# Corpora and Corpus-Based Morpho-Lexical Processing

Dan Tufiş, Ana-Maria Barbu, Vasile Pătraşcu, Georgiana Rotariu, Camelia Popescu

## 1. Corpora and types of corpora

The term *corpus* as used here refers to a collection of spoken or written texts encoded into a specific machine readable format. Corpora are used in language engineering to gather both qualitative and quantitative real language evidence. Qualitative evidence consists of examples which can be used for the construction of computational lexicons, grammars, and multi-lingual lexicons and term banks, for lexicography, etc. Quantitative information consists of statistics indicating frequent or characteristic uses of language. These statistics can also be used to guide preference-based parsers, assist in lexicography, determine translation equivalents, etc. In addition, statistics can be used to drive morphological taggers, POS taggers, alignment programs, sense taggers, etc. Common operations on corpora for the purposes of language engineering include extraction of sub-corpora; sophisticated search and retrieval, including collocation extraction, concordance generation, generation of lists of linguistic elements, etc.; and the determination of statistics such as frequency information, averages, mutual information scores, etc. We do not address corpora intended for other applications, such as stylistic studies, socio-linguistics, historical studies, information retrieval, etc., although these uses are not excluded a priori (in fact, many of the features required for these applications may be the same as those needed for language engineering). The encoding format should be standardised and homogeneous both for reasons concerning interchange and open-ended retrieval tasks [1]. Treating a restricted domain enables development of a tighter standard than that of the TEI, by providing specific encoding solutions rather than general or multiple ones, and, most importantly, by providing standards for elements particularly important in that domain (e.g., linguistic annotation). The texts are selected according to explicit criteria, according to the main purpose the corpus is supposed to serve (lexicographic tasks; terminological data extraction/acquisition; contrastive/comparative studies on parallel texts; grammar induction; machine translation; etc.). Based on several linguistic classification criteria, Sinclair defines in [1] a corpus typology (see also [2] in this volume). From that typology, we are here concerned with the special and parallel type corpora.

A *special corpus* is composed of homogeneous texts, which are not necessarily representative for a language, that is the corpus is not supposed to provide covering evidence for all the registers of the language in case. Of course, since these texts belong to a given language, they will display a set of grammatical and lexical features proper to the language in question. But, on the other hand, the texts contained in special corpora as opposed to the reference corpora [1] also exhibit features (mainly at the level of the lexicon) which may be idiosyncratic with respect to the language (rare words, specific collocation patterns, etc.).

Whereas a reference corpus is required to be as large as possible and based on complete and homogeneous texts, covering all the linguistic registers and dialectal variations, a special one is allowed to have smaller dimensions. It may contain more specialized (and even fragmentary) texts - for instance, texts belonging to an author, or a period - as well as texts obtained under artificial and experimental conditions (e.g., the ones resulted from investigations).

The second corpus type relevant for the work reported here is the *parallel corpus*. It is a collection of original texts, each of which is translated into at least one foreign language. The most simple case is the one where only two languages are involved. In this case, one language translates the other. However, parallel corpora may usually contain many languages. A parallel corpus is inherently a special one, since even if the source texts, taken in isolation, would make a reference corpus, by translation into a foreign language, the representativity with respect to the new language is generally not preserved. Thanks to the alignment between the original and its translation(s), parallel corpora open insightful perspectives into the nature of the translation. Also, they may help to devise more and more appropriate tools of translation. For instance, translation systems that employ probabilistic machines might be increasingly used in working with such corpora. There is a large number of applications a corpus may serve. In [3] we made a plea for computational corpora and discussed some of the methodological benefits (and direct applications) of corpus linguistics as contrasted to the classical approach of introspective theoretical linguistics.

In the context of MULTEXT-EAST Copernicus project (http://www.lpl.univ-aix.fr/MTE and http://nl.ijs.si/ME) it was developed a parallel corpus consisting of 6 integral translations of George Orwell's "1984": besides the original English version, the corpus contains translations in Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene. This corpus is consistently SGML encoded over the seven languages (CES conformant [4] up to level 3), sentence aligned and morpho-syntactically annotated. MULTEXT-EAST also produced a second multilingual and balanced corpus (but not parallel) consisting of original novels and newspaper articles for each language.

A second parallel corpus was developed within the TELRI Copernicus Concerted Action (Trans European Language Resource Infrastructure-http://www.ids-mannheim.de/TELRI) and consists of integral translations of Plato's "Republic" (besides the original Greek version, the sub-corpus contains translations in Bulgarian, Czech, English, German, Lithuanian, Polish, Romanian, Slovak and Slovene).

It is worth mentioning that both "1984" and "Republic" are available in many other languages but at the time of this writing they had not been aligned (in terms of text encoding) to the conventions used for the languages mentioned above. The encoding of Romanian version of "Republic" is described in [5].

There were several reasons for the choice of the source texts ("1984" and Republic") based on which the parallel corpora were built, but the most important were the availability of the translations in all the languages of the partners involved and copyright permissions. The two corpora are complementary as far as the linguistic registers and (human) translation techniques are concerned (whereas "1984" was translated more literary -in the spirit of the book, "Republic" was translated in a rather technical way, providing support for Sinclair's hypothesis on "translation protected elements").

In order to develop reusable resources, it has been essential to establish standardized methods and specifications for these resources. For corpus encoding, MULTEXT-EAST and to some extent TELRI, adopted the SGML-based CES (Corpus Encoding Specifications) schema [4], which on its turn is based on the TEI recommendations [6,7]. While corpus-encoding schema posed no significant problems in accommodating the several language specific renderings of the paralleled texts, harmonising the lexicon encoding was a serious task. Based on the specifications developed in the EAGLES [8] and their extension by the MULTEXT Project [9] to six western European languages (English, French, Dutch, Italian, German, Spanish), MULTEXT-EAST Project further extended these specifications to cover new languages (Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene) with different characteristics. Because these language families include many features and properties not found in western European languages, such as heavy inflection and agglutination, adapting the specifications initially developed for western European languages to these languages posed many interesting and difficult problems and demanded substantial assessment and modification of the pre-existing specifications. The work carried out in MULTEXT-EAST has thus broadened the base and contributed significantly to defining a general mechanism for lexical specification (<URL: http://www.lpl.univ-aix.fr/pub/multext/docs/ME1.1.tex> and<URL: http://www.ijs.si/ME/docs/>). A significant benefit from conforming to corpus and lexicon specifications originating in EAGLES and MULTEXT, was an easy adaptation or extension of corpus-annotation tools developed initially within the MULTEXT project.

Since presentations from a multilingual point of view of MULTEXT-EAST and TELRI corpora are given elsewhere [10,11,12], in the following, we will dwell on the Romanian component.

## 2. Word-form dictionary and morpho-syntactic descriptions

For corpus morpho-lexical processing purposes (fast text lemmatisation, language model constructing, automatic tagging, etc.) the MULTEXT-EAST consortium developed several language specific word-form dictionaries covering at least the words appearing in the corpus. A dictionary entry has the following structure:

word-form <TAB> lemma <TAB> MSD <TAB> comments

where word-form represents an inflected form of the lemma, characterised by a combination of feature values encoded by MSD code (Morpho Syntactic Description); the forth column, comments, which is optional, is currently ignored and may contain either comments or information processable by other tools. The morpho-syntactic descriptions are provided as strings, using a linear encoding. In this notation, the position in a string of characters corresponds to an attribute, and specific characters in each position indicate the value for the corresponding attribute. That is, the positions in a string of characters are numbered 0, 1, 2, etc., and are used in the following way:

- the character at position 0 encodes part-of-speech;
- each character at position 1, 2,...,n, encodes the value of one attribute (person, gender, number, etc.), using the one-character code;
- if an attribute does not apply, the corresponding position in the string contains the special marker '-' (hyphen).

