison with the previous layouts of this standard.

The most important ones are:

1. The attribute of first position (“flexional category”), its values and the corresponding codes have been planned ex novo for CHLT LEMLAT coding, because they were not used in any previous application.

   For example, the SF a:
   - ÿ having the paradigm n1, receives the first position code A (first declension),
   - ÿ having the paradigm n2n (second declension neuter nouns), receives the first position code B (second declension),
   - ÿ having the paradigm n3n1 (third declension neuter nouns with plural genitive –um), receives the first position code C (third declension).

   In the example, let’s note that some codes which, in LEMLAT, identify the morphological paradigms (n1, n2n, n3n1,…), bring informations not only about the declention or the conjugation, but even about the gender (n2n), or about the plural genitive type (n3n1). Nevertheless, different paradigms referring to a same declension have now been unified under a single code. For example, both the paradigms n2n and n2ni (second declension neuter nouns in –ium) receive the first position code B.

2. The values of the second position attribute (“verbal mood”) and their codes have been doubled, in order to code separately the SF related to the active conjugation and the ones related to the passive conjugation.

   In this way, for example, the SF with the paradigm V1sa (first active conjugation, present indicative) receive the second position code a (active indicative), whereas those with the paradigm V1da (first passive/deponent conjugation, present indicative) receive the second position code b (passive/deponent indicative).

1.3.2 Exceptional wordforms

The steps described above in 1.2.3 are those ones accomplished by CHLT LEMLAT to analyze a wordform with the structure LES + SF (ros + as): but in latin, besides this one,
there are other structures. They are analyzed in partially different ways from the one described.

We report below an illustrative example.

In the LEMLAT dictionary, about 2600 exceptional wordforms (forme eccezionali: FE) are coded: these are wordforms that, related to a LES, diverge from its regular morphological paradigm. An example is the wordform \textit{amassint} related to the LES \textit{am} and to the lemma \textit{amo}.

In the LEMLAT dictionary, the FE, differently from the other wordforms, are recorded in their entirety: in fact, whereas the wordform \textit{amat} does not exist in the LEMLAT dictionary (but there are the LES \textit{am} and, in the list of SF compatible with it, the SF \textit{at}), the wordform \textit{amassint} is, on the contrary, reported in full. It implies that, if LEMLAT receives as input an FE, it does not segment it, not allowing the system to identify the LES and the SF in the input wordform. This identification lack prevents the described algorithm to apply.

The solution for this problem has been to make a list of the FE in an archive (FE archive) and to assign each of them its own morphological codes. In this way, CHLT LEMLAT, receiving in input a FE and recognizing it as such, identifies it in the FE archive, draws the FE codes and achieves the steps described in 1.2.3 at points 7. and 8.

1.3.3 Particularities in the application of some SF values

In some cases, the morphological informations of an SF are not applicable to all the wordforms coming into the morphological paradigm to which that SF belongs.

For example, the SF \textit{as} with paradigm \textit{n1e} is bearer of the values plural masculine and feminine accusative (e.g. \textit{poetas} and \textit{plasticas}) and singular nominative of the same genders (e.g. \textit{acontias} and \textit{andremas}). Now, if it is true that \textit{as n1e} transmits the plural accusative in all the wordforms with paradigm \textit{n1e}, it does not count for singular nominative, because the most \textit{n1e} wordforms show the singular nominative in \textit{a} (e.g. \textit{poeta} and \textit{plastica}), whereas only few in \textit{as} (e.g. \textit{acontias} and \textit{andremas}).

It implies that, if the SF \textit{as} with paradigm \textit{n1e} is coded with both the values plural accusative and singular nominative, everytime CHLT LEMLAT receives in input \textit{am n1e} wordform ending in \textit{as} (e.g. \textit{poetas}), it assigns to it both the values plural accusative and singular nominative. This is a wrong analysis: in the example, in fact, \textit{poetas} would be analysed not only as plural accusative, but even as singular nominative.

During the coding, this problem has occurred in many cases: the reason is the existence in LEMLAT of some paradigmatic categories, in which lemmas with different morphological features are recorded: using the same example above, the lemmas \textit{poeta} and \textit{acontias}, even if differently inflected, come into the same paradigmatic category (\textit{n1e}). If this classification, minimizing the number of the LES categories, was suitable for the aims
planned for the previous LEMLAT version, the actual CHLT requirements necessitate smaller and more homogeneous paradigmatic categories.

