EAGLES

Expert Advisory Group on
Language Engineering Standards

# EAGLES

# Recommendations for the Morphosyntactic Annotation of Corpora

EAGLES Document EAG–TCWG–MAC/R

Version of Mar, 1996

# Contents

# 0 Contributors

This document, as with all EAGLES documents, is the result of cooperative work involving a number of individuals, contributing in different ways.

## 0.1 Authors

G. Leech
Department of Linguistics and
Modern English Language
Lancaster University
Lancaster Tel.: +44.524.59.30.36
United Kingdom Fax.: +44.524.84.30.85
LA1 4YT E-mail: G.N.Leech@lancaster.ac.uk

A. Wilson
Department of Linguistics and
Modern English Language
Lancaster University
Lancaster Tel.: +44.524.59.30.25
United Kingdom Fax.: +44.524.84.30.85
LA1 4YT E-mail: A.Wilson@lancaster.ac.uk

## 0.2 Other contributors

| | |
|---|---|
| G. Arrarte | Instituto Cervantes, Madrid |
| F. Bacelar | Centro de Linguística, University of Lisbon |
| A. Braasch | CST, Copenhagen |
| N. Calzolari | ILC-CNR, Pisa |
| K.A.C. Depuydt | University of Leiden |
| P. Guerreiro | ILTeC, Lisbon |
| P. de Haan | University of Nijmegen |
| R. Hatzidaki | University of Birmingham |
| U. Heid | University of Stuttgart |
| J-M. Langé | IBM Paris, Paris |
| P. King | University of Birmingham |
| L. Lemnitzer | University of Münster |
| M. Monachini | ILC-CNR Pisa |
| W. Paprotté | University of Münster |
| A. Schiller | University of Stuttgart |
| A. Spanu | ILC-CNR, Pisa |
| P. Steiner | University of Münster |
| S. Teufel | University of Stuttgart |
| A. Zampolli | ILC-CNR, Pisa |

## 0.3 Editors and Assistants

N. Calzolari
Co-Chief Editor, EAGLES
ILC–CNR
Via della Faggiola 32
Pisa I-56126
Italy
Tel.: +39.50.56.04.81
Fax.: +39.50.58.90.55
E-mail: eagles@ilc.pi.cnr.it

J. M$^{c}$Naught
Co-Chief Editor, EAGLES
Centre for Computational Linguistics
UMIST
Manchester
United Kingdom
M60 1QD
Tel.: +44.161.200.3098
Fax.: +44.200.3099
E-mail: jock@ccl.umist.ac.uk

T. Paskiewicz
EAGLES Editorial Assistant
Centre for Computational Linguistics
UMIST
Manchester
United Kingdom
M60 1QD
Tel.: +44.161.200.3097
Fax.: +44.200.3099
E-mail: teresap@ccl.umist.ac.uk

T. Ayazi
EAGLES Secretariat
ILC–CNR
Via della Faggiola 32
Pisa I-56126
Italy
Tel.: +39.50.56.04.81
Fax.: +39.50.58.90.55
E-mail: eagles@ilc.pi.cnr.it

A. Enea
EAGLES Webmaster
ILC–CNR
Via della Faggiola 32
Pisa I-56126
Italy
Tel.: +39.50.56.04.81
Fax.: +39.50.58.90.55
E-mail: enea@ilc.pi.cnr.it

# 1 Introduction

*Corpus annotation* is the practice of adding interpretative, especially linguistic, information to a text corpus, by coding added to the electronic representation of the text itself. A typical case of corpus annotation is that of *morphosyntactic annotation* (also called *grammatical tagging*), whereby a label or tag is associated with each word token in the text, to indicate its grammatical classification (see 4.2 for more information).

For a written text, it is generally easy to make a distinction between the electronic representation of the text itself and annotations which are added to the text. On the other hand, for a spoken text (i.e. a transcribed representation of a spoken discourse) the difference between the text and its annotations cannot be taken for granted, particularly in the areas of phonemic, phonetic and prosodic transcription. Here the representation of the text itself entails linguistic interpretation at the phonological level. For the purposes of EAGLES, however, features of phonemic / phonetic / prosodic transcription are not considered to be part of the annotation. Consideration of such features may be found in a companion document on Spoken Language.

In principle, annotation can represent any type of analytic information about the language of a text. In practice, so far, the two types of annotation most commonly applied to a text have been:

**Morphosyntactic annotation:** Annotation of the grammatical class of each word-token in a text, also referred to as "grammatical tagging" or "part of speech (POS) tagging";

**Syntactic annotation:** Annotation of the structure of sentences, e.g. by means of a phrase-structure parse or dependency parse.

Other types of annotation which have been applied to text are:

**Semantic annotation:** For example, annotating word-tokens for their dictionary sense, or for their semantic category;

**Discourse annotation:** For example, the marking of discoursal relations such as anaphora in a text;

**Lemma annotation:** Indicating the lemma of each word-token in a text.

Because of their relative feasibility and their obvious application to areas such as lexicon and grammar development, morphosyntactic and syntactic annotation are regarded as the most important kinds of annotation at the present stage of the development of text corpora. They are certainly the best-developed types and those for which there are well-established working practices. Hence, they will be the major topics of EAGLES recommendations. *Morphosyntactic annotation*, in particular, is the subject of recommendations presented here. Syntactic annotation is the subject of a separate document. Other types of annotation, such as semantic tagging, are necessarily given less attention at the present stage, as the work that has been done in these areas is less systematised and more experimental. Lemma annotation is closely related to morphosyntactic annotation, and may be treated as an adjunct to it (see 4.5).

At the current stage, detailed provisional conclusions have been reached on the recommendation of standards for *morphosyntactic annotation* (or *grammatical tagging*, as it is generally called).

# 2 Rationale for the present proposal

The guidelines for morphosyntactic annotation are very similar to those for the morphosyntactic level in the lexicon. Large lexicons are increasingly being used in the annotation of corpora, and corpora are increasingly being used as sources of information to be acquired by lexicons. These processes are increasingly being automated. There is therefore a great advantage in being able to transduce directly from word-class annotations in texts to morphosyntactic information in lexicons, and vice versa. On the other hand, there are reasons for assuming that these two types of word classification need not be identical.

One reason for differences is that morphosyntactic annotation (which has been so far carried out extensively on English, but not on other languages) is at a relatively primitive stage of development. It is typically carried out largely automatically, but without the benefit of a full parse, frequently using simple statistical models of grammar such as Hidden Markov Models (Rabiner 1990).

There is a major problem of automatic tag disambiguation, resulting in a substantial rate of error or of failure to disambiguate (typically of several percent), and although these less-than-ideal results can in principle be corrected by hand, in practice the correction of a large corpus (say, of 100 million words) is a Herculean task. Thus, while annotators might wish to provide as much lexically relevant information as possible in the tagged corpus, in practice they are limited by what current taggers are realistically able to achieve. Some attributes or values routinely entered in lexicons are virtually impossible to mark automatically in a corpus without a prohibitive amount of error (e.g. the distinctions between the different functions of the base form of the English verb — indicative plural, imperative, subjunctive, etc. — are virtually impossible to make without a full parse, which itself would produce unreliable results in the present state of the art).

A second reason is the opposite of the first: just as there are kinds of information which are expected in a lexicon, but cannot be included in tagging, so there are kinds of information which may be useful for tagging, but may be extraneous to morphosyntax in the lexicon. It may be useful, for automatic tagging, to mark some syntactic or semantic distinctions, thereby going beyond the definition of morphosyntax. Examples include the marking of the purely syntactic distinction between attributive-only and predicative-only adjectives, or the marking of small semantic classes such as names of months or names of days, in order to facilitate the identification of dates (which have a distinctive syntactic structure) in certain kinds of texts. While these values are normally excluded from the morphosyntactic level in the lexicon, they can be easy to identify in texts, and may have a valuable syntactic role in disambiguating neighbouring words. Also, in text corpora, one constantly finds the necessity to deal with phenomena which have been regarded as peripheral to a lexicon, such as naming expressions (including proper nouns), acronyms, formulae and special symbols. In all these respects, it would artificially constrain tagging, and often make it less useful, if the tagset had to mirror the attributes and values typically found in lexicons. *Grammatical tagging*, to use the traditional term, is a less clearly definable process than is implied by the stricter term *morphosyntactic annotation*.

The relation between the lexicon guidelines and these morphosyntactic annotation guidelines will be explained in section 3. At this point, it is important to note that the distinctions made in morphosyntactic tagging may usefully correspond to various linguistic levels (morphological, morphosyntactic, syntactic, semantic) in the lexicon. But the level with which they are centrally concerned is that of morphosyntax.

Considering 'levels' in a different sense, it is also essential to distinguish *levels of abstraction* at which the notion of *tagset* may be identified.

**Character-coding level:** This is the least abstract level, where we identify a morphosyntactic tag with a particular sequence of characters in a marked-up text.

**Descriptive level:** This is a more abstract level, where a tag is identified with a set of attribute–value pairs in a morphosyntactic description of a particular language. For a completely explicit description, it is desirable to formalise this description as an attribute–value hierarchy with monotonic inheritance. The tagset may then be termed a *logical tagset*.

**Cross-linguistic level:** This is the most abstract level, where we are examining attributes (e.g. number) and values (e.g. singular, plural) as generically applied to a number of different languages. This is the level we are concerned with in the guidelines which follow (see 4.2).

The Intermediate Tagset (see 4.3) suggested as a way of mapping different language-specific tagsets into a common set of attributes and values is an example of tags considered at this level.

## 3 Harmonisation with proposals of the Lexicon Working Group

Like the lexicon guidelines, the morphosyntactic tagging guidelines

1. Make use of an attribute–value formalism.
2. Do not adhere to a strict attribute–value hierarchy (in terms of monotonic inheritance).
3. Use three levels of constraint (obligatory, recommended and optional) in defining what is acceptable according to the guidelines.

4. Subdivide the optional level into two types of optional extension to tagsets:
   (a) Extensions to deal with phenomena which are marginal to morphosyntactic annotation strictly defined, but common to a number of languages (e.g. the distinction between countable and mass nouns);
   (b) Extensions to deal with phenomena which are specific to particular EU languages.