By convention, trailing hyphens are not included in the lexical MSDs. Such specifications provide a simple and relatively compact encoding, and are in intention similar to feature-structure encoding used in unification-based grammar formalisms.

When the word-form is the very lemma, then the equal sign is written in the lemma-field of the entry ('=').

The attributes and mainly the values of the attributes were chosen considering only word-level encoding. This is why, some values described in grammar text-books, but assuming compounding (such as compound tenses), were not considered in the MULTEXT-EAST encoding.

Concerning the "applicability" of a certain feature/value one has to distinguish at least 2 cases:

1) a specific attribute is not at all applicable to a given language (such as it is the case in Romanian for Referent_type or Owner_gender attributes of the pronouns);
2) a specific combination of feature-values applying to a word-form belonging to a given category makes some attributes of the category irrelevant in the case considered (for instance tense for some non-predicative verbs, or gender for personal pronouns, first and second persons).

These two cases were encoded by using a special attribute-value, namely '-'.

A special case was raised by *token-indeterminable* values. This is the case when a given word form cannot be assigned, without considering the context, a unique value for a given attribute, but a subset or even all of the permissible values of that attribute.

A typical example of such an attribute is the grammatical case. For the indefinite forms of nouns and adjectives nothing precise can be said about case unless preceding word(s) are considered. In such a situation one could only say that all possible values of a considered attribute (in this example, the case) are applicable. Since the MULTEXT-EAST encoding schema does not use a special value "any" (which in some encoding proposals is denoted by a dot '.') two possible solutions could be envisaged:

a) to explicitly encode in the word-form lexicon all the combinations resulting from expanding the would be '*any*' values. This solution is 'clean' but introduces a large degree of redundancy and ambiguity. Consider the 'dotted' encoding of the word '*munte'*

munte munte Ncms.n.

Expanding the two dots (see tables on the next pages) this entry should be replaced by the following 10 fully expanded entries:

| | | | | | |
|---|---|---|---|---|---|
| munte | munte | Ncmsnnn | munte | munte | Ncmsnny |
| munte | munte | Ncmsgnn | munte | munte | Ncmsgny |
| munte | munte | Ncmsdnn | munte | munte | Ncmsdny |
| munte | munte | Ncmsann | munte | munte | Ncmsany |
| munte | munte | Ncmsvnn | munte | munte | Ncmsvny |

In the initial phase of the project, we took this approach, but after a while, due to computational reasons (both in terms of space requirements and lookup performances) we opted for a second solution, that is:

b) to further overload the significance of the special value '-' with the '*any*' interpretation.

A very simple script checked the word-forms of each lemma for attributes instantiated with all the possible values (as in the example above). Then, after indexing these attributes for later recovering, the corresponding positions in the MSD were replaced by '-'. Based on similar motivations, we decided to use two special cases (*direct* and *oblique*) to deal with the nominative-accusative and genitive-dative syncretism, and to eliminate neuter gender from the lexicon encoding. With duplicates eliminated, the word-form lexicon size decreased more than 4 times (from 1895837 to 440363). Eliminating neuter gender from the lexicon encoding needs a few explanations. Traditionally, grammar books distinguish in Romanian three genders: masculine, feminine and neuter. However there are few reasons - if any- not to get rid of neuter value and consider a simpler dual gender system. From the inflectional point of view, neuter nouns/adjectives behave in singular as masculine nouns/adjectives and in plural as feminine ones. Since there is no intrinsic semantic feature specific to neuter nouns (inanimacy is by no means specific to neuter nouns, existing plenty of feminine and masculine nouns denoting inanimate things) preserving the distinction masculine/feminine/neuter creates more problems than it solves [13]. Due to the agreement rules, adjectives can take masculine, feminine and neuter gender. At the lexicon level, this would require to add about 33% more entries in case of adjectives. At the lookup level, considering only gender, any adjective would be two way ambiguous (masculine/neuter in singular and feminine/neuter in plural). However, it is worth mentioning that if needed, the neuter nouns or adjectives can be easily identified: those nouns/adjectives that are tagged with masculine gender in singular and with feminine gender in plural are what the traditional Romanian linguistics calls neuter nouns/adjectives.

This position has recently found adherents among theoretical linguists as well. For instance, in [14] neuter nouns are considered to be underspecified for gender in their lexical entries, having default rules assigning masculine gender for occurrences in singular and feminine gender for occurrences in plural.

The tables on the next pages present the MSD encoding for Romanian in the following format:

**Category (category code)**

| Attribute Position | Attribute | Value | Example |
|---|---|---|---|
| | | | |

The category code is an uppercase letter (N, V, A, P, D, T, R, S, C, M, Q, Y, I, X) identifying one of the 14 parts of speech considered in the MULTEXT-EAST encoding schema. Any MSD code will begin with one of these 14 letters. The **Attribute Position** column specifies the position in the linear MSD encoding of the attribute, the name of which is given in the **Attribute** column. The **Value** column contains the allowable values of the current attribute. Their codes, given between parentheses, may appear in an MSD headed by the appropriate category code, at the position specified by the **Attribute Position** column**.** This linear encoding is a relatively efficient and compact way to represent the flat attribute-value matrices.

**1. Noun (N)**

| Attribute Position | Attribute | Value | Example |
|---|---|---|---|
| 1 | **Type** | *common (c)* | carte |
| | | *proper (p)* | Ion |
| 2 | **Gender** | *masculine (m)* | băiatul |
| | | *feminine (f)* | casa |
| 3 | **Number** | *singular (s)* | fată |
| | | *plural (p)* | fete |
| 4 | **Case** | *direct(r)* | omul |
| | | *oblique (o)* | omului |
| | | *vocative (v)* | omule |
| 5 | **Definiteness** | *yes (y)* | omul |
| | | *no (n)* | om |
| 6 | **Clitic** | *no (n)* | sora$_n$ |
| | | *yes (y)* | soru$_y$-mea |

**2. Verb (V)**

| Attribute Position | Attribute | Value | Example |
|---|---|---|---|
| 1 | **Type** | *main (m)* | a vedea |
| | | *auxiliary (a)* | a avea, a fi, a voi |
| | | *modal (o)* | a putea, a trebui |
| | | *copulative (c)* | a fi, a deveni |
| 2 | **VForm** | *indicative(i)* | vine |
| | | *subjunctive (s)* | vină |
| | | *imperative (m)* | vino |
| | | *infinitive (n)* | veni |
| | | *participle (p)* | venit |
| | | *gerund (g)* | venind |
| 3 | **Tense** | *present (p)* | văd |
| | | *imperfect (i)* | vedeam |
| | | *past (s)* | văzui |
| | | *pluperfect (l)* | văzusem |
| 4 | **Person** | *first (1)* | văd |
| | | *second (2)* | vezi |
| | | *third (3)* | vede |
| 5 | **Number** | *singular (s)*<br>*plural (p)* | bătut<br>bătuşi |
| 6 | **Gender** | *masculine(m)* | bătut |
| | | *feminine(f)* | bătută |
| 10 | **Clitic** | *yes (y)* | ridicându$_y$-l$_y$ |
| | | *no (n)* | am$_n$ ridicat$_n$ |

### 3. Adjective (A)

| Attribute Position | Attribute | Value | Example |
|---|---|---|---|
| 1 | **Type** | *qualificative (f)* | frumos |
| 2 | **Degree** | *positive (p)* | frumos |
| | | *comparative (c)* | ulterior |
| | | *superlative (s)* | extrem |
| 3 | **Gender** | *masculine (m)* | bun |
| | | *feminine (f)* | bună |
| 4 | **Number** | *singular (s)* | bun |
| | | *plural (p)* | buni |
| 5 | **Case** | *direct(r)* | bunul |
| | | *oblique (o)* | bunului |
| | | *vocative (v)* | bunule |
| 6 | **Definiteness** | *yes (y)* | bunul |
| | | *no (n)* | bun |
| 7 | **Clitic** | *yes (y)* | bunu- |
| | | *no (n)* | bun |