This kind of problems must be solved one by one in two possible ways:

1. through a manual check of LEMLAT dictionary, marking with a special code the lemmas which must or must not receive one of the values of a certain SF. This is a suitable solution if a limited number of lemmas is involved; differently, the risk of mistakes in the manual coding process is very high;

2. through an automatic procedure: this is worth searching in order to preserve the work from possible mistakes.

In the case of as nle we have identified an automatic way to solve the problem above-mentioned.

LEMLAT analyzes the input wordforms, associating to them the lemma and its morphological paradigm code: the lemma is created thanks to generation rules based on the CodLes\textsuperscript{9}. However, for exceptional lemmatizations (LE), LEMLAT uses some special informations recorded in its dictionary. These informations can be used to activate automatic procedures in order to solve problems of the same kind of the one described above: in fact, it is possible to make that the system, receiving in input a wordform ending in as (and of LES nle), analyses it both as singular nominative and plural accusative, using the information –AS recorded for the LE.

We report below the informations recorded in the LEMLAT dictionary concerning the LES aconti and poet.

<table>
<thead>
<tr>
<th>LES</th>
<th>CodLes</th>
<th>Infos for LE</th>
</tr>
</thead>
<tbody>
<tr>
<td>aconti</td>
<td>nle</td>
<td>–AS</td>
</tr>
<tr>
<td>poet</td>
<td>nle</td>
<td>–A/-ES</td>
</tr>
</tbody>
</table>

On the row of the LES aconti the information –AS is recorded: using this information, LEMLAT generates the lemma acontias instead of the regular one acontia. But, for the problem described above, it can even imply that CHLT LEMLAT, finding –AS among the informations for the LE of the involved LES, supplies the analysis of the wordform acontias as:

1. singular nominative
2. and plural accusative.

On the contrary, as far as the wordform poetas is concerned, CHLT LEMLAT, not finding –AS (but –A/–ES) among the informations for the LE, analyses the wordform poetas only as plural accusative.

---

\textsuperscript{9} For example, the lemmas of the LES of CodLes n1 are generated automatically applying the SF a to the LES: the LES ros, with CodLes n1, generates the lemma ros-a.
1.3.4 Addition of the gender code to the LES

In LEMLAT, lemmas of a different gender can appear into the same paradigmatic category, because the gender does not come into the morphological informations of the output in the actual LEMLAT system: for example, both masculine (*nauta, pirata*) and feminine nouns (*rosa, absentia*) belong to the category *n1*. This implies that a wordform such as *rosis* (lemma: *rosa*) would be analyzed by CHLT LEMLAT as a first declension plural dative and ablative both masculine and feminine: in fact, in the LES archive, the information related to the gender is not recorded, whereas, as we have seen, we have decided to assign a gender code to the SF. In order to avoid ambiguous analysis, we identified the following solution.

On the LEMLAT dictionary, we will mark with gender codes (*m, f, n*) all the nominal LES. This will be manually done for the nominal LES coming into “ambiguous” category (for example, *n1*), that is to say the ones including LES of more than one gender; on the contrary, it will be possible to assign automatically the gender codes to the nominal LES coming into “unambiguous” categories, that is to say the ones including LES of one gender only (for example, *n2n*).

Thanks to the recording of the gender codes on the dictionary, CHLT LEMLAT, before generating the analysis of the wordform, checks the compatibility between the LES gender code and the SF gender code and stops the analysis in case of incompatibility. For example, let’s have as input the wordform *rosis* (lemma: *rosa*): on the dictionary the LES *ros* has the gender code *f* (feminine), whereas the SF *is* has both *m* and *f* codes in the gender position. Using these informations, CHLT LEMLAT can choose the correct analysis, that is the one where the SF gender code corresponds to the LES gender code (*f*). As a consequence, the wordform *rosis* is correctly analysed as a plural feminine (and not even masculine) dative and ablative.

1.4 Correspondance between planned project progress and actual accomplishments

The progresses done in Workpackage 5 respect what planned in the Project Program. In particular, they are the following ones:

- LEMLAT has been chosen as an automatic lemmatization tool to be developed for CHLT requirements.

  LEMLAT has:
  - a basic dictionary of about 60.000 LES;
  - tables of themes, prefixes, suffixes, endings and enclitics.
the SF (endings) related to the nominal and adjectival inflexions have been coded in view of an analytical lemmatization. This coding has been tested on the LEMLAT results; 

a basic software system has been implemented in language C and the results, produced on a specific benchmark, are in a checking phase. Moreover, a MySQL database for the management of the LES archive has been planned and implemented;
2. Work progress overview

2.1 Specific objectives for the reporting period
In the period covered by this report, the specific aims proposed in Workpackage 5 have been the following:
- identification of an automatic lemmatization tool to be used as a starting point and to be developed and implemented for the aims of Workpackage 5;
- evaluation of this tool;
- evaluation of the adaptation methodologies of the chosen tool to the aims of Workpackage 5;
- first phase of development and implementation of these methodologies.

