Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages

Draft version

Monica Monachini and Nicoletta Calzolari (Coords.)
Istituto di Linguistica Computazionale - ILC - Pisa

Other Contributors:
Anna Braasch and Ole Norling-Christensen (DA)
Tilly Dutilh-Ruitenberg (DU)
Geoffrey Leech and Andrew Wilson (EN)
Jean-Marc Langé, and Jean Veronis, Liliane Khouri and Christine Meunier (FR)
Anne Schiller (GE)
Penny Labropoulou and Maria Gavrilidou (GR)
Paula Guerreiro (PO)
Nuria Bel and Marta Villegas (SP)

Commentators:
Susan Armstrong, Gerardo Arrarte, F. Bacelar de Nascimento
K.A.C. Depuydt, Pieter de Haan, Ramia Hatzidaki,
Ulrich Heid, Dirk Heylen, Philip King,
Lothar Lemnitzer, Wolfgang Paprottee, Helena Soares,
Petra Steiner, Simone Teufel, Annie Zaenen

October 1994

Contents

1 Introduction

The objective of the present document is to propose a common core set of morphosyntactic distinctions to be encoded in lexicons and in corpora of the European languages[1].

A bottom-up approach, looking at existing practices in the main lexical and textual projects world-wide (both in lexical specifications and in corpus tagsets) has been adopted.
The procedure followed was:
– to survey the main encoding practices for morphosyntactic description in lexicons and in corpora with the aim of deriving a consensus from their comparison;
– to work with close cooperation between the specialists both in linguistic annotation of text corpora and morphosyntactic description in computational lexicons, with the aim of working out a compatible set of distinctions;
– to carry out a first testing of the proposal, by applying it concretely to the European languages.

This allowed us to highlight the areas of common ground and some aspects of discrepancy between the different systems for classifying morphosyntactic phenomena, and to provide a first common nucleus of morphosyntactic distinctions. These are proposed in a feature-based notation, in the form of attribute and value pairs.

After testing with respect to all EC languages, the possibility of elaborating common consensual and explicit guidelines for morphosyntactic encoding in lexica and corpora is foreseen.

1.1 The Survey and Proposal Phase

The morphosyntactic descriptions and encoding schemes involved in the comparison are the following:

- on the side of lexica:
    1. the MULTILEX model, - 1st row in the tables - as presented in the the Final Report on Morphology (MULTILEX 1993),
    2. the GENELEX model - 2nd row in the tables - for the encoding of the morphological and syntactic levels in a lexicon, originally studied for French (GENELEX Apr.1993, GENELEX Sept.1993),
    3. the AlethDic application of the GENELEX model (GSI-Erli 1993) - 3rd row in the tables;
- on the side of corpora:

[1] The two EAGLES subgroups working on Morphosyntax in Corpora and in Lexicons agreed on the approach towards morphosyntactcic specifications presented in this document at the Joint Meeting of July 9th and 10th in Pisa. The following members took part in the meeting: Arrarte, Calzolari, Hellwig, Guerreiro, Leech, Monachini, Paprottee, Schiller, Zaenen.

1. the proposal of a consensual nucleus of morphosyntactic information encoded by the most common existing tagging practices - 4th row in the tables - presented in the framework of the NERC Project (Monachini and Oestling 1992a and 1992b), and
2. the scheme proposed by Leech and Wilson - 5th row in the tables - as an outcome of a joint meeting of the Lexicon and Corpus Groups in Pisa (Leech and Wilson 1993). Note that this proposal is a step foreward with respect to that of NERC since, besides corpora, it also takes into account the first requirements for lexica which emerged from the discussion at the joint meeting in Pisa.

Additionally, it should be noted that the NERC proposal also took into account the list of common morphological features proposed within the TEI by the Linguistic Analysis Committee (TEI AI1W2 1991).

The comparison is displayed, category by category, by means of synoptical tables containing the relevant features, followed by some explanatory notes and comments.

1.1.1 Morphosyntactic categories: description of the tables

The tables representing a category are structured as follows:

- the 1st vertical column on the left contains the encoding systems under analysis;
- the top horizontal row displays the morphosyntactic category considered (first item on the left), and the relevant morphosyntactic information, presented as attribute names;
- each column has the name of an attribute and lists the relevant values that are used within each system. If the cell is left empty, this means that the system does not mark the information;
- the bottom box (a set of rows named EAG...) contains the features and values proposed within EAGLES for the category, resulting from the present comparison. The features are articulated on different levels, corresponding to different degrees of "obligatoriness". As already pointed out (see Leech 1992, Monachini and Oestling 1992b, Leech and Wilson 1993b), different levels of constraints can be isolated in the morphosyntactic encoding of a category and, therefore, different levels of standardization can be suggested:
  - Level 0 (L0) contains only the category (or PoS), presented in the draft by Leech and Wilson (1993b) as the only "obligatory features";
  - Level 1 (L1) presents the grammatical features, such as Gender, Number, Person, etc., which are usually encoded in lexica and corpora: these are considered as "recommended features" (in Leech and Wilson defined as "optional") constituting the "minimal common core set of features" for the PoS;
  - Leech and Wilson "Special Extensions" are presented here on two levels; they are in fact exemplary of current practices from two different points of view:

    * Level 2a (L2a) contains information which is pertinent to all or many languages and is considered useful and easy to standardize, but is either not yet usually encoded by current practices or not purely morphosyntactic (e.g. countability for nouns): these are to be considered as "optional features". As a rule of thumb, a value is put here if it is relevant to more than three languages.
    * Level 2b (L2b) presents "language-specific features", not belonging to the set of recognized common features. The values presented in this row are labelled with the initials of the language they are pertinent to.

The table below shows the structure of the synoptical tables and the multi-layered structure of the present EAGLES proposal:

PoS	attributes
MULTILEX	
GENELEX	
AlethDic	
NERC	
EAGLeech	
EAG-L0	obligatory : PoS
EAG-L1	recommended : minimal common core set of features
EAG-L2a	optional: info common to languages, either not usually encoded or not purely morphosyntactic
EAG-L2b	language-specific: language-specific info

Having a multi-layered or hierarchical structure, instead of a flat one, gives more flexibility to the proposal and allows the user to choose the most appropriate level of encoding (which may vary, e.g. according to different applications). The idea is that, going from Level 0 to Level 2 the amount of information is increased and more granularity is achieved. The use of all three levels allows the representation of more information, whereas a description at Level 1 is a subset of a description resulting from the application of all three levels. Moreover, it provides an easy framework for extensions and updating, as well as permitting a comparison of different schemes (at the lowest common level).

1.2 Application of the Proposal to the European languages

For each category, an application of the common proposal (which corresponds, as far as possible, to the union of the analysed schemes) to some European languages has been carried out. Examples from the specific language are given for each attribute value, and, when necessary, criteria for the application of the feature. This proposal may have to be revised, when more applications have been performed and feedback received, in a cyclical process.

1.2.1 Description of the Tables

In the sections "Application to European languages", attributes and values – presented in the preceding tables as the EAGLES proposal for each category – are tested on real corpora and lexica in a number of European languages. The schema applied to specific European languages is, therefore, the multi-layered/EAGLES proposal, which copes not only with the requirements for corpora but also with the requirements for lexica.

In these sections, each morphosyntactic feature pertinent to a category in a specific language is represented in the form of a table: the attribute name is presented in bold characters, the values are listed in italics and examples are provided, together with the information on the corresponding tag in a language-specific lexicon encoding scheme or corpus tagset (if available).

- If a language does not have some values for a given attribute, the cell of the example (and of the tag) is left empty.
- If the attribute has, in a particular language, more values than in the proposed common system, the extra-value is inserted in the second part of the table (with the note *l-spec* to the left).
- If, in a particular language, some features are not at all applicable, this will be dealt with in the "Comments" section; the feature will not be represented as a table.

We give here an example of such a table, i.e. the feature "Number" for Nouns in the Italian section:

Attribute	*value*	It. example	It. tag
Number	*singular*	(il) libro	S/m**s**
	plural	(i) libri	S/m**p**
l-spec	*invariant*	(la/le) attivita'	S/f**n**

1.2.2 Application to Italian

The application to Italian has been carried out by M. Monachini and N. Calzolari (Monachini and Calzolari 1994).

When speaking of the Italian lexicon, we refer here to the Italian Machine Dictionary - DMI (Calzolari et al. 1983, Monachini 1992); when speaking of the Italian Corpus, we refer to the

Italian Reference Corpus (Bindi et al. 1991). The column "It.tag" in the tables displays the actual tag used in the Italian Dictionary and in the Corpus tagset to represent the attribute and the value.

1.2.3 Application to German

Sections on German have been written by A. Schiller (Schiller 1993).

For the application to German we will refer to the tagset developed in the project TC[2] at the IMS Stuttgart (hereafter called the "IMS-Tagset") in collaboration with the project ELWIS[3] at the University of Tübingen.

1.2.4 Application to English

Sections on English have been contributed by G. Leech and A. Wilson (Leech and Wilson 1994b).

The English tagset suggested here does not follow, except in a very broad sense, the practices now established in the tagging of English corpora. There are existing tagging systems - for example, the C5 and C7 tagsets employed for the British National Corpus - which might have been used as a model. However, the difficulty is that the morphosyntax of English is exceptionally simple, and many of the distinctions often represented in English tagsets are not strictly morphosyntactic: for example, the distinction between attributive and predicative adjectives. On the other hand, it is usual practice to assign to the base form of the English verb just one tag, or at most two, rather than to represent all the values of number, tense, person, and mood which would be required by most direct adherence to the EAGLES guidelines. In practice, therefore, the English tagset presented here is something of a compromise solution which will enable English morphosyntactic information to be integrated in the document.

1.2.5 Application to Dutch

The application to Dutch has been carried out by T. Dutilh Ruitenberg (Dutilh Ruitenberg 1994)[4].

Most of the tags described in the sections devoted to Dutch application are selected from two different CELEX lexical databases: some of them are taken from the Syntactical information columns of the Lemmas Lexicon and some from the Inflectional information columns of the Wordforms Lexicon. To understand the term column it is necessary to understand that the databases consist of rows divided into dozens of columns, each column containing a value tag. Each row contains in its columns the information belonging to one Lemma (Lemmas lexicon) or to one word form (Wordforms Lexicon). The Syntactical information in the Lemmas lexicon is represented by Yes/No tags or by abbreviated name tags. As well as such Yes/No tags, the

[2] TC = **T**ext**c**orpora und Erschließungswerkzeuge

[3] ELWIS = korpusunterstützte **E**ntwicklung **l**exikalischer **Wis**sensbasen für die Computerlinguistik.

[4] Special thanks are due to J. v.d. Voort v.d. Kleij for substantial comments.

Wordforms lexicon presents a different kind of tags, called *Flection Type* tags. These are single or composed tags, each element representing one aspect of the inflected wordform.

The examples given in the CELEX tables below are deliberately taken from the *DUTCH LINGUISTIC GUIDE* (Burnage 1990). If there were no examples the author made them up herself. This is always marked.

The CELEX system allows the attribution to one word form of double or triple tags, separated by slashes, such as: 'geo./pers.' and 'hebben/zijn' and 'intrans./trans./wederk.'.

It should be noted that there are two representations of each syntactic tag in the Lemmas lexicon: a numeric one and an alphanumeric one. In the tables below you will only find the alphanumeric tags.

Some categories are not present in the CELEX databases mentioned above. To fill this gap we present here ten 'Proposals for Dutch', either based on the German Tables, or on the MULTILEX table, and always based on Dutch traditional grammar.

There is another CELEX database, containing still more detailed information on subcategorization and subclassification of a syntactic as well as a semantic nature. This database is not yet fully described, but the author did use this source for some tables in the application.

The CELEX tags are in the Dutch language, whereas the tags in the ten 'Proposals' are Latin-derived or English.

For Dutch, a report has also been contributed by Dirk Heylen (Heylen 1994) from Utrecht. In his document Heylen makes useful comments mainly concerning the application to Dutch, but some general considerations and questions are raised as well. (This feedback is taken into account and incorporated in the Dutch section by the Dutch correspondent.)

1.2.6 Application to Spanish

Sections on Spanish have been written by N. Bel and M. Villegas (Bel and Villegas 1993).

When speaking of the Spanish dictionary we refer to the Spanish EUROTRA dictionary. The Spanish tagging is that used in Eurotra's dictionaries.

1.2.7 Application to French

The EAGLES proposal has been analysed from a corpus point of view by Jean-Marc Langé (Langé 1994) of IBM France and its applicability to a French Lexicon has been tested by the Veronis group of CNRS (Veronis et al. 1994a and 1994b). Their reports also constitute a first proposal for encoding a lexicon to be used as a basis for the specific application of corpus tagging within the MULTEXT project.

The sections on corpora describe the application of the EAGLES proposed model to a French corpus, taking as a reference point the tagset developed and used at IBM France Scientific Center, in particular by the speech group. For ease of use, this tagset will be referred to as the *IBMF tagset*.
Despite this focus on a particular tagset, general problems will also be mentioned, even when they do not lead to any discussion of the French language (for example, everybody will agree that nouns do not bear case information in French).
Tagsets such as the IBMF tagset have a certain bias since they are used for the very specific purpose of predicting the exact part-of-speech of words in a corpus; in other words, they are used for *modelling* the language at the morphological level, whereas a lexical tagset would be developed for *describing* the language. For example, this tagset doesn't cover the full set of features for verbs (mood, tense, etc.), for two reasons:

- the graphic form of the verb will help –if necessary– determine the "missing" features
- the addition of tags with these specific features would not improve the language model's essential capability: that of correctly predicting other tags.

For similar reasons, several corpus-oriented tagsets for the same language might differ considerably, depending on the goal pursued (e.g. speech recognition, terminology identification, etc.) or on the type of language modelling used (e.g. stochastic vs. rule-based models, etc.).
So the reader should be careful and consider this contribution for French only as one possible application, not as an attempt to describe a universal solution.

The features are listed in the following order:
1) the different EAGLES attributes/values applicable to the IBMF tagset,
2) then the EAGLES features that are NOT applicable to the tagset or to French,
3) finally the specific items for which there is no attribute/value in EAGLES (those which are not *language-specific* but rather *tagset-specific* and therefore do not need a new EAGLES attribute).

When only a part of the tag name is relevant to the specific attribute-value pair considered, this part will be put in bold font in the Tag column (e.g. in tag SUBSFS, **SUBS**FS applies to substantive, SUBS**F**S to feminine and SUBSF**S** to singular).

1.2.8 Application to Danish

The application to Danish has been contributed by Anna Braasch of CST, Copenhagen (Braasch 1994); the description of the Danish Word Bank has been contributed by Ole Norling-Christensen.

In all the tables in the Danish sections the value names are used in accordance with the Eurotra Dictionary Encoding Manual for Danish (EDEMD) and/or with the appropriate namings used in traditional dictionaries. The names are also harmonised with the tagset planned for the corpus tagging tool for Danish. However, the authors would like to update their input before the editing of the final version of the present document. The reason why they need this possibility is that a project recently begun in Denmark is aiming at the standardisation of lexical data for encoding, storage and exchange purposes. A number of relevant institutions are involved in this project, the outcome of which will be a standard proposal (with delivery planned approximately for the end of 1995.)

General remarks:

A tag within a table printed in **bold face** as the last part of a tag combination is the value relevant to that particular attribute.
The most frequent value for an attribute is often left unmarked by taggers, i.e. it is regarded as the default value. For instance, the most common value of the attribute Case occurring in a corpus is the value ‘ngen’ (non-genitive), thus there no tag will be inserted for non-genitive (nominative and oblique) cases. Similarly, nouns of type ‘common’ will receive only the tag ’sb’ (substantive) and no tag for the Type.

The tags used in the morphosyntactic descriptions and tables mainly follow the proposal for English, because Danish is closely related to the English language and also the proposed annotation tags seem to be well-suited to the description of the features of Danish as well.

The remarks on Danish are also partly based on a description of The Danish Dictionary project that comprises high-level corpus work and corpus-based dictionary work according to advanced lexicographical principles and methods.

Corpus annotation

A general language corpus of 40 million words compiled by The Danish Dictionary is in the process of being tagged with part of speech (word class) and word form codes for each word.

The degree of annotation granularity is based on the need for the disambiguation of homographs and the proper recognition of lemmas, although the tool used does not support the handling of lexical or morphosyntactic ambiguities.
A coherency and consistency between lexical specifications given in the Word Bank and the corpus tagset is ensured by using interrelated tagsets.

Lexicon annotation

The Word Bank, that is the lexical database of The Danish Dictionary, contains morphosyntactic information on Danish compiled from the official Spelling Dictionary and from two major bilingual dictionaries, with necessary corrections, extensions, etc.

Corpus and Lexicon Morphosyntax for Danish

During its first 30 months the Danish Dictionary project has designed and built a text corpus of 40 million tokens, and a lexical database, the Word Bank, of 340,000 entries (lemmas).

The corpus consists of more than 40,000 samples of a wide variety of text types, written as well as spoken, from the last ten years. Each individual sample is tagged with information on provenence and text properties; no linguistic tagging of the text proper has yet been made.

The Word Bank integrates the information on Danish that is found in the official Spelling Dictionary, two major bilingual dictionaries (Danish-English and -French), and in the excerpts collected since 1955 by the Board of the Danish Language. By a boot-strapping procedure, starting from the Spelling Dictionary, and with considerable manual intervention, all of the 340,000 entries have been assigned a PoS and, when relevant, inflectional information according to the Danish norm (which permits quite a number of variants).

Morphosyntactic information in the Word Bank

The following parts of speech are distinguished: adj, adv, art(icle), fork(ortelse [abbreviation), konj (conjunction), lydo(ord [onomatopoeia), pron, prep(osition), prop(rium), sb, talo(rd [numeral), ud(rebs)o(rd [interjection), vb.
Inflectional information is provided for adj, adv (comparison), pron, prop, sb, and vb, in a way that makes it possible to generate the entire set of inflectional forms. For each of 897 different paradigms the inflectional information is supplied as a string like the following:

```
"+"; "+s"; "+en"; "+ens"; "od"dder"; "od"dders";
"od"dderne"; "od"ddernes"; "+e" "+s".
```

This represents the inflection of *fod “foot”: fod; fods; foden; fodens; fodder; fodders; fodderne; foddernes; fode fods.*

For a given PoS there are a fixed number of positions, separated by semicolons; in this case (a noun) there are nine positions, each representing a distinct combination of features. If inflectional variants are allowed, there will be two or more morphemes in one or more of the positions (like the last position of the example). If a form is impossible / never instantiated, the position is empty (no quoted string).

Nouns (sb)

9 different positions are distinguished. The dichotomies sg/pl, def(inite)/indef(inite), and gen(itive)/non(-gen), cause eight of them:
sg-indef-non; sg-indef-gen; sg-def-non; sg-def-gen; pl-indef-non; pl-indef-gen; pl-def-non; pl-def-gen.
The ninth position is intended for fossilized forms, if any, typically found in frozen prepositional phrases like *til fods* “on foot”, *(hjolpe ham) po fode* “(set him) on his feet”.

Proper names (prop)

Neither number nor definiteness apply, and only two positions are distinguished: sg-indef-non;

sg-indef-gen.

Verbs (vb)
Ten different positions:
infinitive; pres-active; pres-pass; past-active; past-pass; pres.ptc; pf.ptc. (neuter); pf.ptc., common gender; pf.ptc. plural/definite; imperative.
Example, *vinde* "win":
vinde; vinder; vindes; vandt; vandtes; vindende; vundet; vunden; vundne; vind!

Adjectives (adj)
Twelve positions. An adjective must agree with the number, definiteness and gender of the noun; there are two genders: neu(ter) and com(mon gender). There is no concord of case; the genitive forms listed are only for adjectives used substantively. Many adj.s have comp(arative) and sup(erlative):
sg-com-indef-non; sg-com-indef-gen; sg-neu-indef-non; adverbial form; pl/def-non; pl/def-gen; comp-non; comp-gen; sup-indef; sup-pl/def-non; sup-pl/def-gen.
Example, *rigtig* "right":
rigtig; rigtigs; rigtigt; rigtig rigtigt; rigtige; rigtiges; rigtigere; rigtigeres; rigtigst; rigtigste; rigtigstes.

Adverbs
This traditional category (or rather: garbage bin) is likely to be subcategorized according to function. As the Word Bank information is to be used for corpus tagging / disambiguation, a major criterion will be the different surface syntactic behaviour of the adverbs. For the moment, three positions are set aside for adverbs: positive, comparative and superlative. However, comparison is only registered for three adverbs.

Pronouns
For words in this category, up to five different inflexional forms have been registered. However, like the adverbs, the entire set of pronominal forms will have to be analysed and recategorized in a manner that makes it suitable for corpus analysis.

Morphosyntactic information in the Corpus

As stated above, preparations are underway for explicit linguistic tagging of the corpus. The tool (CorpuszBench, produced by TEXTware A/S) that is used by the editors for on-line corpus exploration, can handle SGML-tagged words with attributes for word-class and word-form, for example:

```
<W WC{sb WF{pl>words</W>
```

The handling of ambiguous word classes or word forms is not supported by this tool, which means that for the daily dictionary work it is necessary to concentrate on disambiguating word classes and only occasionally use word form tagging.

1.2.9 Application to Greek

The application to Greek has been carried out by P. Labropoulou and M. Gavrilidou, based on (Aglamissis et al. 1994).

The Greek Morphological Lexicon described in this report is the ILSP Morphological Lexicon, which is used for the purpose of morphologically annotating the Greek Reference Corpus.

The entries of the lexicon are classified on the basis of the grammatical category and the information included is:

- for inflected forms: stem, code number of inflectional paradigm, location of stress (number of stressed syllable), optional diaeresis,

- for uninflected forms: word form, number of stressed syllable.

From the above classification, information on the morphological features of each entry is derivable from the inflectional classes (e.g. gender, number and case for nouns). Further types of information depending on the grammatical category are explicitly coded for each entry. More details on these information types are presented in each section of each category.

The columns presenting the tags in the tables show both the tag that is generated for each word form of the lexicon by a special mechanism, and the tag that would accompany this word if found in a text.

Greek examples are transcribed in the Latin alphabet, using the system adopted for Eurotra. This system of transcription is based on the visual similarity of the graphemes used by the Greek language. The correspondences of this transcription with the IPA are as follows:

a	a
v	v
gh	=9A or j
dh	=9B
e	e
z	z
y	i
th	=9F
i	i
k	k or c
l	l
m	m
n	n
x	ks
o	o

p	p
r	r
s	s
t	t
u	i
f	f
h	x
ps	ps
w	o
au	av or af
ou	u
eu	ev or ef
ghk	g

1.2.10 Application to Portuguese

The sections on Portuguese have been written by P.Guerreiro (ILTEC) (Guerreiro 1994).

She reports on a skeleton evaluation of the EAGLES proposal with respect to the Portuguese language. A comparative approach was adopted. As a main basis, the specifications used in the morphological layer developed by ILTEC in the framework of the GENELEX project (Guerreiro (ed.), 1994), instantiated in a demo lexicon of about 5000 entries, were taken into account and compared with those provided by EAGLES.
Ongoing work on a morphological analyser for the Portuguese language was also sometimes taken into account.

1.3 Further application: the MULTEXT experience

The main objectives of the MULTEXT project (MULTEXT Tech.Ann. 1993) are the definition and the implementation of a set of tools for Corpus-based research and applications, and the production of a corpus in a multilingual framework. Tools and resources will be developed on the basis of operational standards and in the light of the conventions which are being defined by the major international projects dealing with the issue of standardization.

One of the MULTEXT tasks (under the responsibility of ILC-Pisa), dealing with annotation conventions and hence strongly connected with the work presented in this document, aims at formulating:
(i) common specifications and a common notation for the MULTEXT lexicon, and
(ii) a tagset for the MULTEXT corpus on which the tools will run.

The MULTEXT partners involved in this task [5] have carefully evaluated the proposal which was defined within EAGLES – the proposal presented in the tables of the February version of the present document – for each PoS at Level-1 (the level of recommended features). After a global evaluation of the EAGLES proposal, also taking into account the different grammatical traditions and the different language requirements, they checked if the features suited the description of their respective languages and added those features and/or values needed at the language specific level.
All the partners have performed the evaluation by translating their existing – or still under development – lexicons into the EAGLES features and values, i.e. applying the proposal concretely to their languages and providing examples of all the admitted combinations of values for each category. In such a way, constraints on the application of the values have emerged (see the application to the French Lexicon, included in the present report).

The MULTEXT experience turned out to be a very important test-bed for the EAGLES Lexicon proposal:
– a large core of lexicon specifications proved to fit the description of all the six MULTEXT languages (Dutch, German, English, French, Italian and Spanish);
– the cycle of testing and concrete application has stressed the need of further specifications at the language-specific level.

The essential change affected the class of Pronouns, which in the preceding version of the EAGLES Lexicon document incorporated the Determiners; the previous merging of the two categories (at least at L1, with the possibility of splitting the two categories at a more fine-grained level) seemed, in the first instance, to be the best solution to cope with the requirements of many corpus practices – that keep the two categories undistinguished – and attempted to reconcile lexicon specifications and corpus tagsets.
This choice, however, was rightly felt to be too corpus-oriented and the MULTEXT partners have expressed their opinion in favour of having, at the lexicon level, two different categories for Pronouns and Determiners.
Lexical descriptions should aim, indeed, at a general and fine-grained description of the language which is independent from particular applications, while, given a set of practical reasons – state-of-the-art tagging techniques and computability (see Calzolari and Monachini, Coords., 1994) –, broader categories are to be preferred for the tagset and many “collapsings” of values are to be made.

The requirements that emerged from all the language specific applications included in the document and from the MULTEXT feedback, therefore, constitute the basis of the lexicon specifications described in the present version of this report.

[5] Istituto di Linguistica Computazionale - Pisa, Italy (Coord.); Laboratoire Parole et Languge CNRS - Aix-en-Provence, France; Fundacion Bosch Gimpera - Barcelona, Spain; Institut Dalle Molle pour les Etudes Semantiques et Cognitives - Geneva, Switzerland; Universitaet Muenster - Muenster, Germany; Siemens - Barcelona, Spain; Stichting Taaltechnology - Utrecht, The Netherlands.

1.4 Some relevant aspects for lexical specifications

We deal in this section with a number of aspects that are of relevance as introductory general remarks with respect to the goal of providing a common core of lexical specifications.

The issues that we want briefly to discuss concern aspects such as the relationship between lexical specifications and corpus tagsets, the concept of "common category", the objects that are described and the level of description, the descriptive approach (general vs. language specific) and its implications for problems like monotonicity, redundancy, etc., the introduction of constraints at the language-specific level, and so on.

1.4.1 Lexical specifications vs. corpus tagsets

The agreed decision of the two Subgroups of Morphosyntax in the Computational Lexicons WG and of Linguistic Annotation in the Text Corpus WG was to prepare two separate documents for lexical specifications and for corpus tagsets (Leech and Wilson 1994), even though the two topics are clearly strongly interrelated. The background motivation to this decision was essentially the view of corpus tagging as just one of the possible applications of a Computational Lexicon, which has to be seen in a more neutral context as an application-independent set of lexical specifications.

The two subgroups have always worked in close cooperation, and much attention was paid to the definition of compatible sets of attributes and values. The outcome of the work has generated two parallel documents, each focusing on the specific areas, and with clear cross-references between the two[6]. The two documents are therefore not to be seen as two independent sets of recommendations for almost the same set of phenomena, but as two complementary sets of recommendations, a more general one capable of being directly mapped into an application oriented one.

This interdependence between lexicon and corpus is a very important aspect for any future action aiming at creating lexicons and/or tagsets to be shared and made available to the community. Corpus tagging is in fact the first obvious application of a Computational Lexicon and cannot be developed on an independent basis: both the lexicon specialists and the corpus specialists feel that it is very important to reconcile the two views.

The separation betwen the lexical specification area and the tagset can be reflected at the level of terminology:
- the term "features" is preferred when talking about lexical descriptions;
- the terms "tag" and "tagsets" are used for the information associated with words in context, i.e. in corpus annotation.

[6] The present Corpus document does not yet reflect the last changes included in this version of the Lexicon proposal after the MULTEXT feedback.

For the sake of re-usability, the lexical descriptions should be (as far as possible) independent from applications, and should aim at a general description of each language.
The actual corpus tags depend on at least the following:
(1) the lexical features, and
(2) the capabilities of the tagger to disambiguate between different lexical descriptions or different types of typical homographies present in different language types.

Therefore, morphosyntax is encoded in a lexicon with fine granularity, while a set of corpus tags usually reflects broader categories.

The corpus tags are, in fact, developed for each language with a particular application in mind, that of producing a corpus tagged for part of speech (and possibly other morphosyntactic information) by means of automatic disambiguation. It would be ideal to tag a corpus with the lexical descriptions for each word themselves. However, it is well known that this is considerably beyond the capabilities of state-of-the art tagging techniques. Corpus tags are, therefore, to be seen as a kind of underspecified lexical tags. There are two reasons why we may want underspecify corpus tags:

1. Experience shows that some distinctions are difficult to get right with a high rate of accuracy. (For example, in some languages, the disambiguation of indicative present and subjunctive present in a corpus is extremely difficult by automatic means).

2. In order to train the tagger, we need statistical tables (based on co-occurrences of tags). If we have a large tagset, we need a very large corpus to train the disambiguator, in order to observe rare co-occurrences. For example, in the proposal for French presented in the MULTEXT document (Monachini and Calzolari, Coords., 1994), there are 249 different lexical descriptions, but only 74 collapsed corpus tags. Experience (Church, Penn Treebank, IBM France, etc.) shows that the tagset should be under 100. In fact the Penn Treebank collapsed many tags compared to the original Brown, and got better results.

Two other observations are of relevance as regards the relation between lexical specifications and corpus tags.

(a) Sometimes tagging classes are in reality different from lexical descriptions. For example, classes for punctuation are needed and certain types of semantic or pragmatic or lexical information can be present in the tags (e.g. the days of the week).

(b) Furthermore, the "collapsing" decisions are sometimes language dependent and therefore it may be not be appropriate to have completely identical tagsets across languages. We must preserve certain language-specific peculiarities (e.g. if certain distinctions can be easily maintained by an automatic tagger, it may be useful to preserve them).

1.4.2 The concept of "common" category

A debate between the MULTEXT partners generated some possible definitions of "common category" on which is worth reflecting:
– the same meaning of a category implies that it enters in the same combinations in different languages (distributional definition of common categories);
– a category is common "if it yields isomorphic partitions of word-lists or lexicons under translation (module exceptions)". A category system is common to languages L0 and L1 if members of class C in L0 translate as members of C in L1 (translational definition of common categories);
– a common category is "a category which conveys the same linguistic information (i.e. stands for the same linguistic phenomenon) in all the languages in which it is used";
– "a word class is common across the languages if it reflects a linguistic category/phenomenon which is either morphologically or lexically expressed in at least two of these languages, even though there may be not a one-to-one lexical relationship for this in translation".

We could define common categories as "those whose members satisfy the same criteria and tests": this crucially implies a clear definition of criteria for the recognition of their members. These criteria will be given by EAGLES in the next phase with the production of a set of Guidelines for the application of the morphosyntactic features.
In absence of such explicit criteria, we can empirically recognize "common" categories which are relevant in the morphosyntactic descriptions of a number of European languages. These common categories are usable and actually used in the largest lexicons and corpora and have in general the same "meaning" in the different languages, even though the property of "commonality" holds more for open classes and poses more problems for function words or closed classes.

Taking in mind this simple equation of "common categories" and "actually used categories", their "adequacy" in terms of user requirements can also be achieved: it is "empirically" obtained through the bottom-up process of looking at the largest and most "used" lexicons and tagging schemata (as was done here). Behind these lexicons and annotation schemata there have been many different types of users (of lexicons and of annotated corpora).

In fact, with the above caveat in mind, it was found from the analysis of the schemes that a lot of "commonalities" proposed here are relevant for many languages. Different schools and traditions of languages can agree on such a simple set of features.

1.4.3 Objects described and level of description

The typical objects that are described in this proposal are lemmas (even though we do not deal here with lexical decisions as to what to consider to be the lemma) and word-forms at the morphosyntactic level. This essentially includes information on the grammatical category or part of speech, their subtypes as found in lexicons, and inflectional phenomena to be encoded in attributes such as gender, number, tense, etc.

What we propose here is the basic set of core features, derived from a detailed analysis of the major European lexical and corpus projects; we do not aim at giving a completely worked out set of specifications ready to be implemented as such. This task is to be left at the level of the language specific development of concrete application lexicons.

One problem which we encounter as far as objects described are concerned is how to deal with the two complementary phenomena of a) grammatical categories split into more than one graphical unit (multiwords or discontinuous words), and b) graphical units composed of more than one grammatical category (e.g. contractions). Examples of the first type are found in section 3, while examples of the second type are found in section 8.
In general EAGLES recommends handling multiwords as belonging to a single grammatical category, and contractions as two separate grammatical categories, but the option of a different treatment is left open.

1.4.4 Descriptive approach over different languages wrt monotonicity

The general approach underlying the EAGLES proposal follows the ET-7 (Heid and McNaught Eds. 1991) proposal of looking for the basic phenomena at each level of linguistic description, going to the more granular level, and providing the more detailed set of features able to encode the relevant phenomena.

This approach is taken here over the set of European languages, trying to reach the same level of granularity for the description of each of them. For each language, the most common practices for lexicons and corpora were considered.