### 4. Pronoun (P)

| Attribute Position | Attribute | Value | Example |
|---|---|---|---|
| 1 | **Type** | *demonstrative(d)* | acesta |
| | | *indefinite (i)* | oricine |
| | | *possessive(s)* | (al) meu |
| | | *int_rel (w)* | ce |
| | | *personal(p)* | eu |
| | | *reflexive(x)* | se |
| | | *negative (z)* | nimeni |
| | | *emphatic (h)* | însumi |
| 2 | **Person** | *first (1)* | eu |
| | | *second (2)* | tu |
| | | *third (3)* | el |
| 3 | **Gender** | *masculine (m)* | el |
| | | *feminine (f)* | ea |
| 4 | **Number** | *singular (s)* | eu |
| | | *plural (p)* | noi |
| 5 | **Case** | *nominative (n)* | el |
| | | *genitive (g)* | (al) lui |
| | | *dative (d)* | lui |
| | | *accusative (a)* | (pe) el |
| | | *vocative (v)* | tu, voi! |
| | | *direct(r)* | acesta |
| | | *oblique (o)* | acestuia |
| 6 | **OwnerNumber** | *singular (s)* | meu |
| | | *plural (p)* | nostru |
| 8 | **Clitic** | *yes (y)* | bunul |
| | | *no (n)* | bun |
| 14 | **Pron-Form** | *strong (s)* | lui |
| | | *weak (w)* | îi, i- |

**5. Determiner (D)**

| Attribute Position | Attribute | Value | Example |
|---|---|---|---|
| 1 | Type | *demonstrative (d)* | acest |
| | | *indefinite (i)* | orice |
| | | *possessive (s)* | meu |
| | | *int_rel (w)* | ce |
| | | *negative (z)* | nici un |
| | | *emphatic (h)* | însuşi |
| 2 | Person | *first (i)* | meu |
| | | *second (2)* | tău |
| | | *third (3)* | său |
| 3 | Gender | *masculine (m)* | meu |
| | | *feminine (f)* | mea |
| 4 | Number | *singular (s)* | meu |
| | | *plural (p)* | mei |
| 5 | Case | *direct(r)* | aceasta |
| | | *oblique (o)* | acestei |
| 6 | | *singular (s)* | meu |
| | | *plural (p)* | nostru |
| 8 | Clitic | *no (n)* | mama$_n$ *mea*$_n$ |
| | | *yes (y)* | maică$_y$- *mea*$_y$ |
| 9 | Modific Type | *prenominal (e)* | acest |
| | | *postnominal (o)* | acesta |

**6. Article (T)**

| Attribute Position | Attribute | Value | Example |
|---|---|---|---|
| 1 | Type | *definite (f)* | lui |
| | | *indefinite (i)* | un |
| | | *possessive (s)* | al |
| | | *demonstrative* | cel |
| 2 | Gender | *masculine (m)* | un |
| | | *feminine (f)* | o |
| 3 | Number | *singular (s)* | un |
| | | *plural (p)* | nişte |
| 4 | Case | *direct(r)* | cel |
| | | *oblique (o)* | celui |
| 5 | Clitic | *yes (y)* | -un |
| | | *no(no)* | un |

**7. Adverb (R)**

| Attribute Position | Attribute | Value | Example |
|---|---|---|---|
| 1 | Type | *general (g)* | bine, acolo |
| | | *particle (p)* | mai, cam |
| | | *negative (z)* | nicăieri |
| | | *modifier (m)* | fireşte, poate |
| | | *int_rel(w)* | cum |
| 2 | Degree | *positive (p)* | bine |
| | | *comparative (c)* | ulterior |
| | | *superlative (s)* | extrem |
| 3 | Clitic | *yes (y)* | -ncet |
| | | *no (n)* | încet |

### 8. Adposition (S)

| Attribute Position | Attribute | Value | Example |
|---|---|---|---|
| 1 | **Type** | *preposition (p)* | la, pe, în |
| 2 | **Formation** | *simple (s)* | la, pe, în |
| | | *compound (c)* | de la |
| 3 | **Case** | *genitive (g)* | înaintea |
| | | *dative (d)* | datorită |
| | | *accusative (a)* | la |
| 4 | **Clitic** | *yes (y)* | într- |
| | | *no (n)* | întru |

### 9. Conjunction (C)

| Attribute Position | Attribute | Value | Example |
|---|---|---|---|
| 1 | **Type** | *coordinating (c)* | şi, dar |
| | | *subordinating (s)* | că, dacă |
| 2 | **Formation** | *simple (s)* | deoarece |
| | | *compound (c)* | de vreme ce |
| 3 | **Coord Type** | *simple (s)* | şi, deoarece |
| | | *repetit (r)* | fie...fie... |
| | | *correlat (c)* | atât...cât şi |
| 4 | **Sub-Type** | *negative (z)* | nici |
| | | *positive (p)* | dar |
| 5 | **Clitic** | *no (n)* | cã$_n$ aşa |
| | | *yes (y)* | c-$_y$-aşa |

### 10. Numeral (M)

| Attribute Position | Attribute | Value | Example |
|---|---|---|---|
| 1 | **Type** | *cardinal (c)* | trei |
| | | *ordinal (o)* | (al) treilea |
| | | *fractal (f)* | treime |
| | | *multiple (m)* | întreit |
| | | *collect(l)* | tustrei |
| 2 | **Gender** | *masculine (m)* | doi, primul |
| | | *feminine (f)* | două, prima |
| 3 | **Number** | *singular (s)* | primul |
| | | *plural (p)* | primii |
| 4 | **Case** | *direct(r)* | primul |
| | | *oblique (o)* | primului |
| 5 | **Form** | *digit (d)* | 1690 |
| | | *letter (l)* | unsprezece |
| | | *both (b)* | 5 mii |
| | | *roman (r)* | XIV |
| 6 | **Definiteness** | *yes (y)* | primul |
| | | *no (n)* | prim |
| 7 | **Clitic** | *yes (y)* | unu- |
| | | *no (n)* | doi |

### 11. Abbreviation (Y)

| Attribute Position | Attribute | Value | Example |
|---|---|---|---|
| 1 | SyntacticType | *nominal(n)* | dna (doamna) |
| | | *verbal (v)* | v. (vezi) |
| | | *adjectival (a)* | ant. (anterior) |
| | | *adverbial(r)* | f. (foarte) |
| | | *pronominal (p)* | d-voastră (dumneavoastră) |
| 2 | Gender | *masculine (m)* | dlui (domnului) |
| | | *feminine (f)* | dna (doamna) |
| 3 | Number | *singular (s)* | dna (doamna) |
| | | *plural (p)* | dnele (doamnele) |
| 4 | Case | *direct (r)* | dna (doamna) |
| | | *oblique (o)* | dnei (doamnei) |
| 5 | Definiteness | *yes (y)* | dnele (doamnele) |
| | | *no (n)* | dne (doamne) |

### 12. Particle (Q)

| Attribute Position | Attribute | Value | Example |
|---|---|---|---|
| 1 | Type | *negation(z)* | nu, n- |
| | | *infinitive (n)* | a |
| | | *subjunctive (s)* | să |
| | | *aspect (a)* | fi |
| | | *future (f)* | o |
| 3 | Clitic | *yes (y)* | nV-am |
| | | *no (n)* | nu am |

## 13. Interjections (I)

In Romanian, as in most other languages, no further refinement of the interjections has been considered.