A few words may be added regarding each of these points:

1. At a descriptive level, morphosyntactic tags are therefore defined as sets of attribute–value pairs, although at a 'visible' character-coding level they may not be symbolised as such.
2. For an individual language, it may be an important step to formalise the tagset as an attribute–value hierarchy. However, this degree of formalisation is not appropriate to the cross-linguistic level of abstraction, where we are specifying guidelines to apply to all EU languages.
3. The obligatory level of constraint is limited to the major categorisations of parts of speech as **Noun**, **Verb**, **Conjunction**, etc. The recommended level of constraint applies to well-known attributes used widely in the description of European languages: e.g. (for nouns) **Number**, **Gender** and **Case**.
4. At the optional level, the guidelines clearly have a weaker import, and should not be regarded as mandatory in any sense, but simply as a presentation of possibilities sanctioned by current practice.

The tagset guidelines should allow mappings to be stated between the coding of morphosyntactic phenomena in a lexicon and their coding in the morphosyntactic annotation of text corpora. However, because of the different perspective and goals of these two activities (see 2) there is no necessary expectation that this will be a straightforward mapping. One suggestion, therefore, is that it should be easier to specify the conversion between lexicon and annotation categories by making use of an *Intermediate Tagset* (see 4.3).

# 4 Recommendations for morphosyntactic categories

## 4.1 Reasonable goals for standardisation

Some kind of standardisation is becoming urgent, particularly in the area of morphosyntactic annotation. This is an area in which most annotation has been done, and morphosyntactic tagging is likely to be undertaken for many different languages in the next few years. In the interests of interchangeability and reusability of annotated corpora, and particularly for the development of multilingual corpora, it is important to avoid a free-for-all in tagging practices.

On the other hand, the varied needs and constraints which govern any annotation project, or which might govern such projects in the future, urge caution in setting out to achieve a rigid standardisation. Where possible, it is important to offer a default specification which can be adopted where there are no overriding reasons for departing from it. In this way, invariance will establish itself across different projects and languages, and a *de facto* standard will progressively come into being.

However, the need to go beyond a preferred standard — a principle of *extensibility* — should also be recognised. There will be a need to extend the specification to new phenomena and sometimes a need to represent different perspectives on the same data. *Extensibility* means, on the one hand, the ability to extend the specification to language-specific phenomena, and on the other, the ability to vary the degree of granularity for this or that annotation task.

The use of the term *guidelines*, in reference to the documentary specification of annotation standards, is salutary in suggesting that there is no absolute normative prescription of annotation practices, but at most a set of recommendations, from which the annotator may justify departures or extensions for particular purposes. Even the term *recommendations* is too strong a word in some cases: often we can only point out the range of practices which exist, without offering advice to prefer one to another.

We consider, in the following three sections, the feasibility of achieving a measure of standardisation in three important areas.

### 4.1.1 Common standards for representation and encoding of annotations of texts

At face value, the most trivial aspect of annotation guidelines is in recommendation of ‘visible’ character-coded means to represent this or that linguistic phenomenon. Any device for encoding a given linguistic phenomenon is (in the last resort) arbitrary, and, so long as it is distinctive, can be automatically converted into a different device. We propose that the criteria of *compactness*, *readability*, and *processability* be given priority, although different degrees of priority may be assigned to such criteria for different projects. On the other hand, we suggest below (see 4.3 for further details) that local specifications should be translatable into a common EAGLES standard automatically, by a regular mapping via an Intermediate Tagset.

### 4.1.2 Common standards in describing and representing linguistic categories or structures

The specification of common standards for linguistic categories/structures is more serious and challenging. If a common standard implies the recognition of invariants across different languages or different descriptions of the same language, then the extent to which this is feasible depends on the extent to which such invariants are recognised by those already working in the field. This may be unproblematic in the case of the grossest categories such as **Noun**, **Prepositional Phrase**, etc., but as one moves toward (a) greater granularity of description, and (b) more abstract levels of linguistic annotation, the degree of consensus is likely to decline. The level of morphosyntactic tagging is the one most favourable to a reasonable degree of standardisation in this sense and is also the level for which the urgency of establishing common standards is greatest. In sections on tagset guidelines and the Intermediate Tagset, this will be dealt with in some detail (4.2–4.3) and in close relation to the standards for morphosyntactic categorisation in the lexicon.

### 4.1.3 Common standards for specifying annotation schemes and their application to texts

The final area for standardisation that we consider, here, appears to be the most difficult to achieve, if it is to be equated with laying down rules for *consistency* in the application of tags to texts. To take the apparently favourable area of morphosyntactic annotation: the ideal need is to specify an annotation scheme so precisely that a different annotator, applying the same annotation scheme to the same text, would arrive at exactly the same result. This would mean that each word-token, in a given text, would end up with the same tag, even if done independently by two analysts. But in practice, there are always “fuzzy boundaries” between categories, such as the uncertainty (in English) of whether to regard *gold* in *a gold watch* as an adjective or a noun. Decisions on such matters have to be specified in the annotation scheme, which should also deal with such general issues as whether functional or formal definitions of the use of tags are to be adopted; or whether both function and form have to be represented in the annotation. Individual words may need to be discussed, where their recognition as members of this or that category is problematic. But new phenomena, not covered by existing guidelines, are always liable to occur, however detailed the annotation scheme.

Such issues as these cannot be decided in the abstract, in a way which generalises across languages and across annotation tasks. This kind of standardisation is best met, not by laying down detailed specifications of how this or that category is applied in the tagging of this or that word, but by recommending that a sufficiently detailed annotation scheme be made available to users of the annotated corpus. There is little possibility of seeking detailed agreement between different annotators on matters of how to apply tags to texts, particularly if different languages are involved. But at least, one can ensure that the user be provided with information, as detailed as possible, about how annotations have been applied to texts.

### 4.1.4 Conclusion: Manageable levels of achievement in specifying standards

The following degrees of standardisation may thus be recommended at the current stage:

**Representation/encoding:** Observance of general principles of transparency, processability, brevity and unambiguity; translatability of annotation devices into a set of language-generic conventions.

**Identifying categories/subcategories/structures:** Agreement on common categories, etc., across different languages, where these can be justified by linguistic analysis and descriptive tradition; allowance for variation, subject to three degrees of constraint: *obligatory*, *recommended* and *optional* specifications.

**Annotation schemes and their application to texts:** Agreement merely on the requirement that annotation schemes should be made available to end-users and to other annotators, and should be as detailed as possible.

## 4.2 Word categories: Tagset guidelines

Four degrees of constraint are recognised in the description of word categories by means of morphosyntactic tags:

1. **Obligatory** attributes or values (4.2.1) have to be included in any morphosyntactic tagset. The major parts of speech (Noun, Verb, Conjunction, etc.) belong here, as obligatorily specified.
2. **Recommended** attributes or values (4.2.2) are widely-recognised grammatical categories which occur in conventional grammatical descriptions (e.g. Gender, Number, Person).
3. **Special extensions** are subdivided to yield two constraints:

    i. *Generic* attributes or values (4.2.3) are not usually encoded, but may be included by anyone tagging a corpus for any particular purpose. For example, it may be desirable for some purposes to mark semantic classes such as temporal nouns, manner adverbs, place names, etc. But no specification of these features is made in the guidelines, except for exemplification purposes. They are purely optional.

    ii. *Language-specific* attributes or values (4.2.4) may be important for a particular language, or maybe for two or three languages at the most, but do not apply to the majority of European languages.

In practice, generic and language-specific features cannot be clearly distinguished.
Type *special extensions* is an acknowledgement that the guidelines are not closed, but allow modification according to need. The four types above correspond to the four types of constraint applied to word categorisation in the lexicon. In general, this document repeats (in a somewhat different form) much of the material dealing with morphosyntactic categorisation in the lexicon, where further information on particular features of the classification can be obtained.

### 4.2.1 Obligatory attributes/values

Only one attribute is considered obligatory: that of the major word categories, or *parts of speech*:

**Major Categories**

| | | | | | |
|---|---|---|---|---|---|
| 1. | N [noun] | 2. | V [verb] | 3. | AJ [adjective] |
| 4. | PD [pronoun/determiner] | 5. | AT [article] | 6. | AV [adverb] |
| 7. | AP [adposition] | 8. | C [conjunction] | 9. | NU [numeral] |
| 10. | I [interjection] | 11. | U [unique/unassigned] | 12. | R [residual] |
| 13. | PU [punctuation] | | | | |

Of these, the last three values are in need of explanation.
The **unique** value (U) is applied to categories with a unique or very small membership, such as *negative particle*, which are 'unassigned' to any of the standard part-of-speech categories. The value **unique** cannot always be strictly applied, since (for example) Greek has three negative particles, $\delta\epsilon\nu$, $\mu\eta(\nu)$, and $o\chi\iota$.
The **residual** value (R) is assigned to classes of textword which lie outside the traditionally accepted range of grammatical classes, although they occur quite commonly in many texts and very commonly in some.

For example: foreign words, or mathematical formulae. It can be argued that these are on the fringes of the grammar or lexicon of the language in which the text is written. Nevertheless, they need to be tagged.

**Punctuation** marks (PU) are (perhaps surprisingly) treated here as a part of morphosyntactic annotation, as it is very common for punctuation marks to be tagged and to be treated as equivalent to words for the purposes of automatic tag assignment.

The symbols used to represent the major categories (above) and the attributes and values of other categories (below) will be used later for a method of language-neutral representation called the *Intermediate Tagset* (see 4.3).

### 4.2.2 Recommended attributes/values

These are specified below under part-of-speech headings. Each numbered heading refers to the number assigned under major category. The set of values for each attribute is definitely not a closed set and will need to be augmented to handle peculiar features of individual languages (4.2, point 3). Not all EU languages will instantiate all attributes or all values of an individual attribute. For each attribute, **0** designates a zero value, meaning "this attribute is not applicable" for the particular language, or for a particular textword in that language. The standard requirement for these *recommended* attributes/values is that, if they occur in a particular language, then it is advisable that the tagset of that language should encode them.

**1. Nouns (N)**

| | | | | | | |
|---|---|---|---|---|---|---|
| (i) | Type: | 1. Common | 2. Proper | | | |
| (ii) | Gender: | 1. Masculine | 2. Feminine | 3. Neuter | | |
| (iii) | Number: | 1. Singular | 2. Plural | | | |
| (iv) | Case: | 1. Nominative | 2. Genitive | 3. Dative | 4. Accusative | 5. Vocative |

Inflection type is omitted as an attribute, as it is purely morphological.