The obvious consequence of the two approaches together (i.e. granularity and many languages) is that a large repository of potentially useful lexical specifications is formed where all the features which are necessary for the description of the basic phenomena in the many languages considered are juxtaposed to each other. Given the differences in the different language-specific systems, where each system has its own set of constraints, the large collections of features – summarized in the Tables at the beginning of the description of each grammatical category – do not, and cannot, constitute a consistent system to be implemented as such, but are a redundant inventory of all the possible features relevant for that category across the different languages (an ET-7 conformant big repository, according to the "data pool" model). Each language-specific system can afterwards be implemented as a specific application of the general redundant set, by picking up the features and values appropriate for its system.

At the general level of the large synoptical tables there is, by definition (i.e. by the very way in which they were constructed), no property of monotonicity (and no necessity of it), but, in contrast, there is redundancy and conflicting values may also be found there.

It is only at the level of the language-specific instantiations – considered as applications where the problem of the representation formalism will also arise – that monotonicity can be looked for. Instantiations of the general tables are given for most of the European languages as a reinforcement of the EAGLES proposal, and here – in particular when these will be detailed for real lexicon building – the different constraints have to be specified, as far as possible, between different features, the range of pertinent values for each feature have to be made explicit and aspects of the hierarchical organisation of the features have to be solved.

Within this approach it is assumed that not all the values presented in the general Tables are relevant for "all" the languages. There will be cases in which some do not apply, and this has to be made clear in the language-specific applications.

1.5 Present Status of the Morphosyntactic Proposal

The present document, therefore, reflects the results obtained after different phases and many cycles of work:
– The first version (December 1993) of the document contained the survey of the various systems, their comparison and a first sketch of the consensual nucleus of morphosyntactic specifications.
– This preliminary proposal has been applied to a number of European languages: Italian, German, Spanish, English, Dutch and French. This phase, which has been developed in a interval of few months – corresponding to different versions of the present document –, has generated a lot of feedback, especially at the language-specific level.
– A revision of the preliminary proposal in the light of the feedback coming from the first cycle of applications, which has caused changes mainly in the tables at the Level-2b, was performed (March 1994 and June 1994).
– The MULTEXT experience (see above) – in the multilingual framework of a concrete project – has been an important test-bed for the proposal which has brought some modifications in the common proposal itself (October 1994).
– Other LRE projects, RENOS and DELIS, have used the EAGLES specifications to encode morphosyntactic information in their lexicons. On the side of corpora, the EAGLES proposal has been applied within the CRATER Project.

All this feedback, incorporated in this document, constitutes the basis of the present version of the document.

A further phase of revision of all the language-specific applications – already included in the document and following an old version of the proposal – is now being performed in the light of the last changes.
The applications which have already been corrected and restructured on the basis of the present proposal are those for German, Italian, French and Spanish. The Danish, Greek and Portuguese applications have been contributed by our correspondents on the basis of the present proposal.

The inclusion of the revised applications in the document could mean a new phase of minor changes of the attributes and values of the general tables and a new version of the document.

Experience teaches that the process of consensus building in order to arrive at a stable and broadly accepted standard, is of necessity a slow process: the proposal must undergo many phases of discussions between experts in the field and many cycles of applications for testing, evaluating and refining the specifications. It is essential that the proposal/applications interaction proceeds in both directions, as is happening now. The groups will, therefore, take into account any relevant comment and feedback generated by the circulation of the present preliminary proposal, which has still to be considered as "work in progress".

In the next phase – as already mentioned above –, within the exercise of trying to agree on a common schema while applying it in concrete to different languages, the EAGLES mophosyntactic subgroup will also concentrate on the drafting of Guidelines providing definitions and explicit criteria for the application of the specifications in the different languages.

Linked to the production of Guidelines for concrete application, more work will also be required on some issues still outstanding in this document, such as:
– different cases of the 'not applicable' value;
– specification of 'any value' in the lexicon (when a value matches all the values foreseen for a given attribute);
– multiword expressions;
– contractions.

2 Noun

N	Type	Gend	Numb	Case	Count	Defin	Inflect
Mult	com prop	m f n	sg pl	nom gen dat acc voc	cou mass		
Gen	com prop	m f	sg pl				
AlD	*	m f n	sg pl				
NERC	com prop	m f m+f	sg pl s+p	nom gen dat acc bas			
Leech	com prop	m f n c	sg pl	nom gen dat acc voc bas	cou mass		
L0	**NOUN**						
L1	com prop	m f n	sg pl	nom gen dat acc			
L2					cou mass		
L2b		It c Du f(m) Du cont Sp trns Sp notr	It n	Gr voc Gr ind		Da def Da indf Da unmk	Da/Ge weak Da/Ge strg Da/Ge mix

2.1 Comments

Nouns are commonly recognized by the systems under analysis, and an immediate core of agreement as to their grammatical features emerges, as is evident from the table above.

2.1.1 Type

This attribute is used to distinguish common and proper nouns. MULTILEX encodes this information among the "syntactic attributes of the LU (Lexical Unit)". AlethDic splits the nouns into two different categories: "nom" and "nom propre" (this diverging treatment is signalled in the table with an asterisk). However, no transduction (see Monachini, Oestling 1992b) and convergence problems arise between these two approaches, since they are easily intertranslatable.

2.1.2 Case

A discrepancy seems to concern this feature; but "Case" does not appear in the GENELEX and AlethDic models amongst the features for Nouns, owing to their monolingual French-based orientation.

As already discussed in NERC (see Monachini and Oestling 1993b), the values proposed under Case are clearly not mutually disjunctive: rather, some of them overlap (see Introductory part 1.3). Given these overlappings, some facts have to be pointed out:

- the values can never appear all together in one language, but have to be associated with a list of permitted values for that particular language
- the signification of a value has to be seen in relation to the other values of the same language

The relationship between the values, as it can be derived from their use in the analysed tagsets, is illustrated in the following tree:

```
   Case
 |    /\
 |    |  oblique
 |    |   |    |
 |    |   |    |
gen   nom  dat  acc
```

It must be stressed that each language system will use its own pertinent set of values.
'Oblique', discussed here as a value of the feature Case, is not pertinent to the category Noun; it is a value marked for Pronouns.

2.1.3 Countability

This attribute, which MULTILEX encodes at the syntactic level, is pertinent to many European languages; it is, therefore, here included on level 2a.

2.1.4 Values on Level 2b - language specific

On the level 2b of language-specific features the following values and attributes appear (for the meaning and the discussion of the language-specific values, see under the relevant applications):
- for **Gender** ‘common’, for Italian; ‘fem(masc)’ and ‘context’, for Dutch; ‘transart’, ‘notransart’ for Spanish;
- for **Number** ‘invariant’;
- for **Case** ‘vocative’ and ‘indeclinable’ marked in Greek;
- **Definiteness** with the values ‘def, indef, unmark’, for Danish;
- **Inflection Type** with ‘weak, strong, mixed’ is also on this level to deal respectively with this Danish and German characteristics of nouns.

2.2 Application to Italian (Dictionary and Corpus)

The following attribute-value sets are applied in Italian to Nouns (It. tag: S for “substantive”):

2.2.1 Type

Attribute	*value*	It. example	It. tag
Type	*common*	(il) libro	**S**
	proper	Mario	**SP**

2.2.2 Gender

Attribute	*value*	It. example	It. tag
Gender	*masculine*	(il) libro	S/**m**
	feminine	(la) casa	S/**f**
	neuter		
	common	(l’)insegnante	S/**n**

The value *common*, i.e. the same morphological form for masculine and feminine, which is typical of Romance languages, is pertinent both to lexica and corpora: unlike in the lexicon, in a corpus the ambiguity can often - though not always - be resolved by the context.

l’insegnante capace insegna (not possible to disambiguate)
*l’insegnante brav***o** *insegna* (possible to disambiguate)

The value *neuter* is not pertinent to Italian. Some nouns present a sort of fossilized neuter suffix, coming from the ancient Latin neuter gender (in particular, plural nouns denoting “parts of the body” such as *membra, braccia, ciglia* are still present in Italian, but they are not classified as neuter: they are considered as exceptions in the morphological inflexional paradigms for nouns, and are classified as “feminine”, alternating with a “masculine” inflexion)[7].

2.2.3 Number

Attribute	*value*	It. example	It. tag
Number	*singular*	(il) libro	S/m**s**
	plural	(i) libri	S/m**p**
l-spec	*invariant*	(la/le) attivita’	S/f**n**

[7] In general, these forms ending in *-a* alternate with the regular form of the plural: *sg. braccio, pl. bracci, braccia; sg. membro, pl. membri, membra; sg. ciglio, pl. cigli, ciglia.* The two forms disambiguate, in the plural, the polysemy contained in the noun in the singular.

In Italian, the value *invariant* (It. tag: n), i.e. the same morphological form for the singular and the plural, is also used. In particular, this value has to be coded in a lexicon, while in a corpus the context usually solves the ambiguity.

*L'uomo svolgeva alcune attivita' particolar***i** (plur.)
L'uomo svolgeva **un'** *attivita' particolar***e** (sing.)

2.2.4 Countability

Attribute	*value*	It. example	It. tag
Countability	*countable*	la penna	-
	uncountable	lo zucchero	-

This feature is neither used in the tagging of the corpus nor is it at present encoded in the Italian lexicon, although its coding is planned.

Application test:

"Countable" is the noun which has the plural form; usually it cannot be involved in the partitive construction, ex. *due penne* but not ** un po' di penna*; a mass noun admits the partitive, ex. *un po' di zucchero*, and if used in the plural *tre zuccheri* means "three different types of sugar".

2.2.5 Features not pertinent to Italian

The feature **Case** does not apply to Italian nouns.

Definiteness and **Inflection Type** are not pertinent to Italian.

2.3 Application to German

2.3.1 Type

Attribute	*value*	example	tag
Type	*common*	(das) Buch	**NN**:Neut.Nom.Sg
	proper	Hans	**NE**:Masc.Nom.Sg

2.3.2 Gender

Attribute	*value*	example	tag
Gender	*masculine*	(der) Mann	NN:**Masc**.Nom.Sg
	feminine	(die) Frau	NN:**Fem**.Nom.Sg
	neuter	(das) Haus	NN:**Neut**.Nom.Sg

2.3.3 Number

Attribute	*value*	example	tag
Number	*singular*	(der) Mann	NN:Masc.Nom.**Sg**
	plural	(die) Männer	NN:Masc.Nom.**Pl**

2.3.4 Countability

This feature is not used in the IMS-Tagset for German.

2.3.5 Case

Attribute	*value*	example	tag
Case	*nominative*	(der) Mann	NN:Masc.**Nom**.Sg
	genitive	(des) Mannes	NN:Masc.**Gen**.Sg
	dative	(dem) Mann	NN:Masc.**Dat**.Sg
	accusative	(den) Mann	NN:Masc.**Akk**.Sg

2.3.6 language-specific features

In German, there are some nouns with adjectival inflection (especially nominalized adjectives). These nouns distinguish *strong*, *weak* and *mixed* inflection, depending on the preceding determiner (cf. section 4.3.6).

Attribute	*value*	example	tag
Inflection	*strong*	(nichts) Gutes	NN:Neut.Nom.Sg.**St**
	weak	(der) Beamte	NN:Masc.Nom.Sg.**Sw**
	mixed	(ein) Beamter	NN:Masc.Nom.Sg.**Mix**

2.4 Application to English

English nouns do not have gender as a morphosyntactic category. Also, case is a dubious category, since the only basis for a case distinction in modern English lies in the *'s* or *s'* ending attached to nouns and to some pronouns. It is arguable, however, that in modern English, this is not a case form, but an enclitic postposition. (This would explain the occurrence of phrases such as *in a month or two's time* or *in someone else's garden*, where the *'s* is clearly suffixed not to the head noun, but to the whole phrase.) Hence, in the present tagset, case is not applied to nouns, and the *'s* is treated as a postposition.

2.4.1 Type

Attribute	*values*	Examples	Tags
Type	*common*	book	NCs
	proper	Martha	NPs

2.4.2 Number

Attribute	*values*	Examples	Tags
Number	*singular*	book	NCs
	plural	books	NCp

2.5 Application to Dutch

2.5.1 Type

Attribute	*value*	Example	Tag
Type	*common*	boek	N
	proper	Piet	N

CELEX has no specific attribute-value tag for Proper Nouns as a separate category, but has a proper noun subclassification tag set distinguishing four values. So each Proper Noun gets two tags: 'N' for Noun and 'geo.' for the type of Proper Noun geographical names, and so on: 'N' and 'pers.' (names of people), 'N' and 'merk.' (Company or brand names), 'N' and 'over.' (other).

Attribute	*value*	Example	Tag
Proper-Noun-Type	*geographic*	Belfast	N **geo.**
	persons	Wilma	N **pers.**
	brand	Droste	N **merk.**
	other	Teleac	N **over.**

The proposed attribute values Noun and Proper Noun are entirely appropriate on the first level for Dutch as well. The Proper Noun subclassification would be introduced as an L2 tag.

2.5.2 Gender

CELEX distinguishes five basic gender tags. Apart from these, a number of combinations of these five tags are also possible (double tags).

Attribute	*value*	Example	Tag
Gender	*masculine*	wijn	m.
	feminine	ruimte	v.
	neuter	vertrek	o.
	fem.(masc)	bierhal	v.(m.)
	context	gelovige	m.-v.

The fourth category is: female nouns which can be treated as male.

The fifth category is: nouns whose gender is context-dependent. These are, for example, Adjectives and Present and Past participles used as nouns. Referential characteristics, such as possessive nouns in the near context or other contextual factors must be used to help decide on the respective gender:

```
De gelovige heeft een kaars in HAAR handen
The worshiper holds a candle in HER hands
```

'gelovige' (worshiper) is feminine.

```
De gelovige heeft een kaars in ZIJN handen
The worshiper hold a candle in HIS  hands
```

'gelovige' (worshiper) is masculine.

See also the subsection Language-specific features.

2.5.3 Number

Attribute	*value*	Example	Tag
Number	*singular*	kanarie	e
	plural	kanaries	m

These attribute values are not present in the LEMMAS lexicon, but are presented in the CELEX Wordforms Lexicon. As will be seen below, 'e' and 'm' also feature in verbal tags.

CELEX distinguishes two other Noun features as well: diminutive singular and diminutive plural. Flection Type tags are : 'de' and 'dm'.

Attribute	*value*	Example	Tag
Diminutive-Number	*singular*	kanarietje	de
	plural	kanariestjes	dm

This could be an L2 tag.

2.5.4 Countability

CELEX does not account for countability, but it is pertinent to Dutch.

2.5.5 Case

CELEX does not treat Case as part of the syntactical system but as part of the inflectional system. In Dutch the genitive 's', directly fixed behind the noun, marks possession, as in English. It also substantivizes adjectives: (iets) moois, (iets) or (het) lekkers etc. Case, other than genitive, is not pertinent to modern Dutch. Only archaic forms have the former case-dependent inflection. This must be the reason why case values are not given in the Syntactical part of the CELEX Lemmas Lexicon but are given in the Wordforms lexicon.

Attribute	*value*	Example	Tag
Case	*genitive*	(dag des) oordeels	G
	dative	(te) gronde (richten)	D

Both tags are always combined with a 'singular' or 'plural' tag.
There are also GP tags (Genitive of Adjective Positive). Some morphologists count such word forms as adjectives, others as nouns. We count them as nouns.

Ge: Genitive singular: hoofds.
GP: Genitive Positive: lekkers, moois, nieuws (nouns deriving from adjectives).
Gm: Genitive plural: only pronouns in the CELEX lexicon!
De: Dative singular: bate, behoeve, berde etc.
Dm: Dative plural: only pronouns in the CELEX lexicon!

See also Case under Articles and Pronouns.

2.5.6 Language-specific features

Concurrent Gender distinction Since in Dutch the distinction between the male and female gender of nouns is disappearing, another appropriate gender distinction is the distinction neuter/non-neuter, which is expressed in the definite article determining the noun: 'de' for female/male words and 'het' for neuter words. As a consequence of the fact that some words can be preceded by both articles, CELEX distinguishes three tags:

Attribute	*value*	Example	Tag
De/Het Woord	*non-neuter*	de deur	de
	neuter	het schip	het
	de or het	de/het jolijt	de/het

We need a special tag to mark up nouns which are part of a separable verb: 'paard' in 'paardrijden', separated word form: 'Ik heb paard gereden'. And 'hout' in 'houthakken', separated word form: 'Wij hebben hout gehakt' etc. See also under *Separability* and compare language-specific features of Adjectives and Adverbs.

2.6 Application to Spanish

2.6.1 Type

Attribute	*Value*	Example	Tag
Type	*common*	libro	common
	proper	Pedro	proper

2.6.2 Gender

Attribute	*Value*	Example	Tag
Gender	*masculine*	(el) libro	masc
	feminine	(la) casa	fem
	neuter		
	common	(el/la) responsable	*
l-spec	*transart*	el area	yes

The value neuter is not pertinent to Spanish: we have no neuter nouns.
There are nouns in Spanish which could be valued as common since they can be both masculine and feminine. However, in Eurotra there is no such value: instead we leave the attribute "gender" un-valued at the lexical level. It is the grammar that fills this value via unification, that is, any noun with its "gender" attribute un-valued will unify with both feminine and masculine articles.
A special case in Spanish with respect to gender is that of feminine nouns beginning with stressed "a" (having "h" before or not). These nouns take, when in the singular, the masculine allomorph of the (un)definite article if this immediately precedes the noun:

el aguila calva (the(masc,sing) bald(fem,sing) eagle)
las aguilas calvas (the(fem,plu) bald(fem,plu) eagles)

la gran aguila (the(fem,sing) great eagle)

In order to avoid problems of concord, we add a new boolean attribute ("transart") which serves to distinguish these nouns: when its value is 'yes', a formation rule performs the marking of the article in order for the translator to change its gender value:

Attribute	*Value*	Example	Tag
transart	*yes*	el area	yes
	no	el niño	no

This "transart" attribute could be treated as a language-specific value for the attribute "Gender".

2.6.3 Number

Attribute	*Value*	Example	Tag
number	*singular*	el libro	sing
	plural	los libros	plu

A special case concerning "number" is that of those nouns with the same morphological form for the singular and the plural:
eg: *la crisis, la dosis, el martes, el chasis, el análisis*
las crisis, las dosis, los martes, los chasis, los análisis

Just as in the case of "el/la responsable" above, we leave these nouns unvalued for "Number". We get the value via unification with the value of the article and/or adjective accompanying the noun. Of course, we could add a language-specific value to mark these nouns.
Another special case is that known as "pluralia tantum". This group comparises nouns which are always plural, where the plurality does not come from any inflectional process but is rather inherent to the noun itself:
eg: *víveres, añicos, entendederas...* (provisions, bits, brains)

2.6.4 Case

Not applicable to Spanish nouns although it is present in the Eurotra dictionary for nouns since we include pronouns in the noun category. To distinguish between nouns and pronouns we have the attribute "Nform":

Attribute	*Value*	Example	Tag
nform	*normal*	casa	norm
	clitic	le	cli
	pronoun	él	pro

This requires us to include new attribute referents to pronouns such as "possessive" in our category Noun.

2.6.5 Countability

This attribute is not included in Eurotra's lexicons at the morphosyntax level. Countability is treated as a semantic attribute with the values "countable" and "uncountable".

2.7 Application to French (Corpus)

2.7.1 Type

Attribute	*value*	Example	Tag
Type	*common*	livre	**SUBS**MS
	proper	UNIX	NPRO

2.7.2 Gender

Attribute	*value*	Example	Tag
Gender	*masculine*	homme	SUBSMS
	feminine	femme	SUBS**F**S

Note that some proper nouns can bear gender information (see below).

2.7.3 Number

Attribute	*value*	Example	Tag
Number	*singular*	bébé	SUBSM**S**
	plural	bébés	SUBSM**P**

Note that some proper nouns can bear number information (see below).

2.7.4 EAGLES features not applicable

Countability is not applicable to the IBMF tagset. Being clearly of a semantic nature, it is not obvious why it should feature in a morphosyntactic description scheme.
Case, Definiteness, Inflexion type are not applicable to French.

2.7.5 IBMF Tagset features not applicable in EAGLES

In the IBMF tagset, it has been found useful for disambiguation purposes to refine the proper noun class into subclasses. It can be noted that some of these classes bear gender/number information.

Tag	Meaning	Examples
XFAMIL	family name	Smith, Dupont
XPAYFP	country name, fem.pl.	(les) Seychelles
XPAYFS	country name, fem.sing.	(la) France
XPAYFP	country name, masc.pl.	(les) U.S.A.
XPAYFP	country name, masc.sing.	(le) Danemark
XPREF	christian name, fem.	Mary, Marie
XPREM	christian name, masc.	John, Jean
XSOC	company name	Olivetti, IBM
XVILLE	city name	Paris, Pise

2.8 Application to French (Lexicon)

2.8.1 Type

Attribute	Value	Example	Code
Type	common	livre	c
	proper	Jean	p

2.8.2 Gender

Attribute	Value	Example	Code
Gender	masculine	homme	m
	feminine	femme	f

2.8.3 Number

Attribute	Value	Example	Code
Number	singular	homme	s
	plural	femme	p

2.8.4 Case

Non applicable to French.

2.8.5 Combinations

Tag	Example
Ncms-	homme
Ncmp-	hommes
Ncfs-	femme
Ncfp-	femmes
Npms-	Jean
Npmp-	Pays-bas
Npfs-	Anne
Npfp-	Pyrenees

2.9 Application to Portuguese

The following attribute-value sets are applied in Portuguese to Nouns :

2.9.1 Type

Attribute	Value	Example	Tag
Type	common	rapaz	
	proper	Lisboa	

2.9.2 Gender

Attribute	Value	Example	Tag
Gender	masculine	rapaz	
	feminine	rapariga	
	neuter		
	common		

The value *common* is not explicitly used in GENELEX, though the nominal inflectional patterns built in the framework of this project implicitly make use of it for nouns such as, for instance, *(o/a) dentista*. For Portuguese this value can be a L2a value.

The value *neuter* is not pertinent to Portuguese. (Although, as in other languages, some nouns from classical languages presenting a kind of neuter suffix [e.g. *lexica*] can be used in Portuguese, they are typically assigned a pattern of masculine inflection.)

2.9.3 Number

Attribute	Value	Example	Tag
Number	singular	rapaz	
	plural	rapazes	
	invariant		

The value *invariant* is not explicitly used in the Portuguese GENELEX model, though the nominal inflectional patterns built in the framework of this project implicitly make use of it for nouns such as, for instance, *(o/os) atlas*, *(a/as) sandes*. For Portuguese this value can be a L2a value.

2.9.4 Countability

Though this feature is not encoded in the GENELEX morphological layer, it is true that nouns in Portuguese can be used sometimes in a “count” sense and other times in a “mass” sense. Considering the principles of granularity supporting the distinction between L1 and L2a, we think that this feature should be considered “Application dependent”.

2.9.5 Features not pertinent to Portuguese

The features **Case**, **Defineteness** and **Inflection Type** don’t apply to Portuguese nouns.

2.10 Application to Danish

For Danish nouns the following features are relevant: Type, Gender, Number, Case (restricted) and Definiteness (language-specific property). The attribute ‘Countability’ may also be useful at the lexicon level.

2.10.1 Type

Attribute	*value*	Example	Tag
Type	*common*	lampe	sb
	proper	Peter	sb_**prop**

2.10.2 Case

The Danish nominal inflectional system comprises only two cases – genitive and non-genitive (oblique) – which are mutually disjunctive. The value ‘ngen’ will not be used by the corpus tagger, because this oblique form is unmarked in Danish.

Attribute	*value*	Example	Tag
Case	*non-genitive*	lampe	sb_sg_**ngen**
	genitive	lampes	sb_sg_**gen**

2.10.3 Countability

This feature is treated within the EDEMD and within traditional Danish dictionaries as defective inflectional patterns like ‘singularia tantum’ and ‘pluralia tantum’, as a parallel to countability, i.e. non-countable or mass words like ‘maelk’ (milk), ‘vand’ (water), and a number of deverbal nouns like ‘drift’ (drift) are non-countable and also singularia tantum. Those words do not have a plural form. On the other hand, words like ‘penge’ (money) only have a plural form, which is also marked in the dictionaries mentioned above. The tags below are examples of the Eurotra encoding, and they are values of the attribute sub-category of nouns in the lexicon. Other useful tags would be ‘count’ and ‘non-count’ as values in the lexicon for the attribute countability. The corpus tagger will recognise the inflected forms occurring in the corpus.

Attribute	*value*	Example	Tag
Countability	*singulare tantum*	(en) kulde	sb_**singtant**
	plurale tantum	penge	sb_**plutant**

2.10.4 Gender

Attribute	*value*	Example	Tag
Gender	*common*	(en) lampe	sb_**comm**
	neuter	(et) bord	sb_**neut**

2.10.5 Number

Attribute	*value*	Example	Tag
Number	*singular*	(en) lampe	sb_**sg**
	plural	lamper	sb_**pl**

2.10.6 Language-specific features: Definiteness

In Danish, nouns can be marked for definiteness with the enclitic article; in singular common gender: -(e)n; singular neuter: -(e)t; and in plural: -(e)ne (no gender distinction).
The order of inflectional endings is: number + enclitic article + genitive suffix, e.g. lampe + s; lampe + r + ne + s, which give the tag combinations sb_comm_sg_indef_gen and sb_comm_pl_def_gen, respectively.

Attribute	*value*	Example	Tag
Definiteness	*indefinite*	(en) lampe	sb_comm_sg_**indef**
	definite	lampen	sb_comm_sg_**def**

2.11 Application to Greek

Nouns in Greek are inflected for number and case. The following set of attribute-value pairs is derivable from information coded in the Morphological Lexicon (either in the stem or the inflectional paradigm).

2.11.1 Type

Attribute	*value*	Gr. example	Gr. tag
Type	*common*	vivlio	No**Cm**
	proper	Ghiannys	No**Pr**

2.11.2 Gender

Attribute	*value*	Gr. example	Gr. tag
Gender	*masculine*	ouranos	NoCm**Ma**
	feminine	karekla	NoCm**Fe**
	neuter	vivlio	NoCm**Ne**
l-spec	*masc-fem*	tamias	NoCm**Co**

The value *masc-fem* is used for nouns that remain the same both in the masculine and in the feminine gender (e.g. nouns denoting professions); note that since these nouns do not have a neuter gender, the label *masc-fem* is considered preferable to *common.* Although the linguistic context may contribute to the resolution of the ambiguity in certain cases (usually on the basis of the modifying article and/or adjective, given that they must agree in gender with the noun), the value remains in the corpus as well, for reasons of uniformity:

O *tamias eina apasholimen*<b>os</b> (masc.)
y *tamias einai apasholimen*<b>y</b> (fem.)
but **Oi** *tamies efughan* (masc. or fem.)

2.11.3 Number

Attribute	*value*	Gr. example	Gr. tag
Number	*singular*	vivlio	NoCmNe**Sg**
	plural	vivlia	NoCmNe**Pl**
l-spec	*invariant*	asanser	NoCmNe**Nv**

The value *invariant* is used in Greek for those nouns that do not have a morphological variant for the singular and plural number. This holds mainly for foreign words that have been incorporated in the Greek language and have been accepted as "Greek words" but which, however, have not been adapted to the Greek inflectional system. The linguistic context (modifying article/adjective) can serve as a disambiguating device in the case of corpus tagging:

To *asanser vrisk*<b>etai</b> *sto isogheio* (sing.)
Einai halasmen<b>a</b> *kai* **ta** *dhuo asanser* (plural)

2.11.4 Case

Attribute	*value*	Gr. example	Gr. tag
Case	*nom*	vivlio	NoCmNeSg**Nm**
	gen	vivliou	NoCmNeSg**Ge**
	acc	vivlio	NoCmNeSg**Ac**
l-spec	*voc*	Ghianny	NoPrMaSg**Vo**
l-spec	*indcl*	asanser	NoCmNeNv**Ic**

The value *indcl*, mutually exclusive from the other four values, is used in Greek for words that retain the same form in whichever case they are found, either in the singular or in the plural. As commented earlier, this value mainly applies to foreign words which have been incorporated in the Greek language without having taken on the morphological characteristics of the Greek inflectional system. The case value can be disambiguated by taking into account the linguistic context in corpora, on the basis of agreement features:

To *asanser vrisketai sto isogheio* (nominative)
Vrika tyn porta **tou** *asanser anoihty* (genitive)

The vocative case (value *voc*) is rarely used in written texts; it can only be found in literary texts, and, in general, in dialogues. The sentence in which it is found is often followed by an exclamation mark:

Ghianny, ela edhw<b>!</b>

The above two values, *indcl* and *voc* are specific to the Greek language, and, therefore, belong to level L2b.

It is often the case (as regards nouns of the feminine or neuter gender) that the same form is used for both the nominative and the accusative case. Although the use of a new value, *nom-acc*, would contribute to economy in the lexicon, we have decided to keep them as distinct values and proceed to ambiguity resolution in the process of corpus tagging. Disambiguation can only be made on the basis of the linguistic context, either simply by inspection of the modifying article/adjective, or, if that is not sufficient, by resort to shallow syntactic analysis (subject/object role):

Y *karekla einai sto dhiplano dhwmatio* (nom.)
Koitazei **tyn** *karekla* (acc.)
but
To *vivlio vrisketai sto trapezi* (subj. - nom.)
Vrika **to** *vivlio* (obj. - acc.)

2.11.5 Countability

Attribute	*value*	Gr. example	Gr. tag
Countability	*countable*	vivlio	-
	mass	plythos	-

This feature is not used at present in the Greek Morphological Lexicon (and, thus, is not used in the corpus tagging either). However, it applies to the Greek linguistic system and it is intended to code it in the future.

2.11.6 Features not pertinent to Greek

Features **Definiteness** and **Inflection Type** are not pertinent to Greek.

3 Verb

V	Type	Fin	Vf-M	Tns	P	N	G	Asp	Vce	Refl	MVf	AuF	Aux	Sep	Clt
Multil	mai mod prf prg pas cop	inf prpt pspt ger sup cond	ind sub impr	pres fut past	1 2 3	s p		ifve perf pfve prog	act pas refl norf				LU	LU	
Genel			ind sub impr cond part inf	pres impf fut past pa-s	1 2 3	s p	m f								
Aleth			ind sub impr cond part inf	pres impf fut past pa-s	1 2 3	s p	m f								
NERC		inf part ger sup fin	ind sub impr cond	pres impf fut past pret	1 2 3	s p sp	m f mf								
Leech	mai mod prf prg pas pph	inf part ger sup fin	ind sub impr cond	pres impf past fut	1 2 3	s p		ifve simp	act pas refl				hv be	sep oth	

L0	**V E R B**														
EAG L1	mai aux mod	fin no-fin	ind sub impr cond inf part ger sup	pres impf fut past	1 2 3	s p	m f								
L2a	s-aux cop							pfve ifve	act pas	rfl norf	trs int imp	prg prf pss pph			
L2b			Ge zu En ing En bas				It c						hv be	Du sep Du nos	clt noc

3.1 Comments

Verbs are consensually recognized at the level of category.

The major problem concerns the different internal organizations of the verbal systems which characterize Romance languages on the one hand as opposed to English on the other hand. Romance languages, but also, e.g., German, are highly inflectional, whilst English presents very few inflections and scarce morphological distinctions for moods and tenses.

Furthermore, the Leech/Wilson and MULTILEX proposals, in particular, introduce new distinctions which in the present proposal appear distributed on different levels (see sections 3.1.3 and following).

3.1.1 Type

This attribute permits the encoding of the nature of a verb as 'main', 'auxiliary', 'modal' (distinctions seen as recommended), and 'copulative' and 'semiauxiliary' (optional, and therefore proposed at level 2a, e.g. *venire* in Italian is used in some cases as a variant to *essere* to form the passive).

3.1.2 Finiteness, Verb-Form/Mood

A different grammatical tradition motivates in English the use of the feature Verb-Form, with values which in Romance languages are typical of the feature Mood. This emerges immediately from the above table, if MULTILEX, NERC and the Leech/Wilson proposal, all designed for a multilingual encoding system, are compared with GENELEX and AlethDic, which are modeled for French verbs, i.e. a system belonging to the Romance tradition.

In a certain sense, the matter can be seen as a problem of the distribution of some values under one or another attribute. As already noted in the NERC survey (Monachini and Oestling 1992b, p.13), the feature Verb-Form appears "strange" from a Romance language point of view, and therefore the problem was left open as an area for further analyses and suggestions. The features Verb-Form and Mood in Monachini and Oestling have been arranged in a tree, with Mood hierarchically dependent on one of the possible values of Verb-Form, i.e. "finite".

The proposal formulated in what follows for reaching a consensus is to introduce, besides "finite", the value "non-finite" as a new value of a new feature **Finiteness**, and to create a feature called **Verb-Form/Mood (Vf-M)** with the relevant values which could be seen in subordinate position with respect to the values 'finite' and 'non-finite'. In such a way, the feature Verb-Form/Mood can be considered totally in hierarchical subordination with respect to Finiteness. For this reason, each language-specific morphological system has to define explicit and strong constraints on which combinations of attributes and values are meaningful for the system (see e.g. the Italian application, section 3.2.1).
The situation can be represented as follows:

```
                    Finiteness
                        /\
       non-finite             finite
           |                     |
     Verb-Form/Mood        Verb-Form/Mood
         /||\                  /||\
   inf part ger sup      ind sub impr cond
```

With this new arrangement, English encoding systems, which do not reach the level of distinction given here for finite Verb-Form/Mood, can stop at the encoding of Finiteness with the value "finite", while Romance languages will not have problems in being compelled to call some Moods Verb-Forms (contrary to their tradition). Furthermore, the above proposal is completely compatible with the Romance verbal systems, since, in those systems, the values of Moods are implicitly finite and non-finite, in the speaker's intuition.