## 14. Residual (X)

As with the interjections, no further refinement of residual has been considered. This class, inherited from EAGLES standard proposal, was meant to cover textual units such as mathematical formulas, alphanumeric strings, and generally everything that cannot go easily into another category.
The table below provides a comparative overview of the number of attributes that were considered in each of the MULTEXT-EAST languages for each category (a '-' in a cell signifies that the corresponding POS is not relevant for the considered language).

| POS | Rom | Bul | Cz | Slo | Est | Hun |
|---|---|---|---|---|---|---|
| Noun | 6 | 5 | 5 | 5 | 3 | 7 |
| Verb | 7 | 8 | 10 | 8 | 8 | 5 |
| Adjective | 7 | 3 | 7 | 5 | 3 | 8 |
| Pronoun | 8 | 8 | 12 | 10 | 4 | 7 |
| Adverb | 3 | 1 | 2 | 2 | 0 | 4 |
| Adposition | 4 | 1 | 3 | 3 | 1 | 1 |
| Conjunction | 5 | 2 | 3 | 2 | 1 | 3 |
| Numeral | 7 | 5 | 7 | 5 | 4 | 7 |
| Interjection | 0 | 1 | 0 | 0 | 0 | 1 |
| Residual | 0 | 0 | 0 | 0 | 0 | 0 |
| Abbreviation | 5 | 0 | 0 | 0 | 3 | 0 |
| Particle | 2 | 2 | 0 | 0 | - | - |
| Determiner | 8 | - | - | - | - | - |
| Article | 5 | - | - | - | - | 1 |

## 3. A Statistical account of the lexicon

Once the harmonised set of morpho-syntactic specifications for the six MULTEXT-EAST languages was developed, lexicons incorporating these specifications were created for each language. The Romanian lexicon was created based on a 35.000-lemma lexicon by means of our EGLU natural language processing platform [15]. Since several words in the corpus were not in the EGLU lexicon, most of them were manually lemmatised, introduced in the unification-based lexicon and later on expanded to the full paradigms of every new lemma. The Romanian word-form lexicon is actually made of two parts: the main one contains only words attested by the Explanatory Dictionary of Romanian (DEX); all the other words, appearing in the corpus were entered an auxiliary lexicon. The auxiliary lexicon contains, among other things, proper names, technical terms and the weird (made-up) words from Orwell's"1984" (*newspeak* dialect).

The table below provides information on the data content of the main dictionary that is used for the corpus analysis. When interpreting the figures, one has to consider the treatment of syncretic forms as previously discussed. The *any* interpretation of the '-' attribute value was not considered in counting these figures since the expansion of the *token-indeterminable* values is external to the lexicon.

Table 1: Lexicon overview

| **Entries** | **Wordforms** | **Lemmas** | **=** | **MSDs** | **AMB_POS** | **AMB_MSD** |
|---|---|---|---|---|---|---|
| 440363 | 347960 | 33259 | 35421 | 674 | 90 | 981 |

The first column (**Entries**) provides the number of dictionary entries, that is triplets:

<word-form lemma MSD>

The distribution of the entries over the parts of speech is shown in Table 1.1. As one would expect, given the large paradigms of verbs, the largest number of entries are verbal (192775). The percentage of noun and adjective word-forms is also very high.

Table 1.1: The distribution of the entries over the parts of speech

| **POS** | **N** | **V** | **A** | **P** | **D** | **M** | **R** | **T** | **S** | **C** | **Q** | **I** | **Y** | **X** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Number of entries** | 124135 | 192843 | 119420 | 347 | 231 | 693 | 1357 | 46 | 176 | 81 | 10 | 196 | 777 | 51 |

The second column (**Wordforms**) in Table 1 gives the number of distinct word-forms appearing in the lexicon, irrespective of their lemma and MSD. For instance, the following entries:

| | | |
|---|---|---|
| vin | = | Ncms-n |
| vin | veni | Vmip1s |
| vin | veni | Vmip3p |
| vin | veni | Vmsp1s |

will contribute the Wordform column with only one item (vin). The word-forms distribution over the parts of speech is shown in Table 1.2.

Table 1.2: The distribution of the word-forms over the parts of speech

| **POS** | **N** | **V** | **A** | **P** | **D** | **M** | **R** | **T** | **S** | **C** | **Q** | **I** | **Y** | **X** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Number of wordforms** | 113285 | 142438 | 111274 | 320 | 224 | 615 | 1336 | 45 | 176 | 77 | 10 | 196 | 665 | 51 |

Summing up the numbers in Table 1.2 gives 370712 which is 22752 more than the figure shown in the **Wordform** column in Table 1. Since several identical word-forms (which were counted only once in the Wordform column of Table 1) could differ by the MSD, by lemma or by both, counting by taking into account these differences makes the noticed difference. Put in other words, figures in Table 1.2 considered the homography[1] (more precisely word-form homography), while those in Table 1 did not. The **Lemma** column gives the number of distinct lemmas in the lexicon, that is eliminating any duplications that might appear due to the lemma homography. The lemmas distribution over the POS is given in Table 1.3. Again, by summing up the figures in this table, one gets a higher figure than in Table 1. This is also due to homography at the lemma level (as is *duce*, *a*, *dar*, *și* etc.). However, since lemma homography is much less frequent than word-form homography, the difference is in this case much smaller (2098). As one can observe from Table 1.3, the highest number of lemmas are nouns (18086) and adjectives (11052).

Table 1.3. The distribution of the lemmas over the parts of speech

| POS | N | V | A | P | D | M | R | T | S | C | Q | I | Y | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of lemmas | 18086 | 4247 | 11052 | 80 | 36 | 115 | 1257 | 7 | 145 | 67 | 7 | 196 | 11 | 51 |

The "=" field in Table 1 provides the number of entries which are themselves lemmas (i.e., have "=" in the lemma field of their entry), **without eliminating the lemma homographs**. Therefore, the difference between the **Lemma** and the "=" fields gives an estimation of the number of homographic lemmas[2].

The "**MSD**s" field gives the total number of distinct MSDs used in the encoding of the lexicon stock. The distribution of the MSDs over parts of speech is given in Table 1.4. As one can see, the highest numbers of MSDs are defined for Pronouns (138), Verbs (135) and Determiners (114).

Table 1.4. MSD distribution over the parts of speech

| POS | N | V | A | P | D | M | R | T | S | C | Q | I | Y | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of MSDs | 57 | 135 | 60 | 138 | 114 | 85 | 11 | 33 | 7 | 8 | 7 | 1 | 17 | 1 |

The last two columns in Table 1 (**AMB_POS** and **AMB_MSD**) provide information about the number of *ambiguity classification clusters*. An ambiguity classification cluster gives the multiple ways a homographic word-form can be classified by. If the classification is based on the part of speech (POS), the ambiguity classification cluster is called POS-ambiguity cluster (AMB_POS) and it is a list of parts of speech that can be associated with a given word-form. The AMB_POS for the word-form *vin* above is: (N, V). If the classification is based on MSD, the ambiguity classification cluster is called MSD-ambiguity cluster (AMB_MSD) and it is a list containing all MSDs associated with a given word-form. Considering again the word-form *vin,* its AMB_MSD is: (Ncms-n, Vmip1s, Vmip3p, Vmsp1s). It is worth mentioning that AMB_POS does not change if the '-' values in an MSD, which are interpretable as *any* values, are expanded. Obviously this is not the case for AMB_MSD. The number of ambiguity classes (defined based either on POS or MSD) is a key figure in estimating the space resources needed in constructing a statistical language model (such as HMM) useful for morpho-syntactic disambiguation of natural language.

---

[1] We consider here a restricted definition of homography. We distinguish *between word-form homography*, meaning identical word-forms tagged with different MSD, *POS homograph*, meaning identical wordforms but tagged with different POSes and *lemma homography*, meaning identical lemma-forms with different POSes (this is a restricted case of the second one). For instance the word *vin* is four-way *wordform homographic* (Ncms-n, Vmip1s, Vmip3p, Vmsp1s), two-way *POS homographic* (N, V) but it is not *lemma homographic* (the lemmas *veni* (V) and *vin* (N) are different). The word *dar* is two-way *wordform homographic*(Ncms-n, Ccssp), two-way *POS homographic* (N, C), and two-way *lemma homographic* (same lemma for both N and C interpretations).

