Standardized Specifications, Development and Assessment of Large Morpho-Lexical Resources for Six Central and Eastern European Languages

Dan Tufis
Romanian Academy
Center for Artificial Intelligence
Bucharest, Romania
[tufis@racai.ro]

Nancy Ide
Department of Computer Science
Vassar College
Poughkeepsie, NY, USA
[ide@cs.vassar.edu]

Tomaz Erjavec
Institute Josef Stefan
Ljubljana, Slovenia
[tomaz.erjavec@ijs.si]

Abstract

This paper provides an overview of the harmonized language specifications for MULTEXT-East's six CEE languages, which include languages from the Romance, Finno-Ugric, and Slavic language families, and considers their use and distribution in lexicons for these languages. Because these language families include many features and properties not found in western European languages, such as heavy inflection and agglutination, adapting the specifications for western European languages to these languages posed many interesting and difficult problems. It describes the form and content of the six CEE language lexicons built on the basis of these specifications and provides quantitative assessment of the content, in order to compare the various languages.

1. Introduction

MULTEXT-East was a project under the European Union Copernicus program whose goal was to develop language resources for six Central and Eastern European (CEE) languages (Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene) and to adapt existing tools and standards to them (Erjavec, Ide, Petkevic, Véronis, 1996). The project built on and extends the MULTEXT project (Ide and Véronis, 1994), which developed a comprehensive set of corpus-annotation tools, including tools for text segmentation, stochastic part of speech tagging, and alignment of parallel texts. MULTEXT-East developed linguistic resources and created a multi-lingual, partially parallel corpus in the six CEE languages, a portion of which is annotated for part of speech and aligned (Erjavec & Ide, 1998).

Because the overall goal of MULTEXT-East was to develop reusable resources, it was essential to establish standardized methods and specifications for the created resources. To this end, a harmonized set of specifications for lexicon encoding was developed for the six MULTEXT-East languages (Erjavec and Monachini, 1997), based on the specifications developed in the EAGLES project (Bel, Calzolari, and Monachini, 1996) and their extension by the MULTEXT project to six western European languages (English, French, Dutch, Italian, German, Spanish) (Monachini and Calzolari, 1996). Accommodating the different language families represented among the MULTEXT-East languages (Romance, Finno-Ugrec, and Slavic) demanded substantial assessment and modification of the pre-existing specifications, due to the need to accommodate features which appear rarely in western European languages, such as heavy inflection and agglutination. To validate the specifications, the MULTEXT-East project built lexicons for each of its six languages based on them, and used the information contained in them for the automatic tagging of a parallel corpus of Orwell's *Nineteen Eighty-Four*.

The availability of a harmonized set of lexical specifications provides a common base for comparison of various statistical properties of lexemes in these languages, which has heretofore been impossible. This paper provides an overview of the harmonized language specifications for MULTEXT-East's six CEE languages and English, and considers their comparative use and distribution in lexicons and corpora for these languages. The paper is organized as follows: Section 2 provides a description of the morpho-syntactic specifications used in the lexicons of the project and discusses some encoding decisions. Section 3 describes the form and content of the six lexicons built on the basis of these specifications and provides quantitative assessment of the content, in order to compare the various languages. Section 4 briefly describes the corpora used in the project and provides information on the lexical items distribution. Finally, Section 5 summarizes our conclusions and suggests directions for future research.

2. Morpho-Lexical Specifications

The MULTEXT-East lexical specifications describe the grammar of the morpho-syntactic descriptions (MSDs) used in the lexicons of the project. The development of harmonized lexical specifications for the six MULTEXT-East languages began with proposals developed in the EAGLES project (Bel, Calzolari, and Monachini, 1995) and the modifications proposed for six western European languages in the MULTEXT project (Monachini and Calzolari, 1996). These proposals were evaluated from the point of view of coverage for the six CEE languages. While adaptation for Romanian, a Romance language, was relatively straightforward, the other CEE languages posed many interesting and difficult problems and demanded substantial assessment and modification of the pre-existing specifications.

The nucleus of common features isolated within MULTEXT for western European languages was assumed as the common ground for extension to the CEE languages. Specifications for information peculiar to the CEE languages were added as required, taking care that similar phenomena in the various (e.g., Slavic) project languages were encoded in a similar manner. This led to the formulation of a common proposal for lexicon specifications of the CEE languages, detailed in (Erjavec and Monachini, 1997). The work carried out in MULTEXT-East has thus broadened the base and contributed significantly to defining a general mechanism for lexical specification.