**2. Verbs (V)**

| | | | | | |
|---|---|---|---|---|---|
| (i) | Person: | 1. First | 2. Second | 3. Third | |
| (ii) | Gender: | 1. Masculine | 2. Feminine | 3. Neuter | |
| (iii) | Number: | 1. Singular | 2. Plural | | |
| (iv) | Finiteness: | 1. Finite | 2. Non-finite | | |
| (v) | Verb form / Mood: | 1. Indicative | 2. Subjunctive | 3. Imperative | 4. Conditional |
| | | 5. Infinitive | 6. Participle | 7. Gerund | 8. Supine |
| (vi) | Tense: | 1. Present | 2. Imperfect | 3. Future | 4. Past |
| (vii) | Voice: | 1. Active | 2. Passive | | |
| (viii) | Status: | 1. Main | 2. Auxiliary | | |

Attribute (v) has two names because of different traditions, for different European languages, regarding the use of the term *Mood*. In fact, the first four values (v) 1–4 are applicable to Finite Verbs and the last four (v) 5–8 to Non-finite Verbs.

Attribute (vii) **Voice** refers to the morphologically-encoded passive, e.g. in Danish and in Greek. Where the passive is realised by more than one verb, this does not need to be represented in the tagset.

The same applies to compound tenses (attribute (vi)). In general, compound tenses are not dealt with at the morphosyntactic level, since they involve the combination of more than one verb in a larger construction.

**3. Adjectives (AJ)**

| | | | | | |
|---|---|---|---|---|---|
| (i) | Degree: | 1. Positive | 2. Comparative | 3. Superlative | |
| (ii) | Gender: | 1. Masculine | 2. Feminine | 3. Neuter | |
| (iii) | Number: | 1. Singular | 2. Plural | | |
| (iv) | Case: | 1. Nominative | 2. Genitive | 3. Dative | 4. Accusative |

Attribute (i) **Degree** applies only to inflectional comparatives and superlatives. In some languages, e.g. Spanish, the number of such adjectives is very small.

### 4. Pronouns and Determiners (PD)

| | | | | | |
|---|---|---|---|---|---|
| (i) | Person: | 1. First | 2. Second | 3. Third | |
| (ii) | Gender: | 1. Masculine | 2. Feminine | 3. Neuter | |
| (iii) | Number: | 1. Singular | 2. Plural | | |
| (iv) | Possessive: | 1. Singular | 2. Plural | | |
| (v) | Case: | 1. Nominative | 2. Genitive | 3. Dative | 4. Accusative |
| | | 5. Non-genitive | 6. Oblique | | |
| (vi) | Category: | 1. Pronoun | 2. Determiner | 3. Both | |
| (vii) | Pron.-Type: | 1. Demonstrative | 2. Indefinite | 3. Possessive | 4. Int./Rel. |
| | | 5. Pers./Refl. | | | |
| (viii) | Det.-Type: | 1. Demonstrative | 2. Indefinite | 3. Possessive | 4. Int./Rel. |
| | | 5.Partitive | | | |

The parts of speech **Pronoun**, **Determiner** and **Article** heavily overlap in their formal and functional characteristics, and different analyses for different languages entail separating them out in different ways. For the present purpose, we have proposed placing Pronouns and Determiners in one ‘super-category’, recognising that for some descriptions it may be thought best to treat them as totally different parts of speech.

There is also an argument for subsuming Articles under Determiners. The present guidelines do not prevent such a realignment of categories, but do propose that articles (assuming they exist in a language) should always be recognised as a separate class, whether or not included within determiners. The requirement is that the descriptive scheme adopted should be automatically mappable into the present one via an Intermediate Tagset (see 4.3).

Attribute (iv) accounts for the fact that a possessive pronoun or possessive determiner may have two different numbers. This attribute handles the number which is inherent to the possessive form (e.g. Italian *(la) mia, (la) nostra* as first-person singular and first-person plural) as contrasted with the number it has by virtue of agreeing with a particular noun (e.g. Italian *(la) mia, (le) mie*).

Under attribute (v) **Case**, the value **Oblique** applies to pronouns such as *them* and *me* in English, and equivalent pronouns such as *dem* and *mig* in Danish. These occur in object function, and also after prepositions.

Under attributes (vi) and (vii), the subcategories **Interrogative** and **Relative** are merged into a single value **Int./Rel.**. It is often difficult to distinguish these in automatic tagging, but they may be optionally distinguished (see 4.2, point 3) at a more delicate level of granularity.

Similarly, under attribute (vi), **Personal** and **Reflexive** pronouns are brought together as a single value **Pers./Refl.**. Again, they may be optionally separated (see 4.2, point 3) at a more delicate level.

### 5. Articles (AT)

| | | | | | |
|---|---|---|---|---|---|
| (i) | Article-Type: | 1. Definite | 2. Indefinite | | |
| (ii) | Gender: | 1. Masculine | 2. Feminine | 3. Neuter | |
| (iii) | Number: | 1. Singular | 2. Plural | | |
| (iv) | Case: | 1. Nominative | 2. Genitive | 3. Dative | 4. Accusative |

### 6. Adverbs (AV)

| | | | | |
|---|---|---|---|---|
| (i) | Degree: | 1. Positive | 2. Comparative | 3. Superlative |

There are many possible subdivisions of adverbs on syntactic and semantic grounds, but these are regarded as optional rather than recommended (see 4.2, point 3).

**7. Adpositions (AP)**

| | | |
|---|---|---|
| (i) | Type: | 1. Preposition |

In practice, the overwhelming majority of cases of adpositions we have to consider in European languages are prepositions. Hence only this one value needs to be recognised at the *recommended* level. Other possibilities, such as **Postpositions** and **Circumpositions** are dealt with at the optional level (see 4.2, point 3).

**8. Conjunctions (C)**

| | | | |
|---|---|---|---|
| (i) | Type: | 1. Coordinating | 2. Subordinating |

**9. Numerals (NU)**

| | | | | | |
|---|---|---|---|---|---|
| (i) | Type: | 1. Cardinal | 2. Ordinal | | |
| (ii) | Gender: | 1. Masculine | 2. Feminine | 3. Neuter | |
| (iii) | Number: | 1. Singular | 2. Plural | | |
| (iv) | Case: | 1. Nominative | 2. Genitive | 3. Dative | 4. Accusative |
| (v) | Function: | 1. Pronoun | 2. Determiner | 3. Adjective | |

In some languages (e.g. Portuguese) this category is not normally considered to be a separate part of speech, because it can be subsumed under others (e.g. cardinal numerals behave like pronouns/determiners; ordinal numerals behave more like adjectives). We recognise that in some tagsets **Numeral** may therefore occur as subcategory within other parts of speech. (Compare the treatment of articles under 5 above). At the same time, it is possible to indicate the part-of-speech function of a word within the numeral category by making use of attribute (v).

**10. Interjections (I)**

No subcategories are recommended.

**11. Unique/Unassigned (U)**

No subcategories are recommended, although it is expected that tagsets for individual languages will need to identify such one-member word-classes as **Negative particle**, **Existential particle**, **Infinitive marker**, etc. (See 4.2, point 3 for more information.)

**12. Residual (R)**

| | | | | | | |
|---|---|---|---|---|---|---|
| (i) | Type: | 1. Foreign word | 2. Formula | 3. Symbol | 4. Acronym | 5. Abbreviation |
| | | 6. Unclassified | | | | |
| (ii) | Number: | 1. Singular | 2. Plural | | | |
| (iii) | Gender: | 1. Masculine | 2. Feminine | 3. Neuter | | |

The **Unclassified** category applies to word-like text segments which do not easily fit into any of the foregoing values. For example: incomplete words and pause fillers such as *er* and *erm* in transcriptions of speech, or written representations of singing such as *dum-de-dum*.

Although words in the Residual category are on the periphery of the lexicon, they may take some of the grammatical characteristics, e.g., of nouns. Acronyms such as *IBM* are similar to proper nouns; symbols such as alphabetic characters can vary for singular and plural (e.g. *How many Ps are there in 'psychopath'?*), and are in this respect like common nouns. In some languages (e.g. Portuguese) such symbols also have gender. It is quite reasonable that in some tagging schemes some of these classes of word will be classified under other parts of speech.

### 13. Punctuation marks (PU)

Word-external punctuation marks, if treated as words for morphosyntactic tagging, are sometimes assigned a separate tag (in effect, an attribute value) for each main punctuation mark:

| | | | | |
|---|---|---|---|---|
| (i) | 1. Period | 2. Comma | 3. Question mark | . . . etc. . . . |

An alternative is to group the punctuation marks into positional classes:

| | | | | |
|---|---|---|---|---|
| (i) | 1. Sentence-final | 2. Sentence-medial | 3. Left-Parenthetical | 4. Right-Parenthetical |

Under 1 are grouped **. ? !**. Under 2 are grouped **, ; : —** . Under 3 are placed punctuation marks which signal the initiation of a constituent, such as **(**, **[** , and **¿** in Spanish). Under 4 are grouped punctuation marks which conclude a constituent the opening of which is marked by one of the devices in 3: e.g. **)**, **]** and Spanish **?** . We make no recommendation about choosing between these two sets of punctuation values. [1]

### 4.2.3 Special extensions — Optional generic attributes/values

Here we deal with aspects of morphosyntactic annotation which are optional, and may be included in the annotation scheme according to need. Many of them go beyond morphosyntax and are of a syntactic or semantic nature. There is decidedly no claim to completeness. We do not recommend any of these features, but simply present them as having illustrative value. This subsection deals with generic optional features, i.e. those which are *application-* or *task-specific*. See section 4.2.4 — language-specific features — for another class of special extension.

### 1. Nouns

One might wish to introduce semantically and syntactically oriented attributes such as countability:

| | | | |
|---|---|---|---|
| (v) | Countability: | 1. Countable | 2. Mass |

### 2. Verbs

Additional optional attributes:

| | | | |
|---|---|---|---|
| (ix) | Aspect: | 1. Perfective | 2. Imperfective |
| (x) | Separability: | 1. Non-separable | 2. Separable |
| (xi) | Reflexivity: | 1. Reflexive | 2. Non-reflexive |
| (xii) | Auxiliary: | 1. Have | 2. Be |

Attribute (ix) is needed for Greek and Slavonic languages. It corresponds also to the Past Simple/Imperfect distinction of Romance languages.
Attribute (x) is relevant for German compound verbs (*fängt . . . an*, *anfangen*) and also to phrasal verbs in Danish and English.
Attribute (xii) is applied to main verbs in French, German, Dutch, etc., and determines the selection of *avoir* or *être*, etc., as auxiliary for the Perfect.
Additional optional value for recommended attribute Status (see 4.2.2, point 2):

| | | |
|---|---|---|
| (viii) | Status: | 3. Semi-auxiliary |

In addition to main and auxiliary verbs, it may be useful (e.g. in English) to recognise an intermediate category of semi-auxiliary for such verbs as *be going to*, *have got to*, *ought to*.