3.1.3 Tense

Tense, as can be observed, does not contain values for "compound tenses", in general past tenses, which in corpora are not usually dealt with by automatic taggers. The possibility of "re-building" and deducing analytic verbal forms, by means of e.g. rules and permitted combinations of "auxiliary-past participle" in the various languages, is provided by values presented here as special extensions, which encode the auxiliary nature of a verb.
This introduces a distinction between the requirements of an encoding scheme for corpora and for lexica: the latter also need specifications for compound tenses.

Another problem for Tense is posed by the value "past", since it changes its "meaning" in the various verbal systems, depending on the opposition relationships among values in a language. The English 'past' does not have the same meaning as the 'past' in the Romance languages: it is not opposed to an 'imperfect' value, but instead it seems to be in opposition to the 'present'. In NERC, a proposal for the disambiguation of this complicated situation was made with the following tree:

```
             Tense
          /    |    \
   future   present   past_E
                         /\
      past_R (perfect)  imperfect
```

In Romance language systems, the values 'past' and 'imperfect' are opposed and designate two different aspects of a past action. Both are opposed to the 'present' with respect to the notion they represent: 'past' is a non-durative action finished in the past and 'imperfect' is a durative action initiated in the past. In order to avoid misunderstandings, a tentative solution could be to rename the Romance 'past' value as 'perfect', since it is opposed to 'imperfect'.

3.1.4 Person, Gender and Number

These features do not pose problems.

3.1.5 Aspect

The values proposed here are 'perfective' and 'imperfective' and, given the syntactic pertinence of this attribute, they appear at level 2a.

3.1.6 Voice

The values 'active' and 'passive' are proposed at the level 2a of optional features. The distinction between 'reflexive' vs. 'non-reflexive', foreseen by MULTILEX under this attribute, is treated here as a separate feature (see below).

3.1.7 Reflexivity

This feature encodes phenomena which in the MULTILEX scheme are collapsed under the feature Voice. The two values of boolean type are arranged on level 2a for the encoding of verbs appearing with the reflexive pronoun as a clitic. The further distinction between true reflexive, pronominal, reciprocal, etc. can be dealt with only at the syntactic level.

3.1.8 Main-Verb Function and Auxiliary Function

These two attributes have to be seen in subordinate position with respect to the two values of Type 'main' and 'auxiliary' and belong to level 2a. Their respective values are:

Main-Verb Function (MVf):
'trans', 'intrans', 'impers'.

Auxiliary Function (AuF):
'prg': for auxiliaries used to form progressive tenses (*be, ...*)
'prf': for auxiliaries used to form perfect tenses (*have, avere, ...*)
'pss': for auxiliaries used to form passives (*etre, worden, werden, ...*)
'pph': periphrastic auxiliary (*do, ...*)

Auxiliary Function, proposed in the present work, encodes phenomena which in the surveyed schemes are collapsed under the feature Type. Main-Verb Function is a completely new feature.

3.1.9 Level-2b specifications

Auxiliary This attribute encodes information concerning the choice of auxiliary for perfect or compound tenses and obviously is language-specific (e.g. *have, be* for English).

Separability This is used in Dutch for verbs including separable particles, such as *faengt ... an.*

Clitic This feature with boolean values is introduced at language-specific level to encode the presence in the verbal form of the clitic.

language-specific values for Tense The values proposed here for English are 'ing-form' and 'base-form'. The value 'infinitive incorporating "zu"' is added for German.

3.2 Application to Italian

The following set of morphological features are pertinent to Verbs in Italian: Mood, Tense, Person, Gender, Number. These features, variously combined, dynamically generate the inflected forms of the Italian verbal system, but not all the features and sometimes not all the values of a feature are applied to an inflected form. The constraints in the application of these features are made explicit in the sections devoted to each feature.

The following attribute-value sets are applied to Italian verbs.

3.2.1 Type

The verbs *avere* and *essere* are marked as auxiliaries. It should be noted that an automatic tagger is not able to disambiguate the cases in which these verbs function as full verbs [*il bambino e' a casa* (the child is at home), *Io ho un cane* (I have a dog)] from the cases when they are auxiliaries.

3.2.2 Finiteness and Verb-Form/Mood

The feature Finiteness is not encoded in itself, i.e. does not have a special mark in the Italian corpus or in the lexicon. It is implicity contained in the various distinctions of Mood, which is hierarchically dependent on the two possible values of Finiteness: on the one hand, infinitive, participle, gerund are non-finite; on the other hand, indicative, subjunctive, conditional, imperative are finite. It can, therefore, be derived automatically from existing codes.
The application of these two features to Italian can be represented as follows:

Attribute	*value*		It. ex.	It. tag
Finiteness	*non-finite*		-	-
	Attribute	*value*	It. example	It. tag
	Verb-Form/Mood	*infinitive*	amare	V/**f**
		participle	amato	V/**p**
		gerund	amando	V/**g**
		supine	-	-
Finiteness	*finite*		-	-
	Attribute	*value*	It. example	It. tag
	Verb-Form/Mood	*indicative*	amo	V/s1**i**
		subjunctive	amasse	V/s3**c**
		imperative	amate	V/p2**m**
		conditional	amerei	V/s1**d**

3.2.3 Tense

It has to be said that not all the values listed for the feature Tense are applied to all the values of Mood.

The table below represents the constraints on the application of the values of Tense to Mood in Italian. Note that, if a value of tense is applied to a value of mood, but is expressed by a compound form, this is noted with (cmpd). All details about compound forms are given below.

Finiteness		*finite*				*non-finite*		
Verb-Form/Mood		*indic*	*subj*	*condit*	*imper*	*inf*	*part*	*ger*
Tense	*pres*	amo	ami	amerei	ama	amare	amante	amando
	impf	amavo	amassi					
	fut	amero'						
	past	amai +(cmpd)	(cmpd)	(cmpd)		(cmpd)	amato	(cmpd)

Considerations about the "meaning" of *past* in Italian have to be added: **past** corresponds here to *passato remoto*.
A problematic aspect of Tense, that of "simple" tenses and "compound" tenses, and the connected difference in requirements for corpora and lexica, has already been dealt with in the general section of comments on the table for verbs.
We give here the list of permitted "auxiliary-past participle" combinations in Italian, in order to obtain compound tenses.

Auxiliary	Past participle	Compound tense	It. ex.
indic pres	past participle	passato prossimo	ho amato
indic impf	past participle	trapassato prossimo	avevo amato
indic past	past participle	trapassato remoto	ebbi amato
indic fut	past participle	futuro anteriore	avro' amato
subj pres	past participle	congiuntivo passato	abbia amato
subj impf	past participle	congiuntivo trapassato	avessi amato
cond pres	past participle	condizionale passato	avrei amato
inf pres	past participle	infinito passato	avere amato
ger pres	past participle	gerundio passato	avendo amato

3.2.4 Person

Attribute	*value*	It. example	It. tag
Person	1	amo	V/s**1**
	2	ami	V/s**2**
	3	ama	V/s**3**

3.2.5 Gender

This feature, among the simple morphological verbal units, is pertinent only to the Mood *participle*: it agrees with nouns as an adjective (note on the agreement with object *avendo letta la lettera, uscii*, note on ablativo assoluto).
The value *common* does not apply to the past participle. The gender of past participle, in

context, helps to disambiguate the gender of the noun itself, when this is morphologically unmarked.

*Insegnanti capaci sono partit***e** *...*
*Insegnanti capaci sono partit***i** *...*

The value *common* is applied only to the present participle, where the gender is undecided in the lexicon, even though it can be disambiguated in the corpus.

Attribute	*value*	It. example	It. tag
Gender	*masculine*	amato (past part) (un) amante(pr part)	V/pr**m** V/pp**m**
	feminine	amata (past part) (un') amante (pr part)	V/pr**f** V/pp**f**
	neuter	-	-
	common	(l') amante (pr part)	V/pp**n**

3.2.6 Number

This has only two values, *singular* and *plural*.

Attribute	*value*	It. example	It. tag
Number	*singular*	amo, ami, ama	V/**s**
	plural	amiamo, amate, amano	V/**p**
	invariant	-	-

3.2.7 Combination of features

Italian verbs have a maximum of 52 different inflected word-forms derived from different combinations of morphological features for simple morphological units (i.e. excluding compound forms).
In summary, the following combinations of morphological features are applied, depending on Verb-form and, within a Verb-form, on different Moods.

SIMPLE FORMS	**Finiteness**	**Verb-Form/Mood**	**gramm. features**
	finite	*ind*	**Tense Person Number**
		sub	**Tense Person Number**
		impr	**Tense Person Number**
		cond	**Tense Person Number**
	non-finite	*inf*	**Tense**
		ger	**Tense**
		part	**Tense Number Gender**

Compound forms have the following combinations:

COMPOUND FORMS	**Finiteness**	**grammatical features**
	finite	**Tense Person Number Gender**
	non-finite	**Tense Gender Number**

Compound forms, because of the presence of the participle, have the feature Gender.

3.2.8 Enclitic Phenomenon

In Italian, pronouns (pronominal particles) can accompany verbs in order to form:
(i) the pronominal form of the verb
(ii) the reflexive form
(iii) the reciprocal form

As far as the lexicon is concerned, the possibility of taking clitics of different types (reflexive, reciprocal, etc.) has to be encoded in the entry, but depends on the syntactic type of the verb; these problems should therefore be dealt with on the syntactic level.

In these forms, the unstressed variant of the pronoun is, in general, separated graphically by the verb (it precedes the verb), except with the infinitive, the gerund and the imperative form and sometimes (rarely) with the past participle.

```
egli si lava
lavarsi, lavandosi, lavati, lavatosi
```

At the corpus level, a clitic pronoun can be attached to the verb
- when it expresses the direct object, *dirlo (dire+lo* = to say it) or the indirect object, *dirgli (dire+gli* = to say to him),
- when it expresses a place adverb: *andarci (andare+ci)* (to go there).

Sequences of a verb with more than one pronoun in a unique graphical form are found if both the direct object and the indirect object of a pronominal verb form are represented by a clitic pronoun: ex. *dandomelo (dando+me+lo)* = giving it to me). These types of compounds can present the phenomenon of epenthesis, i.e. the insertion of a letter for euphonetic reasons: ex. *dandoglielo (dando+gli+e (epenthetic)+lo)* = giving it to him).

As to the dictionary encoding, 'verb-pronoun' compounds have to be represented; the problem of encoding specific 'verb-more-than-one-pronoun' compounds only concerns corpus encoding. The strategy adopted in tagging them is to assign differents tags to the different parts forming the word-token, with a special mark which maintains the graphical links, thus permitting the recovery of the unique graphical form.

3.3 Application to German

3.3.1 Type

The IMS-tagset considers the following 3 types of verbs:

- **modals**: *müssen, können, dürfen, wollen, sollen, mögen*
- **auxiliaries**: *haben, sein, werden*
- **full verbs**: all other verbs

Attribute	*value*	example	tag
Verb-Type	*modal*	wollen	**VM**INF
	auxiliar	haben	**VA**INF
	full verb	gehen	**VV**INF

3.3.2 Finiteness, Verb-Form/Mood

Finite and non-finite verb forms are not distinguished by a single feature in the IMS-tagset. Non-finite verb forms are *infinitive, past participles*. Present participle forms are classified as adjectives.

Example: das lachende/ADJA Kind
er kommt lachend/ADJD herein

Attribute	*value*	example	tag
Finite			
Verb-Form/Mood	*infinitive*	können gehen	VM**INF** VV**INF**
	infinitive with incorporated "zu"	anzukommen	VV**INFZU**
	past participle	gewesen geliebt	VA**PPF** VV**PPF**
Non-Finite			
Verb-Form/Mood	*indicative*	(er) geht	VV**FIN**:3.Sg.Pres.**Ind**
	subjunctive	(er) gehe	VV**FIN**:3.Sg.Pres.**Konj**
	imperative	geh!	VV**IMP**:2.Pl

3.3.3 Tense

The feature *tense* applies only to finite verb forms (but not to imperatives). The values of tense reflect only synthetic tensed verb forms, i.e. present tense (*Präsens*) and simple past (*Präteritum*). Compound tenses (eg. *ich habe gesehen*) are not included.

Attribute	*value*	example	tag
Tense	*present*	(ich) gehe	VVFIN:1.Sg.**Pres**.Ind
	past	(ich) ging	VVFIN:1.Sg.**Past**.Ind

3.3.4 Person

The feature *person* applies only to finite verb forms.
In the IMS-Tagset it is also added to imperatives, where it would not be necessary since German imperatives exist only for 2nd person singular and plural.

Attribute	*value*	example	tag
Person	*first*	(ich) gehe	VVFIN:**1**.Sg.Pres.Ind
	second	(du) gehst	VVFIN:**2**.Sg.Pres.Ind
	third	(er) geht	VVFIN:**3**.Sg.Pres.Ind

3.3.5 Number

The feature *number* applies only to finite verb forms and imperatives.

Attribute	*value*	example	tag
Number	*singular*	(ich) gehe geh!	VVFIN:1.**Sg**.Pres.Ind VVIMP:2.**Sg**.Pres.Ind
	plural	(wir) gehen geht!	VVFIN:1.**Pl**.Pres.Ind VVIMP:2.**Pl**.Pres.Ind

3.3.6 Gender

The feature *gender* does not apply to German verb forms. It occurs only with participles which are used as adjectives (and will be tagged as such).

Example: er hat es getan/VVPPF
nach getaner/ADJA Arbeit

3.4 Application to English

English verbs have no Gender distinction. In addition, the deployment of the Person, Number, Mood and Tense attributes is extremely limited. In the past tense, there are no distinctions of Person and Number, except for the solitary case of the verb *to be*, which has a singular form *was* distinct from the plural form *were*. In the present tense, except again for the verb *to be*, only two forms of each verb occur: the *-s* form in the third person singular, and the base form for all other combinations of person and number. These observations apply to the indicative mood. As for the other moods (subjunctive and imperative), the subjunctive is rarely used, and the imperative is invariable. The subjunctive is also invariable, except for a vestigial past subjunctive of the verb *to be*, in the singular use of *were*. Because of extensive syncretism in the English verb, the historical paradigms of Person, Number, Case, and Mood are barely sustainable in the description of modern English. The base form, if regarded as multi-functional, has at least the following morphological functions for all verbs except *to be*:
(a) singular present tense indicative, 1st person
(b) singular present tense indicative, 2nd person
(c) plural present tense indicative
(d) imperative
(e) present tense subjunctive
(f) infinitive
Since it is impractical, however, given the current capabilities of tagging software, to resolve automatically the ambiguity of these six morphological functions, it is a common practice to assign a single value to the base form, or else to assign two values, one for the finite and one for the non-finite functions. Because of this, the tables below show two tagsets: one tagset representing the 6 attribute-values above, and a reduced tagset ("RTags"), which resembles most tagsets so far used for the English language in reducing the six values to two.

3.4.1 Type

Attribute	*values*	Examples	Tags	RTags
Type	*Main*	eaten	VVPp	VVN
	Auxiliary	has (eaten)	VP(Ind)PrS3	VPZ

The value "auxiliary verb" is not represented directly in the tagset, but is implied by the use of VP (for primary auxiliaries) and VM (for modal auxiliaries). Alongside the principal values of "Main" and "Auxiliary" verbs, it would have been useful to add a third, intermediate category of "semi-auxiliaries", for expressions such as *be going to, have to, have got to,* and *be able to*. However, for simplicity, this category has been omitted from the tagset.

3.4.2 Finiteness

Attribute	*values*	Examples	Tags	RTags
Finiteness	*finite*	eats	VV(Ind)PrS3	VVZ
	nonfinite	eating	VVIng	VVG

The values "finite" and "nonfinite" are not distinguished in the tagset, since they are predictable from other values, viz. those for "verb-form" and for "mood".

3.4.3 Verb-form

Two attributes in this description of English, Verb-form and Mood, are derived from a single attribute (Verb-form/Mood) used elsewhere in this document. Verb-form applies only to non-finite verbs, whereas Mood applies only to finite verbs:

Attribute	*values*	Examples	Tags	RTags
Verb-form	*infinitive*	(to) eat	VVInf	VVI
	ing-form	eating	VVIng	VVG
	past part.	eaten	VVPp	VVN

The value "ing-form" applies to all verb forms ending in the inflectional suffix *-ing*. In modern English, the distinction between present participle and gerund, representing two different functions of the *-ing*-form, is difficult to draw and of questionable validity. Consequently, tagsets for English generally treat the *-ing* form as a unitary category. This is a language-specific value for English.

3.4.4 Mood

This applies only to finite verbs. The value "indicative" can remain implicit in the tagset, being the default or unmarked mood.

Attribute	*values*	Examples	Tags	RTags
Mood	*indicative*	eats	VV(Ind)PrS3	VVZ
	subjunctive	eat	VVSubPr	VVB
	imperative	eat	VVImp	VVB

3.4.5 Tense

Attribute	*values*	Examples	Tags	RTags
Tense	*present*	(they) eat	VV(Ind)PrP	VVB
	past	(they) ate	VV(Ind)Pa	VVD

3.4.6 Person

Attribute	*values*	Examples	Tags	RTags
Person	*1st person*	(I) am	VP(Ind)PrS1	VPM
	2nd person	(you) are	VP(Ind)PrS2	VPR
	3rd person	(she) is	VP(Ind)PrS3	VPZ

3.4.7 Number

Attribute	*values*	Examples	Tags	RTags
Number	*singular*	was (eaten)	VP(Ind)PaS	VPDZ
	plural	were (eaten)	VP(Ind)PaP	VPDR

3.4.8 Auxiliary Type

Attribute	*values*	Examples	Tags	RTags
Aux.Type	*Primary*	had (eaten)	VPPa	VPD
	Modal	would (eat)	VM	VM

The auxiliaries in English subdivide into the primary verbs *be, have*, and *do*, which can also function as main verbs, and the modal auxiliaries such as *can, will,* and *would*, which are uninflected, and always function as auxiliaries. These are language-specific values.

3.5 Application to Dutch

The thirteen Inflectional Verb features distinguished in the CELEX Wordforms lexical database are represented twice, once separately as a Yes or No code and once in combination, in a final composed 'Flection Type' tag containing those features which received a Yes tag in four columns. For example: the word form with final Flection Type tag 'te1s' received Yes four times, once for present tense (t), once for singular (e), once for first person (1) and once for separable word form (s).

3.5.1 Verb-Type

CELEX distinguishes four types of verbs. The first three (full verbs, modals and auxiliaries) are the same as in the German application; the fourth is the group of impersonal verbs which have only one (impersonal) subject, 'het'. Compare French: 'il pleut' (pleuvoir).

Attribute	*value*	Example	Tag
Verb-Type	*full verb*	afwassen	zelfst.
	auxiliary	hebben	hulp.
	modal	lijken	koppel.
	impersonal	regenen	onpers.

Note that many verbs are tagged with double or triple tags because they can have two or three values.

3.5.2 Verb-Form

The CELEX Wordforms Lexicon distinguishes between *infinitive, present participle, past participle, finite* verb forms and some other forms.

Attribute	*value*	Example	Tag
Verb-Form	*infinitive*	gaan	i
	pres participle	gaand	pt
	past participle	gegaan	pv
	finite	(hij) gaat	
	imperative	ga	g

See below for other participle word forms.

Finite verb forms are not marked as such, but the 14 present and past tense Flection Type tags which are distinguished are all finite verb forms.

As the degree of 'language-specificness' of the following five distinctions is not clear, they are not classified as such.

Participle word forms The CELEX Wordforms Lexicon distinguishes three Present participle word forms and seven Past participle word forms.

`pt`	`Present participle`	`lopend`
`ptE`	`Present participle with suffix  e`	`lopende`
`ptEm`	`Pres.part. with suffix  e , plural(de)`	`lopenden`
`pv`	`Past participle`	`gecoordineerd`
`pvC`	`Past participle Comparative`	`gecoordineerder`
`pvE`	`Past participle with suffix  e`	`gecoordineerde`
`pvEe`	`Idem of irregular verbs, singular`	`beschonkene`
`pvEm`	`Past participle with suffix e, plural`	`gedekoreerden`
`pvEs`	`Past part. with suffix e, separated`	`rond gezeilde`
`pvs`	`Past participle, separated`	`rond gezeild`

Only the Past and Present participle word forms 'pt', 'pv' and 'pvs' are verbal word forms. However, they can also function as adjectives (and thus as adverbs) or nouns. The other Participle word forms are only used as adjectives, adverbs or nouns, so they can be covered by adjectival, adverbial or nominal tags. From the perspective of lemmatization, however, it might be important to stress the verbal aspect of participles, since, as long as participles are not yet 'lexicalized' (accepted as real adjectives or nouns with a meaning of their own, differing from the meaning of the verb they are derived from) they do not have a lemma of their own, but are still to be lemmatized as words deriving from a verbal lemma.

The table below is a proposal for Dutch, not a CELEX table !
This table applies to the verb forms Infinitive, Present and Past participle.

Attribute	*value*	Example	Tag
Use (non-finite)	*verbal*	(hij heeft) geleerd	PartVerb
	adjectival	(hij is) geleerd	PartAdj
	nominal	Leren (motivert) (Het) geleerde	Part/Nom

3.5.3 Mood

The feature *mood* only applies to finite verb forms.

Attribute	*value*	Example	Tag
Mood	*indicative*	(hij) gaat	
	subjunctive	kome (wat komt)	a

The *Indicative* is not marked as such, most of the finite verb forms being indicative. Only the few subjunctive verb forms have a special tag: 'a'. In Dutch the subjunctive only occurs in the third person singular. Phrases containing subjunctive verb forms mostly belong to the written language and possess an archaic or idiomatic quality. The subjunctive generally occurs in the Present tense. Only one Past tense form remains (ware).

3.5.4 Tense

In the CELEX Wordforms Lexicon the feature Tense applies to finite verb forms and to participle verb forms. There are 10 different Present tense tags and 4 different Past tense tags. All of them are composed tags starting with 't' (Present time) or 'v' (Past time). See also the list of Present and Past participle tags above.

Attribute	*value*	Example	Tag
Tense	*present*	(ik) ga	t
	past	(ik) ging	v

3.5.5 Person

The three different person-values are only distinguished in the Present tense singular. Present tense plural has only one verb form for the 1st, 2nd and 3rd person.
The Past tense has one singular verb form and one plural verb form. The person-tags '1', '2' and '3' always appear in a composed Flection Type tag starting with: 'te' (present tense singular).

Attribute	*value*	Example	Tag
Person	*first*	(ik) ga	te**1**
	second	(jij) gaat ga (jij)	te**2** te**2I**
	third	(hij) gaat	te**3**

See above the feature *Inversion*.

3.5.6 Number

The number-tags 'e' and 'm' always appear in a composed Flection Type tag in combination with 't' (Present tense) or 'v' (Past tense).

Attribute	*value*	Example	Tag
Number	*singular*	(ik) ga	**te**
		(ik) ging	**ve**
	plural	(wij) gaan	**tm**
		(wij) gingen	**vmI**

3.5.7 Gender

The feature gender does not apply to Dutch verb forms. It occurs only in Present and Past participles used as nouns or as adjectives. In these cases gender will have to be determined from the context (Attribute value *context* of the Gender table in nouns).

3.5.8 Main-Verb Function

CELEX also distinguishes subcategorization values in full verbs (called 'lexical' verbs). The three possible values are intransitive, transitive and reflexive.

Attribute	*value*	Example	Tag
Main-Verb Funct.	*intransit*	vallen	intrans.
	transit	kopen	trans.
	reflex	vergissen	wederk.

3.5.9 Auxiliary

Another syntactical verb feature tagged in the Lemmas Lexicon is the type of auxiliary with which a verb is conjugated. There are three tags:

Attribute	*value*	Example	Tag
Auxiliary	*hebben*	doen	hebben
	zijn	groeien	zijn
	hebben or zijn	volgen	hebben/zijn

Note that many verbs are tagged with double or triple tags because they can have two or three values. This is the common Dutch dictionary practice to mark verb use.

3.5.10 Other Verb-Form features

In the CELEX Lemmas and Wordforms Lexica two further important verbal features have been tagged: *separability* and *inversion*. Another feature, *word order*, is distinguished in the other CELEX database mentioned above. In Dutch different word order brings about different word forms in the case of separable verbs.

3.5.11 Separability

This is tagged in the Syntactical part of the Lemmas Lexicon with a Yes or No tag and in the Wordforms Lexicon with an 's' in the composed final Flection Type tag:

Attribute	*value*	Example	Tag
Separability	*separable*	geef aan	te1**s**
		geeft aan	te2**s**
		geeft aan	te3**s**
		geven aan	tm**s**
		gaf aan	ve**s**
		gaven aan	vm**s**

Separable verbal word forms have a verbal and a non-verbal part and may occur at different places in the sentence. So another solution (in terms of corpus linguistics) would be to tag the verbal part as a separable verb form and the prepositional part as the non-verbal part of a separable verb, this last feature becoming a value of Adpositions.

3.5.12 Inversion

This occurs in interrogative phrases and only brings about a different word form in the Present tense singular, second person. It is an inflectional feature tagged in the Wordforms Lexicon with an 'I'.

Attribute	*value*	Example	Tag
Inversion	*inverted*	ga (jij)?	te2**I**
		geef (jij) aan	te2**Is**

te2**I**= present tense, singular, second person, inverted wordform
te2**Is**= present tense, singular, second person, inverted and separable wordform

The inversion tag is important from the point of view of disambiguation, since the inverted form of the second person singular Present tense is the same as the non-inverted form of the first person singular Present tense!

3.5.13 Word order separable verbs

This distinction comes from a CELEX database; the tag is proposed by the Dutch correspondent.

Attribute	*value*	Example	Tag
Word order - sep.verb	*main clause*	(hij) geeft aan	..MC
	sub-clause	(dat hij) aangeeft	..SC

Instances of separable verbs are written together in sub-clauses and written separately in main clauses!

3.5.14 Politeness

Politeness is not represented in the CELEX tag sets. In Dutch polite verb forms correspond to the 2nd person singular and plural. The personal pronoun expressing politeness is: 'U'.

3.6 Application to Spanish

3.6.1 Verb-Form and Mood

In Eurotra's terminology Verb-form corresponds to the attribute "e_mstype" and its organization is as follows:

Attribute	*Value*	Example	Tag
e_mstype	*finite*	cantamos	finite
	infinitive	cantar	infin
	gerund	cantando	gerund
	pastpart	cantado	pastpart

As for Mood ("e_mood" in Eurotra) we have three possible values:

Attribute	*Value*	Example	Tag
e_mood	*indicative*	canto	indicative
	subjunctive	cantase	subjunctive
	imperative	canta	imperative

It is not difficult to redistribute this organization so that it follows the strategy suggested in the present proposal:

Attribute	*value*		ex.	tag
Verb-Form	*finite*		-	-
	Attribute	*value*	example	tag
	Mood	*indicative*	canto	
		subjunctive	cantase	
		imperative	canta	
Verb-Form	*non-finite*		-	-
	Attribute	*value*	example	tag
	Mood	*infinitive*	cantar	
		gerund	cantando	
		pastpart	cantado	

3.6.2 Tense

Attribute	*Value*	Example	Tag
Tense	*present*	canto	
	past	canté	
	imperative	canta	
	future	cantaré	
	conditional	cantaría	

A controversial point here concerns the interpretation of conditionals as tenses, in contrast to how they are usually considered in traditional grammars. This treatment is based on the assumption that the conditionals have a temporal value that correponds to a "future of the past".

"Compound tenses", as in Italian, are formed with an auxiliary.

3.6.3 Person

Attribute	*Value*	Example	Tag
Person	*first*	(yo) como	
	second	(tu) comes	
	third	(el) come	

3.6.4 Number

Attribute	*Value*	Example	Tag
Number	*singular*	(yo) como	sing
	plural	(ellos) comen	plu

3.6.5 Gender

Attribute	*Value*	Example	Tag
Gender	*feminine*	cansada	fem
	masculine	cansado	masc

Gender is only pertinent to past participle forms.

3.7 Application to French (Corpus)

3.7.1 Type

The IBMF tagset has special tags for the two auxiliaries, *avoir* and *être*, otherwise similar (indication of person only) to the VERB tag used for "main" verbs. The distinction between the two auxiliaries would require a tagset-specific tag in EAGLES, or a special feature LU= indicating the exact lexical unit considered.

Attribute	*value*	Example	Tag
Type	*main*	lisait	VERB3
	auxiliary	avait, était	AUXA3, AUXE3

3.7.2 Finiteness and Verb-Form/Mood

As in the case of Italian, there is no specific encoding for the finiteness feature, which is self-contained in the mood information.

In French, the word forms allow for much of the disambiguation between moods, tenses, etc. When there is ambiguity (e.g. between present indicative and present subjunctive), the type of modelling involved with the IBMF tagset (i.e. stochastic tri-pos) would be unable to decide anyway. This is why the IBMF tagset does not distinguish all main moods/tenses, only Infinitive, Participle, and the rest (tag VERB).

Attribute	*value*	Example	Tag
Verb-Form	*indicative* *subjunctive* *imperative* *conditional*	(je) lis	**VERB**1 " " "
	infinitive	lire	VINF
	participle	lu	**P**PASMS
	gerund	-	-
	supine	-	-

3.7.3 Tense

Only in the case of participles is the present/past distinction made in the IBMF tagset.
Assuming we already have set Fin=*non-finite* and VFM=*part*:

Attribute	*value*	Example	Tag
Tense	*present*	lisant	P**PRE**
	past	lu	P**PAS**MS

3.7.4 Person

In the IBMF tagset, the person is numbered 1 to 6, thus avoiding the need for number information. The coding of an IBMF verb tag into the EAGLES scheme would therefore imply using both the Person and Number attributes.

Attribute	*value*	Example	Tag
Person	*1*	(je) vais, (nous) allons	VERB**1**, VERB**4**
	2	(tu) vas, (vous) allez	VERB**2**, VERB**5**
	3	(il) va, (ils) vont	VERB**3**, VERB**6**

3.7.5 Gender

In French, gender is only marked in the past participle.

Attribute	*value*	Example	Tag
Gender	*masculine*	fermés	PPASMP
	feminine	fermée	PPAS**F**S

3.7.6 Number

As mentioned for the **Person** attribute, the coding of an IBMF verb tag into the EAGLES scheme implies combining Person and Number attributes in EAGLES.
Otherwise, number is used only for Past participles in the IBMF tagset.

Attribute	*value*	Example	Tag
Number	*singular*	fermé	PPASM**S**
	plural	fermés	PPASM**P**

3.7.7 EAGLES features not applicable

Separability does not apply to French. **Voice** and **Reflexivity** are not marked as such in the verb form (they are built by adjunction of auxiliary or pronoun). **Auxiliary** would apply in French since all verbs build their compound forms with one of the two auxiliaries already mentioned.
These and other features (Main-verb fuction, auxiliary function) are not coded in the IBMF tagset.

3.7.8 IBMF Tagset features not applicable in EAGLES

The specification of the exact lexical unit for the auxiliary (tags AUXA AUXE, see above).

3.8 Application to French (Lexicon)

3.8.1 Type

Attribute	Value	Example	Code
Type	main	partir	m
	auxiliary	avoir	a

3.8.2 Finiteness

Finiteness is redundant with mood in French, as shown in the table below:

Verb Form/M.	Finiteness
indicative	finite
subjunctive	finite
imperative	finite
conditional	finite
infinitive	non-finite
participle	non-finite

We therefore decided not to encode it.