[2] It would be the number of homographic lemmas just in case each such lemma would belong exactly

While the AMB_MSD column in Table 1 shows the number of distinct MSD-ambiguity classes, Table 1.5 shows the number of words which are unambiguous[3], two-way ambiguous, etc. The ambiguity is calculated considering word-form homography. The previously exemplified word-form *vin* has the ambiguity level 4.

Table 1.5. Word-forms grouped according to the MSD ambiguity level

| **Ambiguity level** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| **Number of words** | 284549 | 46635 | 10183 | 2570 | 2864 | 725 | 429 | 2 | 2 |

The *hapaxes* labeled with 8 MSDs (2) or 9 MSDs (2) are due to quite rare superposition of paradigmatic classes of different word-forms. For instance, the word *urâţi* is the 9-way wordform homographic word in our lexicon. It can be either a N, an A or a V. This wordform could have as lemma form one of the following: *urât* - ugly (N and A), *urâţi*/to make things ugly (V), *urî*/to hate(V). Similarly, the word *vii* can be either a N (vie/vineyard), an A (viu/alive), a verb (veni/to come) or a numeral written in the roman style (şapte/seven).

When considering lemma homography, the MSD-ambiguity figures are shown in Table 1.6.

Table 1.6. Lemmas grouped according to the MSD ambiguity level

| **Ambiguity level** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| **Number of lemmas** | 31218 | 1927 | 107 | 7 |

The following two tables in this section show the distribution of the MSDs over the MSD-ambiguity classes and of POSes over POS-ambiguity classes, respectively. Table 1.7 reads as follows: a given POS-msd (an MSD belonging to a given part of speech) appears in k MSD-ambiguity classes. For instance, verbal MSDs (V-msd) appear in 490 out of the total 981 MSD-ambiguity classes, while nominal MSDs (N-msd) appear in 445 MSD-ambiguity classes.

Table 1.7. MSD distribution over the MSD-ambiguity classes

| **POS-msd** | **N-msd** | **V-msd** | **A-msd** | **P-msd** | **D-msd** | **M-msd** | **R-msd** | **T-msd** | **S-msd** | **C-msd** | **Q-msd** | **I-msd** | **Y-msd** | **X-msd** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **No. of MSD-amb. classes** | 445 | 490 | 334 | 131 | 93 | 73 | 102 | 15 | 22 | 18 | 7 | 19 | 34 | 9 |

If one considers only the part of speech and the POS-ambiguity classes, the corresponding distribution is shown in Table 1.8. Out of the total number of 90 POS-ambiguity classes, 34 contain the verb (V), 30 contain the noun (N) and so on.

Table 1.8. POS distribution over the POS-ambiguity-classes

| **POS** | **N** | **V** | **A** | **P** | **D** | **M** | **R** | **T** | **S** | **C** | **Q** | **I** | **Y** | **X** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **No. of POS-amb. classes** | 30 | 34 | 18 | 28 | 12 | 14 | 28 | 10 | 16 | 11 | 6 | 11 | 16 | 9 |

Comparing the figures in Table 1.7 and Table 1.8 one may draw some interesting conclusions. For instance, considering the number of word-forms with more than 1

[3] Remember the previous discussion on the encoding of '*any*' values. For instance the entry: munte munte Ncms-n is not MSD-ambiguous unless the case value is considered. With *token-indeterminable* values fully expanded, the number of unambiguous words becomes much smaller. To be precise, the number of unambiguous tokens in "1984" would have been 22% instead of 57%

MSD (63411, i.e. 18,22% of the total number of word-forms) any such ambiguous word-form will have in almost 50% of the cases, one or more verbal readings; if considering only part of speech, in almost than 38% of the cases an ambiguous word-form would have a verb interpretation. Table 1.9 summarises this comparison for all parts of speech.

Table 1.9. POS percentual distribution over the POS and MSD-ambiguity-classes

| % | N | V | A | P | D | M | R | T | S | C | Q | I | Y | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MSD** | 45 | 50 | 34 | 13 | 9 | 7 | 10 | 2 | 2 | 2 | 0.7 | 2 | 3 | 0.9 |
| **POS** | 33 | 38 | 20 | 31 | 13 | 16 | 31 | 11 | 18 | 12 | 7 | 12 | 18 | 10 |

While these figures are hardly useful for disambiguation of a running text, they show that intra-categorial ambiguity is harder than discriminating among categories. Therefore, in designing the corpus tags (the set of codes an automatic tagger is supposed to work with) we concentrated on those attribute-values which would maximally discriminate among the MSDs belonging to the same part of speech. This way, out of the 674 possible morpho-syntactic descriptions of the word-forms in the lexicon, we defined a tagset containing 73 corpus tags. Since there is no generally accepted methodology on designing tagsets, except for the empirical trial and error approach, this tagset is considered just a working hypothesis towards a corpus-supported proposal for a Romanian tagset. The automating tagging literature reports excellent results in statistical disambiguation (accuracy above 95-96%) but none of such successful works considered highly inflectional languages as Romanian (or any other language in the MULTEXT-EAST project). Moreover, the number of tags we selected for the beginning is approximately four times higher than the number of tags used for English, the language with the most successful results in automatic tagging. That is why, we expect the final proposal for a tagset applying to Romanian to be the result of intensive experiments on large volumes of data. For these experiments, large training data is needed, prepared by semi-automatic annotation and manual validation. The next Section of the paper addresses this very issue.

## 4. A Statistical account of the Romanian sub-corpus of the multilingual corpus

From our collection of plain texts (about 15 million words) belonging to different genres and registers and more than 100 authors (poetry-Eminescu's complete poetic work, original Romanian fiction-14 novels, translations - Orwell's "1984" and Plato's "Republic", journalism, technical reports in computer science and linguistics, informal notes, electronic mails, etc.) we selected about 220.000 words for standardised corpus encoding with the purpose of research in statistical natural language processing (lexicographic statistics, automatic morpho-syntactic tagging, stochastic parsing, parallel texts alignment, statistics-based machine translation).

The European projects MULTEXT (LRE) and EAGLES (in particular, the EAGLES Text Representation subgroup), together with the Vassar/CNRS collaboration (supported by the U.S. National Science Foundation), have joined efforts to develop a Corpus Encoding Specification (CES) optimally suited for use in language engineering, which can serve as a widely accepted set of encoding standards for corpus-based work. The overall goal is the identification of a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation (marking of structural and linguistic information) as well as general architecture (so as to be maximally suited for use in a text database). It also provides encoding conventions for more extensive encoding and for linguistic annotation. The CES is an application of SGML (ISO-8879, Standard Generalized Markup Language), conformant to the TEI Guidelines for Electronic Text Encoding and Interchange of the Text Encoding

Initiative. [6]. The TEI Guidelines are expressly designed to be applicable across a broad range of applications and disciplines and therefore treat not only a vast array of textual phenomena, but are also designed with an eye toward the maximum of generality and flexibility. Most applications will use only those parts of the TEI that are required to meet their needs. The CES is such an application that have utilized the TEI modular DTD and the TEI customization mechanisms to select those pieces of the TEI that are appropriate for corpus encoding. The TEI is an ongoing project and for some areas it is not complete; and as a result, there are some areas of importance for corpus encoding that the TEI Guidelines do not cover. Therefore, developing the CES has involved not only selecting from, but also in some cases extending the TEI Guidelines to meet the specific needs of corpus-based work in language engineering. All results and specifications developed for the CES are fed back to the TEI as input for further revisions of the Guidelines. The current version of the CES is a first draft of the standard. It has not been widely implemented, and the intention is to continue to develop the CES on the basis of input and feedback from users after it is put to greater use. Therefore, this document will continue to evolve and should not be regarded as "final". All current CES documents and DTDs will continue to be available at the following site: http://www.cs.vassar.edu/CES/. Anyone actively implementing the standard should consult this site regularly.