For each part of speech that is distinguished in the morpho-syntactic descriptions, the specifications give a table detailing the features used for that part of speech, the names and one-character codes for the values these features can take, and the applicability of the attribute/values to the six languages. The tables distinguish two types of attributes:

- minimal core features, i.e., those shared by most of the languages. These are common to all the MULTEXT and MULTEXT-East languages. This facilitates comparability of the information encoded in the lexical lists for the six MULTEXT-East and six western European languages treated in MULTEXT.
- language-specific features, which apply only to (one or more) MULTEXT-East languages.

The cross-language tables provide a concise summary of language differences and similarities. Table 1 provides an overview of the kinds of information contained in the cross-language tables by providing information about how many attributes each of the six languages distinguishes for the various parts of speech. A hyphen in the table means that the particular part of speech is not valid for the language in question, while a zero denotes that the language distinguished no features for that part of speech.

	Romance	Slavic			Finno-Ugric	
PoS	ROM	BUL	CZE	SLO	EST	HUN
Noun	6	5	5	5	3	7
Verb	7	8	10	8	8	5
Adjective	7	3	7	7	3	8
Pronoun	8	8	12	11	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	7	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	3	-	-	-	-	-
Article	5	-	-	-	-	1

Table 1. Number of attributes distinguished for each part of speech, by language

The summary in Table 1 shows the marked distinction between the language families and the languages themselves. The table in the Appendix shows the attributes and all all their applicable values for NOUN across the MULTEXT-East languages.

The grammar of the morpho-syntactic descriptions is realized in the lexical MSDs. These are provided as strings, using a linear, term-like 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 from the tables.

§ 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.

For example, the specification

```
Vmm-2s
```

stands for "Verb main imperative (no Tense) second singular".

The “does not apply” marker (‘-’) in the MSD encoding is slightly different from the one used in Table 1. Besides the basic meaning that the attribute is not valid for the language in case, it also indicates that a certain combination of other morpho-syntactic attributes makes the current one irrelevant. For instance, non-finite verbal forms are not specified for Person.

The MSD encoding provides a simple and relatively compact representation, and is in intention similar to the feature-structure encoding used in unification-based grammar formalisms. So, for example, **Vmm-2s** (standing for *Verb main imperative (no Tense) second singular*) can be represented as the attribute-value matrix:

```
Verb
Type:   main
VForm:  imperative
Tense:  -
Person: second
Number: singular
```

Here, "Verb" stands for the type of the feature structure, in the sense of (Carpenter, 1992), which determines the appropriate attributes of the feature structure of this type. Attributes and values follow.

The EAGLES recommendations provide another special attribute value, the dot ("."), for cases where an attribute can take any value in its domain. The ‘any’ value is especially relevant in situations where wordforms are under-specified for certain attributes, which, however, can be recovered from the immediate context (by grammatical rules such as agreement). The MULTEXT-East Specifications did not originally include the "any" value, which was not necessary for most of the CEE languages. However, because of the peculiarities of the Romanian case system,[1] the "any" value was included to avoid

[1] In Romanian, the case is relevant for Nouns, Adjectives, (proclitic) Articles, Determiners, Pronouns, Numerals and it is subject to the agreement rule (it is also valid for adjectival participles and some abbreviations. However, it is morphologically marked only on the first word of the sequence subject to the agreement relation:

ArticleCASE Adjective Noun; DeterminerCASE Adjective Noun;

NounCASE Adjective; AdjectiveCASE Noun; etc.

redundancy in the Romanian wordform lexicon (Tufis *et al.*, 1997), which we estimate would have been almost four times larger (for the same informational content) without it. Therefore, to be syntactically conformant with the MULTEXT-East specifications the Romanian encoding loaded the semantics of the "-" value with the additional meaning of "any value from the domain of the corresponding attribute".[2]

3. The Wordform Lexicons

Once the harmonized set of morpho-syntactic specifications for the six MULTEXT-East languages was developed, lexicons incorporating these specifications were created for each language. The lexicons were created by adapting dictionaries and lexicons where they existed and via automatic and semi-automatic generation of wordforms and association of wordforms with MSDs. Because the lexicons were used to automatically tag texts in the MULTEXT-East corpus, they provide full coverage of all corpus texts. Token lists for the texts were automatically generated and then fed through morphological analyzers in order to produce the lemma list (and associated morpho-syntactic information). In the next step, these lemmas were fed back to the morphological generators (except for the agglutinative languages-see below) in order to produce the complete inflected list, i.e., the full paradigms of the lemmas, which constitute the final lexicons of the project. The creation process and lexicon contents for each language are described in (Ide, 1996).