3.8.3 Verb Form/Mood

Attribute	Value	Example	Code
Verb Form/M.	indicative	viens	i
	subjunctive	vienne	s
	imperative	viens	m
	conditional	viendrais	c
	infinitive	venir	n
	participle	venu	p

3.8.4 Tense

Attribute	Value	Example	Code

Tense	present	viens	p
	imperfect	venais	i
	future	viendrai	f
	past	vins	s

3.8.5 Person

Attribute	Value	Example	Code
Person	first	suis	1
	second	es	2
	third	est	3

3.8.6 Number

Attribute	Value	Example	Code
Number	singular	viens	s
	plural	venons	p

3.8.7 Gender

Attribute	Value	Example	Code
Gender	masculine	venu	m
	feminine	venue	f

3.8.8 Combinations

Tag	Example
V1-s-ip-m	viens
V2-s-ip-m	viens
V3-s-ip-m	vient
V1-p-ip-m	venons
V2-p-ip-m	venez
V3-p-ip-m	viennent
V1-s-ii-m	venais
V2-s-ii-m	venais
V3-s-ii-m	venait
V1-p-ii-m	venions
V2-p-ii-m	veniez
V3-p-ii-m	venaient
V1-s-if-m	viendrai
V2-s-if-m	viendras
V3-s-if-m	viendra
V1-p-if-m	viendrons
V2-p-if-m	viendrez
V3-p-if-m	viendront
V1-s-is-m	vins
V2-s-is-m	vins
V3-s-is-m	vint
V1-p-is-m	vinmes
V2-p-is-m	vintes
V3-p-is-m	vinrent
V1-s-ip-a	suis, ai
V2-s-ip-a	es, as
V3-s-ip-a	est, a
V1-p-ip-a	sommes, avons
V2-p-ip-a	etes, avez
V3-p-ip-a	sont, ont
V1-s-ii-a	etais, avais
V2-s-ii-a	etais, avais
V3-s-ii-a	etait, avait
V1-p-ii-a	etions, avions
V2-p-ii-a	etiez, aviez
V3-p-ii-a	etaient, avaient
V1-s-if-a	serai, aurai
V2-s-if-a	seras, auras
V3-s-if-a	sera, aura
V1-p-if-a	serons, aurons
V2-p-if-a	serez, aurez
V3-p-if-a	seront, auront
V1-s-is-a	fus, eus
V2-s-is-a	fus, eus
V3-s-is-a	fut, eut
V1-p-is-a	fumes, eumes
V2-p-is-a	futes, eutes
V3-p-is-a	furent, eurent
V1-s-sp-m	finisse
V2-s-sp-m	finisse

V3-s-sp-m	finisse
V1-p-sp-m	finissions
V2-p-sp-m	finissiez
V3-p-sp-m	finissent
V1-s-si-m	finisse
V2-s-si-m	finisse
V3-s-si-m	finit
V1-p-si-m	finissions
V2-p-si-m	finissiez
V3-p-si-m	finissent
V1-s-sp-a	sois, aie
V2-s-sp-a	sois, aies
V3-s-sp-a	soit, ait
V1-p-sp-a	soyons, ayons
V2-p-sp-a	soyez, ayez
V3-p-sp-a	soient, aient
V1-s-si-a	fusse, eusse
V2-s-si-a	fusses, eusses
V3-s-si-a	fut, eut
V1-p-si-a	fussions, eussions
V2-p-si-a	fussiez, eussiez
V3-p-si-a	fussent, eussent
V2-s-mp-m	viens
V1-p-mp-m	venons
V2-p-mp-m	venez
V2-s-mp-a	sois, aie
V1-p-mp-a	soyons, ayons
V2-p-mp-a	soyez, ayez
V1-s-cp-m	viendrais
V2-s-cp-m	viendrais
V3-s-cp-m	viendrait
V1-p-cp-m	viendrions
V2-p-cp-m	viendriez
V3-p-cp-m	viendraient
V1-s-cp-a	serais, aurais
V2-s-cp-a	serais, aurais
V3-s-cp-a	serait, aurait
V1-p-cp-a	serions, aurions
V2-p-cp-a	seriez, auriez
V3-p-cp-a	seraient, auraient
V----n--m	venir
V----n--a	etre, avoir
V-ms-ps-m	venu
V-fs-ps-m	venue
V-mp-ps-m	venus
V-fp-ps-m	venues
V-ms-ps-a	eu
V-fs-ps-a	eue
V-mp-ps-a	eus
V-fp-ps-a	eues
V----pp-m	venant
V----pp-a	etant, ayant
V----ps-m	semble'
V----ps-a	ete'
---------	------------------------

Note:

We have decided to encode the past participle of the auxiliary verb être, the copulative verbs (sembler etc.) and the impersonal verbs (e.g. falloir), with the "not applicable" feature (-) instead of the "neutral" or "masculine" one. The notion of neutrality, though used for pronouns in the present model, is not a very traditional notion in French grammars and in any case the neutral feature does not bring any further information than the not applicable one. As for the masculine feature, syntactic considerations led us to drop it in order to maintain consistent agreement marks inside a sentence.

3.9 Application to Portuguese

The following set of morphological features are pertinent to Verbs in the Portuguese model GENELEX: Mood, Tense, Person, Gender, Number. The generation of the inflected forms of a given verb system is based on the various combinations of these features. Some features are not pertinent to the characterization of some wordforms, as we will try to explain below.

3.9.1 Type

The value *auxiliar* is not encoded in the morphological layer of the Portuguese model of GENELEX, but it is required on its syntactic layer, where it is applied to the verbs *ter* and *haver*. The value *auxiliar* is therefore pertinent to the Portuguese language.

3.9.2 Finiteness

This feature can be unproblematically inferred from the feature Mood, as the schema below clearly shows. The redundant application of these two features to Portuguese could be represented as follows:

Verb Form/Mood	Finiteness
indicative	finite
subjunctive	finite
imperative	finite
conditional	finite
infinitive	non-finite
participle	non-finite

The feature Finiteness is therefore not required by the Portuguese language.

3.9.3 Mood

From the values proposed for Mood, the following values are used in the Portuguese application:

Attribute	Value	Example	Tag
Mood	indicative	como	
	subjunctive	coma	
	imperative	come	
	conditional	comeria	
	infinitive	comer	
	participle	comido	
	supine		

3.9.4 Tense

Given that Plus-quam-Perfect is also a simple tense in the Portuguese language, we stress the need for integrating a specific value for this tense, so that the relevant simple wordforms can be handled in morphology.

Attribute	Value	Example	Tag
Tense	present	como	
	imperfect	comia	
	perfect	comi	
	future	comerei	
	past	comido	
l_spec	plus_quam_perfect	comera	

3.9.5 Person

The values proposed in the tables are pertinent to Portuguese. However, in order to cope with typical aspects of Politeness reflected in verbal inflection, the GENELEX model built for the Portuguese language made a particular use of this attribute - instead of a feature *person*, a decomposition of this is used: *person-deixis* and *person-conc*. Since in Portuguese polite verb forms correspond to the 3rd person, these two features allow an enhanced specification of verbal wordforms where the value ‘person=3’ could give rise to ambiguity (remark that ambiguity may be not solved by context, because in Portuguese you can drop the np subject.)
In the GENELEX application such verbal wordforms (e.g. *come* (‘eats’ or ‘eat’)) are thus marked in two ways: polite verbal wordforms (as in ‘voce’2 *come*’ (‘you eat’)) bear the combination of values ‘person-deixis=2’ and ‘person-conc=3’; true ‘person=3’ wordforms (as in ‘ele *come*’ (‘he eats’)) bear the combination of values ‘person-deixis=3’ and ‘person-conc=3’.

3.9.6 Gender

This feature is pertinent only to the Mood *participle*.

Attribute	Value	Example	Tag
Gender	masculine	comido	
	feminine	comida	

	neuter		
	common		

Only the values *masculine* and *feminine* are used.

3.9.7 Number

Attribute	Value	Example	Tag
Number	singular	como	
	plural	comemos	
	invariant		

This has only two values, *singular* and *plural*.

3.9.8 Combination of features

In the GENELEX application, Portuguese verbs have a maximum of 91 different inflected word-forms derived from different combinations of morphological features for simple morphological units (i.e. excluding compound forms).
They are the following:

	Example
01 Ind-Pr-1-1---S	como
02 Ind-Pr-2-2---S	comes
03 Ind-Pr-2-3---S	come
04 Ind-Pr-3-3---S	come
05 Ind-Pr-1-1---P	comemos
06 Ind-Pr-2-2---P	comeis
07 Ind-Pr-2-3---P	comem
08 Ind-Pr-3-3---P	comem
09 Ind-Im-1-1---S	comia
10 Ind-Im-2-2---S	comias
11 Ind-Im-2-3---S	comia
12 Ind-Im-3-3---S	comia
13 Ind-Im-1-1---P	comi’1amos
14 Ind-Im-2-2---P	comi’1eis
15 Ind-Im-2-3---P	comiam
16 Ind-Im-3-3---P	comiam
17 Ind-Fu-1-1---S	comerei
18 Ind-Fu-2-2---S	comera’1s
19 Ind-Fu-2-3---S	comera’1
20 Ind-Fu-3-3---S	comera’1
21 Ind-Fu-1-1---P	comeremos
22 Ind-Fu-2-2---P	comereis
23 Ind-Fu-2-3---P	comera’3o
24 Ind-Fu-3-3---P	comera’3o
25 Ind-Pa-1-1---S	comi
26 Ind-Pa-2-2---S	comeste
27 Ind-Pa-2-3---S	comeu
28 Ind-Pa-3-3---S	comeu
29 Ind-Pa-1-1---P	comemos
30 Ind-Pa-2-2---P	comestes
31 Ind-Pa-2-3---P	comeram
32 Ind-Pa-3-3---P	comeram
33 Ind-Pq-1-1---S	comera
34 Ind-Pq-2-2---S	comeras
35 Ind-Pq-2-3---S	comera
36 Ind-Pq-3-3---S	comera
37 Ind-Pq-1-1---P	come’2ramos
38 Ind-Pq-2-2---P	come’2reis
39 Ind-Pq-2-3---P	comeram
40 Ind-Pq-3-3---P	comeram
41 Con-Pr-1-1---S	comeria
42 Con-Pr-2-2---S	comerias
43 Con-Pr-2-3---S	comeria
44 Con-Pr-3-3---S	comeria
45 Con-Pr-1-1---P	comeri’1amos
46 Con-Pr-2-2---P	comeri’1eis
47 Con-Pr-2-3---P	comeriam
48 Con-Pr-3-3---P	comeriam
49 Sub-Pr-1-1---S	coma
50 Sub-Pr-2-2---S	comas
51 Sub-Pr-2-3---S	coma
52 Sub-Pr-3-3---S	coma
53 Sub-Pr-1-1---P	comamos
54 Sub-Pr-2-2---P	comais
55 Sub-Pr-2-3---P	comam
56 Sub-Pr-3-3---P	comam
57 Sub-Im-1-1---S	comesse
58 Sub-Im-2-2---S	comesses
59 Sub-Im-2-3---S	comesse

```
60 Sub-Im-3-3---S    comesse
61 Sub-Im-1-1---P    come'2ssemos
62 Sub-Im-2-2---P    come'2sseis
63 Sub-Im-2-3---P    comessem
64 Sub-Im-3-3---P    comessem
65 Sub-Fu-1-1---S    comer
66 Sub-Fu-2-2---S    comeres
67 Sub-Fu-2-3---S    comer
68 Sub-Fu-3-3---S    comer
69 Sub-Fu-1-1---P    comermos
70 Sub-Fu-2-2---P    comerdes
71 Sub-Fu-2-3---P    comerem
72 Sub-Fu-3-3---P    comerem
73 Imp----2-2---S    come
74 Imp----2-3---S    coma
75 Imp----1-1---P    comamos
76 Imp----2-2---P    comei
77 Imp----2-3---P    comam
78 Inf----1-1---S    comer
79 Inf----2-2---S    comeres
80 Inf----2-3---S    comer
81 Inf----3-3---S    comer
82 Inf----1-1---P    comermos
83 Inf----2-2---P    comerdes
84 Inf----2-3---P    comerem
85 Inf----3-3---P    comerem
86 Inf-----------    comer
87 Par-Ps-----M-S    comido
88 Par-Ps-----F-S    comida
89 Par-Ps-----M-P    comidos
90 Par-Ps-----F-P    comidas
91 Ger-----------    comendo
```

We stress that there is no need to define any specification labelling the typically Portuguese 'inflected infinitive'. In fact, it is suitably handled by combining 'mood=infinitive' with the values of attributes 'number' and 'person' (the last is decomposed in the Portuguese GENELEX model, as we explained under the subsection 'Person' above).

3.9.9 Clitics

This feature is pertinent to Portuguese verbs, but the assignment of a feature concerning clitics to verbs depends on the syntactic classification of the verb.
For the time being, verbs are not yet encoded wrt clitics in the Portuguese GENELEX morphological layer.

3.10 Application to Danish

In Danish, the verbal system comprises a very limited number of inflectional forms. There is no morphological distinction for person and number; moods and tenses are limited as well. Thus, the set of morphological features pertinent to Danish verbs are: Type, Tense, Mood (distinction between indicative and imperative only) and Voice (distinction of active and -s passive).

3.10.1 Type

Attribute	*value*	Example	Tag
Type	*Main*	skriver	vb_**mainv**
	Auxiliary	har (skrevet)	vb_**aux**
	Passive auxiliary	bliver (skrevet)	vb_**auxpass**
	Modal	skulle (skrive)	vb_**modv**

The type 'passive auxiliary' can be regarded as a language-specific feature; in Danish passive forms can be generated by means of the morphological suffix '-(e)s' or by means of the passive auxiliaries 'blive' and 'vaere', respectively. This feature is used within the EDEMD.

3.10.2 Finiteness

In general, the finiteness is implicitly given by the particular tag of the verb-form occuring in the corpus, e.g. 'skriver' is indicative present active which is a finite form. Thus, in corpus tagging the distinction finite/non-finite is not needed. However, in the Eurotra description of Danish finite and non-finite forms have been distinguished.

Attribute	*value*	Example	Tag
Finiteness	*finite*	skriver	vb_**pres_act**
	non-finite	skrivende	vb_**pres_ptc**

3.10.3 Verb-form

Here we follow the tagging for English: the attribute Verb-form applies to non-finite verb forms, Mood applies to finite verb forms only.

Attribute	*value*	Example	Tag
Verb-form	*infinitive*	skrive	vb_**inf**
	perfect participle	skrevet	vb_**pf.ptc**
	present participle	skrivende	vb_**pres.ptc**

A perfect participle remains uninflected when used in verbal function; in adjectival (attributive) function perfect participle forms are inflected like adjectives (without inflectional comparison).

3.10.4 Mood

In corpus tagging the mood 'indicative' will be left unmarked, because this is the most frequently used - and thus the default - mood in texts. In corpora a few archaic examples of other moods do occur (e.g. subjunctive) but they are not taken into account here because of their very low frequency.

Attribute	*value*	Example	Tag
Mood	*indicative*	(vi) skriver	vb_pres_**ind**
	imperative	skriv	vb_**imp**

3.10.5 Tense

Compound tenses (e.g. auxiliary + past participle for past tenses) are not included.

Attribute	*value*	Example	Tag
Tense	*present*	(de) skriver	vb_**pres**
	past	(du) skrev	vb_**past**

3.10.6 Number

In modern Danish, there is no difference between singular and plural verb-forms.

3.10.7 Person

In modern Danish, there is full syncretism within the person paradigm of verb-forms, i.e. one single morphological form covers all Person + Number combinations of a given tense/mood. Therefore, the attributes Number and Person are not applied to Danish verbs. In a corpus a few archaic forms of the plural may occur.

3.10.8 Gender

Danish verbs have no Gender distinction.

3.10.9 language-specific feature: Voice

In Danish, there is a morphological difference between active and passive forms in the present and the past. (The inflectional ending of the passive is always **-(e)s**.) Passive forms can also be composed by means of an auxiliary (**auxpass**) + perfect participle. Perfect participle forms of a few verbs may also occur as -(e)s passive forms.

Attribute	*value*	Example	Tag
Voice	*active*	skriver	vb_pres_**act**
	-s passive	skrives	vb_pres_**s-pas**

3.11 Application to Greek

Potentially, Greek verbal entries are characterised for the following features: Finiteness, Verb-form/Mood, Tense, Person, Number, Gender, Aspect, Voice.

Depending on the combination of these features, all inflected forms of the Greek verbal system can be generated. The combination of the features is dependent upon constraints, resulting, in certain cases, in the total absence of certain features or the exclusion of certain values. More information on these constraints is given in the following sections.

The Greek verbal system presents both simple and compound forms. Compound forms are the result of two different combinations:

- two verbs, the auxiliary and the main verb (the main verb is in the “infinitival” or past participial form) for the formation of certain tenses,

- a particle and the verb, for the formation of the subjunctive mood and the conditional tenses.

In the following tables, compound forms are put in parentheses, and no tag is given for them as their recognition would require a multi-word tagger.

3.11.1 Type

This feature is not currently coded in the Greek Morphological Lexicon. However, it can be applied to the Greek language. In this case, the table of the attribute-value set could take the following form:

Attribute	*value*	Gr. example	Gr. tag
Type	*main*	ghrafw	-
	aux	ehw	-
	cop	eimai	-

The verb “eimai” in Greek acts as a main, copular or auxiliary verb. In the majority of the cases, the linguistic context contributes to the resolution of the ambiguity:

To vivlio einai **panw** *sto trapezi* (prep. - main)
To vivlio einai **teleiwmeno** (past. part. - aux)
To vivlio einai **katharo** (adj. - cop.)

The verb “ehw” acts both as an auxiliary and as a main verb. Again, the linguistic context serves for disambiguation purposes:

O Ghiannys ehei **spiti** (noun - main)
O Ghiannys ehei **teleiwsei** (inf. - aux)

3.11.2 Finiteness and Verb-Form/Mood

In its present form, the Greek Morphological Lexicon does not code the feature **Finiteness**, while the features *Verb-Form* and *Mood* are kept distinct, the first one taking as values *finite* and *participle*, and the latter *indicative* and *imperative*. However, a combination of these two features gives information on finiteness. Apart from participles, all other simple forms in Greek are finite and may take the values *indicative* and *imperative* as regards mood.

(For the formation of the subjunctive mood a compound form is used, as presented in the following table.)

The application of the features *finiteness* and *verb-form/mood* to Greek can be represented as follows:

Attribute	*value*		Gr. ex.	Gr. tag
Finiteness	*non-finite*		-	-
	Attribute	*value*	Gr. example	Gr. tag
		participle	eisaghomenos	Vb**Pp**
Finiteness	*finite*		-	-
	Attribute	*value*	Gr. example	Gr. tag
	Verb-Form/Mood	*indicative*	eisaghw	VbFi**Id**
		imperative	ghrapse	VbFi**Mp**
		subjunctive	(na ghrapsw)	-

Although there exists one “infinitival” form in Greek, it is not used as a value, given that it is only used for the generation of certain compound tenses and cannot be found on its own; this form is morphologically the same as the third person singular used for the formation of the future tense or the subjunctive mood (again formed periphrastically, with the combination of a particle):

O Ghiannys tha **teleiwsei** (future - untns 3rd sing.)
O Ghiannys ehei **teleiwsei** (pres.perf. - inf.)

3.11.3 Tense

The tense system in Greek is traditionally perceived as similar to the one described for Romance languages. However, in the current codification system we have adopted a system more similar to the English one. Thus the system is based on a combination of the *Tense* and *Aspect* features (for *Aspect*, see relevant section). Further constraints are applied to this attribute by the value of the mood. The tense system in Greek includes both simple and compound forms.

Attribute	*value*	Gr. example	Gr. tag
Tense	*pres*	ghrafw	VbFiId**Pr**
	past	eghrapsa	VbFiId**Pa**
	fut	(tha ghrapsw)	-
l-spec	*untns*	ghrapsei	VbFiId**Un**

The value *untns*, specific to Greek, is used for the codification in the Lexicon of the “infinitival” form of the verb, a form which, as already presented, never appears on its own but is used for the formation of compound tenses. It is also used for the form that combined with the appropriate particle forms the future tense and the subjunctive mood.

3.11.4 Person

Attribute	*value*	Gr. example	Gr. tag
Person	1	ghrafw	VbFiIdPr**01**
	2	ghrafeis	VbFiIdPr**02**
	3	ghrafei	VbFiIdPr**03**

For the codification of the “infinitival” form (e.g. *ghrapsei*), the value of Person is left unspecified, given that the person of the compound form is provided by the auxiliary verb:

*O Ghiannys eh***ei** *fughei.* - 3rd person
*Eh***w** *fughei.* - 1st person

The value is left unspecified for the participial forms as well.

3.11.5 Number

This has only two values, *singular* and *plural*.

Attribute	*value*	Gr. example	Gr. tag
Number	*singular*	ghrafw	VbFiIdPr01**Sg**
	plural	ghrafoume	VbFiIdPr01**Pl**
	invariant	-	-

The attribute of Number is left unspecified for the “infinitival” form, because the number of the compound form is marked on the auxiliary verb:

*O Ghiannys eh***ei** *fughei.* - singular
*Oloi eh***oun** *fughei.* - plural

3.11.6 Gender

This feature applies only when the value of finiteness is marked as *participle*. Participles in Greek behave as adjectives, and thus agree in gender, number and case with the nouns they modify. Therefore, apart from the value *masc-fem*, which does not apply here, all values are those presented in the section on nouns, and are shown in the following table:

Attribute	*value*	Gr. example	Gr. tag
Gender	*masculine*	eisaghwn (act. part.) eisaghomenos (pass.pr. part) eisahtheis (pass.past. part)	VbPpPrSg**Ma** VbPpPrSg**Ma** VbPPPaSg**Ma**
	feminine	eisaghousa (act. part.) eisaghomeny (pass. pr. part.) eisahtheisa (pass.past part.)	VbPpPrSg**Fe** VbPpPrSg**Fe** VbPpPaSg**Fe**
	neuter	eisaghon (act. part.) eisaghomeno (pass. pr. part.) eisahthen (pass. past part.)	VbPpPrSg**Ne** VbPpPrSg**Ne** VbPpPaSg**Ne**

3.11.7 Aspect

This attribute is used in combination with the attribute of Tense for the appropriate and unique characterisation of tenses of all forms (simple and compound) in Greek.

Attribute	*value*	Gr. example	Gr. tag
Aspect	*imperf*	eghrafa	VbFiIdPa01Sg**Im**
	perf	eghrapsa	VbFiIdPa01Pl**Pe**

Below, we give a table showing the formation of various tenses of the indicative mood with the morphological features of tense and aspect. Compound forms are in parentheses, and no tags are given for them as their tagging requires the use of a multi-word tagger.

Attribute	**Attribute**	Gr. example	Gr. tag
Tense	**Aspect**		
pres	*imperf*	ghrafw	VbFiIdPr01SgIm
past	*imperf*	eghrafa	VbFiIdPa01SgIm
past	*perf*	eghrapsa	VbFiIdPa01SgPe
pres	*perf*	(ehw ghrapsei)	-
fut	*imperf*	(tha ghrafw)	-
fut	*perf*	(tha ghrapsw)	-

3.11.8 Voice

The Greek language recognises two values for the attribute of voice, the same as proposed by EAGLES at Level 2a.

Attribute	*value*	Gr. example	Gr. tag
Voice	*active*	ghrafw	VbFiIdPr01SgIm**Ac**
	passive	ghrafomai	VbFiIdPa01PlPe**Ps**

4 Adjective

A	Type	Degree	Gen	Num	Case	Use	Mod.T	Infl.T	Defin	Pos	Prs
Mult		pos comp sup	m f n	s p	nom gen dat acc voc	attr pred advb nomn	pren postn	det indt			
Genelex	qualf dem poss ord card indf inter excl	comp+ comp– comp= sup+ sup– supabs	m f	s p							
Aleth	indf dem poss ord card		m f	s p						sg pl	1 2 3
NERC		pos comp sup	m f mf	s p sp	nom gen dat acc bas						
Leech		pos comp sup	m f n c	s p	nom gen dat acc bas	attr pred prem postm weak strg					
L0	**ADJECTIVE**										
L1	qual poss indf card ord	pos comp sup	m f n	s p	nom gen dat acc						
L2a		comp+ comp– sup+ sup– supabs				attr pred	prem postm				
L2b			It c Sp c	It n Sp c	Gr voc Gr ind			Ge wek Ge str Ge mxd	Da def Da indf Da unmk		

4.1 Comments

A large core of agreement emerges as to the Adjective category.

4.1.1 Type

The GENELEX and AlethDic models reflect a distinction which is also made in traditional grammars of other Romance languages (e.g. Italian), i.e. a first disjunction between the so-called ‘qualificatif’ and ‘indicatif’ Adjectives.
The latter are those Adjectives which also have pronominal function, and in different grammatical traditions are called Determiners in their adjectival function. The former are all the other Adjectives.
Indicative Adjectives are further divided into possessives, demonstratives, relatives, indefinites, numerals (ordinals and cardinals), interrogatives, exclamatories.
Moreover, the GENELEX model distinguishes 3 possible functions, i.e.:

(i) le chien est nôtre
(ii) nôtre chien
(iii) le nôtre

as (i) Adjective, (ii) Determiner, (iii) Pronoun respectively.

In a preceding version of this proposal we did not propose the feature Type at the recommended level and we had suggested the inclusion of indicative adjectives among Pronouns and/or Determiners, in order to make possible a comparison of the Romance and English traditions. However, after the first cycle of tests and practical applications carried out in the framework of the MULTEXT project, the French partners strongly recommended having the possibility of marking the three possible functions of e.g. possessives, and hence their inclusion also in the Adjective category.
Furthermore, the majority of the partners have been in favour of the inclusion in this feature of values such as ‘cardinal’ and ‘ordinal’.
Hence, the values proposed for **Type** to be marked at common level are: ‘qualificative’, ‘possessive’, ‘ordinal’, ’cardinal’ and ‘indefinite’.

The fact of having indicative adjectives motivates the presence in this table the of features such as **Number of Possessor** and **Person**.

Each language-specific application should specify clearly in which category the indicative adjectives are treated in order for cross-linguistic comparisons to be made possible.

4.1.2 Degree

GENELEX only foresees, in addition to the values commonly agreed for this attribute, further specifications such as ‘comp+’, ‘comp-’, ‘comp=’ (*comparatif-superiorite, comparatif-inferiorite, comparatif-egalite* ‘sup+’, ‘sup-’, ‘subabs’, (*superlatif-superiorite, superlatif- inferiorite, superalltif-absolu* (GENELEX, Sept. 1993), which are proposed here at level 2a, given that they are relevant for many languages.

4.1.3 Gender, Number and Case

These are arranged in the same way as for Nouns.

4.1.4 Use

This is an attribute introduced by MULTILEX among the syntactic specifications, to specify how an adjective can be used:
- ‘attributive’: the adjective modifies a noun inside an NP.
- ‘predicative’: the adjective can be used as a subject complement of a copular verb, as an object complement or as a secondary predicate.
- ‘adverbial’: the adjective modifies a verb or a VP.
- ‘nominalized’: the adjective can be used attributively in an NP without head (Dutch-dependent).

The value here is not to intended to specify whether a noun can be derived from an adjective.

In the EAGLES rows only the first two are proposed in the common core, Level-2a: they are marked in German and in Dutch.

4.1.5 Modification Type

This attribute, called Order in MULTILEX, specifies whether an attributive adjective precedes or follows the noun. The default value can change depending on languages: ‘prenominal’ for English, ‘postnominal’ for Romance languages. The different position often determines distinctions in sense (see section on Italian).

These values are very important in tagging.

In the Leech/Wilson proposal, the values of **Uses** and **Modification Type** are collapsed in the attribute **Uses** (in their tagset called **Position**): the two values ‘weak’ and ‘strong’ are added for German adjective inflection. In the EAGLES proposal these two values are put under the attribute **Inflection Type**, with the value ‘mixed’ for dealing with the two possibilities.
All these values are only relevant for tagging, but are left in the Lexicon proposal as an example of something one might want to record, e.g. in frequency lexicons.

4.1.6 Inflection Type

This feature is given for languages such as German.

4.1.7 Definiteness

This feature is introduced with the values 'definite', 'indefinite', 'unmark' for Danish.

4.2 Application to Italian

In Italian, the Adjective agrees in Number and Gender with the noun to which it refers. Another pertinent feature is Degree.

4.2.1 Type

Two types are distinguished: 'qualificatives' and 'determinatives'. The former are tagged 'A'; the latter constitute the class of Pronominal Adjectives (see the sections Pronouns and Determiners) and are recognized by the tag 'D'. In the Italian Corpus and Dictionary the tag **A** contains, therefore, the value 'qualificative' by default.

4.2.2 Degree

The default value is 'positive'; adjectives can also have comparative and superlative degree.

– "Analytical forms":
The comparatives are, in general, expressed by analytical forms: *il piu' bravo, il meno onesto, alto quanto me.*

The relative superlative indicates the highest or the least degree of a quality in relation to something (humans or things): *il piu' bravo della classe.*

These analytical forms present encoding problems from a corpus perspective: they are not dealt with in corpus tagging practice with a word-by-word approach, since they are constructed with more than one word. They belong to the set of phenomena to be codified with the strategy of multi-word expression tagging (see Leech and Wilson Invitation Draft).

–"Synthetic forms".
The absolute superlative is constructed by adding the suffixes *-issimo, -errimo, -entissimo* or the prefixes *super-, extra-, iper-* etc. to the stem of the adjective: *dolcissimo, acerrimo, munificentissimo, ipercritico, ultrarapido.*

There are a restricted number of adjectives which have the so-called 'organic' comparative and superlative, which are considered as exceptions, e.g. *maggiore, migliore, ...*

Attribute	*value*	It. example	It. tag
Degree	*positive*	grande (uomo)	A/ms
	comparative	(fratello) maggiore	A/ms**c**
	superlative	massimo (dolore) grandissimo (dolore)	A/ms**s**

4.2.3 Gender and Number

Within this category, two groups are recognized:
- I group: *ver-o/-a, ver-i/-e.*
- II group: *dolce, dolci.*

This second group is given the value *n* (common) for gender in the lexicon, whereas in the corpus it can sometimes be disambiguated by the gender of the noun (see under nouns).

```
amico vero, amiche vere       A/ms   A/fp
parola dolce, biscotto dolce  A/fs   A/ms
insegnante capace             A/ns
```

Some adjectives are completely invariable: they have the tag *nn* for Gender and Number in lexicon. Adjectives derived from Adverbs (*dappoco, dabbene*, etc.) belong to this group.

Attribute	*value*	It. example	It. tag
Gender	*masculine*	vero	A/**m**
	feminine	vera	A/**f**
	neuter		
l-spec	*common*	dolce	S/**n**

Attribute	*value*	It. example	It. tag
Number	*singular*	caro	A/m**s**
	plural	cari	A/m**p**
l-spec	*invariant*	pari	A/n**n**

4.2.4 Use

Adjectives can have 'attributive' and 'predicative' value: these are not encoded in corpus tagging practice, but they are clearly retrievable from the observation of category sequences in a tagged corpus.

However, these two different uses influence the agreement in Number if the adjective refers to more than one noun (Gender is assigned by the Gender of the nouns: if the coordinated nouns are masculine and feminine the adjective selects the value masculine):
- predicative use: singular nouns with masculine and feminine gender, *mio figlio e mia figlia sono studios***i** (Gend: masc.; Numb: plur.).
- in attributive use, the agreement in gender is not very precise: with singular nouns which are very similar in meaning, the adjective can agree with the closest noun, i.e. it can be singular: *un carattere e una condotta onest***a**.

4.2.5 Modification Type

This information is not marked (but as already pointed out above, it is extractable from sequences).

In prenominal position the adjective looses intensity:

un amico caro
un caro amico

Always depending on position, some adjectives also change sense:

un uomo povero (i.e. a not rich man)
un pover'uomo (i.e. a man to be sympathyzed)

4.2.6 Features not applicable in Italian

Case, Inflection Type and Definiteness are not applicable to Italian adjectives.

4.3 Application to German

4.3.1 Degree of comparison

Attribute	*value*	example	tag
Degree	*positive*	(das) große (Haus) (es ist) groß	ADJA:**Pos**.Neut.Nom.Sg.St ADJNA:**Pos**
	comparative	(das) größere (Haus) (es ist) größer	ADJA:**Comp**.Neut.Nom.Sg.St ADJNA:**Comp**
	superlative	(das) größte (Haus) (es ist am) größten	ADJA:**Sup**.Neut.Nom.Sg.St ADJNA:**Sup**

4.3.2 Gender

The feature *gender* applies only to attributive adjectives.

Attribute	*value*	example	tag
Gender	*masculine*	(ein) großer (Mann)	ADJA:Pos.**Masc**.Nom.Sg.Mix
	feminine	(eine) große (Frau)	ADJA:Pos.**Fem**.Nom.Sg.Mix
	neuter	(ein) großes (Haus)	ADJA:Pos.**Neut**.Nom.Sg.Mix

4.3.3 Number

The feature *number* applies only to attributive adjectives.

Attribute	*value*	example	tag
Number	*singular*	(der) große (Mann)	ADJA:Pos.Masc.Nom.**Sg**.Sw
	plural	(die) großen (Männer)	ADJA:Pos.Masc.Nom.**Pl**.Sw

4.3.4 Case

The feature *case* applies only to attributive adjectives.

Attribute	*value*	example	tag
Case	*nominative*	(der) große (Mann)	ADJA:Pos.Masc.**Nom**.Sg.Sw
	genitive	(des) großen (Mannes)	ADJA:Pos.Masc.**Gen**.Sg.Sw
	dative	(dem) großen (Manne)	ADJA:Pos.Masc.**Dat**.Sg.Sw
	accusative	(den) großen (Mann)	ADJA:Pos.Masc.**Akk**.Sg.Sw

4.3.5 Use

The IMS-Tagset considers only the *attributive* and *non-attributive* use of adjectives. Adjectives which are used predicatively or as adverbs do not differ in their inflectional shape and are therefore not distinguished. Nominalized adjectives are annotated as common nouns[8].

Attribute	*value*	example	tag
Use	*attributive*	(ein) schnelles (Auto)	ADJ**A**:Pos.Neut.Nom.Sg.Mix
	non-attributive	(es ist) schnell (es fährt) schnell	ADJ**NA** ADJ**NA**

4.3.6 Inflection

The feature *inflection* applies only to attributive adjectives.