[1] The punctuation category is clearly the most peripheral of the above categories as regards relevance to morphosyntax. There is also a scale of peripherality within the punctuation category. For example, on the boundary between punctuation and the mark-up of a text are such features as highlighting, whether realised by italics, bold-face, or capitals, which according to one view, should be included within tagging schemes.

### 3. Adjectives

Additional optional attributes:

| | | | | |
|---|---|---|---|---|
| (v) | Inflection-type: | 1. Weak-Flection | 2. Strong-Flection | 3. Mixed |
| (vi) | Use: | 1. Attributive | 2. Predicative | |
| (vii) | NP Function: | 1. Premodifying | 2. Postmodifying | 3. Head-function |

**Weak** and **Strong** (attribute (v)) are values for adjectival inflection in the Germanic languages German, Dutch and Danish. The syntactic attribute (vi) makes a distinction, for example, between *main* (Attributive) and *asleep* (Predicative) in English.

### 4. Pronouns and Determiners

Additional optional attributes:

| | | | | |
|---|---|---|---|---|
| (ix) | Special Pronoun Type: | 1. Personal | 2. Reflexive | 3. Reciprocal |
| (x) | Wh-Type: | 1. Interrogative | 2. Relative | 3. Exclamatory |
| (xi) | Politeness: | 1. Polite | 2. Familiar | |

Attribute (xi) is limited to second-person pronouns. In some languages (e.g. French) it is possible to treat **Polite** and **Familiar** simply as pragmatic values encoded through other attributes — especially person and number. In languages where there are special polite pronoun forms (e.g. Dutch *u* and Spanish *usted*), the additional **Politeness** attribute is required.

### 6. Adverbs

| | | | | |
|---|---|---|---|---|
| (ii) | Adverb-Type: | 1. General | 2. Degree | |
| (iii) | Polarity: | 1. Wh-type | 2. Non-wh-type | |
| (iv) | Wh-type: | 1. Interrogative | 2. Relative | 3. Exclamatory |

Attribute (ii) allows the tagset to distinguish degree adverbs, which have a distinctive syntactic function, (such as *very, so, too*) from others. Attribute (iv) enables the tagset to mark separately the **Wh-** or **Qu-** adverbs which are interrogative, relative or exclamatory in function. The relevant adverbs (in English) are *when*, *where*, *how* and *why*.

### 7. Adposition

| | | |
|---|---|---|
| (i) | Type: | 2. Fused prep-art |

The additional value **Fused prep-art** is for the benefit of those who do not find it practical to split fused words such as French *au* (= *à* + *le*) into two textwords. This very common phenomenon of a fused preposition + article in West European languages should preferably, however, be handled by assigning two tags to the same orthographic word (one for the preposition and one for the article).

### 8. Conjunctions

| | | | | | |
|---|---|---|---|---|---|
| (ii) | Coord-Type: | 1. Simple | 2. Correlative | 3. Initial | 4. Non-initial |

This attribute subclassifies coordinating conjunctions. It is easier to assign one tag to one orthographic word and it is therefore suggested that the four values are assigned as follows: **Simple** applies to the regular type of coordinator occurring between conjuncts: German *und*, for example. When the same word is also placed before the first conjunct, as in French *ou... ou...*, the former occurrence is given the **Correlative** value and the latter the **Simple** value. When two distinct words occur, as in German *weder... noch...*, then the first is given the **Initial** value and the second the **Non-initial** value.

### 4.2.4 Special extensions — Optional language-specific attributes/values

The following are examples of special extensions of the tagset which may be needed for *specific languages*. As in 4.2.3 above, the examples are purely illustrative and there is certainly no claim to completeness. Thus, we do not recommend any of these features. In some cases, they derive from work already done on tagsets and their applications to texts. In other cases, they derive from specialist research, or from comments on an early draft of these guidelines, supplied by specialists in particular languages.

**1. Nouns**

An additional language-specific attribute is:

| | | | | | |
|---|---|---|---|---|---|
| (vi) | Definiteness: | 1. Definite | 2. Indefinite | 3. Unmarked | [Danish] |

This is to handle the suffixed definite article in Danish: e.g. *haven* (‘the garden’); *havet* (‘the sea’)
Additional values:

| | | | | |
|---|---|---|---|---|
| (ii) | Gender: | 4. Common | | [Danish, Dutch] |
| (iv) | Case: | 6. Vocative | 7. Indeclinable | [Greek] |

The **Common** gender contrasts with **Neuter** in a two-gender system.

**2. Verbs**

An additional attribute:

| | | | | |
|---|---|---|---|---|
| (xiii) | Aux.-function: | 1.Primary | 2.Modal | [English] |

The primary (non-modal) auxiliaries are *be*, *have* and *do*.
An additional value to the non-finite category of verbs is arguably needed for English, because of the merger in that language of the gerund and participle functions. The -ing form does service for both and the two traditional categories are not easily distinguishable.

| | | | |
|---|---|---|---|
| (v) | Verb-form / Mood: | 9. -Ing form | [English] |

**3. Adjectives**

Additional values for **Case**:

| | | | | |
|---|---|---|---|---|
| (iv) | Case: | 5. Vocative | 6. Indeclinable | [Greek] |

**4. Pronouns and Determiners**

An additional value for **Gender** and for **Case**:

| | | | |
|---|---|---|---|
| (ii) | Gender: | 4. Common | [Danish] |
| (v) | Case: | 7. Prepositional | [Spanish] |

An additional attribute:

| | | | | |
|---|---|---|---|---|
| (xii) | Strength | 1. Weak | 2. Strong | [French, Dutch, Greek] |

**Weak** and **Strong** distinguish, for example, *me* from *moi* in French, and *me* from *mij* in Dutch.

### 5. Articles

Again, additional values for **Article-Type**, **Gender** and **Case** are:

| | | | | |
|---|---|---|---|---|
| (i) | Article-Type: | 3. Partitive | | [French] |
| (ii) | Gender: | 4. Common | | [Danish] |
| (iv) | Case: | 5. Vocative | 6. Indeclinable | [Greek] |

### 6. Adverbs

Additional values for **Adverb-Type**:

| | | | | |
|---|---|---|---|---|
| (ii) | Adverb-Type: | 3. Particle | 4. Pronominal | [English, German, Dutch] |

In some tagging schemes, especially for English, a particle such as *out, off* or *up* counts as a subclass of adverb. In other tagging schemes, the particle may be treated under **Residual** as a special word-class. German and Dutch have pronominal adverbs such as German *daran*, *davon*, *dazu*.

### 7. Adpositions

Values for **Adposition-Type**, in addition to 1. Preposition and 2. Fused-preposition:

| | | | | |
|---|---|---|---|---|
| (i) | Type: | 3. Postposition | 4. Circumposition | [German, English] |

German *entlang* is a **Postposition**, and arguably, the *'s* which forms the genitive in English is no longer a case marking, but an enclitic postposition, as in *the Secretary of State's decision*, *in a month or so's time*. German *(auf. . . ) hin* is an example which can be analysed as a **Circumposition**.

### 8. Conjunctions

An additional attribute, applying to subordinating conjunctions only:

| | | | | | |
|---|---|---|---|---|---|
| (iii) | Subord.-type: | 1. With-finite | 2. With-infin. | 3. Comparative | [German] |

For example, in German, *weil* introduces a clause with a finite verb, whereas *ohne (zu. . . )* is followed by an infinitive, and *als* is followed by various kinds of comparative clause (including clauses without finite verbs).

### 11. Unique/Unassigned

The following miscellaneous values may occur:

| | | | |
|---|---|---|---|
| (i) | Unique-type: | 1. Infinitive marker | [German *zu*, Danish *at*, Dutch, English] |
| | | 2. Negative particle | [English *not*, *n't*] |
| | | 3. Existential marker | [English *there*, Danish *der*] |
| | | 4. Second negative particle | [French *pas*] |
| | | 5. Anticipatory *er* | [Dutch] |
| | | 6. Mediopassive voice marker *se* | [Portuguese] |
| | | 7. Preverbal particle | [Greek] |

## 4.3 Intermediate Tagset

For any tagset designed for the annotation of texts in a given language, the guidelines do not impose any particular set of choices to be used in distinguishing and representing grammatical categories. But it is important that the tagset should be mappable (if possible automatically) on to a set of attribute–value pairs in conformity with the guidelines presented in 4.2. This includes the possibility (indeed the probability) that the annotator will need to define optional values other than the special extensions (see 4.2, point 3).

This mapping will have the additional value that it will enable the annotator to transfer the information in a morphosyntactically-tagged corpus to the morphosyntactic component of a lexicon (e.g. in order to record frequencies of word-tag pairs). It will also enable a lexicon of the given language to be used as a major input to automatic tagging.

To aid this mapping, and to test out its efficacy, we suggest that an *Intermediate Tagset* can be used as a language-neutral representation of a set of attribute–value pairs, based on the word categorisation presented in 4.2. This can act as an intermediate stage of mapping between the tags assigned to textwords in corpus annotation and the labels assigned to words in a lexicon. Another important function of this Intermediate Tagset is to act as a basis for interchange between different local tagsets for particular corpora and particular languages.

A convenient linear method of representation is arrived at as follows:

(i) Represent the obligatory part-of-speech attribute value by using one or more letters, as indicated in 4.2.1:

| | | |
|---|---|---|
| N = noun | AV = adverb | I = interjection |
| V = verb | AP = adposition | U = unique/unassigned |
| AJ = adjective | C = conjunction | R = residual |
| PD = pronoun/determiner | NU = numeral | PU = punctuation |
| AT = article | | |

(ii) Represent the whole tag as a linear string of characters, each attribute (roman number (i), (ii), (iii), (iv), ...) representing the first, second, third, fourth,. . . place in a string of digits.

(iii) Represent each value of each attribute by employing the arabic digits used in the recommended attributes and values 4.2.2. Thus, the interpretation of the string of digits will vary according to the part-of-speech category. (The optional attributes and values in 4.2, point 3 may also be used, but have to be specially defined for each tagset).