The statistics given in the previous chapter referred to the dictionary. In running texts, as one would expect, they are different, reflecting the actual usage of the lexical stock. The procedure for evaluating the numbers in this Section involved automatic segmentation of the text MSD annotating the segmented text and manual disambiguation of approximately 220.000 lexical tokens.

Please note that a token (a lexical unit as identified by the segmenter) is not necessarily a word: one orthographic word may be split into several tokens - the Romanian "dă-mi-l" (give it to me) is split into 3 tokens") or several orthographic words may be combined into one token (the Romanian words "de la" are combined into one token "de_la"). In Figure 1 is a fragment of text from our corpora, Figure 2 shows the segmented text, Figure 3 gives the MSD ambiguously annotated text and Figure 4 presents the MSD disambiguated text.

The selected texts from the corpora were segmented by means of a tokenizer, part of the tool-set implemented within the MULTEXT project with the language resources we developed within the MULTEXT-EAST. The segmenter is a language independent and configurable processor used to tokenize an input text, given in one of the three possible formats: plain text (as in Figure 1), a normalized SGML form (nSGML) as output by another MULTEXT tool (MTSgmlQl) and a tabular format (also specific to MULTEXT processing chain). The output of the segmenter is a tokenized form of the input text, with paragraph and sentence boundary marked-up. Punctuation, lexical items, numbers and several alpha numeric sequences (such as dates and hours) are annotated with various tags out of a hierarchy class structured tagset. The language specific behavior of the segmenter is driven by several language resources (abbreviations, compounds, clitics, etc.). The general behavior of the segmenter (valid over several languages) can also be parametrised by means of external resources such as definition for space and punctuation, number orthography, sentence delimiters, etc.

Once the input text is tokenized (Figure 2), a dictionary look-up procedure can be invoked to assign each lexical token all its possible morpho-lexical interpretations (Figure 3). This procedure was incorporated into a special XEMACS mode (mtems-mode, due to T. Erjavec from the Josef Stefan Institute in Ljubliana, modified and extended by Şt. Bruda from RACAI in Bucharest) in order to take advantage of the editing facilities of XEMACS. By knowing the significance of the morpho-syntactic codes we used, the mtems-mode allows a user to manually disambiguate the segmenter's output (see Figure 4).

*Într-o zi senin` [i friguroas` de aprilie, pe când ceasurile băteau ora treisprezece, Winston Smith, cu bărbia înfundată în piept pentru a scăpa de vântul care-l lua pe sus, se strecură iute prin uşile de sticlă ....*

Figure 1 Plain text

Într-<br>
o<br>
zi<br>
senină<br>
şi<br>
friguroasă<br>
de<br>
aprilie<br>
,<br>
pe<br>
când<br>
ceasurile<br>
băteau<br>
ora<br>
treisprezece<br>
,<br>
Winston<br>
Smith<br>
,<br>
cu<br>
bărbia<br>
înfundată<br>
în<br>
piept<br>
pentru<br>
a<br>
scăpa<br>
de<br>
vântul<br>
care<br>
-l<br>
lua<br>
pe<br>
sus<br>
,<br>
se<br>
strecură<br>
iute<br>
prin<br>
uşile<br>
de<br>
sticlă<br>
…

Figure 2 Segmented text

| | |
|---|---|
| Într- | Spsay |
| o | Qf Pp3fsa--------w Mcfsrl Tifsr Va--3s |
| zi | Vmip1s Ncfsrn Vmm-2s |
| senină | Afpfsrn |
| şi | Ccssp Px3--d--------w Rp |
| friguroasă | Afpfsrn |
| de | Csssp Spsa |
| aprilie | Ncms-n |
| , | |
| pe | Spsa |
| când | Rw |
| ceasurile | Ncfpry |
| băteau | Vmii3p |
| ora | Ncfsry |
| treisprezece | Mc-p-l |
| , | |
| Winston | Npms-n |
| Smith | Np |
| , | |
| cu | Spsa |
| bărbia | Ncfsry |
| înfundată | Vmp--sf Afpfsrn |
| în | Spsa |
| piept | Ncms-n |
| pentru | Spsa |
| a | I Qn Spsa Tsfs Va--3s |
| scăpa | Vmnp Vmii3s Vmm-2s |
| de | Csssp Spsa |
| vântul | Ncmsry |
| care | Pw3--r Vmsp3 Ncfp-n Dw3--r---e |
| -l | Pp3msa--y-----w |
| lua | Vmp--sm Vmnp Vmii3s Vmm-2s |
| pe | Spsa |
| sus | Ncms-n Rgp |
| , | |
| se | Px3--a--------w |
| strecură | Vmis3s |
| iute | Afpfsrn Afpms-n Rgp Ncfson |
| prin | Spsa |
| uşile | Ncfpry |
| de | Csssp Spsa |
| sticlă | Ncfsrn |
| … | |

Figure 3 Ambiguously annotated text

| | |
|---|---|
| Într- | Spsay |
| o | Tifsr |
| zi | Ncfsrn |
| senină | Afpfsrn |
| şi | Ccssp |
| friguroasă | Afpfsrn |
| de | Spsa |
| aprilie | Ncms-n |
| , | |
| pe | Spsa |
| când | Rw |
| ceasurile | Ncfpry |
| băteau | Vmii3p |
| ora | Ncfsry |
| treisprezece | Mc-p-l |
| , | |
| Winston | Npms-n |
| Smith | Np |
| , | |
| cu | Spsa |
| bărbia | Ncfsry |
| înfundată | Afpfsrn |
| în | Spsa |
| piept | Ncms-n |
| pentru | Spsa |
| a | Qn |
| scăpa | Vmnp |
| de | Spsa |
| vântul | Ncmsry |
| care | Dw3--r---e |
| -l | Pp3msa--y-----w |
| lua | Vmii3s |
| pe | Spsa |
| sus | Rgp |
| , | |
| se | Px3--a--------w |
| strecură | Vmis3s |
| iute | Rgp |
| prin | Spsa |
| uşile | Ncfpry |
| de | Spsa |
| sticlă | Ncfsrn |
| … | |

Figure 4. Disambiguated text

Table 2 shows the global figures for the sub-corpora processed as described above. Since in the first phase of this research we were concerned with parallel corpora, the selected texts were Orwell's "1984" and Plato's "Republic". These two integral texts belong to two different registers (fiction and philosophy) and, as will be shown, have different distributional properties. The **Occurrences** column contains the number of word-forms occurrences in each of the two books without eliminating the duplicates. The **Wordforms** column shows the number of distinct word-forms in the two books with the third row giving the total number of distinct word-forms. The **MSDs** column shows the number of distinct MSD *used* in each of the books and the total. The **AMB_MSD** and **AMB_POS** give, respectively, the numbers of MSD-ambiguity clusters and POS-ambiguity classification clusters *found* in the corpora. The **Common** row displays for each category item, the commonalties among the two texts. As one can see from Table 2, although "Republic" contains more occurrences than "1984", the number of different word-forms and lemmas is significantly smaller. This is not surprising given the two registers of the considered texts. In a philosophical text, most words are used in a rather technical way with as little as possible stylistic variations. Although the two texts contain altogether more than 200.000 words, the number of different words is just 20371 (less than 10%) with 4.016 word-forms occurring in both texts. Therefore, from the lexical coverage point of view the two texts are far from being sufficient in drawing significant conclusions on Romanian language word-usage frequencies. However, for the distributional analysis and morpho-syntactical disambiguation purposes the selected texts offer enough evidence to extract reliable data and draw realistic conclusions. Out of the 674 MSDs defined in the lexicon, these texts contain 500 MSDs (more than 74%) which is a very good figure, considering that the number of distinct word-forms occurring in the selected texts represents approximately 5% of the lexical stock. This is to say that most of the words raising problems (from the ambiguity point of view) appeared at least in one of the two texts.