Attribute	*value*	example	tag
Inflection	*strong*	(welch) großer (Mann)	ADJA:Pos.Masc.Nom.Sg.**St**
	weak	(der) große (Mann)	ADJA:Pos.Masc.Nom.Sg.**St**
	mixed	(ein) großer (Mann)	ADJA:Pos.Masc.Nom.Sg.**Mix**

[8] cf. feature *inflection* of adjectives (section 4.3.6)

4.4 Application to English

4.4.1 Degree

Attribute	*values*	Examples	Tags
Degree	*positive*	big	AJ
	comparative	bigger	AJR
	superlative	biggest	AJT

There are no distinctions of Gender, Number or Case among English adjectives. Most adjectives which are disyllabic or polysyllabic, (e.g. *dreadful, beautiful*) are invariable, and do not take inflection for degree. These adjectives form comparison by the use of the adverbs *more* and *most*. There are many exceptions, however, including in particular disyllabic adjectives ending in *-y*, for which both the inflectional and the non-inflectional comparison are acceptable.

4.5 Application to Dutch

4.5.1 Degree of comparison

These attribute values also apply to Dutch. In the Lemmas Lexicon the degree of comparison is marked with Yes tags for Positive, Comparative and Superlative. The tags in the table below, however, are single or composed Flection Type tags from the Wordforms tag set. The single tags mark non-inflected forms, the composed tags inflected adjectival forms. Inflection is expressed by the 'E' tag, which means: 'with suffix 'e''.

Attribute	*value*	Example	Tag
Degree	*positive*	(een) groot (huis)	P
		(het) groot (huis)	PE
	comparative	(een) groter (huis)	C
		(het) grotere (huis)	CE
	superlative	(dat huis is het) grootst	S
		(het) grootste (huis)	SE

CELEX also attributes the comparative and superlative tags 'C' and 'S' to some adverbs: 'vaker', 'vaakst'.

4.5.2 Gender

Not applicable in Dutch. (See the contextual Gender tag for Adjectives in nominal use.)

4.5.3 Number

Not applicable in Dutch.

4.5.4 Case

Not applicable in Dutch.

4.5.5 Use

Concerning the attributive–non-attributive use of adjectives, CELEX does not have such tags. But if they existed, they would not be distinctive enough because, according to Dutch grammar, an adjective used predicatively (non-attributively) can still be an adjective or an adverb. In the German non-attributive examples: 'er ist schnell' schnell would be an adjective, but in: 'es fährt schnell' schnell would be an adverb according to Dutch grammar. CELEX, however, makes a distinction, which is more pertinent to Dutch, between the adverbial and non-adverbial use of adjectives.

The examples in the table below are not taken from the *Dutch Linguistic Guide*, because they are lacking.

Attribute	*value*	Example	Tag
Use	*adverbial*	(hij rijdt) snel	adv
	non-adverbial	(een) snelle (auto)	nonadv

This can be an L2 tag.

However, a more complete usage table for Dutch Adjectives would be the MULTILEX tag set, presented in the synoptical table of Adjectives. This tagset distinguishes between: Attributive, Predicative, Adverbial and Nominal use:

The table below is a proposal for Dutch, not a CELEX table!

Attribute	*value*	Example	Tag
Use	*attributive*	(de) snelle auto	AdjAttrib
	predicative	(de auto is) snel	AdjPred
	adverbial	(de auto rijdt) snel	AdjAdv
	nominal	(de/het)snelle	AdjNom

4.5.6 Inflection

There is no distinction in Dutch between strong, weak and mixed inflection as exists in German. But the Inflection values *determined* and *indetermined*, as proposed by MULTILEX, are applicable to Dutch, since the inflection of Dutch adjectival word forms partly depends on whether the article preceding the Adjective is definite or indefinite. A Dutch adjective, used *attributively*, only has no inflection when preceded by an indefinite article or pronoun and followed by a neuter noun. In all other cases it has inflected forms. Used *predicatively*, adjectives are always without inflection (with the exception of a certain use of superlatives: 'Zijn huis is het mooiste' !).
However there is no such Inflection value tag in the CELEX tag set, probably because in the Wordforms Lexicon the P and PE tags of Positive and Positive with suffix 'e' cover this domain (see the Degree table).

The table below is a proposal for Dutch, not a CELEX table!

Attribute	*value*	Example	Tag
Inflection	*determined*	(het) groote (huis)	AdjDet
	indetermined	(een) groot (huis)	AdjIndet

This can be an L2 tag.

4.5.7 Language-specific feature: non-verbal part of separable verb

We need a special tag to mark adjectives which are part of a separable verb: 'fijn' in fijnhakken and 'dwars' in dwarsliggen etc.

4.6 Application to Spanish

4.6.1 Type

Adjectives are not subtyped in ET-ES dictionaries.

4.6.2 Degree

Attribute	*value*	Example	Tag
Degree	*positive*	bueno	
	comparative	mejor	
	superlative	buenísimo	

In Spanish most comparatives are expressed by analytical forms; however, we have a few "organic" comparative such as "mejor" or "mayor". Only a small group of adjectives can take superlative suffixes the *ísimo, érrimo: eg. interesantísimo celebírrimo.*

4.6.3 Gender

Attribute	*value*	Example	Tag
Gender	*m*	bueno	masc
	f	buena	fem
	c	responsable	

There are no neuter adjectives in Spanish. We have adjectives which do not inflect for gender (eg. *responsable*). Just as in the case of nouns, in the Eurotra dictionaries we leave them unvalued for the attribute Gender. We can, however, add an extra language-specific value (L2b) "common" for these cases.

4.6.4 Number

Attribute	*value*	Example	Tag
Number	*s*	buena	sing
	p	buenas	plu
	i	gratis	inv

Most Spanish adjectives inflect for number. Only a small group of them are common or invariant.

4.6.5 Modification Type

Attribute	*value*	Example	Tag
Mod.T.	*premod*	un buen libro	adjpos=pren
	postmod	un libro bueno	adjpos=postn
	indif	largo	adjpos=none

Some Spanish adjectives are restricted wrt order (in some cases order determines the meaning of adjectives: *un pobre hombre vs. un hombre pobre*). We have adjectives which:

a) can only occur as postmodifiers: "una entrada accesible"
b) can only occur as premodifiers: "un buen/gran hombre" (most of these adjectives have their postmodifier form: "un hombre bueno/grande", (apocope)).
c) can occur as both pre- and postmodifiers: "un (difícil) trabajo (difícil).

These facts are coded in ET-ES grammars under the feature "adjpos".

4.7 Application to French (Corpus)

4.7.1 Type

As indicated before, French has many more adjective types than just qualificative. These will be dealt with in the pronoun-determiner section.

4.7.2 Gender

Attribute	*value*	Example	Tag
Gender	*masculine*	petit	ADJEMS
	feminine	petite	ADJE**F**S

4.7.3 Number

Attribute	*value*	Example	Tag
Number	*singular*	grand	ADJEM**S**
	plural	grands	ADJEM**P**

4.7.4 EAGLES features not applicable

Use: As mentioned for other languages, attributive/predicative use is a syntactic distinction which applies to French. Since all adjectives can be used either in attributive or in predicative position and would not allow for other disambiguations, the distinction is not relevant in a tagset.
Degree applies to French, but is marked by external premodifiers and therefore does not require a special adjective class in the tagset.
Modification type applies to French; in some specific cases, the position even implies a semantic distinction: e.g. *un grand homme* (a great man) vs. *un homme grand* (a tall man). Although the distinction could be of interest in a tagset for its prediction potential, it is not used in the IBMF tagset.
Case and **inflection type** do not apply to French. **Definiteness** does not apply, although it applies for other types of French ‘adjectives” (see determiners).

4.7.5 IBMF Tagset features not applicable in EAGLES

The tagset has a special feature for indefinite adjectives. This will be put in the pronoun-determiner class.

4.8 Application to French (Lexicon)

4.8.1 Type

Attribute	Value	Example	Code
Type	qualificat.	bon	f
	ordinal	deuxi'eme	o
	cardinal	deux	c
	indefinite	quelconque	i
	possessive	mien	s

4.8.2 Degree

Attribute	Value	Example	Code
Degree	positive	bon	p
	comparative	meilleur	c

Note:

The distinction positive/comparative applies only to two adjectives in French: bon and mauvais. All other adjectives form their comparatives with *plus* + adjective (e.g., *plus grand*). Superlative is also a compound form (*le* + comparative, e.g. *le plus grand*).

4.8.3 Gender

Attribute	Value	Example	Code
Gender	masculine	bon	m
	feminine	bonne	f

4.8.4 Number

Attribute	Value	Example	Code
Number	singular	bon	s
	plural	bons	p

4.8.5 Case

Not applicable to French.

4.8.6 Combinations

Tag	Example

Lexique	Corpus	Example
Afcfp-	AFP	meilleures
Afcfs-	AFS	meilleure
Afcmp-	AMP	meilleurs
Afcms-	AMS	meilleur
Afpfp-	AFP	bonnes
Afpfs-	AFS	bonne
Afpmp-	AMP	bons
Afpms-	AMS	bon
Ai-fp-	AFP	certaines, memes, quelconques
Ai-fs-	AFS	certane, meme, quelconque
Ai-mp-	AMP	certain, memes, quelconques
Ai-ms-	AMS	certain, meme, quelconque
Ac-fp-	AFP	deux
Ac-fs-	AFS	une
Ac-mp-	AMP	deux
Ac-ms-	AMS	un
Ao-fp-	AFP	premieres
Ao-fs-	AFS	premiere
Ao-mp-	AMP	premiers
Ao-ms-	AMS	premier
As-fp-	AFP	leurs, miennes, tiennes, siennes, notres, votres
As-fs-	AFS	leur, mienne, tienne, sienne, notre, votre
As-mp-	AMP	leurs, miens, tiens, siens, notres, votres
As-ms-	AMS	leur, mien, tien, sien, notre, votre

4.9 Application to Portuguese

4.9.1 Type

Attribute	Value	Example	Tag
Type	qualificative	azul	
	possessive	seu	
	indefinite		
	cardinal	dois	
	ordinal	primeiro	

4.9.2 Degree

Attribute	Value	Example	Tag
degree	positive		
	comparative	melhor	
	superlative	paupe’rrimo	

In the Portuguese model the ‘positive’ value is not explicitly marked because it is considered a default value.

4.9.3 Gender

Attribute	Value	Example	Tag
Gender	masculine	bom	
	feminine	boa	
	neuter		

The value *neuter* does not apply to Portuguese adjectives.

4.9.4 Number

Attribute	Value	Example	Tag
Number	singular	bom	
	plural	bons	

4.9.5 Case

Not applicable to Portuguese.

4.9.6 Other features required to encode Possessive Adjectives

The following features are also required to encode the Portuguese possessive adjectives:

4.9.7 Possessor

Attribute	Value	Example	Tag
Number-possessor	singular	meu	
	plural	nosso	

4.9.8 Person

As we explained in the section on Verbs, the Portuguese lexicon adopted the decomposition of the feature *person* into the features *person-deixis* and *person-conc*. These features are also used to encode Portuguese possessive adjectives.

4.10 Application to Danish

For Danish adjectives the following features are relevant: Type, Degree, Gender, Number, Use and Definiteness (language-specific property). The feature Case applies only to nominalised adjectives (cf. section Noun, item Case.)

Type

The attribute Type, including cardinals, ordinals and quantifiers, seems to be difficult to treat cross-linguistically. Basically, in Danish these behave in a similar way to adjectives, but within the Eurotra framework they were treated as separate classes (although they were previously regarded as sub-categories of the adjective category.)

Degree

The default value of the attribute Degree is 'positive' (called 'base' in EDEMD). The absolute superlative can be regarded as an additional degree; it is composed of the prefix **aller-** and the superlative of the adjective.
There are a huge number of adjectives which have periphrastic instead of inflectional comparison; in such cases the comparative is comprised of **mere** + the positive form of the adjective, and the superlative of **mest** + the positive form. In corpus annotation the analytical forms occurring in the text are recognised word-by-word, i.e. not recognised as comparative or superlative.

The EDEMD applies the attribute 'Comparison type' to all adjectives; this feature is relevant to lexicons only. Traditional dictionaries of Danish indicate deviations from the regular inflectional patterns.

Attribute	*value*	Example	Tag
Degree	*positive*	dyr	adj_**pos**
	comparative	dyrere	adj_**comp**
	superlative	dyrest	adj_**sup**

Gender, Number and Definiteness

Agreement applies to adjectives in gender, number and definiteness, when the adjective is used in attributive function. Definiteness does not apply in predicative function.
A number of adjectives have defective inflection. Another group have no possibility of inflection (gender, number and definiteness) or comparison: they are completely invariable.

Gender

Attribute	*value*	Example	Tag
Gender	*common*	(en) **dyr** (bil)	adj_**com**
	neuter	(et) **dyrt** (hus)	adj_**neut**

Number

The adjective has no gender and definiteness agreement in the plural (neither in attributive nor in predicative functions). All regular adjectives receive **-e** as the plural ending, both in definite and in indefinite use.

Attribute	*value*	Example	Tag
Number	*singular*	(en) **dyr** (bil)	adj_com_**sg**
	plural	**dyre** (biler)	adj_**pl**

Definiteness

The morphological distinction between definite and indefinite only exists in the singular in attributive function. As mentioned above (item Number), the definite and indefinite plural forms of adjectives are identical. However, a parser-based corpus tagger can distinguish between the two forms on the basis of the immediately preceding context of the adjective.

Attribute	*value*	Example	Tag
Definiteness	*indefinite*	(en) **dyr** (bil)	adj_com_sg_**indef**
	definite	(den) **dyre** (bil)	adj_com_sg_**def**
Definiteness	*indefinite*	**dyre** (biler)	adj_pl_**indef**
	definite	(de) **dyre** (biler)	adj_pl_**def**

Use

Corpus tagging devices recognise the morphological forms; however, in the case of syncretic forms (e.g. singular definite and plural) it is necessary to parse the sentence to disambiguate the form or else the tagger may insert more than one single tag.

Attribute	*value*	Example	Tag
Use	*attributive*	(en) **dyr** bil	adj_ ...._**attr**
	predicative	(bilen er) **dyr**	adj_...._**pred**
	adverbial	(pigen taler) **smukt**	adj_...._**adv**
	nominal	(den/det/de) **dyre**	adj_...._**nom**

Synoptic table of agreement relations

The value of Use may also be inserted into the tag combination if needed for special purposes.

Attribute	*value*	Example	Tag
Agreement A	*Gender, Number, Def.ness*	en **dyr** bil et **dyrt** hus	adj_com_sg_indef adj_neut_sg_indef
		den **dyre** bil	adj_com_sg_def
		det **dyre** hus	adj_neut_sg_def
	Number, Def.ness	**dyre** biler/huse	adj_pl_indef
	Number, Def.ness	(de) **dyre** biler/huse	adj_pl_def
Agreement P	*Gender, Number*	en bil/bilen er **dyr** et hus/huset er **dyrt**	adj_com_sg adj_neut_sg
	Number	biler/husene er **dyre**	adj_pl

A word-by-word tagger is able to annotate the plural adjective form with the definiteness feature on the basis of the immediate context (i.e. presence or absence of articles, possesives, etc.)

4.11 Application to Greek

Features that apply to Greek adjectives are: Degree, Gender, Number and Case.

4.11.1 Degree

Attribute	*value*	Gr. example	Gr. tag
Degree	*basic*	psylos	Aj**Ba**
	comparative	psyloteros	Aj**Cp**
	superlative	psylotatos	Aj**Su**

All adjectives have a form for the value *basic*. Some of them form the comparative and the superlative degree in one of the following ways:

(a) periphrastically, from the basic form and a word/phrase denoting comparison:

Comparative: *pio kalos, lighotero kalos, to idhio kalos*
Superlative: *o pio kalos, o lighotero kalos*

This type of formation can only be dealt in corpora by multi-word tagging; no information on this is to be coded in the lexicon.

(b) Certain adjectives can also form the comparative degree (denoting superiority) by the addition of -ter- between the stem and the ending, and the superlative by the addition of -tat- in the same position:

Basic: *puknos, psylos*
Comparative: *puknoteros, psyloteros*
Superlative: *puknotatos, psylotatos*

(c) A few adjectives form the superlative degree monolectically, using a different stem or a different ending from the normal one:

Basic: kalos, meghalos, aplos
Comparative: *kaluteros, meghaluteros, aplousteros*
Superlative: *kallistos, meghistos, aploustatos*

4.11.2 Gender, Number and Case

In Greek, adjectives must agree in gender, number and case with the nouns they modify.

The majority of Greek adjectives are inflected according to paradigms that form distinct forms for all three genders. A few adjectives (ending in -ys, -ys -es) share the same form for the masculine and the feminine gender:

kalos, kaly, kalo
but
eutuhys, eutuhys, eutuhes

Although for nouns we use the value *masc-fem*, we have decided not to use it for adjectives, given that in the majority of cases they can be disambiguated in the corpus on the basis of the noun they refer to:

O Ghiannys *einai eutuhys* - masc.
Y Maria *einai eutuhys* - fem.

In relation to Number and Case, the values *invariant* and *indeclinable* respectively are used for adjectives that retain the same form irrespective of gender, number and case. As in the case of nouns, these are foreign words which have entered the Greek language without having adopted the inflectional system, and their disambiguation relies on context.

Attribute	*value*	Gr. example	Gr. tag
Gender	*masculine*	kalos	AjBa**Ma**
	feminine	kaly	AjBa**Fe**
	neuter	kalo	AjBa**Ne**

Attribute	*value*	Gr. example	Gr. tag
Number	*singular*	kalos	AjBaMa**Sg**
	plural	kaloi	AjBaMa**Pl**
l-spec	*invariant*	roz	AjBaMaSg**Nv**

Attribute	*value*	Gr. example	Gr. tag
Case	*nom*	kalos	AjBaMaSg**Nm**
	gen	kalou	AjBaMaSg**Ge**
	acc	kalo	AjBaMaSg**Ac**
l-spec	*voc*	kale	AjBaMaSg**Vo**
l-spec	*indcl*	roz	AjBaMaNv**Ic**

4.11.3 Use

Although this feature (belonging to level 2a) is not currently used in corpus tagging, it is applicable to Greek with the values proposed by EAGLES, i.e. *attributive* and *predicative*.

4.11.4 Comments

Type is not used for Greek, given that cardinal and ordinal adjectives are included under the category of *num*. Details on the reasons that dictate this practice are given in the relevant section.

4.11.5 Features not applicable in Greek

Modification Type, Inflection Type and **Definiteness** are not applicable to Greek adjectives.

5 Pronoun

P	Type	wh-T	Pers	G	N	Case	Pos	Pol	Funct
M U L T	dem indf pers poss rel		1 2 3	m f n	s p	nom gen dat acc voc			
G E N E L	pers rel poss indf excl imprs		1 2 3	m f n	s p		sp pl		
A l e t h	dem poss rel pers indf imprs		1 2 3	m f n	s p		sg pl		
N E R C	poss dem indf int/rel pers refl	int rel	1 2 3	m f mf	s p sp	nom gen dat acc obl bas			
L e e c h	poss dem indf int/rel pers refl	int rel	1 2 3	m f n c	s p	nom gen dat acc obl bas			
E-L0	**PRONOUN**								
E A G L 1	dem indf poss int/rel pers refl recp excl		1 2 3	m f n	s p	nom gen dat acc obl obj			
L 2 a		int rel						pol fam	
L 2 b			It c	It n		Po pobj			Du att Du prd Du adv

5.1 Comments

The NERC and the Leech/Wilson schemes propose a multilayered treatment of Pronouns, thus permitting different levels of granularity in annotation. The reader should note that the table above describes the most fine-grained level of linguistic description of the two proposals, where Pronouns are recognized as a category of their own.

At a less granular level, in these two systems Pronouns appear merged together with Determiners. The background reason for this merging in the NERC survey was the necessity of meeting the requirements, on the one hand, of a number of English tagsets (e.g. Penn Treebank, Brown, Lancaster) where, for example, Demonstratives are undistinguished as to their pronominal and determiner functions and receive a unique tag, and, on the other hand, of tagsets which include Articles among the Determiners. Hence, a multilayered approach was adopted, where three different fine-grained levels of linguistic distinctions offer the possibility for each existing practice to be placed at the appropriate level, thus permitting its reusability (see Monachini and Oestling 1992b).
This solution has also been adopted by the EAGLES Corpus group for linguistic annotation (Leech and Wilson 1994).

A first version of the present document also proposed the same treatment of Pronouns and Determiners. However, after the first cycle of applications, and in particular after the MULTEXT concrete testing, it seemed better to distinguish between different functions and, therefore, to have different categories for Pronouns and Determiners, at least at the lexical level. Lexical descriptions should be independent from applications and should aim at a general description of each language; corpus tags, depending on the capabilities of state-of-art tagging techniques, may underspecify lexical specifications, collapsing many distinctions and presenting broader categories (Calzolari and Monachini 1994).
Furthermore, following the TEI proposal, it has also been decided to have Articles as a separate category.

5.1.1 Type, Wh-Type

The column **Type** shows how Pronouns are split into different subclassifications, recommended at Level-1.
The subfeature **wh-Type** is for the further distinction of the double value ‘interrogative/relative’ and is proposed at Level-2a.

5.1.2 Person, Case and Possessor

Other information, such as **Person**, **Case** and **Possessor** (i.e. the Number of Possessor), are clearly not applicable to all the Types. This tabular representation does not permit the indication of constraints on the application of some features in the presence of others. These constraints have to be explicitly specified in the language-specific applications.

5.1.3 Politeness

This encodes the polite usage of personal pronouns.

5.1.4 Function

This feature has been added for the encoding of ‘attributive’, ‘predicative’ and ‘adverbial’ use (see Dutch specific application.)

5.1.5 Other values on Level 2b

Finally, on level 2b, the value ‘prepositional obj’ is foreseen among the values of the attribute **Case**, for the Portuguese language.

5.2 Application to Italian

In Italian, both corpus and dictionary distinguish Pronouns according to the feature Type (e.g. indefinite, demonstrative, possessive, etc.).

5.2.1 Type

The table below shows the different types pertinent to the category of Pronouns.
In the following sections, each Type will be discussed in detail.

Attribute	*value*	It. example	It. tag
Type	*dem*	questo	PD/ms
	poss	mio	PP/ms
	indf	ognuno	PI/ms
	pers	io	PP/ns1s
	refl	si	PF/nn
wh-Type	*int*	che	PT/ns
	rel	che	PR/ns
	excl	quanto!	PE/ns

5.2.2 Personal Pronoun

Personal Pronouns are inflected for Person and Number, as shown in the following table.

Personal		example	It.tag
Person	**Gend.-Numb.**		
1	*s*	io	PP/ns1s
2	*s*	tu	PP/ns2s
3	*s*	egli	PP/ms3s
1	*p*	noi	PP/np1s
2	*p*	voi	PP/np2s
3	*p*	essi	PP/mp3s

As far as the pronominal paradigm is concerned, Case is not encoded at present in our DMI and corpus. Personal pronouns are not lemmatized: 'gli' is not considered the dative form of the base pronoun 'egli' (he), but constitutes a separate entry.

The Italian pronominal paradigm is described below:

forme toniche: subj ('io, egli'), compl ('me, lui')

ama me / da' a me (dir-obj/prep-obj)
(he loves me / he gives to me)

ama lui / da' a lui (dir-obj/prep-obj)
(she loves him / she gives to him)

forme atone: —, compl ('mi, gli/lo')

mi da' / mi ama (ind-obj/dir-obj)
(he gives me / he loves me)
gli da' (ind-obj) (he gives him)
lo ama (dir-obj) (she loves him)

This paradigm can be mapped on the proposed Case system in the following way:

io, egli	subj	nom
mi/me	dir-obj/ind-obj/prep-obj	obj = acc, dat, prep+obj
lui	dir-obj/prep-obj	obj = acc, prep+obj
gli	ind-obj	dat
lo	dir-obj	acc

Polite form
In Italian, the modern system of Personal Pronoun for addressing a person is bipartite: *tu* and *lei*, feminine singulal used for the polite form (also for addressing masculine persons).
Polite usages are very interesting from a corpus perspective, but, at present, are not encoded in our tagger.

In polite usages, two different types of agreement can be used:

- agreement by nature: the participle has masculine gender, i.e. agrees with the masculine noun and not with the personal pronoun.
Professore, Lei si e' occupatO.
Professor (man), She has been interested in ...
Professor, you have been interested in ...

- agreement by grammar: the participle is feminine, i.e. agrees with the pronoun, even though referred to a masculine noun.
Lei, Professore, l'ho sempre ascoltatA
Professor (man), I always have listened to Her
Professor, I always have listened to you

Non-tonic personal pronouns always agree by grammar:

Professore, vorrei dirLE, spero di rivederLA presto
Professor (man), I want to tell Her, I hope to see Her soon.
Professor, I want to tell you ... , I hope to see you soon.

5.2.3 Reflexive (Pronoun)

In Italian, the reflexive pronouns are represented by *mi, ti, si, se', ci, vi, si.*
They are inflected for Person and Number, while Gender is not pertinent.

For the third singular and plural persons, in addition to *si*, the tonic form *se'* can be used: *egli* **si** *lava; egli lava* **se'** *(stesso); esse aiutano solo* **se'** *(stesse). Se'* can be 'reinforced' by the ajdective *stesso*: in this case the bigram *se stesso* should be encoded on the basis of multiword expression tagging strategy.

Reflexive			
Gender	**Number**	example	It.tag
1	*s*	mi	PF/nn1s
2	*s*	ti	PF/nn2s
3	*s*	si, se'	PF/nn3s
1	*p*	ci	PF/np1p
2	*p*	vi	PF/np2p
3	*s*	si	PF/nn3p

Formally, reciprocal pronouns are the same as reflexive pronouns.

5.2.4 Possessive

Possessives are inflected for Number and Gender and agree with the nouns to which they refer; they are distinguished according to the Person to which they refer.

Possessive							
Person		*1s*	*2s*	*3s*	*1p*	*2p*	*3p*
Gender	**Number**						
m	*s*	mio	tuo	suo	nostro	vostro	loro
m	*p*	miei	tuoi	suoi	nostri	vostri	loro
f	*s*	mia	tua	sua	nostra	vostra	loro
f	*p*	mie	tue	sue	nostre	vostre	loro

Scrivo con la tua penna, perche' non ho la mia P**P**/fs

Italian has two more possessives: *altrui* (of other people) and *proprio* (own), which can both function as pronouns:

spende il denaro altrui, non il proprio PP/nn
occorre dare del proprio denaro, non dell'altrui PP/nn

5.2.5 Possessor

Information about the possessor is not encoded in the Italian lexicon or corpus, but it can be inferred from the lemma:

queste sono le mie (owned things plural, but possessor singular), is tagged PP/fp
questa e' la nostra (owned thing singular, but possessor plural) is tagged PP/fs

5.2.6 Demonstrative

Demonstrative	example	It.tag
	questo	PD/ms
	quello	PD/ms

Among the Demonstratives, the most used are: *questo, codesto, quello.* Deixis is implicitly contained in demonstratives, but not presently encoded:

usa queste PD/fs, *non prendere quelle* PD/fp

While the others can be used as pronouns or pronominal adjectives, *cio'* only has pronoun function:

cio' e' vero PD/nn

Questi, quegli, costui, colui are pronouns used only for humans; *questi* and *quegli*, scarcely used and archaic, refer to the singular:

Questi e' un bell'uomo PD/ms

Stesso, medesimo can be demonstrative pronouns.

5.2.7 Indefinite

Indefinite	example	It.tag
	ognuno	PI/ms
	chiunque	PI/ns

Italian Indefinites are inflected for Gender and Number. Among the indefinites, the following can only be pronouns: *uno, ognuno, qualcuno, chiunque, chicchessia, nulla, niente.*

ognuno deve fare ... PI/ns

Some can have both the pronoun and the pronominal adjective/determiner functions: *alcuno, ciascuno, taluno, nessuno, tutto, alquanto, poco, molto, troppo, tanto.*

ho comprato il libro e l'ho letto tutto PI/ms

5.2.8 Interrogative

Interrogative	example	It.tag
	chi	PT/ns
	quale	PT/ns

Interrogatives are inflected for Gender and Number.

Chi is pronoun only.

Chi viene oggi? PT/ns

Che, quale, quanto can be either pronouns or pronominal adjectives/determiners:

A che pensi? PT/nn

5.2.9 Exclamatory

Exclamatory	example	It.tag
	quanto!	PE/ms
	quale!	PE/ns

Che, quale, quanto can also have exclamatory value:

quanti sono venuti! PE/mp

5.2.10 Relatives

Relative	example	It.tag
	che	PR/nn

The Relatives comprise *che, il quale, cui.*
Che is used both for singular and plural, and for masculine and feminine.

5.3 Application to German

5.3.1 Type

Attribute	*value*	example	tag
Type	*personal*	ich	**PPER**:1.Sg.Nom
	reflexive	sich	**PRF**:3.Sg.Akk
	possessive	meins	**PPOSS**:Neut.Nom.Sg
	demonstrative	dieses	**PDEMS**:Neut.Nom.Sg
	relative	, das	**PRELS**:Neut.Nom.Sg
	indefinite	irgendeines	**PROS**:Neut.Nom.Sg
	interrogative	was?	**PWS**:Neut.Nom.Sg

5.3.2 Gender

The feature *gender* applies to all pronouns – except personal pronouns (where it is relevant only for 3rd person singular) and reflexive pronouns. It is not defined for a number of indefinite pronouns such as *jemand, niemand, man etc.*

Attribute	*value*	example	tag
Gender	*masculine*	er	PPER:3.Sg.Nom.**Masc**
	feminine	meine	PPOSS:**Fem**.Nom.Sg
	neuter	das	PRELS:**Neut**.Nom.Sg

5.3.3 Number

The feature *number* applies to all pronouns.

Attribute	*value*	example	tag
Number	*singular*	keiner	PROS:Masc.Nom.**Sg**
	plural	welche	PWS:Neut.Nom.**Pl**

5.3.4 Case

The feature *case* applies to all pronouns.

Attribute	*value*	example	tag
Case	*nominative*	keiner	PROS:Masc.**Nom**.Sg
	genitive	niemandes	PROS:**Gen**.Sg
	dative	jemandem	PROS:**Dat**.Sg
	accusative	wen?	PWS:**Akk**.Sg

5.3.5 Possessor

The feature *possessor* is not encoded in the IMS-Tagset. It can be determined from the lemma.

5.4 Application to English

5.4.1 Pronoun-type

Attribute	*values*	Examples	Tags
Pronoun-type	*personal*	it, she	PPs3, PPs3NF
	reflexive	myself	PRs1
	possessive	yours	PV2
	demonstrative	this	PDs
	wh-type	what	PW
	indefinite	anyone	PI

5.4.2 Wh-subtype

Attribute	*values*	Examples	Tags
Wh-subtype	*relative*	which	PWR
	other	which	PWQ

The relative and "other" wh-type pronouns may be distinguished by the following examples: *the shoes which she bought* (PWR); *Which shoes did she buy?* (PWQ).

The "other" (non-relative) category includes interrogative and exclamatory pronouns and determiners. As these subtypes are difficult to distinguish automatically, it is convenient to bring them together under a single value "other".

5.4.3 Person

Attribute	*values*	Examples	Tags
Person	1st person	I	PPs1N
	2nd person	you	PP2
	3rd person	she	PPs3NF

5.4.4 Number

Attribute	*values*	Examples	Tags
Number	*singular*	someone, it	PIs, PPs3
	plural	few, they	PIp, PPp3N

5.4.5 Gender

Attribute	*values*	Examples	Tags
Gender	*masculine*	him	PPs3OM
	feminine	her	PPs3OF
	neuter	its	PVs3U
	common	anyone	PIs

The common gender pronouns such as *anyone* have personal reference, but are neutral between masculine and feminine.

5.4.6 Case

Attribute	*values*	Examples	Tags
Case	*nominative*	they	PPp3N
	oblique	them	PPp3O

Personal pronouns in the oblique case (*me, them*, etc) are used as objects, as prepositional complements, and also in some other functions.

5.5 Application to Dutch

5.5.1 Type

CELEX distinguishes 10 subclasses of pronouns:

Attribute	*value*	Example	Tag
Type	*personal*	ik	PRON **pers**
	reciprocal	elkaar	PRON **wkg.**
	reflexive	zich	PRON **wknd.**
	possessive	mijn	PRON **bez.**
	demonstrative	dit	PRON **aanw.**
	relative	wat	PRON **betr.**
	indefinite	geen	PRON **onbep.**
	interrogative	wie?	PRON **vraag.**
	exclamatory	wat!	PRON **uitr.**

5.5.2 Gender

A gender value tag set is applicable to Dutch for Personal, Demonstrative, Possessive, Interrogative, Reflexive and Relative Pronouns, but is not found in the CELEX tagset. *The table below is a proposal for Dutch, not a CELEX table* !

Attribute	*value*	Example	Tag
Gender	*masculine*	zijn (vader)	Masc
	feminine	haar (vader)	Fem
	neuter	welk (boek)	Neut

Gender of pronouns is very important for contextual disambiguation of gender ambiguities. See Noun.

5.5.3 Number

Number is applicable to Dutch Articles and Demonstrative, Reflexive, Relative and Interrogative Pronouns, but is not in the CELEX tagset.

The table below is a proposal for Dutch, not a CELEX table !