Examples:

- A common noun, feminine, plural, countable, is represented: **N122010**
- A 3rd person, singular, finite, indicative, past tense, active, main verb, non-phrasal, non-reflexive, verb is represented: **V3011141101200**
- A comparative, general adjective is represented: **AJ2000000**
- A coordinating conjunction, simple, is presented: **C110**
- An interjection is represented: **I**
- A plural symbol (as in *two Bs*) is represented: **R320**

Wherever an attribute is inapplicable to a given word in a given tagset, the value **0** fills that attribute's place in the string of digits. (See further, for the use of **0**, the section on underspecification in 4.4). When the **0**s occur in final position, without any non-zero digits following, they could be omitted without loss of information. Thus a comparative general adjective could simply be represented: **AJ2**. However, for clarity, the **0**s should be added.

There may be cases where a category needed for tagging in a specific language (given current limitations of automatic tagging) cuts across two or more values in the optional categories of the guidelines, and may even cut across different attributes as well. It is necessary to define what this value means by using the **OR** operator ( | ), and brackets to identify the arguments of this operator. Another operator we can use is the negative operator, signalled by the minus (-), so that **-4** means "all values of this attribute except the 4th".
A good example is the base form of the English verb. The finite base form in English can be specified by using a disjunction "[finite indicative present tense [plural or [first person or second person] singular] or imperative or subjunctive]". This is spelled out, using the intermediate tagset, as follows:

**V[[-301|002]111|000121|000130]0100000**

Even this leaves out the non-finite use of the base form, as an infinitive. This example, awkward as it is, has an explanatory value: the relation between tagsets and a language-neutral representation can be very indirect. Although such cases as this are unusual, they show that the mapping between a lexicon and a tagged corpus is not always an easy one to automate.
To illustrate the method of converting a tagset into this type of language-neutral labelling, we present in Appendix A a rendering into an *Intermediate Tagset* of a tagset for English and in Appendix B of a set of dictionary codes for Italian; the former are based on the English implementation of the lexicon guidelines and the latter on the codes of the *DMI* (Calzolari et al. 1980). (For English, with its simple morphology, we find the most complex interrelation between the morphosyntactic guidelines and the requirements of a particular language. With other languages, the mapping from the language-specific tagset to the Intermediate tagset is likely to be more straightforward.)

## 4.4 Underspecification and ambiguity in tagging

Underspecification and ambiguity are two descriptively incomplete phenomena, where some information which the tagset could in principle provide is not provided. But the reasons for this lack of information are quite different in the two cases, which should therefore be clearly distinguished.

### 4.4.1 Dealing with underspecification

*Underspecification* is the phenomenon (sometimes called *neutralisation*) illustrated by the use of **0** in the Intermediate Tagset. It means that the distinction between the different values of an attribute is not relevant in this instance. One could also say that the particular attribute marked **0** is not applied to the textword under consideration. The possible reasons for this are threefold:

**Language underspecifies:** The attribute does not apply to the part-of-speech in the language under consideration. For example, Gender does not apply to Nouns in English. Case does not apply to Adjectives in French.

**Tagset underspecifies:** Although the attribute does apply to the part-of-speech in the language under consideration, the tagset is not fine-grained enough to represent it. For example, a particular tagset for English may omit representation of Gender (*he*, *she*) for pronouns.

**Word underspecifies:** Although the attribute does apply to the language, and is represented in the tagset, it is not marked on this particular word, because it is neutralised. For example, in French, the plural article *les* is unspecified for Gender. Invariable adjectives, such as German *prima*, are unspecified for Gender, Case and Number.

There is room for different viewpoints on whether morphological syncretism should lead to underspecification of values, or whether values, even where they are not morphologically signalled, should be specified on the basis of context. There is also room for difference of opinion about whether the unmarked value of a binary attribute should be applied to the absence of the marked value. (E.g. should we mark all verbs which are not passive in Danish as active? Or should we leave Voice unspecified, except with those verbs for which the passive is an option?)

However, the important point to make here is that underspecification is normally signalled, in a tagset, simply by the absence of any indicator of the attribute. Alternatively, as in the Intermediate Tagset (see 4.3), a **0** is used to make the absence of an attribute explicit.

### 4.4.2 Dealing with ambiguity

*Ambiguity*, as contrasted with underspecification, is the phenomenon of lack of information, where there is uncertainty between two or more alternative descriptions. Four different senses of ambiguity can be distinguished in morphosyntactic tagging.

#### 4.4.2.1 Grammatical homonymy

The English word *round* has five potential tags: it can be

1. A preposition;
2. An adverb/particle;
3. An adjective;
4. A noun; or
5. A verb.

Normally, this type of ambiguity, if it is considered such, does not occur in an annotated corpus, since the ambiguity is resolved.

#### 4.4.2.2 Portmanteau tags

However, with large corpora, tagging is done automatically, and there may be no need or opportunity for the manual post-editing of the whole corpus. It can be practical, in such cases, to retain more than one tag in the annotated corpus, where the automatic tagging algorithms have not provided strong enough evidence for disambiguation. For example, in the British National Corpus, a set of *portmanteau tags* is used in recording such ambiguities. One of them is the tag **VVD-VVN**, which means “either the past tense or the past participle of a lexical verb”. The portmanteau tag appears in the annotated British National Corpus in the TEI format of an entity reference appended to the word, e.g.: *liked***&VVD-VVN;**. Other formats of presentation would also be reasonable. A portmanteau tag signals uncertainty about the appropriate tag, for reasons of fallible automatic processing. It is assumed that a trained human post-editor would in general have no difficulty in resolving the ambiguity.

#### 4.4.2.3 Human uncertainty ambiguities

A further type of ambiguity may arise where the human annotator cannot decide on a single appropriate tag. There may be good reasons for this type of indecision:

- The annotation scheme may fail to provide criteria for disambiguation;
- Two or more human annotators may have different opinions, or different theoretical perspectives, on the same data;
- The categories may themselves have unclear boundaries — not because of any human failing, but because that is what linguistic categories are like.

In the present stage of development of morphosyntactic tagging, the ability to deal with this kind of ambiguity is not a matter of great priority — but it may become more important in the future.

#### 4.4.2.4 Genuine textual ambiguities

By this we mean cases where the text does not provide enough information for disambiguation between two or more clearly defined categories. For example, it may be unclear whether in a given case the exclamatory word *Fire!* is a verb or a noun. Ideally, in such cases, more than one tag should be attached to the same textword.

The encoding of ambiguity in morphosyntactic annotation has so far received little attention, and we make no recommendations except to propose that in principle, all the kinds of ambiguity listed above should be distinguishable by different mark-up.

### 4.5 Multiple tagging practices: Form-function and lemmatisation

Ambiguity is just one of a number of phenomena for which some kind of multiple tagging of the same textword may be required. Other cases of multiple tagging which should be mentioned are:

1. **Form-function tagging:** Sometimes the need is felt to assign two different tags to the same word: one representing the formal category, and the other the functional category, e.g.:

   - A word with the form of a past participle but the function of an adjective;
   - A word with the form of an adjective but the function of an adverb.

   In principle, it can be argued that two tags should be assigned to each of these word types, and should be distinctly encoded. In practice, tagging schemes up to the present have tended to give priority of one criterion over another (i.e. giving priority to function over form or *vice versa*). The annotation scheme for a given tagged corpus should clearly state the use of such criteria.

2. **Lemma tagging:** A morphosyntactically tagged corpus is generally supposed to specify the grammatical form of a textword, rather than to recover the lemma. However, in transfer of information from a corpus to a lexicon or *vice versa*, it is assumed that a lemmatisation algorithm will have an important role. There is also a case (especially as a preliminary to syntactic and semantic annotation) for a type of annotation which specifies the lemma, as well as the grammatical form, for each textword. *Lemma tagging*, as this process may be called, has so far not been widely undertaken. Once again, the need is for independent ways of representing the lemma tag and the grammatical form tag.

For both the above cases of multiple tagging, as well as for the tagging of ambiguity, there is a need for assigning more than one morphosyntactic tag to the same word. There is a case for preference for a vertical format for presenting such a multiply-tagged annotated corpus. The combination of different kinds of word tagging in the same annotated corpus can then be managed, without confusion, by associating each kind of tag with a different field or column alongside the vertical text.

## 5 References

Calzolari, N., Ceccotti, M.L. & Roventini, A. (1980) Documentazione sui tre nastri contenenti il DMI. Technical Report. Pisa: ILC-CNR.

Rabiner, L.R. (1990) A tutorial on Hidden Markov Models and selected applications in speech recognition. In Waibel, A. & Lee, K. (eds) Readings in speech recognition. New York: Morgan Kaufman. 267–96.

# A English tagset, with intermediate tags

## Part 1

| *Tag* | *Description of word category* | *Example(s)* | *Intermediate Tag* |
|---|---|---|---|
| AJ | (Positive) adjective, general | big | AJ10000000 |
| AJR | Comparative adjective | bigger | AJ20000000 |
| AJT | Superlative adjective | biggest | AJ30000000 |
| APR | Preposition | at, of | AP1 |
| APO | Postposition | ’s | AP3 |
| ATD | Definite article | the | AT1000 |
| ATIs | Indefinite article, singular | a, an | AT2010 |
| AV | (Positive) adverb, general | soon | AV1120 |
| AVD | (Positive) adverb of degree | very, so | AV1220 |
| AVDR | Comparative adverb of degree | more, less | AV2220 |
| AVDT | Superlative adverb of degree | most, least | AV3220 |
| AVDWQ | Adverb of degree, other wh-type | how | AV021[1\|3] |
| AVR | Comparative adverb, general | sooner | AV2120 |
| AVT | Superlative adverb, general | soonest | AV3120 |
| AVWQ | General adverb, other wh-type | when, why | AV011-2 |
| AVWR | General adverb, relative | where, why | AV0112 |
| CC | Coordinating conjunction, simple | and | C110 |
| CCI | Coordinating conjunction, initial | both (. . . and) | C130 |
| CCM | Coordinating conjunction, medial | (neither . . . ) nor | C140 |
| CSC | Subordinating conjunction, comparative | than | C203 |
| CSF | Subordinating conjunction, with finite | if, while | C201 |
| CSN | Subordinating conjunction, with nonfinite | in order (to) | C202 |
| DDs | Singular demonstrative determiner | this, that | PD001002010000 |
| DDp | Plural demonstrative determiner | these | PD002002010000 |
| DI | Indefinite det., neutral for number | no, some | PD000002020000 |
| DIs | Indefinite determiner, singular | every, much | PD001002020000 |
| DIp | Indefinite determiner, plural | both, many | PD002002020000 |
| DVs1 | Possessive det., 1st pers. sing | my | PD100102030000 |
| DV2 | Possessive det, 2nd person | your | PD200002030000 |
| DV3sF | Possessive det, 3rd pers. sg. fem. | her | PD320102030000 |
| DV3sM | Possessive det, 3rd pers. sg. masc. | his | PD310102030000 |
| DV3sU | Possessive det, 3rd pers. sg. neut. | its | PD330102030000 |
| DVp1 | Possessive det, 1st pers. plur | our | PD100202030000 |
| DVp3 | Possessive det, 3rd pers. plur | their | PD300202030000 |
| DWR | Relative determiner | which | PD000002040200 |
| DWQ | Other wh-determiner | what | PD000002040-200 |
| IJ | Interjection | Oh, Yes | I |
| NCs | Singular common noun | book, man | N101000 |
| NCp | Plural common noun | books, men | N102000 |
| NPs | Singular proper noun | Mary | N201000 |
| NPp | Plural proper noun | Rockies | N202000 |