This statement is even better supported if one considers the last column in Table 2: "1984" contains all but 4 POS-ambiguity clusters while "Republic" misses only 10. The MSD-ambiguity clusters are also very well represented in the selected texts (more than 6% of all MSD-ambiguity clusters defined in the lexicon).

Table 2: Corpora overview

| Sub-Corpus | Occurrences | Wordforms | Lemmas | MSDs | AMB_MSD | AMB_POS |
|---|---|---|---|---|---|---|
| '1984' | 101460 | 14037 | 7019 | 444 | 529 | 86 |
| 'Republic' | 114720 | 10350 | 4697 | 400 | 498 | 80 |
| **Common** | 56804 | 4016 | 2527 | 344 | 402 | 78 |

Table 2.1: The distribution of the word-form occurrences over the parts of speech in “1984” and “Republic”

| POS | N | V | A | P | D | M | R | T | S | C | Q | I | Y | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **‘1984’ occrs** | 22570 | 19031 | 6202 | 10967 | 2376 | 1228 | 10633 | 4236 | 12621 | 6389 | 4967 | 99 | 72 | 69 |
| **Republic occrs** | 19936 | 24890 | 6302 | 13577 | 3069 | 633 | 12965 | 4079 | 12141 | 10506 | 6383 | 199 | 37 | 3 |
| **Common occrs** | 5345 | 7768 | 2045 | 8294 | 1506 | 416 | 6556 | 2730 | 10542 | 5797 | 4482 | 36 | 0 | 1 |

The distribution of the word-forms (without counting duplicates) over the parts of speech shows some differences that are motivated by the different registers. The difference due to the linguistic registers the texts belong to, shows up, as one would expect in the counts for common content words (N, V, A, R). For the functional words (S, C, Q, T, P) the number of common occurrences is very high. This consideration is better supported when comparing the word-forms counts (Table 2.2). The abbreviations and residuals are also very different in the two texts. This can be easily understood, if one takes into account that abbreviations in most cases stand for content words. The residual class is poorly represented in Plato’s “Republic”, but it has several instances in Orwell’s “1984” (mainly when the newspeak jargon is in use).

Table 2.2: The distribution of the word-forms over the parts of speech

| POS | N | V | A | P | D | M | R | T | S | C | Q | I | Y | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **‘1984’ wordforms** | 6177 | 4518 | 2473 | 189 | 137 | 114 | 683 | 29 | 127 | 62 | 9 | 34 | 16 | 42 |
| **‘Republic’ wordforms** | 4537 | 3603 | 1682 | 179 | 128 | 111 | 424 | 22 | 101 | 49 | 7 | 12 | 4 | 1 |
| **Common wordforms** | 1467 | 1344 | 607 | 136 | 105 | 48 | 277 | 21 | 74 | 40 | 7 | 5 | 0 | 1 |

Comparing the figures in the previous tables, the discursive nature of Plato’s book is quite obvious when contrasted with “1984”: the smaller number of Nouns is compensated by an increased number of pronouns and the large number of verbal occurrences ensures an alert argumentative rhythm.

By considering lemmas distributions over the parts of speech, one could notice some interesting facts: for instance, in spite of a quite productive inflectional potential for nouns, verbs and adjectives the average number of inflected word-forms from a given lemma that are used in the two texts is really small: less than 2 Noun or Adjective word-forms per Noun or Adjective lemma, less than 4 Verb word-forms per Verb lemma. Table 2.3 outlines also the lexical diversity in the two books, with “1984” having a lexical broader coverage in spite of its smaller number of word occurrences.

Table 2.3: The distribution of the lemmas over the parts of speech

| POS | N | V | A | P | D | M | R | T | S | C | Q | I | Y | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **‘1984’ lemmas** | 3441 | 1370 | 1514 | 63 | 44 | 66 | 659 | 10 | 109 | 55 | 7 | 34 | 16 | 42 |
| **‘Republic’ lemmas** | 2324 | 996 | 940 | 62 | 47 | 65 | 422 | 6 | 93 | 46 | 5 | 11 | 4 | 1 |
| **Common lemmas** | 1015 | 656 | 451 | 43 | 36 | 32 | 277 | 5 | 70 | 37 | 5 | 5 | 0 | 1 |

The figures in Table 2.4 provide arguments for the adequacy of the two texts with respect to the morpho-syntactic descriptions of the lexical stock. By comparing data in Table 1.4 (lexicon) with data in Table 2.4, one could see that out of the 53 MSDs defined for Nouns, the two texts provide contexts for 47 MSDs (39+35-27=47). For the verbal MSDs, out of the 134 defined in the lexicon, the two texts cover 99. With functional words, the coverage is practically 100%.

Table 2.4: MSD distribution over the parts of speech

| POS | N | V | A | P | D | M | R | T | S | C | Q | I | Y | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **'1984' MSDs** | 39 | 92 | 27 | 102 | 86 | 34 | 11 | 26 | 7 | 8 | 6 | 1 | 4 | 1 |
| **'Republic' MSD** | 35 | 74 | 20 | 105 | 83 | 32 | 7 | 21 | 7 | 6 | 6 | 1 | 2 | 1 |
| **Common MSD** | 27 | 67 | 17 | 85 | 73 | 25 | 7 | 20 | 7 | 6 | 6 | 1 | 2 | 1 |

Tables 2.5 and 2.6 show the ambiguity level for the word-forms, computed by first considering distinct MSDs that are attached by lexical lookup to the word-forms (Table 2.5) and second, by considering distinct lemmas a word-form may be attributed to. For instance, the word-form "urâţi" which appears in both texts is 9-way MSD ambiguous. However, it is only 4-way ambiguous when lemma ambiguity is considered (there are four lemmas to which the word-form "urâţi" may be attributed). Similarly, the word-form "VII" appearing in "Republic" (as such, that is in upper case letters) is 9-way MSD ambiguous but only 4-way lemma ambiguous (since it might not be obvious, here there are the MSDs and lemmas: Afpfp-n, Afpfson, Afpmp-n, Mc-p-r, Mo-s-r, Ncfp-n, Ncfson, Vmip2s, Vmsp2s and "şapte/seven", "vie/vineyard", "viu/alive", "veni/come").

Table 2.5: Word-forms grouped according to the MSD ambiguity level

| Ambiguity level | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| **'1984' wordforms** | 9322 | 2703 | 1425 | 301 | 189 | 56 | 44 | 0 | 1 |
| **'Republic' wordforms** | 6191 | 2357 | 1312 | 243 | 174 | 49 | 31 | 0 | 2 |
| **Common wordforms** | 2237 | 895 | 641 | 133 | 81 | 24 | 15 | 0 | 1 |

Table 2.6: Word-forms grouped according to the lemma ambiguity level

| Ambiguity level | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **'1984' wordforms** | 12432 | 1524 | 77 | 6 |
| **'Republic' wordforms** | 9121 | 1153 | 75 | 7 |
| **Common wordforms** | 3384 | 580 | 57 | 4 |

If we distinguish only between ambiguous word-forms (ambiguity level greater than 1) and unambiguous word-forms[4] (ambiguity level equal to 1), the two texts considered here reveal slightly different ambiguity figures: 33.5% ambiguous words in "1984" and 40.2% ambiguous words in "Republic", the explanation resides in the use of a higher number of definite nouns and adjectives in "1984" as compared to "Republic". The definite forms of the nouns, adjectives and to some extent numerals are in the vast majority of cases unambiguous. Judging the ambiguity in terms of lemma ambiguity (as shown in Table 2.6) the two texts exhibit practically the same ambiguity level 11.5% in "1984" and 12% in "Republic".