Attribute	*value*	Example	Tag
Number	*singular*	het (huis)	Sing
		(het huis) dat	Sing
	plural	de (huizen)	Plur
		(de huizen) die	Plur

5.5.4 Case

Case is only applicable to Dutch in some archaic forms still surviving in the language in idiomatic expressions (Articles, Relative, Personal and Interrogative Pronouns). It is only present in the CELEX Wordforms Lexicon. The tags are the same as for Nouns:

Ge: Genitive singular: *des, dezer, harer* etc.
Gm: Genitive plural: *aller, hunner, onzer*
De: Dative singular: *aller, den, der, dien*
Dm: Dative plural: *haren, hunnen, mijnen, onzen, uwen, zijnen.*

5.5.5 Function

There is no Function value tag set in CELEX. The table of the German application is not applicable to Dutch.

The table below is a proposal for Dutch, not a CELEX table !

The feature Function does not apply to all pronouns.

Attribute	*value*	Example	Tag
Function	*attributive*	deze (auto)	Attrib
	predicative	(Dat is) het	Pred
	adverbial	Hoe (werkt dat?)	Adv
	nominal	(Het) mijine	Nom

The value *Predicative* might prove redundant since predicatively used pronouns might turn out to be only nominal. This can be an L2 tag.

5.6 Application to Spanish

The Eurotra dictionaries count pronouns as nouns. Among all the attributes used to describe nouns the following refer to pronouns:
– "Nform", which serves to distinguish between pronouns (valued pro or cli) and "normal" nouns (valued norm).

– "dtype", which helps to distinguish between possesive and non-possesive pronouns.

– "whmor", used to distinguish between relative pronouns (valued as rel) and interrogative pronouns (valued as int) from the rest of the nominals (valued none).

By means of these three attributes we can distinguish between:

Ex.	Nform	Dtype	Whmor	"class"
casa	normal	non-poss	none	"normal" noun
yo	pro	non-poss	none	personal pro
este	pro	non-poss	none	demonstrative pro
algún	pro	non-poss	none	indefinite pro
mi	pro	poss	none	possesive pro
mio	pro	poss	none	possesive pro
cuyo	pro	poss	rel	relative pro
quien	pro	non-poss	rel	relative pro
quién	pro	non-poss	int	interrogative pro
me	cli	non-poss	none	clitic

5.6.1 Type

Demonstrative: PD

Only the attributes Number and Gender are pertinent to Spanish pronominal demonstratives.

Pers	Number	Gender	Pos	Case	Funct	Pol	Infl	Ex.
3	sg	masc						este
3	pl	fem						aquellas
3	sg	n						eso

These pronominals reflect deictic degree of remoteness, which is not, however, coded in ET-ES dictionaries.

As for "person", these pronominals (coded in ET-ES dictionaries as "nouns") are forced to have a person attribute valued as "third", due to verbal agreement.

Possessive: PP

The attributes pertinent to Spanish pronominal possesives are: Number, Gender, Person and Possessor.

Pers	Number	Gender	Pos	Case	Funct	Pol	Infl	Ex.
3/1	sg	masc	sg					(el) mio
3/2	pl	fem	sg					(las) tuyas
3/3	sg	masc	pl					(los) suyos

“Number” refers to possedee referent. Agreement is established w.r.t this attribute. References to possessor number are made via lexical variation (except when the possessor is third person: ”suyo”).

As for the attribute “person”, these pronominals (coded in ET-ES dictionaries as “nouns”) have a person attribute valued as “third” by default since they agree with third person verb forms despite the grammatical person of their referent.

Indefinite: PI

These are: “alguien”, “nadie”, “quienquiera”, “qualquiera”, “uno”, “alguno”, “ninguno”, “algo”, “nada”, and “otro”. All indefinite determiners can be also used as pronouns. In ET-ES dictionaries this category is split into nouns, nform=pronouns, and the category “quantifier”, e.g.: “nada”, “nadie”, “todos/as”.

These are variable in gender and number except for “alguien”, “nadie” and “nada”.

5.6.2 WH-Type

Interrogative

These are: “qué, cuál, quién and cuánto”.
All of them inflect for number except for “qué”; and only “cuánto/a” inflects for gender and number: “cuántos/as”.
In Eurotra dictionaries they are valued as 3rd for person since they agree, when they are the subject of a sentence, with 3rd person verb forms.

Relative

These are “que” and “quien”. Only “quien” inflects for number.

Personal

In Spanish the attributes pertinent to personal pronouns are: person, gender (except for dative pronouns), number, case and politeness.

Case is normally related to grammatical functions, thus nominative corresponds to subject, accusative to first object, dative to second object and oblique to those oblique arguments obligatorily introduced by a preposition. Note that Spanish has oblique synthetic forms such as “conmigo” (with me).

In the following lines we only give examples for the first person singular forms:

Pers	Number	Gender	Pos	Case	Funct	Pol	Infl	Ex.
1	sg	c		nom				yo
1	sg	c		acc				me
1	sg	c		dat				me
1	sg	c		obl				mi
1	sg			obl				conmigo

Reflexive Pronouns

Pers	Number	Gender	Pos	Case	Funct	Pol	Infl	Ex.
1	sg							me
2	sg							te
3	sg							se
1	pl							nos
2	pl							os
3	pl							se

Reciprocal Pronouns

Formally, reciprocal pronouns are the same as reflexive pronouns.

5.7 Application to French (Corpus)

5.7.1 Type

Attribute	*value*	Example	Tag
Type	*dem*	celle-ci	P**DET**FS
	indf	quiconque	P**IND**MS
	poss		(*)
	int/rel	lequel	P**INT**MS, P**REL**MS
	pers	tu	P**PER**2
	refl	se	P**REF**MS

(*) although possessive pronouns exist in French, they are multiword forms (*le nôtre*), and therefore are coded as article + noun with the IBMF tagset.

5.7.2 wh-Subtype

Attribute	*value*	Example	Tag
wh-Type	*int*	lequel	P**INT**MS
	rel	qui	P**REL**MS

Note that most interrogative pronouns are also relative pronouns in French.

5.7.3 Person

In the IBMF tagset, as in the case of verbs, the person is numbered 1 to 6. The coding of an IBMF verb tag into the EAGLES scheme therefore implies using both the Person and Number information.

Attribute	*value*	Example	Tag
Person	*1*	je, nous	PPER**1**, PPER**4**
	2	tu, vous	PPER**2**, PPER**5**
	3	il, ils	PPER**3**M, PPER**6**M

5.7.4 Gender

Gender applies to all pronouns and determiners in French, although it can be without a morphological realization, as in the case of the personal pronouns of the 1st and 2nd persons (*je, tu, nous, vous*).

Attribute	*value*	Example	Tag
Gender	*masculine*	celui	PDET**M**S
	feminine	celle	PDET**F**S

5.7.5 Number

Number applies to all pronouns and determiners in French.

Attribute	*value*	Example	Tag
Number	*singular*	lequel	PINTM**S**
	plural	lesquels	PINTM**P**

5.7.6 Case

Pronouns seem to be the only class where the Case attribute can be used in French. As in the case of Italian, subject, object and oblique personal pronouns can be distinguished. Thus the EAGLES classes nom, acc and obl can be used. The oblique pronoun is not coded specifically in the IBMF tagset.

Attribute	*value*	Example	Tag
Case	*nom*	tu	PP**ER**2
	acc	te	PP**OB**MS

As in the case of reflexive personal pronouns, the IBMF tagset does not code the person in the object personal pronoun.

5.7.7 EAGLES features not applicable

Possessor (the number of) does not apply to French. What could apply, as in other Romance languages, is the person of the possessor.
Function and **Inflection type** do not apply to French.
Politeness applies to French (the polite “vous”) but is not coded specifically in the tagset.

5.8 Application to French (Lexicon)

5.8.1 Type

Attribute	Value	Example	Code
Type	personal	je	p
	demonstrat.	celui	d
	indefinite	certain	i
	possessive	le_mien	s
	interrog.	lequel	t
	relative	quel	r

Possessive pronouns are compound forms only (“le mien”). The form “mien” is an adjective.

5.8.2 Person

Attribute	Value	Example	Code
Person	first	je,ma	1
	second	tu,ta	2
	third	il,sa	3

5.8.3 Gender

Attribute	Value	Example	Code
Gender	masculine	cet,il	m
	feminine	cette.elle	f
	neutre	ce	n

5.8.4 Number

Attribute	Value	Example	Code
Number	singular	certain	s
	plural	certains	p

5.8.5 Case

Attribute	Value	Example	Code
Case	nominative	je	n
	object	me	a
	oblique	moi	o

For the French pronominal system, we could use the following mapping to the EAGLES system:

Nominative = subject ‘il’
Accusative = direct object ‘le’
Dative = indirect object ‘lui’
Oblique = other

The category “other” corresponds to the reinforcement of subject or object (“Moi, je le dis”), attribute (“C’est moi”), etc.

However, this solution splits “object” into “direct” and “indirect”, and this distintion is valid only for the 3rd person pronouns in French (direct: “le, la, les”; indirect: “lui, leur”). Encoding this distinction would duplicate all other forms (direct: “me, te” etc.; indirect: “me, te” etc.). The following mapping applies readily to French personal pronouns:

subject	je, tu, il, elle, nous, vous, ils, elles
object	me, te, le, la, lui, se, nous, vous, les, leur, se
oblique	moi, toi, lui, elle, soi, nous, vous, eux, elles, soi

5.8.6 Possessor

Attribute	Value	Example	Code
Possessor	singular	mon	s
	plural	nos	p

5.8.7 Combinations

Code	Example

```
Ps1fs-s  la_mienne (mienne is not a pronoun)
Ps1fs-p  la_no>tre
Ps1fp-s  les_miennes
Ps1fp-p  les_no>tres
Ps1ms-s  le_mien
Ps1ms-p  le_no>tre
Ps1mp-s  les_miens
Ps1mp-p  les_no>tres

Ps2fs-s  la_tienne
Ps2fs-p  la_vo>tre
Ps2fp-s  les_tiennes
Ps2mp-p  les_vo>tres
Ps2ms-s  le_tien
Ps2ms-p  le_vo>tre
Ps2mp-s  les_tiens
Ps2fp-p  les_vo>tres

Ps3fs-s  la_sienne
Ps3fs-p  la_leur
Ps3fp-s  les_siennes
Ps3fp-p  les_leurs
Ps3ms-s  le_sien
Ps3ms-p  le_leur
Ps3mp-s  les_siens
Ps3mp-p  les_leurs

Pp1-sn-   je
Pp2-sn-   tu
Pp3msn-   il, on
Pp3fsn-   elle
Pp1-pn-   nous
Pp2-pn-   vous
Pp3mpn-   ils
Pp3fpn-   elles

Pp1-sj-   me (-moi after imperative)
Pp2-sj-   te (-toi after imperative)
Pp3msj-   le, se, lui
Pp3fsj-   la, se, lui
Pp3n-j-   en, y
Pp1-pj-   nous
Pp2-pj-   vous
Pp3mpj-   les, se, leur
```

```
Pp3fpj-   les, se, leur

Pp1-so-   moi
Pp2-so-   toi
Pp3mso-   lui, soi
Pp3fso-   elle, soi
Pp1-po-   nous
Pp2-po-   vous
Pp3mpo-   eux, soi
Pp3fpo-   elles, soi

Pd-fp--   celles, celles-ci, celles-la'
Pd-fs--   celle, celle-ci, celle-la'
Pd-mp--   ceux, ceux-ci, ceux-la'
Pd-ms--   celui, celui-ci, celui-la'
Pd-n---   ce, ceci, cela,

Pi-fp--   quelques-unes, certaines...
Pi-fs--   aucune, nulle, certaine...
Pi-mp--   quelques-uns, certains...
Pi-ms--   aucun, nul, quelqu'un, certain...

Pr-fp--   lesquelles, desquelles, auxquelles, qui, que, quoi, dont,
Pr-fs--   laquelle, qui, que, quoi, dont, ou
Pr-mp--   lesquels, desquels, auxquels, qui, que, quoi, dont
Pr-ms--   lequel, duquel, auquel, qui, que, quoi, dont

Pt-----   quoi
Pt-fp--   lesquelles, desquelles, auxquelles, qui, que
Pt-fs--   laquelle, qui, que
Pt-mp--   lesquels, desquels, auxquels, qui
Pt-ms--   lequel, duquel, auquel, qui, que
===============================================================
```

5.9 Application to Portuguese

5.9.1 Type

Attribute	Value	Example	Tag
Type	demonstrative	isso	
	indefinite		
	possessive	meu	
	interrogative/relative	qual,que	
	personal	eu	
	refl	me	
	recp		
	excl	que	

We assume that *possessive* applies to Portuguese pronouns. (Nevertheless, in the Portuguese GENELEX demo lexicon, the value *possessive* was included only at category *adjective*. This was due to the following assumption: if a candidate pronominal item occurs both as a noun phrase and as a modifier to a full np, it will be categorized as an adjective, presuming that its np-like realization is describable as an instance of nominal ellipsis.)

5.9.2 Person

As we explained in the section on Verbs, the Portuguese lexicon adopted the decomposition of the feature *person* into the features *person-deixis* and *person-conc*. These features are also used to encode Portuguese personal pronouns.

5.9.3 Gender

Attribute	Value	Example	Tag
Gender	masculine	ele	
	feminine	ela	
	neuter		

The *neuter* value is not used in Portuguese.

5.9.4 Number

Attribute	Value	Example	Tag
Number	singular	ele	
	plural	eles	

5.9.5 Case

Attribute	Value	Example	Tag
Case	nominative	eu	
	genitive	me	
	dative	me	
	accusative	me	
	oblique	mim, comigo	

5.10 Application to Danish

The values used in the lexical specification within the tables below are based on the EDEMD; however, minor changes have been made here to adapt a few value names to the EAGLES proposal.

Type

Attribute	*value*	Example	Tag
Type	*personal*	jeg	pron_**pers**_sg1_nom
	reflexive	sig	pron_**refl**_sg3/pl3
	reciprocal	hinanden	pron_**rec**
	possessive	min	pron_**poss**_sg1_com_sg
	demonstrative	denne	pron_**dem**_com_sg
	indefinite	nogen	pron_**indef**
	relative	hvem	pron_**rel**
	wh-type	hvilken	pron_**wh**_sg_com

Wh-subtype

All pronouns beginning with **hv-** (except 'hver' (each)) may function both as relative and as interrogative pronouns. The EDEMD specifies each *hv-* pronoun as both relative and interrogative types without a common supertype like Wh-subtype.

Attribute	*value*	Example	Tag
Wh-subtype	*relative*	hvad	pron_**rel**
	interrogative	hvad	pron_**interr**

Gender

The feature Gender applies to the following types of pronouns: personal pronoun (only 3rd person singular), possessive pronouns (see below) and to the relative/interrogative *hvilken*. The only cases were Danish distinguishes the feminine/masculine genders are the personal and possessive pronouns in the third person singular. (In the latter case it refers to the gender (sexus) of the owner.)

Attribute	*value*	Example	Tag
Gender	*feminine*	hun	pron_pers_sg3_**fem**
	masculine	han	pron_pers_sg3_**mas**
	neuter	det	pron_pers_sg3_**neut**
	common	denne	pron_dem_sg_**com**

Number

The feature Number applies to demonstrative, personal, possessive and reflexive pronouns and to types which can be used in adjective-like functions, e.g. *hvilken*.

Attribute	*value*	Example	Tag
Number	*singular*	denne	pron_dem_com_**sg**
	plural	disse	pron_dem_**pl**

Case

The feature Case applies in different ways to personal pronouns and to the other non-adverbial pronouns. Personal pronouns have two distinct cases: subjective and objective, i.e. nominative and oblique cases. The genitive case is covered by the possessive pronouns.
All other non-adverbial pronouns have non-genitive and genitive forms.

Attribute	*value*	Example	Tag
Case	*nominative*	hun	pron_pers_sg3_fem_**nom**
	oblique	hende	pron_pers_sg3_fem_**obl**
	non-genitive	denne	pron_dem_com_sg_**ngen**
	genitive	dennes	pron_dem_com_sg_**gen**

Politeness

Language-specific feature.

The polite form is 'De', used for addressing one or more persons, similar to the personal pronoun of third person plural ('de'), but it is always spelled with a capital D. However, the appropriate reflexive pronoun used in addressing is identical with the oblique case 'Dem', thus deviating from the regular form of the third person plural 'sig'.

Function

The attribute Function is not specified as a lexical property within the EDEMD and no tagset has yet been elaborated for the corpus annotation of this feature.

Attribute	*value*	Example	Tag
Function	*nominal*	hendes (er)	pron_pers_..._**nomin**
	attributive	denne (bil)	pron_dem_..._**attrib**
	predicative	(bilen er) min	pron_poss_..._**pred**
	adverbial	hvor	pron_wh-type_**adverbial**

5.11 Application to Greek

The Greek Morphological Lexicon distinguishes pronouns on the basis of the feature **Type** (e.g. personal, indefinite, possessive, etc.).

It may be noted here that a special category, that of **pou**, has been created for the codification of the word *pou*, which is highly ambiguous in Greek. In fact, it may function as an uninflected relative pronoun, a subordinating conjunction, an adverb or a particle, while the context does not suffice for absolutely correct disambiguation purposes. Therefore, to avoid multiple tagging that would not be resolved, this unique word has been elevated into a category of its own.

In the same way, the category of **enas** includes all uses of the ambiguous word *enas*, one of which is that of the indefinite pronoun. It is also a numeral and the indefinite article in Greek.

In the following paragraphs, we present all values for the types of pronouns and the codification relevant to each type in more detail.

5.11.1 Pronoun Types

The following table shows the attribute Type and the values it takes in the Greek Morphological Lexicon.

Of the values recognised by the Greek Morphological Lexicon, the majority are foreseen by the Eagles Level 1 codification scheme. These are the values *dem, pers, poss, indef* and *reflex*. The distinction between interrogative and relative pronouns is made at the same level, while three language-specific types of pronouns are coded, namely, definite, clitics and relative-indefinite pronouns. More details on each of these types are given in the sections that describe them.

Attribute	*value*	ex.	tag
Type	*dem*	ekeinos	Pn**Dm**03MaSgNm
	poss	mou	Pn**Po**01CoSgGe
	indef	kapoios	Pn**Id**03MaSgNm
	pers	eghw	Pn**Pe**01CoSgNm
	interr	poios	Pn**Ir**03MaSgNm
	rel	opoios	Pn**Re**03MaSgNm
	reflex	eautou	Pn**Rf**03MaSgGe
l-spec	*relindef*	osos	Pn**Ri**03MaSgNm
l-spec	*def*	idhios	Pn**Df**03MaSgNm
l-spec	*clitic*	me	Pn**Cl**01CoSgAc

5.11.2 Personal Pronouns

Personal Pronouns in Greek are morphologically marked for Person, Number and Case; third person pronouns are also marked for Gender.

Two kinds of forms are distinguished : strong and weak pronouns. Under this type, we only include strong forms; weak forms are coded as having type "clitic" (see relevant section below).

Personal			example	tag
Person	**Number**	**Gender**		
1	*sg*	-	eghw	PnPe01SgNm
2	*sg*	-	esu	PnPe02SgNm
3	*sg*	*masc*	autos	PnPe03MaSgNm
3	*sg*	*fem*	auty	PnPe03FeSgNm
3	*sg*	*neuter*	auto	PnPe03NeSgNm
1	*pl*	-	emeis	PnPe01PlNm
2	*pl*	-	eseis	PnPe02PlNm
3	*pl*	*masc*	autoi	PnPe03MaPlNm
3	*pl*	*fem*	autes	PnPe03FePlNm
3	*pl*	*neuter*	auta	PnPe03NePlNm

In the above table, we have given personal pronouns only in the nominative case. All of them are inflected in the genitive case (emena, esena, autou, autis, autou, emas, esas, autwn, autwn, autwn), and the accusative (emena, esena, auton, auti(n), auto, emas, esas, autous, autes, auta). Only two forms appear in the vocative case, namely the second person singular (esu) and plural (eseis).

The value of **Gender** is left unspecified for the first and second persons, which do not code this feature.

5.11.3 Politeness form

In Greek, two forms of the second person of the personal pronoun are used when addressing someone:

- the second singular (*esu*) in the case of familiarity between addresser and addressee, and

- the second plural (*eseis*) for the politeness form.

The pronouns naturally agree in number with the verb. As regards participial constructions, however, agreement with the verb is overruled by agreement with the sex and number of the addressee(s):

*Eseis eiste haroumen***y** *me ta nea;* - fem., sing.

The feature of Politeness does not correspond to a specific attribute in the ILSP Morphological Lexicon.

5.11.4 Reflexive (Pronoun)

In Greek, the reflexive pronoun is formed periphrastically by the pronoun "eautos" preceded by the definite article and followed by the weak form of the personal pronoun in the genitive case (e.g. *tou eautou mou*). It is inflected only for two cases, genitive (*eautou*) and accusative (*eauto*).

In the Greek Morphological Lexicon, we have coded the relevant forms of the lemma "eautos" as a noun; when encountered in the above compound form, it must be recognised as a reflexive pronoun. The following table shows examples for the values it must be assigned when found in the genitive case.

Reflexive			example	tag
Person	**Number**	**Case**		
1	*sg*	*gen*	(tou eautou mou)	PnRf01MaSgGe
2	*sg*	*gen*	(tou eautou sou)	PnRf02MaSgGe
3	*sg*	*gen*	(tou eautou tou)	PnRf03MaSgGe
1	*pl*	*gen*	(twn eautwn mas)	PnRf01MaPlGe
2	*pl*	*gen*	(twn eautwn sas)	PnRf02MaPlGe
3	*pl*	*gen*	(twn eautwn tous)	PnRf03MaPlGe

5.11.5 Possessive pronouns

The weak forms of the genitive case of personal pronouns serve as possessive pronouns for Greek. They are distinguished according to the Person they refer to and code the Number of the possessor (one or more). The third person in the singular number also distinguishes the Gender of the possessor, while in the plural number common form covers all three genders.

Possessive				
Person	**Number**	**Gender**	example	tag
1	*sg*	*masc-fem*	mou	PnPo01CoSgGe
1	*pl*	*masc-fem*	mas	PnPo01CoPlGe
2	*sg*	*masc-fem*	sou	PnPo02CoSgGe
2	*pl*	*masc-fem*	sas	PnPo02CoPlGe
3	*sg*	*masc*	tou	PnPo03MaSgGe
3	*sg*	*fem*	tys	PnPo03FeSgGe
3	*sg*	*neut*	tou	PnPo03NeSgGe
3	*pl*	*masc*	tous	PnPo03MaPlGe
3	*pl*	*fem*	tous	PnPo03FePlGe
3	*pl*	*masc*	tous	PnPo03NePlGe

In certain cases, most usually for emphasis, the possessive pronoun is formed periphrastically on the basis of the above pronoun preceded by the adjective "dhikos" which must agree in gender, number and case with the possessed objects:

*ta dhik***a** *mas vivlia* - neuter, pl., nom/acc
*oi dhik***es** *mas eikones* - fem., pl., nom

Information on the possessor is not specifically coded by an attribute in the Morphological Lexicon.

5.11.6 Demonstrative pronouns

Five demonstrative pronouns are recognised in Greek, all of which are inflected for gender, number and case, agreeing with the object referred to. The following table offers examples for all of them in the singular nominative case of the masculine form.

Demonstrative	example	tag
	autos	PnDm03MaSgNm
	toutos	PnDm03MaSgNm
	ekeinos	PnDm03MaSgNm
	tetoios	PnDm03MaSgNm
	tosos	PnDm03MaSgNm

The first three of the above demonstratives may occur either on their own or followed by the nouns to which they refer, in which case the nouns are preceded by an article. The last two serve only as modifiers of nouns and do not have an independent nominal status - they never appear alone:

autos (o anthrwpos)
but
tetoios anthrwpos

The type of deixis is not encoded, although it is implicit in each pronoun.

One of the pronouns, *toutos*, is more common in everyday use in spoken language, rather than in the written/formal usage of Greek.

5.11.7 Indefinite pronouns

Indefinite	example	tag
	kanenas	PnId03MaSgNm
	kapoios	PnId03MaSgNm
	merikoi	PnId03MaPlNm
	kati	PnId03NeSgIc
	katiti	PnId03NeSgIc
	tipote	PnId03NeSgIc
	kamposos	PnId03MaSgNm
	kathe	PnId03MaSgNm
	kathenas	PnId03MaSgNm
	katheti	PnId03NeSgIc
	allos	PnId03MaSgNm

The first indefinite pronoun, *kanenas*, does not form a plural and has two forms for the masculine gender:

kanenas but also *kaneis*

Both *kanenas* and *tipote* have a negative as well as an indefinite meaning:

kanenas = nobody or somebody
tipote = nothing or something

They have only the indefinite interpretation in affirmative and interrogative sentences, and they have both interpretations in negative sentences and answers to interrogatives:

Eheis tipote;
but
dhen ehei tipote pou na tou pighainei

The two above pronouns, along with *kati*, *katiti*, *tipote*, *kathe* and *katheti* are uninflected. The pronoun *merikoi*, on the other hand, does not form the singular, but inflects for case.

All other pronouns are inflected as adjectives (i.e. they mark gender, number and case).

Most of them function both as pronouns and pronominal adjectives. The pronoun *kathe* functions only as a pronominal adjective; its "respective" pronoun is *kathenas*.

5.11.8 Interrogative Pronouns

Three interrogative pronouns are distinguished for Greek.

Interrogative	example	tag
	ti	PnIr03NeSgIc
	poios	PnIr03MaSgNm
	posos	PnIr03MaSgNm

As indicated by the values of its tag, the first pronoun, *ti*, is indeclinable, while the other two are inflected for gender, number and case, agreeing with the object they refer to. They are all used as pronouns and pronominal adjectives.

5.11.9 Relative pronouns

Two relative pronouns exist for Greek, one of which, namely the indeclinable *pou*, is encoded in the Morphological Lexicon as a separate category. The other pronoun is *opoios*, which is always preceded by the definite article and agrees in gender, number and case with its referent; it is considered more "formal" than "pou". Again, in the following table, only the nominative case of the singular of the masculine is given.

Relative	example	tag
	opoios	PnRe03MaSgNm

5.11.10 Relative-Indefinite Pronouns

Relative-Indefinite	example	tag
	osos	PnRi03MaSgNm
	o,ti	PnRi03NeNvIc

Contrary to *o,ti*, which is indeclinable, *osos* is inflected for gender, number and case.

5.11.11 Definite Pronouns

Definite	example	tag
	idhios	PnDf03MaSgNm
	monos	PnDf03NeSgNm

The definite pronouns are actually adjectives which function as pronouns when the following conditions are fulfilled:

- *idhios* must be preceded by the definite article, and it must follow or precede the noun it refers to, which must have the definite article:

Eghw i idhia tha to kanw
or *O idhios o Takys tha to kanei*

- *monos* must appear without an article and must be followed by one of the possessive pronouns:

Efughe monos tou

5.11.12 Clitic Pronouns

These are actually the weak forms of the personal pronouns. They are morphologically marked for Person, Number and Case. For certain cases, there exists no weak form, as observed in the following table.

Personal				example	tag
Person	**Number**	**Case**	**Gender**		
1	*sg*	*gen*	-	mou	PnCl01SgGe
1	*sg*	*acc*	-	me	PnCl01SgAc
2	*sg*	*gen*	-	sou	PnCl02SgGe
2	*sg*	*acc*	-	se	PnCl02SgAc
3	*sg*	*nom*	*masc*	tos	PnCl03MaSgNm
3	*sg*	*gen*	*masc*	tou	PnCl03MaSgGe
3	*sg*	*acc*	*masc*	ton	PnCl03MaSgAc
3	*sg*	*nom*	*fem*	ty	PnCl03FeSgNm
3	*sg*	*gen*	*fem*	tys	PnCl03FeSgGe
3	*sg*	*acc*	*fem*	ty(n)	PnCl03FeSgAc
3	*sg*	*nom*	*neuter*	to	PnCl03NeSgNm
3	*sg*	*gen*	*neuter*	tou	PnCl03NeSgGe
3	*sg*	*acc*	*neuter*	to	PnCl03NeSgAc
1	*pl*	*gen*	-	mas	PnCl01PlGe
1	*pl*	*acc*	-	mas	PnCl01PlAc
2	*pl*	*gen*	-	sas	PnCl02PlGe
2	*pl*	*acc*	-	sas	PnCl02PlAc
3	*pl*	*nom*	*masc*	toi	PnCl03MaPlNm
3	*pl*	*gen*	*masc*	tous	PnCl03MaPlGe
3	*pl*	*acc*	*masc*	tous	PnCl03MaPlAc
3	*pl*	*nom*	*fem*	tes	PnCl03FePlNm
3	*pl*	*gen*	*fem*	tous	PnCl03FePlGe
3	*pl*	*acc*	*fem*	tis/tes	PnCl03FePlAc
3	*pl*	*nom*	*neuter*	ta	PnCl03NePlNm
3	*pl*	*gen*	*neuter*	tous	PnCl03NePlGe
3	*pl*	*acc*	*neuter*	ta	PnCl03NePlAc

As shown in the above table, the **Gender** feature is left unspecified for the first and second persons, given that it is not applicable to them.

6 Determiner

P	Type	whType	P	G	N	Case	Pos
M U L T			1 2 3	m f n	s p		
G E N E L E X	dem int poss card indf def rel part excl		1 2 3	m f n	s p		sg pl
A l D			1 2 3	m f n	s p		
N E R C	poss dem indf int/rel	 int rel	1 2 3	m f mf	s p sp	nom gen dat acc	
L e e c	dem indf poss int/rel	int rel	1 2 3	m f n c	s p	nom gen dat acc	sg pl

E-L0	**DETERMINER**						
E L 1	dem indf poss int/rel		1 2 3	m f n	s p	nom gen dat acc	sg pl
2 a		int rel excl					
2 b				It c	It n		

6.1 Comments

One of the major problems presented by this category is that the Romance tradition makes use of the label Pronominal Adjective, whereas other languages (English) use Determiner.

The two classes are not mappable one on to the other. This is a crucial problem, involving the different behaviour of the Determiners and Pronominal Adjectives.

The choice of calling e.g. *some* and *all* respectively Determiner and Predeterminer in English (both Indefinite Pronominal Adjectives in the Italian tradition) rests on their particular distribution in context: *some children, all the children*, and this can also work in Italian: *alcuni ragazzi, tutti i ragazzi*. However, if calling Determiners the possessives in English works for their complementary distribution wrt the article (*my book, the book*), this is not the case in Italian, because *il libro* does not have correspondence in **mio libro*. (As far as Italian possessives are concerned, a complementary distribution with article is found only with a closed number of family nouns: *mio/il padre, mia/la madre, mio/il fratello, mia/la sorella*).

The particular behaviour of possessives perhaps influenced the GENELEX choice to opt for a separate Determiner category, as well as the Adjective (where Pronominal Adjectives are also included) and Pronoun categories. The situation in the GENELEX model is as follows:

le nôtre PRON
nôtre chien DET
le chien est nôtre ADJ

This fact implies the presence of the values typical of Pronominal Adjective in the table of the Adjective category, among the values of the feature Type.

In NERC, the problem of the use of different tags, i.e. Determiner and Pronominal Adjective, in different traditions has been raised, and the proposal has been to opt for the label Determiner, without any further functional distinction.

The TEI work-group on annotation (TEI AI1W2 1991) has made the choice of inserting Determiners in the Adjective category, with the feature 'pronominal'.

6.2 Application to Italian

In Italian, both corpus and dictionary distinguish Determiners according to the feature Type (e.g. indefinite, demonstrative, possessive, etc.).

In the following sections, each Type will be discussed in detail.

6.2.1 Type and wh-Type

Attribute	*value*	ex.	tag
Type	*dem*	questo (libro)	DD/ms
	poss	mio (padre)	DP/ms
	indf	ogni (uomo)	DI/ms
wt-Type	*int*	quale (domanda)	DT/ns
	rel	il quale (uomo)	DR/ms
	excl	quanto (sole)!	DE/ms

6.2.2 Possessive

Possessive	example	It.tag
	(i) miei (amici) mio (cugino)	DP/mp DP/ms
	(la) nostra (casa) (i) nostri (cugini)	DP/fs DP/mp

In English, Possessives are classified as Determiners on the basis of their complementary distribution wrt the article (*my book, the book*). This kind of distribution in Italian works only with a closed number of family nouns, used in the singular: *mio/il padre, mia/la madre, mio/il fratello, mia/la sorella* (see the table above). Other nouns do not show this correspondence, e.g *il libro / *mio libro.*
At the level of encoding, this different behaviour is not represented.
Possessives are inflected for Number and Gender and agree with the noun; they are distinguished according to the Person which is referred to (see the table for Pronouns above).

Scrivo con la tua DP/fs *penna, perche' non ho la mia*

The two Italian possessives: *altrui* (of other people) and *proprio* (own) can be used in determiner function:

spende il denaro altrui DP/nn, *non il proprio*
occorre dare del proprio DP/ms *denaro*, non dell'altrui

As already mentioned for Possessive Pronouns, information about the possessor is not encoded in the Italian lexicon or corpus, but it can be inferred from the lemma.

6.2.3 Demonstrative

Demonstrative	example	It.tag
	questo (uomo)	DD/ms
	quelle (donne)	DD/fp

6.2.4 Indefinite

Indefinite	example	It.tag
	ogni (uomo)	DI/ms
	alcune (donne)	DI/fp

Italian Indefinites are inflected for Gender and Number.