2.2 Achievements

2.2.1 List of Deliverables
No deliverables have been produced so far.

2.2.2 Progress by Workpackage/task
According to the specific appointed targets, the first phase of the work in Workpackage 5 has produced the following results:
- the latin lemmatizer LEMLAT has been chosen as the tool to be used and implemented;
- the LEMLAT system and formalism has been tested and evaluated:
  - the LEMLAT output is missing of detailed morphological informations (such as gender, noun, mood...): it produces a synthetic morphological analysis;
  - LEMLAT has a very rich dictionary, powerful rules for managing the graphic variants and a sofisticated wordform segmentation system;
- linguistic development of LEMLAT, for an analytical morphological analysis of the input wordforms;
- writing of a new analysis algorithm of the input wordforms;
- coding of the wordform segments recognized by the segmentation module:
  - definition of the codes to be used, according to EAGLES conventions;
  - definition of the foundamental coding guidelines;
  - LEMLAT SF (endings) coding related to the nominal and adjectival inflexions;
  - recovering and analysis of the coding problems and their solutions;
- comparison between LEMLAT dictionary and the ones of other existing systems (such as Nomen, Words, Perseus Morphological Analysr);
software development for the wordforms-segmentations and management of linguistic roots (LES) within an open-source suitable DBMS;

the DBMS MySQL has been chosen for the management of the LEMLAT LES archive;

a MySQL database for the management of the LES archive has been planned and implemented;

some procedures for the use of the database by LEMLAT modules have been implemented;

some applications for specific handlings of the informations contained in the LES archive have been implemented.

2.3 Project Reviews

The work planned for the next reporting period is the following:

adding of the gender codes to the LES belonging to ambiguous morphological categories;

coding of SF related to the verbal conjugation;

implementation of the algorithm for the analysis of the wordforms with structure LES + SF;

implementation of new LE rules;

revision and settlement of code C already written to make it really usable by other programmers: reorganisation of the data structures, production of the documents for the different procedures;

implementation of new applications for the management of the LES archive: insertion, erasing and modification of LES.
3. Project Management

3.1 Contractual Issues
See annexed letter.

3.2 Co-operation within the consortium
Project meetings:
Ø September 19th, 2002: Roma, Università Gregoriana
    Marco Passarotti
Ø November 15th, 2002: Boston, Tufts University
    Andrea Bozzi, Marco Passarotti

3.3 Participation in workshops and/or conference, publications…
Writing of an article draft: Cappelli G., Passarotti M., "LEMLAT: uno strumento computazionale per l'analisi linguistica del latino. Sviluppo e prospettive", forthcoming in Euphrosyne, XXXI, 2003

Exploratory workshop on Computer texts: documentation, linguistic analysis and interpretation, organized by the Standing Committee for the Humanities of the European Science Foundation (A. Bozzi and A. Raggioli, Strasbourg, 14-15/6/02)

XIV Round Table on Computer-aided Egyptology (A. Bozzi, Pisa, 8-10/7/02)

Seminars at the Classical Studies Dpt., Faculty of Letter, Lisboa University, on e-philology (A, Bozzi, Lisbona, 29 e 30/7/02)

Seminar on Progettare il digitale. Tecnologie per i beni librari: conservazione e fruizione in una biblioteca digitale (A. Bozzi, Firenze, 1/10/02)

International congress on Francesco Maurolico e le matematiche del Rinascimento: l'edizione critica dei testi scientifici e la sfida delle nuove tecnologie (A. Bozzi, Messina, 16-19/10/02)
Seminar on *Gestione e fruizione di immagini digitali per le biblioteche e gli archivi*, organized by Centro di Ateneo per le Biblioteche dell'Università degli Studi di Padova (A. Bozzi and A. Raggioli, Padova, 28/10/02)

Seminar at Istituto Nazionale di Studi sul Rinascimento (A. Bozzi and A. Raggioli, Firenze, 4/11/02)

Lesson at the Romanisches Seminar, Berlin Freie Universitaet (A. Bozzi and M.S. Corradini, Berlin, 27/11/02)

(See also DUP)