## Part 2

| *Tag* | *Description of word category* | *Example(s)* | *Intermediate Tag* |
|---|---|---|---|
| NUC | Cardinal numeral, neutral for number | two | NU10000 |
| NUCs | Singular cardinal numeral | one | NU10100 |
| NUCp | Plural cardinal numeral | fifties | NU10200 |
| NUOs | Singular ordinal numeral | seventh | NU20100 |
| NUOp | Plural ordinal numeral | sevenths | NU20200 |
| PDs | Singular demonstrative pronoun | this | PD001001100000 |
| PDp | Plural demonstrative pronoun | those | PD002001100000 |
| PI | Indefinite pronoun, neutral for number | all, none | PD000001200000 |
| PIs | Singular indefinite pronoun | anyone | PD001001200000 |
| PIp | Plural indefinite pronoun | few, many | PD002001200000 |
| PPs1N | Personal pronoun, 1st pers. sg. nom. | I | PD101011501000 |
| PPs1O | Personal pronoun, 1st pers. sg. obl. | me | PD101061501000 |
| PP2 | Second person personal pronoun | you | PD2000[1\|6]1501000 |
| PPs3NF | Pers. pron., 3rd pers.sg.nom.fem. | she | PD321011501000 |
| PPs3NM | Pers. pron., 3rd pers.sg.nom.masc. | he | PD311011501000 |
| PPs3U | Pers. pron., 3rd pers.sing.neuter | it | PD3310[1\|6]1501000 |
| PPs3OF | Pers. pron., 3rd pers.sg.obl.fem. | her | PD321061501000 |
| PPs3OM | Pers.pron., 3rd pers.sg.obl.masc. | him | PD311061501000 |
| PPp1N | Personal pronoun, 1st pers. pl. nom. | we | PD102011501000 |
| PPp1O | Personal pronoun, 1st pers. pl. oblique | us | PD102061501000 |
| PPp3N | Personal pronoun, 3rd pers. pl. nom. | they | PD302011501000 |
| PPp3O | Personal pronoun, 3rd pers. pl. oblique | them | PD302061501000 |
| PRs1 | Reflexive pronoun, 1st person singular | myself | PD101001502000 |
| PRs2 | Reflexive pronoun, 2nd person singular | yourself | PD201001502000 |
| PRs3F | Reflexive pronoun, 3rd pers. sg. fem. | herself | PD321001502000 |
| PRs3M | Reflexive pronoun, 3rd pers. sg. masc. | himself | PD311001502000 |
| PRs3U | Reflexive pronoun, 3rd pers. sg. neut. | itself | PD331001502000 |
| PRp1 | Reflexive pronoun, 1st person plural | ourselves | PD102001502000 |
| PRp2 | Reflexive pronoun, 2nd person plural | yourselves | PD202001502000 |
| PRp3 | Reflexive pronoun, 3rd person plural | themselves | PD302001502000 |
| PVs1 | Possessive pronoun, 1st person singular | mine | PD100101300000 |
| PV2 | Possessive pronoun, 2nd person | yours | PD200001300000 |
| PVs3F | Possessive pronoun, 3rd person fem. | hers | PD320101300000 |
| PVs3M | Possessive pronoun, 3rd person masc. | his | PD310101300000 |
| PVs3U | Possessive pronoun, 3rd person neut. | its | PD330101300000 |
| PVp1 | Possessive pronoun, 1st person plural | ours | PD100201300000 |
| PVp3 | Possessive pronoun, 3rd person plural | theirs | PD300201300000 |
| PWQ | Other wh-pronoun, neutral for case | what, which | PD000001400-200 |
| PWQN | Other wh-pronoun, nominative | who | PD000011400-200 |
| PWQO | Other wh-pronoun, oblique | whom | PD000061400-200 |
| PWR | Relative pronoun, neutral for case | which | PD000001400200 |
| PWRN | Relative pronoun, nominative | who | PD000011400200 |
| PWRO | Relative pronoun, oblique | whom | PD000061400200 |

## Part 3

| *Tag* | *Description of word category* | *Example(s)* | *Intermediate Tag* |
|---|---|---|---|
| RFO | Formula | X/21 | R200 |
| RFW | Foreign word | mawashi | R100 |
| RSY | Symbol, neutral for number | £, ’ | R300 |
| RSYs | Symbol, singular | A, b | R310 |
| RSYp | Symbol, plural | As, b’s | R320 |
| RUN | Unclassified | er, um | R600 |
| UI | infinitive marker | to (eat) | U1 |
| UN | negative particle | not, -n’t | U2 |
| UX | existential *there* | there | U3 |
| VM | Modal auxiliary verb | can, will | V0001100200002 |
| VPB | Finite base form, primary auxiliary | be, do, have | V[[-301\|002]111\|000121\|000130]0200001 |
| VPD | Past tense, primary auxiliary | did, had | V0001140200001 |
| VPDR | Past tense -re form, primary auxiliary | were | V[[201\|002]11\|00012]40200001 |
| VPDZ | Past tense -s form, primary auxiliary | was | V-2011140200001 |
| VPG | -Ing form, primary auxiliary | being, having | V0002900200001 |
| VPI | Infinitive, primary auxiliary | (to) be/have | V0002500200001 |
| VPM | Pres. tense 1st pers. sg, primary aux. | am | V1011110200001 |
| VPN | Past participle, primary auxiliary | been | V0002640200001 |
| VPR | Pres. tense -re form, primary auxiliary | are | V[201\|002]1110200001 |
| VPZ | Pres. tense -s form, primary auxiliary | is, does, has | V3011110200001 |
| VVB | Finite base form, main verb | eat, have | V[[-301\|002]111\|000121\|000130]0100000 |
| VVD | Past tense, main verb | ate, had | V0001140100000 |
| VVDR | Past tense -re form, main verb | were | V[[201\|002]11\|00012]40100000 |
| VVDZ | Past tense -s form, main verb | was | V-2011140100000 |
| VVG | -Ing form, main verb | leaving, being | V0002900100000 |
| VVI | Infinitive, main verb | (to) leave/do | V0002500100000 |
| VVM | Present tense 1st pers. sing, main verb | am | V1011110100000 |
| VVN | Past participle, main verb | eaten, left | V0002640100000 |
| VVR | Present tense -re form, main verb | are | V[201\|002]1110100000 |
| VVZ | Present tense -s form, main verb | is | V3011110100000 |

# B Italian DMI codes, with intermediate tags

## Part 1

| *Code* | *Description of word category* | *Example(s)* | *Intermediate Tag* |
|---|---|---|---|
| AFN | Adj.pos.femm.inv. | carta/e valore | AJ12[1\|2]0 |
| ANS | Adj.pos.comm.sing. | dolce | AJ1410 |
| ANP | Adj.pos.comm.plur. | dolci | AJ1420 |
| AMN | Adj.pos.masc.inv. | complemento/i oggetto | AJ11[1\|2]0 |
| AFSS | Adj.sup.femm.sing. | grandissima, massima | AJ3210 |
| AFPS | Adj.sup.femm.plur. | grandissime, massime | AJ3220 |
| AMPS | Adj.sup.masc.plur. | grandissimi, massimi | AJ3120 |
| AMSS | Adj.sup.masc.sing. | grandissimo, massimo | AJ3110 |
| ANSC | Adj.com.comm.sing. | maggiore | AJ2410 |
| ANPC | Adj.pos.comm.plur. | maggiori | AJ2420 |
| ANNC | Adj.pos.comm.inv. | meglio, peggio | AJ24[1\|2]0 |
| ANN | Adj.pos.comm.inv. | pari, dappoco | AJ14[1\|2]0 |
| AFS | Adj.pos.femm.sing. | vera | AJ1210 |
| AFP | Adj.pos.femm.plur. | vere | AJ1220 |
| AMP | Adj.pos.masc.plur. | veri | AJ1120 |
| AMS | Adj.pos.masc.sing. | vero | AJ1110 |
| B | Adv.pos. | forte | AV1000 |
| BC | Adv.com. | maggiormente | AV2000 |
| BS | Adv.pos.mann. | fortemente | AV1600 |
| BSS | Adv.sup.mann. | fortissimamente | AV3600 |
| C | Conj.subord. | perché | C200 |
| CC | Conj.coord. | e | C100 |
| DDMS | PrAdj.dem.masc.sing. | quello, quel | PD01100201 |
| DDMP | PrAdj.dem.masc.plur. | quelli | PD01200201 |
| DDFS | PrAdj.dem.femm.sing. | quella | PD02100201 |
| DDFP | PrAdj.dem.femm.plur. | quelle | PD02200201 |
| DDNS | PrAdj.dem.comm.sing. | ciò | PD04100201 |
| DDNP | PrAdj.dem.comm.plur. | costoro | PD04200201 |
| DIMS | PrAdj.ind.masc.sing. | alcuno, alcun | PD01100202 |
| DIMP | PrAdj.ind.masc.plur. | alcuni | PD01200202 |
| DIFS | PrAdj.ind.femm.sing. | qualcuna | PD02100202 |
| DIFP | PrAdj.ind.femm.plur. | poche | PD02200202 |
| DINS | PrAdj.ind.comm.sing. | ogni | PD04100202 |
| DINP | PrAdj.ind.comm.plur. | tali, altrui | PD04200202 |