They cover the English class of Quantifiers; some English practices (Brown) usually encode some quantifiers as Predeterminers, (e.g. *all the girls*), on the basis of their peculiar behaviour, i.e. they precede the determiner. *Tutto* in Italian has the same behaviour, which is not annotated:

tutte le ragazze DI/fp

Some can only be pronominal adjectives, i.e. determiners in this proposal:
ogni, qualche, qualunque, qualsiasi, qualsivoglia.

ogni colpa si sconta DI/fs

Alcuno, ciascuno, taluno, nessuno, tutto, alquanto, poco, molto, troppo, tanto can have both the pronoun and pronominal adjective/determiner function:

ho letto tutto il libro DI/ms

6.2.5 Interrogative

Interrogative	example	It.tag
	quanto (zucchero)?	DT/ms
	quale (libro)?	DT/ns

Interrogatives are inflected for Gender and Number.

Che, quale, quanto can be either pronouns or pronominal adjectives/determiners:

Quale vestito scegli? DT/ns

6.2.6 Exclamatory

Exclamatory	example	It.tag
	quanto (vento)!	DE/ms
	che (orrore)!	DE/nn

Che, quale, quanto can also have exclamatory value:

quante persone hanno aderito! DE/fp

6.2.7 Relatives

Relative	example	It.tag
	il quale	DR/ms

Il quale is inflected for Gender and Number, and can be used in adjectival function:

..., il quale film mi e' piaciuto molto DR/ms

It should be noted that it constitutes a bigram and should be encoded as a multiword expression.

6.3 Application to German

6.3.1 Type

Attribute	*value*	example	tag
Type	*demonstrative*	dieses (Buch)	**PDEMAT**:Neut.Nom.Sg
	indefinite	irgendein (Buch)	**PROAT**:Neut.Nom.Sg
	possessive	mein (Buch)	**PPOSAT**:Neut.Nom.Sg
	interrogative	welches (Buch) ?	**PWAT**:Neut.Nom.Sg

6.3.2 Gender

Attribute	*value*	example	tag
Gender	*masculine*	dieser (Mann)	PDEMAT:3.Sg.Nom.**Masc**
	feminine	meine (Mutter)	PPOSAT:**Fem**.Nom.Sg
	neuter	welches (Buch)	PWAT:**Neut**.Nom.Sg

6.3.3 Number

Attribute	*value*	example	tag
Number	*singular*	kein (Mensch)	PROAT:Masc.Nom.**Sg**
	plural	welche (Kinder)	PWAT:Neut.Nom.**Pl**

6.3.4 Case

Attribute	*value*	example	tag
Case	*nominative*	kein (Mensch)	PROAT:Masc.**Nom**.Sg
	genitive	dieses (Mannes)	PDEMAT:Masc.**Gen**.Sg
	dative	keinem (Menschen)	PROAT:Masc.**Dat**.Sg
	accusative	welchen (Mann)?	PWS:Masc.**Akk**.Sg

6.3.5 Possessor

The feature *possessor* is not encoded in the IMS-Tagset. It can be determined from the lemma.

6.4 Application to English

6.4.1 Type

Attribute	*values*	Examples	Tags
Type	*possessive*	your	DV2
	demonstrative	those	DDp
	wh-type	what	DW
	indefinite	every	DIs

6.4.2 Person

Attribute	*values*	Examples	Tags
Person	*first*	my, our	DV1ps, DV1pp
	second	your	DV2
	third	his, her, their	DV3psM, DV3psF, DV3ppN

For determiners, as for pronouns, distinctions of person, gender and number sometimes apply. Person applies to possessive determiners, and gender applies to third person singular possessive determiners.

6.4.3 Gender

Attribute	*values*	Examples	Tags
Gender	*masculine*	his	DV3psM
	feminine	her	DV3psF

6.4.4 Number

Number applies to demonstrative determiners and, in some cases, to indefinite determiners:

Attribute	*values*	Examples	Tags
Number	*singular*	this, much	DDs, DIs
	plural	these, many	DDp, DIp

The number of the possessor is also distinguishable for possessive determiners. However, for the present tagset, this attribute is omitted, being unimportant as an indicator of syntactic function.

6.5 Application to Spanish

Case does not apply to Spanish determiners. As for the other features, each Type is discussed in detail in the following sections.

Demonstrative: DD

Pers	Number	Gender	Pos	Case	Ex.
	sg	masc			este (libro)
	pl	fem			esas (casas)
	sg	masc			aquel (chico)

Only Number and Gender are pertinent attributes for Determiner demonstratives.

These determiners reflect deictic degree of remoteness, which is not, however, coded in ET-ES dictionaries.

Possessive: DP

*Possessive Determiners need the same attributes as the Pronominal Possessives. They differ in that in this case we have determiners valued as common wrt gender and possessor:

Pers	Number	Gender	Pos	Case	Ex.
1	sg	c	sg		mi (libro)
2	sg	fem	pl		vuestra (casa)
3	sg	c	c		su (libro)

In this case, the attribute "person" refers to semantic person. This attribute is not coded in ET-ES dictionaries.

Indefinite DI

The main ones are: *algún, ningún, cierto, varios, cualquier, poco, bastante, mucho, demasiado, demás, cuanto, todo, cada, tal,otro, un, tanto.*
They are variable in gender except for *bastante*, *cualquiera*, *cada*, *demás* and *tal*, and variable in number except for *cada*, *demas* and *varios*.

Interrogative

These are: *qué (libros), cuál (caja), and cuánto (dinero).*
"Qué is invariant in gender and number. Only *cuánto* inflects for gender.

Relative

The Spanish relative determiner inflects for number and gender (*cuyo, cuya, cuyos and cuyas*).

6.6 Application to French (Corpus)

6.6.1 Determiner type

Attribute	*value*	Example	Tag
Type	*dem,poss*	cette,ma (maison)	**DETR**FS
	indf	aucune (maison)	**ADJI**FS
	int	quelle (maison)	**DINT**FS

Note that the IBMF tagset only distinguishes two determiner types: interrogative and others (demonstrative, possessive, and articles). The indefinite determiner is classified among the adjectives in the said tagset.

6.7 Application to French (Lexicon)

6.7.1 Det.-Type

Attribute	Value	Example	Code
Type	demonstrat.	ce	d
	indefinite	certain	i
	possessive	mon	p
	interrogat.	quel	t

6.7.2 Person

Attribute	Value	Example	Code
Person	first	ma	1
	second	ta	2
	third	sa	3

6.7.3 Gender

Attribute	Value	Example	Code
Gender	masculine	cet,il	m
	feminine	cette,elle	f

6.7.4 Number

Attribute	Value	Example	Code
Number	singular	certain	s
	plural	certains	p

6.7.5 Possessor

Attribute	Value	Example	Code
Possessor	singular	mon	s
	plural	nos	p

6.7.6 Combinations

Tag	Example
Ds1fss--	ma (tasse)
Ds1fsp--	notre (tasse)
Ds1fps--	mes (tasses)
Ds1fpp--	nos (tasses)
Ds1mss--	mon (livre)
Ds1msp--	notre (livre)
Ds1mps--	mes (livres)
Ds1mpp--	nos (livres)
Ds2fss--	ta (tasse)
Ds2fsp--	votre (tasse)
Ds2fps--	tes (tasses)
Ds2fpp--	vos (tasses)
Ds2mss--	ton (livre)
Ds2msp--	votre (livre)
Ds2mps--	tes (livres)
Ds2mpp--	vos (livres)
Ds3fss--	sa (tasse)
Ds3fsp--	leur (tasse)
Ds3fps--	ses (tasses)
Ds3fpp--	leurs (tasses)
Ds3mss--	son (livre)
Ds3msp--	leur (livre)
Ds3mps--	ses (livres)
Ds3mpp--	leurs (livres)
Dd-fs---	cette
Dd-ms---	cet, ce
Dd-fp---	ces
Dd-mp---	ces
Dn-fs---	aucune, nulle, certaine, toute, chacune...
Dn-ms---	aucun, nul, certain, tout, chacun...
Dn-fp---	certaines, toutes...

Dn-mp---	certains, tous...
Dt-fs---	quelle
Dt-ms---	quel
Dt-fp---	quelles
Dt-mp---	quels
Di-fs---	aucune, nulle, certaine, toute, chacune...
Di-ms---	aucun, nul, certain, tout, chacun...
Di-fp---	certaines, toutes...
Di-mp---	certains, tous...

6.8 Application to Portuguese

6.8.1 Type

As is easily observable, articles are included under the category *determiner* in the Portuguese application:

Attribute	Value	Example	Tag
Type	demonstrative		
	indefinite	um	
	possessive		
	interrogative/relative		
l-specif	definite	o	
	quantifier	algum	
	cardinal	um	

6.8.2 Wh-Type

Not used in the Portuguese application.

6.8.3 Person

Not used in the Portuguese application.

6.8.4 Gender

Attribute	Value	Example	Code
Gender	masculine	um	
	feminine	uma	
	neuter		

6.8.5 Number

Attribute	Value	Example	Code
Number	singular	algum	
	plural	alguns	

6.8.6 Case

Not used in the Portuguese application

6.8.7 Possessor

Not used in the Portuguese application.

6.9 Application to Danish

The label Determiner as used within the present EAGLES proposal obviously covers a number of lexical items classified as pronouns and quantifiers. Traditional Danish grammars and dictionaries do not operate with a common designation for the determination function. The application of the label Determiner as a member of the Danish tagset for corpus annotation has not yet been clarified.
In Danish, from a functional point of view, members of certain word classes, like pronoun or noun (viz. subclasses specifier and classifier), can be considered as determiners:

Determiner type

Attribute	*value*	Example	Tag
Type	*demonstrative*	denne	pron_dem_com_ sg
	possessive	min	pron_poss_sg1_com_sg
	quantifier	begge	quant
	ordinal	tredie	ord
	cardinal	tre	card
	specifier	halvdelen (af)	spec
	classifier	en gruppe	class

Although articles and possessive genitives also function as determiners, they have not been included in the table above. We follow here an approach wherein articles (as a separate category) are distinguished from determiners, and possessive genitives do not constitute a category as such.

6.10 Application to Greek

For the classification of the items that fall under the categories of Pronouns and Determiners, the ILSP Morphological Lexicon followed the traditional grammar distinctions. Although, as mentioned in the relevant section, "absolute" pronouns can be considered to be the only personal and demonstrative pronouns, we have included (in a broad interpretation of the term) all adjectival pronouns as well.

Furthermore, it is the case that certain "indefinite pronouns" have a strong quantificational character; such cases are "arketoi" (several), "merikoi" or "kamposoi" (some), "kanenas" (none). Based on their distributional behaviour, we could classify them as Determiners: they appear immediately before the Noun, or alone, they are never modified by degree adverbials, they are never introduced by articles, etc. However, we opted for the traditional classification of all these items under the category of Pronouns. Thus, since there is no attribute available for their codification (no tag assigned), no tables are presented in this section.

7 Article

ARTICLE	Type	Gen	Num	Case
MULTILEX	def indf	m f n	s p	nom gen dat acc voc
GENELEX	(*in Det)			
AlethDic	def indf	m f n	s p	
NERC	def indf	m f mf	s p sp	nom gen dat acc
Leech	def indf	m f n c	s p	nom gen dat acc

EAG-L0	**ARTICLE**			
EAG-L1	def indf	m f n	s p	nom gen dat acc
EAG-L2a				
EAG-L2b		It c	It n	

7.1 Comments

Articles constitute a separate category, as is found in many lexicons and annotation practices, even though, as already pointed out in section 5, in others (e.g. Penn) they are found incorporated into the Determiner category.

The separation of the two categories is not a problem, since, as Articles form a very restricted closed class, they can also easily be distinguished from Determiners in the latter cases, and no serious mapping problem arises.

7.2 Application to Italian

Article traditionally constitutes a class of its own.

7.2.1 Article

The class of Articles has a finite list of members. The two possible Types 'definite' and 'indefinite' are marked in the lexicon (RD and RI respectively), while in the corpus they are not encoded. The other pertinent features are Number and Gender.

Article				It.tag
		Type		
Gender	**Number**	*definite*	*indefinite*	
m	*s*	il, lo	un, uno	RD/ms
n	*s*	l'		RD/ns
m	*p*	i, gli		RD/mp
f	*s*	la	una/un'	RI/fs
f	*p*	le		RI/fp

Un, il are graphical variants of *uno, lo*, used before words beginning with simple consonant, or beginning with a consonantal group plus *l, r : un trono, il cloro* (not if the the group of consonants begins with *s*).

Uno, lo are used before the consonantal digram beginning with *s*, before the digram *gn* and before *z, x* and consonantic groups whose second element is not *l, r.* They are also used before words beginning with a vowel and in this case they are elided.

The article helps in general in disambiguating other words in the context: *l'*, sometimes, is not sufficient to disambiguate: *l'insegnante capace parte domani* (can be both masculine and feminine).

7.3 Application to German

7.3.1 Types

Attribute	*value*	example	tag
Type	*definite*	der (Mann)	ART:**Def**.Masc.Nom.Sg
	indefinite	ein (Mann)	ART:**Indef**.Masc.Nom.Sg

7.3.2 Gender

Attribute	*value*	example	tag
Gender	*masculine*	der (Mann)	ART:Def.**Masc**.Nom.Sg
	feminine	die (Mutter)	ART:Def.**Fem**.Nom.Sg
	neuter	ein (Buch)	ART:Indef.**Neut**.Nom.Sg

7.3.3 Number

Attribute	*value*	example	tag
Number	*singular*	der (Mann)	ART:Def.Masc.Nom.**Sg**
	plural	die (Dinge)	ART:Def.Neut.Nom.**Pl**

7.3.4 Case

Attribute	*value*	example	tag
Case	*nominative*	der (Mann)	ART:Def.Masc.**Nom**.Sg
	genitive	des (Vaters)	ART:Def.Masc.**Gen**.Sg
	dative	einem (Mann)	ART:Indef.Masc.**Dat**.Sg
	accusative	den (Mann)	ART:Indef.Masc.**Akk**.Sg

7.4 Application to English

The attributes of Gender and Case do not apply to articles in English. Only two words belong to this category: *the* and *a*, (*an* being the variant of *a* occurring before vowels.) *The* is invariable.

7.4.1 Type

Attribute	*values*	Examples	Tags
Type	*definite*	the	ATD
	indefinite	a, an	ATIs

7.4.2 Number

Attribute	*values*	Examples	Tags
Number	*singular*	a, an	ATIs
	plural	–	–

7.5 Application to Spanish

7.5.1 Article Type

Attribute	*value*	Example	Tag
Type	*definite*	el	msdefs
	indfefinite	un	msindef

7.5.2 Gender

Attribute	*value*	Example	Tag
Gender	*m*	el	
	f	la	
	n	lo	

7.5.3 Number

Attribute	*value*	Example	Tag
Number	*s*	un	
	p	unos	

There are different positions on the “neuter” article in the Spanish grammatical literature. It is only identifible in singular, since in the plural the form would coincide with the masculine plural. Normally it is used for nominalizations:

lo bueno es ... (the good thing is...)

7.5.4 Case

Case does not apply to Spanish articles.

7.6 Application to French (Corpus)

7.6.1 Definiteness

French has definite and indefinite articles (*le, la les, un, une, des*). However, there is no distinction in the IBMF tagset because it would have no predictive power and it can be retrieved from the graphic form.
In the tagset, articles are coded as regular determiners.

7.6.2 Gender

Attribute	*value*	Example	Tag
Gender	*masculine*	le, un	DETRMS
	feminine	la, une	DETR**F**S

7.6.3 Number

Attribute	*value*	Example	Tag
Number	*singular*	le, un	DETRM**S**
	plural	les, des	DETRM**P**

7.6.4 EAGLES features not applicable

Case does not apply to French.

7.7 Application to French (Lexicon)

7.7.1 Type

Attribute	Value	Example	Code
Type	definite	le	d
	indefinite	un	i

7.7.2 Gender

Attribute	Value	Example	Code
Gender	masculine	le	m
	feminine	la	f

7.7.3 Number

Attribute	Value	Example	Code
Number	singular	le	s
	plural	les	p

7.7.4 Case

Case is not applicable to French.

7.7.5 Combinations

Tag	Example
Tdms-	le
Tdfs-	la
Tdmp-	les
Tdfp-	les
Tims-	un
Timp-	une
Tifs-	des
Tifp-	des

7.8 Application to Danish

Gender, number and definiteness are pertinent to Danish articles. The table below covers the whole set of articles. The corpus tagger for Danish will treat articles as incorporated into the pronoun category.

Attribute	*value*	Example	Tag
Type	*indefinite*	en	art_com_sg_indef
	indefinite	et	art_neut_sg_indef
	definite	den	art_com_sg_def
	definite	det	art_neut_sg_def
	definite	de	art_pl_def

language-specific property

The enclitic article

Definiteness of nouns must be expressed in certain cases by means of enclitic articles which function as suffixes: *-(en)* singular common, *-(e)t* singular neuter and *-(e)ne* in plural.
In corpus tagging this phenomenon will be annotated as a value of the noun definiteness feature.

7.9 Application to Greek

Two types of articles are distinguished in Greek: definite and indefinite, each class including one member:
- definite article: o, y, to,
- indefinite article: enas, mia, ena.

In the first version of the ILSP Morphological Lexicon and the corresponding Tagset, these two were both coded under the grammatical category of *article*, further coded for the feature Type with the values *definite* and *indefinite*. At a later stage, however, this was abandoned and currently only the definite article is characterised as *article*. The indefinite article has been elevated to a category of its own, *enas*, owing to its high ambiguity: besides indefinite article, it can also be a cardinal and an indefinite pronoun. The resolution of the ambiguity, even with the use of the linguistic context, is extremely difficult to perform automatically:

Mono enas anthrwpos ytan ekei. - card.
Ton eidha mia mera. - ind. art.
Perase enas apo edhw kai se zytaghe. - ind. pron.

The following tables present the values and tags of the attribute **Type** as used in the first version, in accordance with the EAGLES proposal.

Attribute	*value*	Gr. example	Gr. tag
Type	*definite*	o	At**Df**
	indefinite	enas	At**Id**

Articles are further coded for Gender, Number and Case.

7.9.1 Gender, Number and Case

Articles agree in gender, number and case with the nouns they modify. These attributes have almost all the values presented in the relevant sections of Nouns, with the exception of *masc-fem* for Gender, *invariant* for Number and *voc* and *indcl* for Case. Certain forms are homographs, necessitating the use of a disambiguation procedure during corpus tagging:

tou **vivliou** - neut. gen.
tou **anthrwpou** - masc. gen.
but
to **exwfullo** - neut. nom./acc.

Attribute	*value*	Gr. example	Gr. tag
Gender	*masculine*	o	AtDf**Ma**
	feminine	y	AtDf**Fe**
	neuter	to	AtDf**Ne**

Attribute	*value*	Gr. example	Gr. tag
Number	*singular*	to	AtDfNe**Sg**
	plural	ta	AtDfNe**Pl**

Attribute	*value*	Gr. example	Gr. tag
Case	*nom*	o	AtDfMaSg**Nm**
	gen	tou	AtDfMaSg**Ge**
	acc	ton	AtDfMaSg**Ac**

8 Adverb

ADVERB	Type	Degree	Polarity	wh-Type
MULTILEX	(14 val.s)		true false	
GENELEX	time place manner quant degree compar superl	comp= comp+ comp– sup+ sup– supabs		
AlethDic				
NERC		posit compar superl		
Leech	general degree particle*	posit compar superl	Q-Type other	inter relat

EAG-L0	**ADVERB**			
EAG-L1	general particle	pos comp sup		
EAG-L2a				
EAG-L2b	En degr Fr ne Fr pas Du conj Du pron Du sep1 Du sep2		En-Sp wh En-Sp no-w	En-Sp int En-Sp rel

8.1 Comments

8.1.1 Type

The MULTILEX standard proposes among the “syntactic-properties of LU’s” the attribute ‘Advsubclass’, which corresponds to Type, with 13 values based on distinctions of semantic type, plus the value ‘other’ for adverbs other than those indicated by the 13 possible values. For each value application tests and examples are provided (MULTILEX 1993, p.69).
In the GENELEX model, the values for Type are proposed at the level of syntactic specifications.
The TEI also proposes values on semantic grounds.
The feature Type was not proposed in the nucleus of minimal standard specifications by NERC, since most of the analysed tagsets do not make such a distinction. It appears difficult to map existing corpus tagsets onto the subclassifications of adverbs proposed by the values of this feature.
Another interesting fact is the inclusion of Particles among the values of the feature Type by the Leech/Wilson proposal. The treatment of Particles appears to be somewhat complicated: NERC suggests their inclusion in the preposition category; in GENELEX and MULTILEX they form a separate category without any attribute.

In the present EAGLES proposal, the values proposed for the feature Type at recommended level reflect a first distinction between general adverbs and particles. All further language-specific or application-dependent distinctions can be made at level 2b.

8.1.2 Degree

The values of this feature are ‘comparative’, ‘superlative’, ‘positive’. ‘Positive’ can be seen as the default. GENELEX has three more distinctions for each of them.

8.1.3 Polarity

This is a language-dependent feature: most English tagsets distinguish between interrogative/relative adverbs (wh-adverbs) and the others. This distinction is also relevant for the Spanish EUROTRA Lexicon.

8.1.4 Other Distinctions

We must observe that, even though many richer classifications for Adverbs exist and are used both in lexicons and in tagging practices, these are based on semantic grounds and there is too much disagreement between them. They are, therefore, left as a matter for individual systems.

8.2 Application to Italian

Adverbs are very similar to the adjective in that, in general, they behave similarly wrt the verb as the adjective does wrt the noun.

Una parola dolce A/ns vs. *Parlare dolcemente* B/

In traditional Italian grammar, adverbs are subdivided into different subcategories on the basis of their semantic value: manner, time, place, quantity, comparative, affirmative, negative, relative/interrogative, etc.

In corpus tagging and dictionary encoding not all these types are distinguished.

8.2.1 Degree

Attribute	*value*	Example	Tag
Degree	*positive*	bene	B
	comparative	meglio	BC
	superlative	benissimo	BS

8.3 Application to German

8.3.1 Type

There are three types of adverbs:

- **ADV**: *general* (i.e. *non-adjectival*, cf. section refgerman:adj-use)
- **PWAV**: *interogative or relative*
- **PROAV**: *pronominal* adverbs. These are *da* or *hier* + preposition (such as *damit, dafür, dagegen, daneben etc*), which replace a PP with the specific preposition.

Example: er kommt heute/ADV
er kommt schnell/ADJNA
wo/PWAV wohnt er?
der Ort, wo/PWA er wohnt
er besteht darauf/PROAV

Semantic distinctions like *time, place,* etc. are not expressed in the annotations.

Attribute	*value*	example	tag
Type	*general*	oft	**ADV**
	int/rel	wo	**PWAV**
	pronominal	hierbei	**PROAV**

8.3.2 Degree of comparison

(This feature is not used for adverbs in the IMS-Tagset as it applies only to a very few non-adjectival adverbs.)

Attribute	*value*	example	tag
Degree	*positive*	oft	ADV:**Pos**
	comparative	öfter	ADV:**Comp**
	superlative	(am) öftesten	ADV:**Sup**

8.4 Application to English

In English, adverb particles are an important category because of their role in the forming of phrasal verbs (compare separable verb forms in German and Dutch: 3.5.11).

8.4.1 Adverb-type

Attribute	*values*	Examples	Tags
Adverb-type	*general*	quickly	AV
	degree	very	AVD
	particle	up	AVP

8.4.2 Degree

Attribute	*values*	Examples	Tags
Degree	*positive*	soon	AV
	comparative	sooner	AVR
	superlative	soonest	AVT

8.4.3 Polarity

Attribute	*values*	Examples	Tags
Polarity	*wh-type*	where?	AVWQ
	other	here	AV

8.4.4 Wh-subtype

Attribute	*values*	Examples	Tags
Wh-subtype	*relative*	when	AVWR
	other	when?	AVWQ

(Compare Pronoun and Determiner above.) (Similar distinctions are made under Pronoun and Determiner above, 5.4.2 and 6.3.1).

8.5 Application to Dutch

8.5.1 Type: Language-specific features

Dutch grammar distinguishes two subcategories of adverbs. One category is called *pronominal adverb*, because it has the same referential function as a pronoun. The other is called *conjunctival adverb*, because it functions as a conjunction.

The table below is a proposal for Dutch, not a CELEX table !

Attribute	*value*	Example	Tag
Type	*normal*	(hij kmot) vaak	Adv
	pronominal	(Ik reken) erop daar (houk ik van)	PronAdv SepPronAdv
	conjunctival	bovendiven (is hij ziek) waar (het betref)	ConjAdv

As pronominal adverbs are separable word forms, Dutch needs a way of marking up the following three possibilities:

The pronominal adverb as a whole: Tag = PronAdv
The first part of a separable pronominal adverb: Tag = SepPronAdv
The second part of a separable pronominal adverb: Tag = SepAdp (see Adpositions).

This can be an L2b tag.

8.5.2 Degree of comparison

CELEX does not have this distinction at adverbial level. It does have this information at adjectival level.

8.6 Application to Spanish

8.6.1 Type (semantic)

Attribute	*value*	Example	Tag
Type	*general*		
	degree		

The Spanish Eurotra dictionaries include "semantic typing" at the syntactic level in order to deal with several phenomena related to the attachment of modifiers. The feature used and its values are as follows:

vadv = time, manner, deg(ree), loc(ation), free, manconcret.

Also, in relation to overgeneration in attachment, some features control the attachment of adverbs in order to prevent sentences as:

* Todo termin'o desgraciadamente
(Everything finished unfortunately)

Desgraciadamente aquello era verdad
(Unfortunately that was true)

8.6.2 Degree

Attribute	*value*	Example	Tag
Degree	*positive*	muy	
	comparative	mejor	
	superlative		

In the ET-ES dictionaries there is no attribute to code the values of "degree".

In Spanish, the comparative of adverbs is formed analytically with *tan*, *más*, or *menos*:

más rápido que yo (faster than I)
más pronto (soonner)

Only four adverbs are have irregular comparative morphology:

bien/mejor (well/better)
mal/peor (badly/worse)
mucho/más (much/more)
poco/menos(little/less)

8.6.3 Polarity

The Eurotra dictionaries have the attribute "whmor", with the three possible values "int", "rel" and "none", which serves to distinguish between interrogative, relative and other adverbs:

Attribute	*value*	Example	Tag
Polarity		muy	
Wh-Type	*int*	dónde, cuándo, cómo	whmore=int
	rel	donde, cuando,	whmore=rel
non-Wh		aquí, ahora, así	whmore=none

8.6.4 Apocope

Some adverbs (just as we saw wrt adjectives) change their form when they occurr before an adjective or another adverb:

tanto/tan (ricamente)
cuanto/cuan (largo)
mucho/muy (despacio)

8.7 Application to French (Corpus)

As in other languages, French has a tradition of distinguishing classes of adverbs (place, manner, etc.) on the basis of semantic considerations. Degree is one of these classes. However, the IBMF tagset has a unique tag for adverbs, **ADVE** (with the exception of negative particles, see below).

8.7.1 EAGLES features not applicable

Polarity and **wh-type** do not apply to French.

8.7.2 IBMF Tagset features not applicable in EAGLES

The tagset has two tags for specific negation adverbs: **NE** and **PAS** (the latter class containing *pas* and *plus*).

8.8 Application to French (Lexicon)

8.8.1 Type

Attribute	Value	Example	Code
Type	general	fortement	g
	particle	ne, pas	p

It seems necessary in French to distinguish the two parts of the negation (“ne ... pas”), because they play an important role in disambiguation.

8.8.2 Degree

Attribute	Value	Example	Code
Degree	positive	fortement	p
	comparative	davantage	c

We encode Degree for compatibility with other languages, but, as with adjectives, the comparative feature is not very productive in French. It applies only to beaucoup (comp.= “davantage”), bien (comp.= “mieux”), “mal” (comp.= “pis”) and “peu” (comp. = “moins”). The comparative for other adverbs is marked by “plus” + adverb (e.g. “plus fortement”). The superlative is usually marked by “le”+ comparative (e.g. “le plus fortement”).

8.8.3 Combinations

Tag	Example
Rgp	beaucoup
Rgc	davantage
Rpn	ne
Rpn	pas, plus

8.9 Application to Portuguese

8.9.1 Type

Not used in the Portuguese morphological application.

8.9.2 Degree

Attribute	Value	Example	Code
Degree	positive comparative superlative		
l-specif	comparative+ comparative- comparative=	mais menos ta~o	

In the Portuguese model the 'positive' value is not explicitly marked because it is considered a default value.

8.10 Application to Danish

In traditional Danish grammars the word class of adverbs is divided into subclasses according to their semantic content and possible position within a sentence (field grammar). Dictionaries do not deal with semantic subclasses.

The EDEMD distinguishes between real adverbs and derived adverbs. Furthermore, distinctions are made on the basis of the syntactic properties of adverbs (modifying role, scope and position in sentence). Generally, in corpus tagging and dictionary encoding the above mentioned distinctions are not used.

For real adverbs (i.e. not derived from adjectives) there are only a few relics of a system of degree comparison. Most grammars regard adverbs as uninflectable.

Attribute	*value*	Example	Tag
Degree	*positive*	ofte	adv_pos
	comparative	oftere	adv_comp
	superlative	oftest	adv_sup

8.11 Application to Greek

INSERT SECTION ON GREEK ADVERBS HERE.

9 Adposition

ADPOSITION	Type	Formation
MULTILEX	preposition postposition circumposition	
GENELEX	Preposition*	
AlethDic	Preposition*	
NERC	Preposition*	
Leech	preposition postposition	

EAG-L0	**ADPOSITION**	
EAG-L1	preposition postposition circumposition	simple fused
EAG-L2a		
EAG-L2b	Du 2nd-p Du sepadv	

9.1 Comments

9.1.1 Type

The values of this feature allow one to distinguish between 'preposition' and 'postposition'. MULTILEX also has a value for circumposition (see MULTILEX). In NERC, GENELEX and AlethDic, Prepositions constitute a separate category without any attribute at the morphosyntactic level; in GENELEX the distinction based on a semantic classification between 'temporal', 'manner' and 'place' is foreseen at the syntactic level.

In NERC, Prepositions also include Particles. In English there is a problem of distinguishing the verb particles from prepositions, but many tagsets provide a separate label for particles.

9.1.2 Formation

In some languages, e.g. in Italian (see below), prepositions may appear fused with articles in a unique graphical form. EAGLES encourages the treatment of these forms as separable, therefore having two tags corresponding to the two different categories, but also allows the use of a unique tag.

This general recommendation, i.e. to separate different categories and to encode them with different tags, is also applicable to other cases of contractions, such as negation, cliticisation, etc.

In general, there will be language-specific options for the treatment of these phenomena.

9.2 Application to Italian

9.2.1 Type

Attribute	*value*	example	It.tag
Type	*preposition*	di, a, da	E

In Italian, simple prepositons are distinguished in two categories:

(i) a closed set, *di, a, da, in, con, su, per, fra, tra* which are used to express a number of syntactic relations.
(ii) a larger set with more specific meanings, e.g. *sopra, sotto, prima, dopo, senza,* etc.. These constitute a crux in tagging because they are involved in transcategorization phenomena (see Conjunctions).
The tag is E.

Most of the prepositions in category (ii) require the presence of another preposition before the argument, *sopra di noi*: this raises the problem of which strategy to choose in multiword expression tagging (see Leech and Wilson Invitation Draft).

9.2.2 Formation

Prepositions in category (i) are often fused with following articles, forming the so-called *preposizioni articolate*: ex. *agli (a+gli), ai (a+i), degli (di+gli), dello (di+lo).* These compounds, which can be found instantiated only in a corpus, are tagged with the tag (E) plus the morphological features of the fused article.

Attribute	*value*	example	It.tag
Formation	*simple*	di, a, da	E
	fused	dagli, agli	E/mp

9.3 Application to German

Three types of adpositions are distinguished: *preposition, postposition* and *circumposition*. There are a few cases where prepositions and a following definite article are contracted:

Example: zum = zu + dem
ans = zu + das

9.3.1 Type

Attribute	*value*	example	tag
Type	*preposition*	ohne (mich)	AP**PR**:Akk
	postposition	(ihm) zuliebe	AP**PO**:Dat
	circumposition	von (dem Tag) an	AP**ZR**[9]

9.3.2 Formation

Attribute	*value*	example	tag
Formation	*simple*	in (das Haus)	APPR:Akk
		zu (den Leuten)	APPR:Akk
	fused	ins (Haus)	APPR**ART**:Neut.Akk.Sg
		zur (Schule)	APPR**ART**:Fem.Dat.Sg

[9] Only the second part of the circumposition is actually annotated with APZR: *von/APPR:Dat dem Tag an/APZR.*

9.4 Application to English

9.4.1 Adposition-type

Attribute	*values*	Examples	Tags
Adposition-type	*preposition*	at	APR
	postposition	's	APO

On the classification of the genitive particle *'s* as a postposition, see Nouns above.

9.5 Application to Dutch

9.5.1 Type

CELEX has no Adpositions of type preposition. It only has straightforward Prepositions as a part of speech. However the German table with attribute values: preposition, postposition and circumposition is applicable to Dutch as well. But, as mentioned above, we need a special language-specific attribute value to mark up the non-verbal parts of separable verb forms and another to mark up the second part of separable pronominal adverbs: *The table below is a proposal for Dutch, not a CELEX table* !