The next two tables provide further evidence on inter and intra categorial ambiguity in Romanian. The intra-categorial ambiguity -"horizontal" homography, is much more significant than the inter-categorial ambiguity -"vertical" homography. This is in sharp contrast with English and might be considered a characteristic of highly inflectional languages.

Contrast for instance the verb horizontal ambiguity in Table 2.7 (561, 555, 459) and the verb vertical ambiguity in Table 2.8 (35, 34, 33). A natural expectation is that in highly inflectional languages, the horizontal ambiguity is much more difficult to solve than the vertical ambiguity.

[4] Remember the discussion in section 2 on ambiguity and the conventions we used in encoding the

Table 2.7: MSD distribution over the MSD-ambiguity classes

| POS-msd | N-msd | V-msd | A-msd | P-msd | D-msd | M-msd | R-msd | T-msd | S-msd | C-msd | Q-msd | I-msd | Y-msd | X-msd |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| '1984' MSD-amb. classes | 317 | 561 | 246 | 140 | 87 | 41 | 89 | 15 | 21 | 21 | 9 | 18 | 14 | 10 |
| 'Republic' MSD-amb. classes | 289 | 555 | 211 | 142 | 84 | 47 | 87 | 14 | 20 | 19 | 10 | 14 | 14 | 7 |
| Common MSD-amb. classes | 229 | 459 | 169 | 133 | 79 | 29 | 77 | 14 | 19 | 19 | 9 | 14 | 11 | 7 |

Table 2.8: POS distribution over the POS-ambiguity-classes

| POS | N | V | A | P | D | M | R | T | S | C | Q | I | Y | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| '1984' POS-amb. classes | 28 | 35 | 17 | 27 | 11 | 12 | 29 | 12 | 17 | 12 | 8 | 12 | 11 | 9 |
| 'Republic' POS-amb. classes | 25 | 34 | 17 | 26 | 10 | 14 | 27 | 11 | 16 | 10 | 8 | 11 | 10 | 6 |
| Common POS-amb. classes | 24 | 33 | 16 | 26 | 10 | 12 | 27 | 11 | 16 | 10 | 8 | 11 | 9 | 6 |

## 5. Conclusions

The paper described the results in constructing TEI-conformant encoded corpora in a multilingual environment, a wide coverage lexicon and a large training corpus. By several counts and statistics we showed that the training corpus covers most of the morpho-syntactic classes defined for Romanian, thus offering a solid basis for further projects on automatic disambiguation and inductive natural language processing. We have shown that given the fact that in Romanian (as in other highly inflectional languages) the intra-categorial ambiguity is by far more significant than the inter-categorial ambiguity, the tagsets used for statistical disambiguation as developed for western languages (especially those used for English) are not appropriate and future trial- and error experiments are needed to design useful tagsets and tagging procedures.

We plan to investigate, based on the MSD ambiguity classes which we identified, the possibility to construct in a principled way language specific tagsets. Given that the morpho-syntactic specification of the lexicon was done in a multilingual environment (7 languages), the prospect of development of the language specific tagsets with a common methodology, starting from these common morpho-syntactic specification, is challenging. At present, a semi-automatic procedure allows for generating from the MSD several tagsets, according to criteria specified by the designer. This is a first step towards the tagset design. Since modifying the training corpus is not desirable (actually annotated at the MSD level, that is at the finest granularity as provided by the lexicon), the automatic tagging procedure should be able to consider an external resource file mapping the MSDs onto a corpus experimental tagset. This way, one should be able to experiment with several tagsets without modifying training corpus, but just the small mapping file.

**Note:** At the time this article went into print some extremely good results were obtained in automatic tagging of Romanian texts. The tagger, a brand new version of the Birmingham tagger authored by Oliver Mason, was trained on about 180.000 hand tagged texts extracted from "1984" and "Republic" and tested on about 50.000 unseen words from the same books. The tagset contained 82 tags (including 9 tags for punctuation) and the accuracy of the tagging process was well beyond our expectations, namely 95.5% for "Republic" and 97.1% for "1984".

According to our knowledge these are the best results obtained up to now for an Eastern European Language.

The details on the tagger, the tagset design as well as other experimental data are presented in [16].

## References


1. J. SINCLAIR, *Corpus Typology*, EAGLES DOCUMENT EAG-CWG-IR-2, Version of October 1994.
2. W. TEUBERT, *Language Resources for Language Technology*, in Recent Advances in Romanian Language Technology, D. Tufiş, P. Anderson (eds.), Edit. Academiei Române, Bucureşti, 1997.
3. ANA-MARIA BARBU, D. TUFIŞ, C. DIACONU, LIDIA DIACONU, *A Plea for Corpus-based Linguistics*, in Language and Technology, D. Tufiş (ed.), Edit. Academiei Române, Bucureşti, 1996 (in Romanian).
4. N. IDE, J. VERONIS, *Corpus Encoding Standard,* MULTEXT/EAGLES Report, :http//www.lpl.univ-aix.fr/projects/multext/CES/CES1.html, 1995.
5. D. TUFIŞ, Şt. BRUDA, *Structure Markup in CES and Preliminary Statistics on Romanian Translation of Plato's "Republic"*, in TELRI Newsletter, no.5, April 1997.
6. C.M. SPERBERG-MCQUEEN, L. BURNARD, *Text Encoding Initiative: Guidelines for Electronic Text Encoding and Interchange*, Chicago and Oxford Press, 1994.
7. N. IDE, J. VERONIS (eds.), *Text Encoding Initiative: Background and Context*, Dordrecht, Kluwer Academic Publishers, 1995.
8. M. MONACHINI, N. CALZOLARI (eds.), *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages*, EAGLES Report, October 1994.
9. N. CALZOLARI, M. MONACHINI (eds.), *Common Specifications and Notation for Lexicon Encoding and Preliminary Proposal for the Tagsets*, MULTEXT Report, March 1995.
10. T. ERJAVEC, N. IDE, D. TUFIŞ, *Development of Common Lexical Specifications for Six Eastern European Languages and Their Application to Stochastic Part of Speech Tagging*, in Proceedings of ACH/ALLC'97 Conference, Kingston, Ontario, Canada, June 3-7, 1997.
11. T. ERJAVEC, N. IDE, V. PETKEVIC, J. VERONIS, *MULTEXT-EAST: Multilingual Text Tools and Corpora for Central and Eastern European Languages,* in Proceedings of the Trans European Language Resource Infrastructure First Conference, Tihany, September 1996.
12. D. TUFIŞ, *Morpho-lexical processing in a multilingual environment*, Tutorial Notes at EUROLAN'97 Summer School on Corpus Linguistics, Tu°nad, July 1997, http//www.infoiasi.ro/eurolan/annl and also http//www.racai.ro/~tufis.

13. D. TUFIŞ, *A Unification-Based Description of Gender-Agreement in Romanian*, (forthcoming).

14. D. FARKAS, D. ZEC, *Agreement and Pronominal Reference,* in Gugliermo Cinque and Giuliana Giusti (eds.), *Advances in Romanian Linguistics*, John Benjamin Publishing Company, Amsterdam -Philadelphia, 1995.

15. D. TUFIŞ, *A Generic Platform for Developing Language Resources and Applications*, in Proceedings of the Trans European Language Resource Infrastructure Second Conference, Kaunas, April 1997.

16. O. MASON, D. TUFIŞ, *Probabilistic Tagging in a Multilingual Environment: Making an English Tagger Understand Romanian*, in Proceedings of the Third TELRI International Seminar, Tuscan World Center, Montecatini, October 1997.