**Part 2**

| *Code* | *Description of word category* | *Example(s)* | *Intermediate Tag* |
|---|---|---|---|
| DEMS | PrAdj.escl.masc.sing. | quanto! | PD0110020003 |
| DEMP | PrAdj.escl.masc.plur. | quanti! | PD0120020003 |
| DEFS | PrAdj.escl.femm.sing. | quanta! | PD0210020003 |
| DEFP | PrAdj.escl.femm.plur. | quante! | PD0220020003 |
| DENS | PrAdj.escl.comm.sing. | quale! | PD0410020003 |
| DENP | PrAdj.escl.comm.plur. | quali! | PD0420020003 |
| DENN | PrAdj.escl.comm.inv. | che! | PD04[1\|2]0020003 |
| DPMS1 | PrAdj.poss.1p.masc.sing. | mio | PD11100201 |
| DPMP1 | PrAdj.poss.1p.masc.plur. | miei | PD11200201 |
| DPFS1 | PrAdj.poss.1p.femm.sing. | mia | PD12100201 |
| DPFP1 | PrAdj.poss.1p.femm.plur. | mie | PD12200201 |
| DPMS2 | PrAdj.poss.2p.masc.sing. | tuo | PD21100201 |
| DPMP2 | PrAdj.poss.2p.masc.plur. | tuoi | PD21200201 |
| DPFS2 | PrAdj.poss.2p.femm.sing. | tua | PD22100201 |
| DPFP2 | PrAdj.poss.2p.femm.plur. | tue | PD22200201 |
| DPMS3 | PrAdj.poss.3p.masc.sing. | suo | PD31100201 |
| DPMP3 | PrAdj.poss.3p.masc.plur. | suoi | PD31200201 |
| DPFS3 | PrAdj.poss.3p.femm.sing. | sua | PD32100201 |
| DPFP3 | PrAdj.poss.3p.femm.plur. | sue | PD32200201 |
| DPMS1 | PrAdj.poss.1p.masc.sing. | nostro | PD11100201 |
| DPMP1 | PrAdj.poss.1p.masc.plur. | nostri | PD11200201 |
| DPFS1 | PrAdj.poss.1p.femm.sing. | nostra | PD12100201 |
| DPFP1 | PrAdj.poss.1p.femm.plur. | nostre | PD12200201 |
| DPMS2 | PrAdj.poss.2p.masc.sing. | vostro | PD21100201 |
| DPMP2 | PrAdj.poss.2p.masc.plur. | vostri | PD21200201 |
| DPFS2 | PrAdj.poss.2p.femm.sing. | vostra | PD22100201 |
| DPFP2 | PrAdj.poss.2p.femm.plur. | vostre | PD22200201 |
| DPNP3 | PrAdj.poss.3p.comm.plur. | loro | PD34200201 |
| DPNN | PrAdj.poss.comm.inv. | altrui | PD04[1\|2]00201 |
| DTMS | PrAdj.int.masc.sing. | quanto? | PD0110020001 |
| DTMP | PrAdj.int.masc.plur. | quanti? | PD0120020001 |
| DTFS | PrAdj.int.femm.sing. | quanta? | PD0210020001 |
| DTFP | PrAdj.int.femm.plur. | quante? | PD0220020001 |
| DTNN | PrAdj.int.comm.inv. | che? | PD04[1\|2]0020001 |
| DTNS | PrAdj.int.comm.sing. | quale? | PD0410020001 |
| DTNP | PrAdj.int.comm.plur. | quali? | PD0420020001 |
| DRNN | PrAdj.rel.comm.inv. | che | PD04[1\|2]0020002 |
| DRNS | PrAdj.rel.comm.sing. | quale | PD0410020002 |
| DRNP | PrAdj.rel.comm.plur. | quali | PD0420020002 |
| I |  | oh! | I |

## Part 3

| *Code* | *Description of word category* | *Example(s)* | *Intermediate Tag* |
|---|---|---|---|
| SFN | Noun comm.femm.inv. | attività (la/le) | N12[1\|2] |
| SFP | Noun comm.femm.plur. | case | N122 |
| SFS | Noun comm.femm.sing. | casa | N121 |
| SMN | Noun comm.masc.inv. | re, caffè (il/i) | N11[1\|2] |
| SMP | Noun comm.masc.plur. | libri | N112 |
| SMS | Noun comm.masc.sing. | libro | N111 |
| SNN | Noun comm.comm.inv. | sosia (il/la, i/le) | N14[1\|2] |
| SNP | Noun comm.comm.plur. | insegnanti (gli/le) | N142 |
| SNS | Noun comm.comm.sing. | insegnante (un/una) | N141 |
| SPFP | Noun prop.femm.plur. | Marie | N222 |
| SPFS | Noun prop.femm.sing. | Maria | N221 |
| SPMP | Noun prop.masc.plur. | Borboni | N212 |
| SPMS | Noun prop.masc.sing. | Mario | N211 |
| PDMS3 | Pron.dem.masc.sing.3 | costui | PD31100110 |
| PDMS | Pron.dem.masc.sing. | quello | PD01100110 |
| PDMP | Pron.dem.masc.sing. | quelli | PD01200110 |
| PDFS | Pron.dem.femm.sing. | quella | PD02100110 |
| PDFP | Pron.dem.femm.plur. | quelle | PD02200110 |
| PDNS | Pron.dem.comm.sing. | ciò | PD04100110 |
| PDNP | Pron.dem.comm.plur. | tali | PD04200110 |
| PEMS | Pron.escl.masc.sing. | quanto! | PD0110010003 |
| PEMP | Pron.escl.masc.plur. | quanti! | PD0120010003 |
| PEFS | Pron.escl.femm.sing. | quanta! | PD0210010003 |
| PEFP | Pron.escl.femm.plur. | quante! | PD0220010003 |
| PENS | Pron.escl.comm.sing. | che (vedo!) | PD0410010003 |
| PENN | Pron.escl.comm.inv. | chi! | PD04[1\|2]0010003 |
| PIMS | Pron.ind.masc.sing. | uno | PD01100120 |
| PIMP | Pron.ind.masc.plur. | alcuni | PD01200120 |
| PIFS | Pron.ind.femm.sing. | una | PD02100120 |
| PIFP | Pron.ind.femm.plur. | alcune | PD02200120 |
| PINS | Pron.ind.comm.sing. | chiunque | PD04100120 |
| PINP | Pron.ind.comm.plur. | tali, quali | PD04200120 |
| PPMS1 | Pron.poss.1p.masc.sing. | mio | PD11100130 |
| PPMP1 | Pron.poss.1p.masc.plur. | miei | PD11200130 |
| PPFS1 | Pron.poss.1p.femm.sing. | mia | PD12100130 |
| PPFP2 | Pron.poss.1p.femm.plur. | mie | PD12200130 |
| PPMS2 | Pron.poss.2p.masc.sing. | tuo | PD21100130 |
| PPMP2 | Pron.poss.2p.masc.plur. | tuoi | PD21200130 |
| PPFS2 | Pron.poss.2p.femm.sing. | tua | PD22100130 |
| PPFP2 | Pron.poss.2p.femm.plur. | tue | PD22200130 |

## Part 4

| *Code* | *Description of word category* | *Example(s)* | *Intermediate Tag* |
|---|---|---|---|
| PPMS3 | Pron.poss.3p.masc.sing. | suo | PD31100130 |
| PPMP3 | Pron.poss.3p.masc.plur. | suoi | PD31200130 |
| PPFS3 | Pron.poss.3p.femm.sing. | sua | PD32100130 |
| PPFP3 | Pron.poss.3p.femm.plur. | sue | PD32200130 |
| PPMS1 | Pron.poss.1p.masc.sing. | nostro | PD11100130 |
| PPMP1 | Pron.poss.1p.masc.plur. | nostri | PD11200130 |
| PPFS1 | Pron.poss.1p.femm.sing. | nostra | PD12100130 |
| PPFP1 | Pron.poss.1p.femm.plur. | nostre | PD12200130 |
| PPMS2 | Pron.poss.2p.masc.sing. | vostro | PD21100130 |
| PPMP2 | Pron.poss.2p.masc.plur. | vostri | PD21200130 |
| PPFS2 | Pron.poss.2p.femm.sing. | vostra | PD22100130 |
| PPFP2 | Pron.poss.2p.femm.plur. | vostre | PD22200130 |
| PPNP3 | Pron.poss.3p.comm.plur. | loro | PD34200130 |
| PTNS | Pron.int.comm.sing. | chi? | PD0410010001 |
| PTNN | Pron.int.comm.inv. | che? | PD04[1\|2]0010001 |
| PTMS | Pron.int.masc.sing. | quanto? | PD0110010001 |
| PTMP | Pron.int.masc.plur. | quanti? | PD0120010001 |
| PTFS | Pron.int.femm.sing. | quanta? | PD0210010001 |
| PTFP | Pron.int.femm.plur. | quante? | PD0220010001 |
| PRNN | Pron.rel.comm.inv. | che, chi, cui | PD04[1\|2]0010002 |
| PRNS | Pron.rel.comm.sing. | quanto | PD0410010002 |
| PRMS | Pron.rel.masc.sing. | quanto | PD0110010002 |
| PRMP | Pron.rel.masc.plur. | quanti | PD0120010002 |
| PRFP | Pron.rel.femm.plur. | quante | PD0220010002 |
| PQNS1 | Pron.pers.comm.sing.1 | io | PD141001001 |
| PQNS2 | Pron.pers.comm.plur.2 | tu | PD241001001 |
| PQMS3 | Pron.pers.masc.sing.3 | egli, lui, esso | PD311001001 |
| PQFS3 | Pron.pers.femm.sing.3 | ella, lei, essa | PD321001001 |
| PQNP1 | Pron.pers.comm.plur.1 | noi | PD142001001 |
| PQNP2 | Pron.pers.comm.plur.2 | voi | PD242001001 |
| PQNP3 | Pron.pers.comm.plur.3 | loro | PD342001001 |
| PQMP3 | Pron.pers.masc.plur.3 | essi | PD312001001 |
| PQFP3 | Pron.pers.femm.plur.3 | esse | PD322001001 |
| PQNS1 | Pron.pers.comm.sing.1 | me | PD141001001 |
| PQNS2 | Pron.pers.comm.sing.2 | te | PD241001001 |
| PQMS3 | Pron.pers.masc.sing.3 | lui, esso | PD311001001 |
| PQFS3 | Pron.pers.femm.sing.3 | lei, essa | PD321001001 |
| PQNP1 | Pron.pers.comm.plur.1 | noi | PD142001001 |
| PQNP2 | Pron.pers.comm.plur.2 | voi | PD242001001 |
| PQNP3 | Pron.pers.comm.plur.3 | loro | PD342001001 |
| PQMP3 | Pron.pers.masc.plur.3 | essi | PD312001001 |
| PQFP3 | Pron.pers.femm.plur.3 | esse | PD322001001 |