Attribute	*value*	Example	Tag
Type	*preposition*	zonder (mij)	PreP
	postposition	(de berg) op	PostP
	circumposition	van (die dag) af	CircumP
	2nd part of pron.adv.	(daar houd ik) van	SepAdp
	non-verb. part of sep.adv.	(gaven) aan	SepVAdp

9.6 Application to Spanish

9.6.1 Prepositions

In our dictionaries there are no special attributes for prepositions at the morphological level. At the configurational level, a semantic typing of prepositions helps to control attachment of modifiers.

ptype: norm, loc, orig, dest, nil

In Spanish some contractions of articles and prepositions can be found:

a + el = al
de + el = del

For analysis purposes, these contractions are treated at the morphological level in ET-ES grammars as samples of a special category "portmanteau", and are decomposed before reaching the configurational level ECS.

9.7 Application to French (Corpus)

Attribute	*value*	Example	Tag
Type	*preposition*	dans	PREP

Attribute	*value*	Example	Tag
Formation	*simple*	dans	PREP
	fused	au	PAU

The IBMF tagset has 2 classes of preposition:

PDEA for *à* and *de*

PREP for other prepositions

and 4 classes of preposition-determiner:

PAU for *au* (*à* + *le*)

PAUX for *aux* (*à* + *les*)

PDES for *des* (*de* + *les*)

PREPMS for *du* (*de* + *le*)

9.8 Application to French (Lexicon)

9.8.1 Type

Attribute	Value	Example	Code
Type	preposition	en, de	p

9.8.2 Combinations

Tag	Example
Sp	en

9.9 Application to Portuguese

9.9.1 Type

Attribute	Value	Example	Code
Type	preposition postposition circumposition	em	

9.9.2 Formation

This feature is not used in the Portuguese lexicon, but it is pertinent to the Portuguese language, where contraction phenomena occur. Following the GENELEX multilingual model, in the Portuguese lexicon contraction phenomena occuring typically with preposition plus determiner are dealt with as a special kind of morphological unity (the *unité morphologique agglutineé*.

9.10 Application to Danish

In automatic corpus tagging the recognition and annotation of circumpositions is a special task to be dealt with.

At present there does not exist a commonly agreed tagset for the three types of adpositions: simple preposition e.g. *til*, complex preposition e.g. *uden for* and circumposition e.g. *for ... skyld*. Furthermore, as in English, the genitive *'s* may be considered as an enclitic postposition. The table below may be regarded as a proposal for Danish.

Attribute	*value*	Example	Tag
Type	*preposition, simple*	til	p_s
	preposition, complex	uden for	p_c
	circumposition	for (denne sags) skyld	p_cc

9.11 Application to Greek

9.11.1 Type

Attribute	*value*	Gr. example	Gr.tag
Type	*preposition*	me, kata	Pp

As indicated by the above table, prepositions in Greek form a category of their own.

In Greek we distinguish between two kinds of prepositions:

(a) simple ones, included in this category, and

(b) complex ones, formed by the combination of an adverb and a preposition: e.g. *panw apo, mazi me, etc.*. These can only be recognised at text level by the use of a tagger that looks above the word level.

The category of simple prepositions constitutes a closed set, while the latter is productive.

9.11.2 Formation

One of the simple prepositions, namely *se*, when followed by the definite article, is fused with it and forms what we call *prepart*: ston (se + ton), styn (se + tyn), sto (se + to), stous (se + tous), stis (se + tis), sta (se + ta). These forms are further coded for the morphological features of the article (i.e. Gender, Number and Case).

Attribute	*value*	example	Gr.tag
Formation	*simple*	se	Pp**Sp**
	prepart	ston	Pp**Pa**

10 Conjunction

CONJUNCTION	Type	Coord-T	Subord-T	Iter	Sent.Intr
MULTILEX	Coord* Subord*	simple repetit correlat		yes no	yes no
GENELEX	coord subord				
AlethDic	coord subord				
NERC	coord subord				
Leech	coord subord	simple repetit correlat sent.init			
EAG-L0	**CONJUNCTION**				
EAG-L1	coord subord				
EAG-L2a			+infve compar +fin		
EAG-L2b		En simple En init En no-init Sp correl			

10.1 Comments

Conjunctions link syntactically two or more words or two or more syntagms: phrases, sentences, etc.

Coordinating: determines a syntactic equivalence between conjuncts:

Mary and John

Subordinating: links two sentences in a dependency relation:

I do not answer, since I was ...

A large core of agreement emerges as regards Conjunctions. All the analysed systems agree as to the distinction between coordinating and subordinating Conjunctions. The only difference is the MULTILEX position which splits the two types of Conjunctions into two different categories, 'coordinators' and 'subordinators': however, no mapping problems arise. The same choice was made by the TEI.

10.1.1 Coordination Type

Within the class of coordinators MULTILEX foresees some features pertaining to the syntactic level, which in the Leech/Wilson proposal are collected under a unique feature Coord-Type:
– 'simple': between conjuncts, (*John and Mery*),
– 'repetitive': before each conjunct, (*both John and Mary*),
– 'correlative': before a conjoined phrase, it requires specific coordinators between conjuncts, (*either John or Mary*).
'Repetitive' and 'correlative' imply the specification of **Expected Coordinator**, i.e. *both* requires *and* etc., and the specification of **Iteration**, i.e. whether a coordinator can coordinate more than two conjuncts, *and ... and ... and* (more than two: yes), *but* (two: no). These last two attributes are introduced by MULTILEX only.
Sentence Introducer is used for coordinators which can introduce a sentence, linking it by coordination to the preceding sentence: *And Mary left.* (*Either* cannot introduce a sentence in this way).

In the EAGLES Proposal these further distinctions concerning the Type of Coordination are treated as language-specific features.

10.1.2 Subordination Type

Further information on the Type of Subordination, i.e. if the conjunction requires a finite verb "+fin", a non-finite verb "+infve" or introduces a comparison "compar" are proposed on the Level-2b, since they are shared by three different languages.

10.2 Application to Italian

Conjunctions in Italian constitute an "open class": the same elements can often have the function of conjunctions, prepositions, adverbs.

Conjunction: *Ti sentirai tranquillo, dopo aver sostenuto l'esame*
Preposition: *Ti sentirai tranquillo dopo l'esame*
Adverb: *Sostenuto l'esame, dopo ti sentirai tranquillo*

For these classes, as far as possible, application tests have to be employed:
- preposition: the argument is an NP;
- conjunction: the sentence introduced by the elements is in a dependency relation with the governing predicate;
- adverb: it semantically modifies the whole sentence and the predicate in particular.

In grammars, no agreement is reached as to the inclusion of some elements in the conjunction or the adverb class (e.g. *anche, pure, dunque, pertanto*).

In some recent linguistic theories, all these elements go into the wider class of "connettivi" (connectives), which link various parts of a text (Berretta 1984).
In tagging Italian the problem of subordinating "locutions" (multi-words) has to be dealt with, e.g. *dal momento che*.

10.2.1 Type

Type	*value*	example	It.tag
	subordinating	perche'	C
	coordinating	e	CC

The dictionary proposes the distinction between the coordinating and subordinating function:

Maria e Giovanni CC
non ho risposto perche' ero disattento C

10.3 Application to German

10.3.1 Type

Attribute	*value*	example	tag
Type	*coordinating*	(ich) und (du)	KO**N**
	subordinating	(er will,) daß	KO**US**
	comparison	(kleiner) als	KO**KOM**

10.3.2 Subord-Type

Attribute	*value*	example	tag
Subord-Type	*finite*	(er kommt,) wenn	KO**US**
	with infinitive	ohne (zu wollen)	KO**UI**

10.4 Application to English

10.4.1 Conjunction-type

Attribute	*values*	Examples	Tags
Conjunction-type	*coordinating*	and, or	CC
	subordinating	if	CSF

10.4.2 Coord.-type

Attribute	*values*	Examples	Tags
Coord.-type	*simple*	and, or	CC
	initial	both (...and)	CCI
	non-initial	(neither...) nor	CCM

10.4.3 Subord.-type

Attribute	*values*	Examples	Tags
Subord-type	*with finite verb*	because	CSF
	with non-finite verb	in order (to)	CSN
	comparative	than	CSC

10.5 Application to Dutch

The German scheme is applicable to Dutch as well, but there is no such application in CELEX.

Attribute	*value*	Example	Tag
Type	*coordinat*	(ik) en (jij)	CoorConj
	postposition	(hij wil) dat	SubConj
	subord + inf.ve	zonder (te willen)	InfConj
	comparison	(kleiner) dan	CompConj

10.6 Application to Spanish

In Eurotra coordinating and subordinating conjunctions are treated as two different categories. While at the morphological level they have no special attributes, at the constituent level we have three attributes for the subordinating type:

– sconjtype (valued as circ, compar, compl), which serves to distinguish between:

circumstantial, e.g.: *MIENTRAS yo trabajo, el juega (while I work, he plays)*
comparative, e.g.: *Juan es tan lento COMO su hermano (John is as slow as his brother)*
complementizers, e.g.: *Juan dice QUE es verdad (John says that it is true)*

– ftranc, to lexically encode the value of the conjunction to make up a transconstructional modifier. [10]
– comp (valued as q, nonq). Complementizers are thus distinguished with respect to their interrogative value: "que/si" (that/whether)

10.6.1 Type

Attribute	*value*	Example	Tag
Type	*coord*	Juan Y Maria	cconj
	subord	no se SI vendré	sconj

cconj: *y, pero, ni, o*

sconj: *cuando, si,que, porque, aunque, mientras, etc.*

10.6.2 Coord-Type

Attribute	*value*	Example	Tag
Coord-T	*simple*		
	correlat	NI Juan NI María	
	initial		
	non-initial		

Coord-T distinctions are not coded in the ET-ES dictionaries. There are no examples for "non-initial" coordinating conjunctions.

[10] Definition of Transconstructionals: The Eurotra Reference Manual 7 defines transconstructionals as those constituents which do not modify the governor of the construction to which they belong, but provide information about the sentence as a whole.

10.6.3 Subord-Type

Attribute	*value*	Example	Tag
Subord-T	*+infve*	para	
	+fin	mientras	
	comparison	tan	

In the ET dictionaries, there exists the feature "sform" with values infin, finite, pastpart and gerund, which is also used to characterise subordinating conjunctions.
For those subordinating conjunctions which lexically require finite verbal forms, Spanish also requires information on the mood of the subordinate verb. To that end, the ET-ES lexica encode a feature ”exig-mood", valued "indicative/subjunctive", mainly for synthesis purposes.

10.7 Application to French (Corpus)

Attribute	*value*	Example	Tag
Type	*coord*	et	CCOO
	subord	que	CSUB

All EAGLES attributes apply in principle to French. Only the coordination/subordination distinction is coded in the tagset.

10.8 Application to French (Lexicon)

10.8.1 Type

Attribute	Value	Example	Code
Type	coordinat.	et	c
	subordinat.	que	s

10.8.2 Combinations

Tag	Example
Cc	mais, ou
Cs	que

10.9 Application to Portuguese

10.9.1 Type

Attribute	Value	Example	Tag
Type	coordination	e	
	subordination	que	

10.10 Application to Danish

In general, the EAGLES attributes apply also to Danish, thus the scheme for English is applicable to Danish as well. Traditional dictionaries do not distinguish between conjunction types, but grammars do. The EDEMD uses the following additional features, as compared with EAGLES, within the description of coordinating conjunctions: conjunction (simple coordination) e.g.*og*, disjunction e.g.*eller* and adversative conjunction e.g.*men*. Furthermore, a negation inherent to the conjunction is also catered for in the encoding guidelines.
At present, there is no common definition of appropriate tags for corpus annotation purposes.

Attribute	*value*	Example	Tag
Conj.Type	*coordinating*	og	coconj
	subordinating	fordi	subconj

Attribute	*value*	Example	Tag
Coord.Type	*simple*	og	coconj
	initial	hverken (... eller)	coconj_init
	non-initial	(hverken ...) eller	coconj_n-init

Attribute	*value*	Example	Tag
Subord. Type	*with finite verb*	fordi	subconj_vfin
	with infinitive	for at	subconj_inf
	comparative	end	subconj_comp

10.11 Application to Greek

As with prepositions, we can distinguish between two kinds of conjunctions in Greek:

(a) a finite set of simple conjunctions: e.g. *kai, alla, myn, oute*, and
(b) a productive set of conjunctions formed periphrastically by some word/phrase combined with a conjunction: e.g. *dhedhomenou oti, an kai*, etc. The tagging of these multi-word forms must be performed by a special mechanism.

10.11.1 Type

Type	*value*	Gr.example	Gr.tag
	coord	kai	CjCo
	subord	oti	CjSb

Conjunctions in Greek can be further subcategorised depending on their semantic value: causal, temporal, enumerating, etc. However, this distinction was not considered necessary for the present codification.

10.11.2 Coordination Type

This feature is not presently coded in the Greek Lexicon, but applies to the language.

We distinguish two cases:

- conjunctions that are always found alone in the text: e.g. *alla*,
- conjunctions that can be found alone or combined with another conjunction before the first conjunct: e.g. *kai, y*.

O Ghiannys **kai** *y Maria efughan.*
or
Kai *o Ghiannys* **kai** *y Maria efughan.*

In this case, the values *simple* and *correl* proposed at level L2b could be used for Greek.

11 Numeral

NUMERAL	Type	Gender	Number	Case	Function
MULTILEX	Cardin*				
GENELEX	cardin* ordin*				Det
AlethDic	cardin* ordin*				Adj
NERC	cardin ordin				
Leech	cardin ordin denominat frequency	m f n c	s p	nom gen dat acc	

EAG-L0	**NUMERAL**				
EAG-L1	cardin ordin	m f n	s p	nom gen dat acc	
EAG-L2a					
EAG-L2b	Du quantif denomin	It c	It n		Pron Det Adj

11.1 Comments

The analysed systems show a diverging treatment of the Numeral category. Not all languages and systems may want to handle them as a separate category. EAGLES leaves open the option of handling them in the respective relevant classes of Pronouns, Determiners and Adjectives.

11.1.1 Type

In MULTILEX cardinals form a separate category without any attribute.

GENELEX and AlethDic propose a distinction between cardinals and ordinals within the Adjective category; furthermore, GENELEX has a cardinal value in the 'Determinant' category.

11.1.2 Function

In EAGLES, Numerals are proposed as a separate category and the attribute **Function** is also introduced in order to deal with systems which presuppose a further distinction between Pronouns and Pronominal Adjectives (Italian), and Determinant and Adjectives (French, GENELEX).

11.2 Application to Italian

In Romance language practices (French, Italian, etc.) Numerals constitute a subclass of Pronouns and Pronominal Adjectives.

They are subdivided into cardinal and ordinal:
- cardinal indicates absolute numeral quantity: *zero, cento, mille*;
- ordinal indicates order and classification: *primo, secondo*

11.2.1 Type

Attribute	*value*	It. example	It. tag
Type	*cardinal*	zero	N
	ordinal	secondo	PN/ms

11.2.2 Gender

Attribute	*value*	It. example	It. tag
Gender	*masculine*	primo (anno)	DN/**m**
	feminine	(la) seconda	PN/**f**

The value *common* does not apply to Italian ordinals, as they behave as adjectives with *due uscite* (two endings).

11.2.3 Number

Attribute	*value*	It. example	It. tag
Number	*singular*	prima	PN/f**s**
	plural	prime	PN/f**p**

11.2.4 Function

Attribute	*value*	It. example	It. tag
Function	*pronoun*	(la) prima	PN/fs
	determiner	primo amore	DN/ms

The following are examples of both the pronominal and determiner functions in Italian:

non voglio la prima fetta DN/fs, *dammi la seconda* PN/fs

11.3 Application to German

In the IMS-Tagset only cardinals are distinguished from other POS-Types. Ordinals are classified as adjectives.

Example: der dritte/ADJA Mann
drei/CARD Männer
im Jahre 2000/CARD

11.3.1 Numerals: Type

Attribute	*value*	example	tag
Type	*cardinal*	27	CARD
	ordinal	–	–

11.4 Application to English

Gender and Case do not apply to English numerals.

11.4.1 Numeral-type

Attribute	*values*	Examples	Tags
Numeral-type	*cardinal*	two, 55	NUC
	ordinal	second, 55th	NUO

11.4.2 Number

Attribute	*values*	Examples	Tags
Number	*singular*	one	NUCs
	plural	fifty	NUCp

Note that when a numeral word occurs with a plural inflection (*firsts, fifties*), this is considered a clear indicator of the word's status as a noun, rather than as a numeral in the normal sense. All numerals can be converted into nouns for special functions.

11.5 Application to Dutch

CELEX has a morphological word class tag Q, used in derivational word analysis, which covers Numerals and Quantifiers. Apart of that, CELEX has a syntactical word class tag NUM (Numerals) divided in two subclasses: cardinals and ordinals.

Attribute	*value*	Example	Tag
Type	*Quantifier*	veel	NUM quant
	Cardinal	zeventien	NUM hoofd
	Ordinal	zeventiende	NUM rang

This is not a CELEX tag. We decided to include it in the table, though, because Dutch traditional grammarians and lexicographers distinguish a subclass, called 'onbepaalde (hoofd)telwoorden' in Dutch. Some call them 'indefinite numerals', others 'indefinite cardinals'. This class concerns words like English 'many', 'more' , 'most' and the like.

11.6 Application to French (Corpus)

The IBMF tagset has only one tag for numerals, **CHIF**. However, most of the attributes distinguished in EAGLES apply to French, except **case**, as usual.

11.7 Application to French (Lexicon)

11.7.1 Type

Attribute	Value	Example	Code
Type	cardinal	deux	c
	ordinal	deuxieme	o

11.7.2 Gender

Attribute	Value	Example	Code
Gender	masculine	premier	m
	feminine	premiere	f

11.7.3 Number

Attribute	Value	Example	Code
Number	singular	premier	s
	plural	premiers	p

11.7.4 Case

Not applicable to French.

11.7.5 Combinations

Tag	Example
Mcms-	un
Mcfs-	une
Mc-s-	zero
Mc-p-	deux, trois

Note:

Traditionnal grammars usually distinguish un/article and un/numeral. However, it is very difficult to find linguistic tests that enable the two to be discriminated. It is not certain that this distinction will be kept.
Cfr.
J'ai vu un chat (article)
J'ai vu un chat et deux chiens (numeral)

11.8 Application to Portuguese

Numerals were included as types of adjectives in the Portuguese demo lexicon.

11.9 Application to Danish

In traditional dictionaries the word class Numeral covers cardinals and ordinals without any subclassification. Some traditional grammars focus on the adjective-like properties of numerals and treat them as subclasses of the adjective. The EDEMD classifies quantifier, cardinal and ordinal as separate categories. However, the EAGLES attributes are applicable to Danish as well.

Attribute	*value*	Example	Tag
Type	*cardinal*	fem	card
	ordinal	femte	ord

11.10 Application to Greek

Numerals in Greek include both numerals presented in digits (Arabic numbers, Greek or Latin letters) and full words. To code this, a specific feature is used (not included in the EAGLES proposal), namely **form** (see relevant section).

11.10.1 Type

Both cardinal and ordinal numerals are recognised in Greek.

Attribute	*value*	Gr. example	Gr. tag
Type	*cardinal*	mydhen, 0	Nm**Cd**
	ordinal	tritos, 3os	Nm**Od**

11.10.2 Gender, Number and Case

Given that numerals in Greek act as nouns and/or adjectives, they are further coded for Gender, Number and Case. When they function as adjectives, they must agree in these three features with the noun they refer to.

Not all values of these three features are applicable to numerals, as shown in the following tables. In fact, the values *masc-fem* for Gender and *invariant* for Number are not used at all.

Attribute	*value*	Gr. example	Gr. tag
Gender	*masculine*	prwtos	NmOd**Ma**
	feminine	prwty	NmOd**Fe**

Attribute	*value*	Gr. example	Gr. tag
Number	*singular*	prwtos	NmOdMa**Sg**
	plural	prwtoi	NmOdMa**Pl**

Attribute	*value*	Gr. example	Gr. tag
Case	*nom*	prwtos	NmOdMaSg**Nm**
	gen	prwtou	NmOdMaSg**Ge**
	acc	prwto	NmOdMaSg**Ac**
l-spec	*voc*	prwte	NmOdMaSg**Vo**
l-spec	*indcl*	dhuo	NoCdMaPl**Ic**

11.10.3 Function

This attribute is used in Greek in order to distinguish between numerals functioning as nouns and numerals that function as adjectives. Adjectival numerals may appear on their own, i.e. the noun they refer to may be ommitted:

Dhwse mou dhwdheka vivlia.
Dhwse mou dhwdheka.

Attribute	*value*	It. example	It. tag
Function	*adj*	dhwdeka	NmCdMaPlIc**Aj**
l-spec	*nom*	dhwdekadha	NmCdFeSgNm**No**

11.10.4 Form

This feature is used to code the form of the numeral (i.e. type of digits or full word-form). All the values are language-specific.

Attribute	*value*	Gr. example	Gr. tag
Form	*word*	dhwdeka	NmCdMaPlIcAj**Wd**
l-spec	*number*	12	NmCdMaPlIcAj**Nu**
l-spec	*gr-num*	ib	NmCdMaPlIcAj**Gn**
l-spec	*lat-num*	XII	NmCdMaPlIcAj**Ln**

12 Interjections

No subcategories are foreseen.

12.1 Application to Italian

In the Italian corpus and lexicon, Interjections are tagged **I**. There are no subcategories.

12.2 Application to German

The IMS-Tagset provides the tag **ITJ** for interjections. There are no subcategories.

12.3 Application to English

There are no subcategories of interjections. To avoid confusion over the use of "I", this word-class is tagged "Ij".

12.4 Application to Dutch

CELEX has no distinctive tags for interjections.

12.5 Application to French (Corpus)

Although they are a valid class in French, interjections are not coded as such in the IBMF tagset.

12.6 Application to French (Lexicon)

Tag	Example
I	eh

12.7 Application to Portuguese

This category is not included in the Portuguese demo lexicon.

12.8 Application to Danish

There are no subcategories in Danish. The EDEMD has no description of interjections.

12.9 Application to Greek

In the Greek Morphological Lexicon as well as the Corpus, interjections are coded with the tag "Ij". No subcategories or other features are used for this category.

13 Unique membership class

13.1 Comments

This category should contain language-specific phenomena.

EAGLES recommends keeping this class as small as possible.
All the members of this class are to be seen at the level 2b of language-specific distinctions, e.g. 'infinitive marker' for English, German, Dutch, Danish), 'existential' for English, etc.

AlethDic has a special class "Mot not autonome", in which e.g. *hui* is included.

13.2 Application to German

Categories with unique members are particles like negation (*nicht*), infinitive marker (*zu*), superlative marker (*am*).
The IMS-tagset includes another particle type for separable prefixes. It covers prefixes like *an, aus, ein, ...* (mostly identical to prepositions), but also other (non-adverbial) prefixes like *rad, statt, instand*.

Example: er kommt an/PTKVZ
er kauft ein/PTKVZ
er fährt rad/PTKVZ *(not: Rad !)*
es findet statt/PTKVZ
er hält das Haus instand/PTKVZ

13.3 Application to English

13.3.1 Particle-Type

Attribute	*values*	Examples	Tags
Particle-Type	*infinitive*	to (+ Infin.)	UI
	negative	not, n't	UN
	existential	there (is/are)	UX

13.4 Application to Dutch

We have recently sent to the EAGLES group of Professor G. Leech the proposal of attributing a value for Dutch in this group: *infinitive-marker* to mark up the prepositions 'om' and 'te' and 'om te'.

13.5 Application to French (Corpus and Lexicon)

We see no language specifics that would be coded in the IBMF tagset and would not fit in other EAGLES classes.

13.6 Application to Greek

This class could be used for the codification of particles, which form a distinct category in Greek. The only feature pertinent to this category is that of **Type**.

13.6.1 Particle-Type

Particles can be used to form compound forms of tense (e.g. future) or mood (e.g. subjunctive), or as introducers of negation.

Attribute	*value*	Gr. Example	Gr. Tag
Type	*fut*	tha	Pt**Fu**
	neg	dhen	Pt**Ng**
	subj	na	Pt**Su**
	other	as	Pt**Ot**

The value *other* is used to group together various functions of particles, such as ascertaining or hesitating, which are semantically motivated and cannot be distinguished unless the linguistic and extra-linguistic context is taken into account.

14 Residual

RESIDUAL	Type	Gender	Number
MULTILEX	Misc*		
GENELEX			
AlethDic			
NERC	letters symbols formulae abbreviations Foreign Words*		
Leech	foreign words alphabetic symbols acronyms formulae abbreviations unclassified	m f n	s p
EAG-L0	**RESIDUAL**		
EAG-L1	foreign words alphabetic symbols formulae acronyms toponyms abbreviations unclassified	m f n	s p
EAG-L2a			
EAG-L2b		c	n

14.1 Comments

EAGLES recommends keeping this class as small as possible and the specification of links to other classes when they are clear.
Some annotation practices prefer to treat phenomena included here as belonging to other parts of speech (e.g. foreign words treated as nouns having number and gender).

14.1.1 Type

MULTILEX proposes two indistinct classes Misc and None for words for which it is unclear what category should be assigned and for words to which no category should be assigned. It can be argued that these two categories should be mapped onto the Residual class proposed here.

In NERC Foreign Words were kept apart from the so-called Rest Group and treated as a separate category, since the former may behave in a different way syntactically.

14.2 Application to Italian

14.2.1 Type

Attribute	*values*	Examples	Tags
Type	*foreign word*	polis	FW
	abbreviation	Sig.	abbr
	sigla	CNR	sigla
	toponyms	Milano	T

14.3 Application to German

Only punctuation is included in the IMS-tagset; no tags are provided yet for symbols, formula, foreign words etc.

Abbreviations are classified according to the "full" word form, and are marked by an additional feature **ABK**. Acronyms are generally classified as proper names.

14.3.1 Type

Attribute	*values*	Examples	Tags
Type	*punctuation*	?	**IP**NORM
	abbreviation	Tel. dt.	NN:**ABK** ADJA:**ABK**
	acronyms	USA	NE:**ABK**

14.4 Application to English

14.4.1 Type

Attribute	*values*	Examples	Tags
Type	*foreign word*	mawashi	RFW
	symbol	£, |	RSY
	formula	X/21=	RFO
	unclassified	la-la-la	RUN

14.4.2 Number

Attribute	*values*	Examples	Tags
Number	*singular*	A, b	RSYs
	plural	As, b's	RSYp

14.5 Application to Dutch

We have recently sent to the EAGLES group of Professor G. Leech the proposal of attributing an additional value for Dutch sub Residual-Type: *Acronym*.

14.6 Application to French (Corpus)

The IBMF tagset codes punctuation marks as specific tags. Indeed, they are morphological manifestations and can help predict other tags.
The tagset distinguishes weak punctuation (tag **AAAA**), strong punctuation (**YAAA**) and sentence boundary (**ZTRM**).

14.7 Application to French (Lexicon)

Tag	Example
X	IBM

14.8 Application to Greek

14.8.1 Foreign Word

In the Greek Morphological Lexicon, we currently keep foreign words apart from the rest of the elements included in EAGLES under the general category of "residual". This distinction allows us to code further information on foreign words necessitated by the morpho-syntactic system of Greek. In this category, we include only those foreign words that have not been adapted to the morphological system of Greek and, furthermore, are not considered "Greek" words.

However, the inclusion of foreign words under the "Residual" category can easily be performed, given the tag "Fw" used for this category.

Two additional features are used for foreign words: **form** and **for-cat**. The first attribute is used to distinguish between foreign words that are transliterated in the Greek alphabet and those kept in their original Latin form, and the second attribute is used to code the grammatical category of the foreign word in order to facilitate further syntactic parsing of the texts.

Attribute	*value*	Gr. example	Gr. tag
Form	*translit*	Tzeikomp	Fw**Tr**
	original	Jacob	Fw**Or**

Attribute	*value*	Gr. example	Gr. tag
For-cat	*noun*	Tzeikomp	FwTr**No**
	adverb	(ad hoc)	FwOr**Av**
	verb	cogito	FwOr**Vb**

14.8.2 Type

In this section we include all other elements included in the EAGLES Residual category. These are coded in the ILSP Morphological Lexicon as belonging to the category "Rest-group". The basic feature for these elements is that of **Type**.

Attribute	*values*	Examples	Tags
Type	*foreign word*	Tzeikomp	Fw
	abbr	forol.	Rg**Ab**
	formula	F(x)	Rg**Fo**
	symbol	¡	Rg**Sy**
l-spec	*date*	(17/03/92)	Rg**Da**

No distinction is made at present between abbreviations and acronyms, while the extra value of *date* is used for the codification of dates, whether written in digits or in letters.

15 References

Aglamissis, Y., M. Gavrilidou, P. Labropoulou, H. Papageorgiou, S. Piperidis and C. Raptis (1994): "The ILSP Morphological Tagset". NERC-2 Document.

Bel N., M. Villegas (1993): "Morphosyntax for Spanish", EAGLES Input Document, Barcelona.

Bindi R., M. Monachini, P. Orsolini (1991): "Italian Reference Corpus" NERC Technical Report, ILC, Pisa.

Burnage G. (1990): *CELEX. A guide for users.*

Dutilh-Ruitenberg T. (1994): "A Comparative Report on Morphosyntactic Categories in Dutch as encoded in the CELEX Dutch Lexical Databases. Augmented with ten proposals for Dutch, EAGLES Input Document, Leiden.

Calzolari N., Ceccotti M.L., Roventini A. (1983): "Documentazione sui tre nastri contenenti il DMI", ILC Technical Report, Pisa.

Calzolari N., Monachini M. (Coords.) (1994): "Commons Specifications and Notation For Lexicon Encoding", MULTEXT WP1.6 Deliverable.

GENELEX Consortium (Apr 1993): "Couche Morphologique", Version 3.0, ASSTRIL, Gsi-Erli, IBM France, SEMA GROUP.

GENELEX Consortium (Sept 1993): "Couche Syntaxique. Les Unites Syntaxique Simple", Tome 1, Version 3.0, ASSTRIL, Gsi-Erli, IBM France, SEMA GROUP.

Gsi-Erli (Jun 1993): "Le Dictionnaire AlethDic", Gsi-Erli, Paris.

Guerreiro, P. (1994): "Morphosyntactic Phenomena Encoded in a Portuguese NLP Lexicon. Application of the EAGLES Specificications", EAGLES input document, Lisbon.

Guerreiro, P. (ed.), (1994): "Linguistic Specifications for Portuguese Computational Lexica", Final Report of GENELEX-PT, Eureka: EU 524.

Heid, U., McNaught, J. (eds.) (1991): "Eurotra-7 Study: Feasibility and Project Definition Study on the Reusability of Lexical and Terminological Resources in Computerised Applications". Eurotra-7 Final Report, Stuttgart.

Heylen, D. (April 1994): "Comments on a Common tagset", EAGLES Input Document, Utrecht.

Langé, J.M. (March 1994): “Application of EAGLES morphosyntactic guidelines to a French tagset”, EAGLES Input Document, IBM, Paris.

Leech, G. (1992): “Corpus Annotation Schemes”, paper presented at the Pisa Corpus Workshop (24-26 January 1992), to be published in Literary and Linguistic Computing, OUP.

Leech G., A. Wilson (1993a): “Invitation Draft”, Draft EAGLES Document, Lancaster.

Leech G., A. Wilson (1993b): “Tagset Guidelines”, Draft EAGLES Document, Lancaster.

Leech G., A. Wilson (1994a): “MSAL21”, Draft EAGLES Document, Lancaster.

Leech G., A. Wilson (1994b): “A Morphosyntactic Tagset for English, making use of the guidelines of the Pisa Document”, EAGLES Input Document, Lancaster.

Monachini (1992): “Core set of PoS tags for Italian”, ILC Technical Report, Pisa.

Monachini M., A. Östling (1992a): “Morphosyntactic Corpus Annotation – A Comparison of Different Schemes”, NERC-WP8-60, Pisa.

Monachini M., A. Östling (1992b): “Towards a Minimal Standard for Morphosyntactic Corpus Annotation”, NERC-WP8-61, Pisa.

Monachini M., N. Calzolari (1994): “Application of EAGLES Proposal for Morphosyntactic encoding to Italian Lexicon and Corpus”, EAGLES Input Document, ILC, Pisa.

MULTILEX Consortium (Apr 1993): “Standards for Multifunctional Lexicon”, CAP GEMINI, Philips, Univ. of Surrey, Univ. of Bochum, Univ. of Muenster.

Schiller A. (1993): “Tagsets für Deutsch”, Draft EAGLES input Document, Stuttgart.

Schiller A. (1994): “Guidelines für das Tagging deutscher Textcorpora (Kleines und erweitertes Tagset)”, Univ. Stuttgart, Internal Document.

Schiller A., Thielen C. (forthc.): “Ein kleines und erweitertes Tagset fürs Deutsche”, *to appear in:* Tagungsberichte des Arbeitstreffens Lexikon + Text, 17./18. Februar 1994, Schloß Hohentübingen, Niemeyer: Lexicographica Series Maior, Tübingen, Spring 1995.

TEI AI 1W2 (June 1991): “List of Common Morphological Feature for Inclusion in TEI Starter Set of Grammatical-Annotation Tags”.

Veronis J., Khouri L., Meunier C. (1994a): “Application of the EAGLES-L1 Proposal to French”, Draft EAGLES input Document, Aix-en-Provence.

Veronis J., Khouri L., Meunier C. (1994): “Proposal for Morphosyntactic Encoding in MULTEXT”, Aix-en-Provence.