**Part 5**

| *Code* | *Description of word category* | *Example(s)* | *Intermediate Tag* |
|---|---|---|---|
| PQNS1 | Pron.pers.comm.sing.1 | mi | PD141001001 |
| PQNS2 | Pron.pers.comm.sing.2 | ti | PD241001001 |
| PQMS3 | Pron.pers.masc.sing.3 | gli | PD311001001 |
| PQNP1 | Pron.pers.comm.plur.1 | ci | PD142001001 |
| PQNP2 | Pron.pers.comm.plur.2 | vi | PD242001001 |
| PQNP3 | Pron.pers.comm.plur.3 | loro | PD342001001 |
| PQMP3 | Pron.pers.masc.plur.3 | li | PD312001001 |
| PQFP3 | Pron.pers.femm.plur.3 | le | PD322001001 |
| PFNS1 | Pron.refl.comm.sing.1 | mi (me stesso) | PD141001002 |
| PFNS2 | Pron.refl.comm.sing.1 | ti (te stesso) | PD241001002 |
| PFNN3 | Pron.refl.comm.inv. 3 | sè, si | PD311001002 |
| PFNP1 | Pron.refl.comm.plur.1 | ci | PD142001002 |
| PFNP2 | Pron.refl.comm.plur.2 | vi | PD242001002 |
| PFNP3 | Pron.refl.comm.plur.3 | loro | PD342001002 |
| VFY | Verb aux. inf.pres. | avere | V00025101 |
| VGY | Verb aux. ger.pres. | avendo | V00027102 |
| VF | Verb main inf.pres. | amare | V00025101 |
| VG | Verb main ger.pres. | amando | V00027102 |
| VP1IFY | Verb aux. 1pl.ind.fut. | avremo | V10211302 |
| VP2IFY | Verb aux. 2pl.ind.fut. | avrete | V20211302 |
| VP3IFY | Verb aux. 3pl.ind.fut. | avranno | V30211302 |
| VS1IFY | Verb aux. 1sg.ind.fut. | avrò | V10111302 |
| VS2IFY | Verb aux. 2sg.ind.fut. | avrai | V20111302 |
| VS3IFY | Verb aux. 3sg.ind.fut. | avrà | V30111302 |
| VP1IF | Verb main 1pl.ind.fut. | ameremo | V10211301 |
| VP2IF | Verb main 2pl.ind.fut. | amerete | V20211301 |
| VP3IF | Verb main 3pl.ind.fut. | ameranno | V30211301 |
| VS1IF | Verb main 1sg.ind.fut. | amerò | V10111301 |
| VS2IF | Verb main 2sg.ind.fut. | amerai | V20111301 |
| VS3IF | Verb main 3sg.ind.fut. | amerà | V30111301 |
| VP1CIY | Verb aux. 1pl.subj.impf. | avessimo | V10212202 |
| VP2CIY | Verb aux. 2pl.subj.impf. | aveste | V20212202 |
| VP3CIY | Verb aux. 3pl.subj.impf. | avessero | V30212202 |
| VS1CIY | Verb aux. 1sg.subj.impf. | avessi | V10112202 |
| VS2CIY | Verb aux. 2sg.subj.impf. | avessi | V20112202 |
| VS3CIY | Verb aux. 3sg.subj.impf. | avesse | V30112202 |
| VP1CI | Verb main 1pl.subj.impf. | amassimo | V10212201 |
| VP2CI | Verb main 2pl.subj.impf. | amaste | V20212201 |
| VP3CI | Verb main 3pl.subj.impf. | amassero | V30212201 |
| VS1CI | Verb main 1sg.subj.impf. | amassi | V10112201 |
| VS2CI | Verb main 2sg.subj.impf. | amassi | V20112201 |
| VS3CI | Verb main 3sg.subj.impf. | amasse | V30112201 |

## Part 6

| *Code* | *Description of word category* | *Example(s)* | *Intermediate Tag* |
|---|---|---|---|
| VP1IIY | Verb aux. 1pl.ind.impf. | avevamo | V10211202 |
| VP2IIY | Verb aux. 2pl.ind.impf. | avevate | V20211202 |
| VP3IIY | Verb aux. 3pl.ind.impf. | avevano | V30211202 |
| VS1IIY | Verb aux. 1sg.ind.impf. | avevo | V10111202 |
| VS2IIY | Verb aux. 2sg.ind.impf. | avevi | V20111202 |
| VS3IIY | Verb aux. 3sg.ind.impf. | aveva | V30111202 |
| VP1II | Verb main 1pl.ind.impf. | amavamo | V10211201 |
| VP2II | Verb main 2pl.ind.impf. | amavate | V20211201 |
| VP3II | Verb main 3pl.ind.impf. | amavano | V30211201 |
| VS1II | Verb main 1sg.ind.impf. | amavo | V10111201 |
| VS2II | Verb main 2sg.ind.impf. | amavi | V20111201 |
| VS3II | Verb main 3sg.ind.impf. | amava | V30111201 |
| VP1CPY | Verb aux. 1pl.subj.pres. | abbiamo | V10212102 |
| VP2CPY | Verb aux. 2pl.subj.pres. | abbiate | V20212102 |
| VP3CPY | Verb aux. 3pl.subj.pres. | abbiano | V30212102 |
| VS1CPY | Verb aux. 1sg.subj.pres. | abbia | V10112102 |
| VS2CPY | Verb aux. 2sg.subj.pres. | abbia | V20112102 |
| VS3CPY | Verb aux. 3sg.subj.pres. | abbia | V30112102 |
| VP1CP | Verb main 1pl.subj.pres. | amiamo | V10212101 |
| VP2CP | Verb main 2pl.subj.pres. | amiate | V20212101 |
| VP3CP | Verb main 3pl.subj.pres. | amino | V30212101 |
| VS1CP | Verb main 1sg.subj.pres. | ami | V10112101 |
| VS2CP | Verb main 2sg.subj.pres. | ami | V20112101 |
| VS3CP | Verb main 3sg.subj.pres. | ami | V30112101 |
| VP1DPY | Verb aux. 1pl.cond.pres. | avremmo | V10214102 |
| VP2DPY | Verb aux. 2pl.cond.pres. | avreste | V20214102 |
| VP3DPY | Verb aux. 3pl.cond.pres. | avrebbero | V30214102 |
| VS1DPY | Verb aux. 1sg.cond.pres. | avrei | V10114102 |
| VS2DPY | Verb aux. 2sg.cond.pres. | avresti | V20114102 |
| VS3DPY | Verb aux. 3sg.cond.pres. | avrebbe | V30114102 |
| VP1DP | Verb main 1pl.cond.pres. | ameremmo | V10214101 |
| VP2DP | Verb main 2pl.cond.pres. | amereste | V20214101 |
| VP3DP | Verb main 3pl.cond.pres. | amerebbero | V30214101 |
| VS1DP | Verb main 1sg.cond.pres. | amerei | V10114101 |
| VS2DP | Verb main 2sg.cond.pres. | ameresti | V20114101 |
| VS3DP | Verb main 3sg.cond.pres. | amerebbe | V30114101 |
| VP1IPY | Verb aux. 1pl.ind.pres. | abbiamo | V10211102 |
| VP2IPY | Verb aux. 2pl.ind.pres. | avete | V20211102 |
| VP3IPY | Verb aux. 3pl.ind.pres. | hanno | V30211102 |
| VS1IPY | Verb aux. 1sg.ind.pres. | ho | V10111102 |
| VS2IPY | Verb aux. 2sg.ind.pres. | hai | V20111102 |
| VS3IPY | Verb aux. 3sg.ind.pres. | ha | V30111102 |

## Part 7

| *Code* | *Description of word category* | *Example(s)* | *Intermediate Tag* |
|---|---|---|---|
| VP1IP | Verb main 1pl.ind.pres. | amiamo | V10211101 |
| VP2IP | Verb main 2pl.ind.pres. | amate | V20211101 |
| VP3IP | Verb main 3pl.ind.pres. | amano | V30211101 |
| VS1IP | Verb main 1sg.ind.pres. | amo | V10111101 |
| VS2IP | Verb main 2sg.ind.pres. | ami | V20111101 |
| VS3IP | Verb main 3sg.ind.pres. | ama | V30111101 |
| VP2MPY | Verb aux. 2pl.imp.pres. | abbiate | V20213102 |
| VS2MPY | Verb aux. 2sg.imp.pres. | abbi | V20113102 |
| VP2MP | Verb main 2pl.imp.pres. | amate | V20213101 |
| VS2MP | Verb main 2sg.imp.pres. | ama | V20113101 |
| VNPPPY | Verb aux. comm.pl.part.pres. | aventi | V04226102 |
| VNSPPY | Verb aux. comm.sg.part.pres. | avente | V04126102 |
| VNPPP | Verb main comm.pl.part.pres. | amanti | V04226101 |
| VNSPP | Verb main comm.sg.part.pres. | amante | V04126101 |
| VP1IRY | Verb aux. 1pl.ind.past | avemmo | V10211402 |
| VP2IRY | Verb aux. 2pl.ind.past | aveste | V20211402 |
| VP3IRY | Verb aux. 3pl.ind.past | ebbe | V30211402 |
| VS1IRY | Verb aux. 1sg.ind.past | ebbi | V10111402 |
| VS2IRY | Verb aux. 2sg.ind.past | avesti | V20111402 |
| VS3IRY | Verb aux. 3sg.ind.past | ebbe | V30111402 |
| VP1IR | Verb main 1pl.ind.past | amammo | V10211401 |
| VP2IR | Verb main 2pl.ind.past | amaste | V20211401 |
| VP3IR | Verb main 3pl.ind.past | amarono | V30211401 |
| VS1IR | Verb main 1sg.ind.past | amai | V10111401 |
| VS2IR | Verb main 2sg.ind.past | amasti | V20111401 |
| VS3IR | Verb main 3sg.ind.past | amò | V30111401 |
| VFPPRY | Verb aux. femm.pl.part.past | avute | V02226402 |
| VFSPRY | Verb aux. femm.sg.part.past | avuta | V02126402 |
| VMPPRY | Verb aux. masc.pl.part.past | avuti | V01226402 |
| VMSPRY | Verb aux. masc.sg.part.past | avuto | V01126402 |
| VFPPR | Verb main femm.pl.part.past | amate | V02226401 |
| VFSPR | Verb main femm.sg.part.past | amata | V02126401 |
| VMPPR | Verb main masc.pl.part.past | amati | V01226401 |
| VMSPR | Verb main masc.sg.part.past | amato | V01126401 |