While the inclusion of full paradigms in the lexicons is still manageable for the Romance and Slavic languages, it is not feasible for the agglutinative languages of the project, namely Estonian and Hungarian. First, automatic generation for agglutinative languages produces a prohibitively large number of unacceptable wordforms. More importantly, even if it were possible to generate correct paradigms for these languages automatically, the number of possible wordforms of a lemma for these languages is so large (estimated at 20 million for Hungarian) as to preclude the possibility of including them all in a wordform lexicon. This problem was bypassed within the project because time and budget constraints did not allow the implementation of a generative solution. As a result, only the wordforms (with their relevant MSD interpretations) that actually occur in the corpus of the project are included for these two languages.

Entries in the lexicons are of the following format:

```
wordform <TAB> lemma <TAB> MSD
```

For example (Estonian):

```
aega        aeg      Nc-s1
```

Note that the same word-form may be associated with different MSDs (or lemmas) and therefore may appear in the first column of two or more entries. For example, the word-form in the entry cited above appears in the first column of the following entries as well:

```
aega        =        St
aega        aeg      Nc-s7
```

When the word-form is its own lemma, the "=" notation is placed in the lemma field. In the example above, for the entry "aega" where the MSD is "Adposition postposition" (St), the lemma is the word-form itself; however, for "aega" as "Noun common singular additive" (Nc-s7), the lemma is "aeg".

Table 2 summarizes the major characteristics of the six CEE lexicons, and includes data for a lexicon of English encoded using the same MSD formalism for comparative purposes. The languages are grouped by language family (Romance, Slavic, Finno-Ugric, plus the Germanic English).

Language	Eng	Rom	Bul	Cze	Slo	Est	Hun
Entries	66473	419869	333721	133803	542773	130409	59614
Word forms	43564	347126	284211	41601	193058	89180	46886
Lemmas	12622	33515	17972	14458	15806	22054	15838
=	25816	35669	19064	14684	16008	23384	17380
MSD	134	611	185	915	2080	563	603
PoS AmbCls	47	83	42	35	51	63	62
MSD AmbCls	328	869	680	698	1210	1012	890

Table 2. Table summarizing the six CEE language lexicons, plus English

- **Entries** : the number of triplets of the form wordform <TAB> lemma <TAB> MSD
- **Wordforms** : the number of distinct words (eliminating duplicates)
- **Lemmas** : the number of distinct lemmas (eliminating duplicates)
- **=** : the number of lemmas (preserving duplicates)
- **MSD** : the number of MSD codes in the lexicon
- **POSAmbCls** : the number of part of speech (PoS) ambiguity classes
- **MSDAmbCls** : the number of MSD ambiguity classes

The "=" field provides the number of entries which are themselves lemmas (i.e., have "=" in the lemma field of their entry). Thus, the arithmetic difference between the "Lemma" and the "=" fields gives (except for Estonian and Hungarian) the number of non-inflecting words in the lexicons. The "MSDs" field gives the total number of distinct MSDs used in the lexicon. Finally, "Ambig" provides the number of ambiguity classes in the lexicon: i.e., the number of different groupings of MSDs associated with any one word-form in the lexicon.

PoSAmbCls and MSDAmbCls provide information about the number of ambiguity classes in each dictionary. Each ambiguous wordform in the lexicon belongs to such an ambiguity class. If the ambiguity is considered in terms of the MSD or PoS, the ambiguity classes are called MSD or PoS ambiguity classes respectively. The ambiguity classes or genotypes (Tzoukermann, Radev, 1997)

[2] A simple Perl program is used to replace those hyphens having the ‘any’ interpretation into ‘.’ and to expand a MSD containing one or more dot-equivalent hyphens into the appropriate set of MSDs.

provided useful information for designing the tagsets appropriate for probabilistic disambiguation (Mason & Tufis, 1997; Tufis, 1998; Tufis & Mason 1998).

The Romanian language has consistent case syncretism for nouns (as well as adjectives) between nominative and accusative and between genitive and dative, and therefore the syncretic cases were collapsed as "direct" (nominative and accusative) and "oblique" (genitive and dative). Interestingly, while nominative/accusative syncretism also exists in Czech and Slovene, it is not an across-the-board phenomenon, and thus could not be sensibly reduced by using such "vague" values.

Table 2 reveals some expected statistics, but also reflects decisions taken during the development of the MSDs. For example, the entry/wordform ratio is similar (in the 1.2 - 1.5 range) for all languages except Czech and Slovene (3.21 and 2.81 respectively). This can be explained in part because these two languages are more inflected than Romanian, Bulgarian, and English. Another explanation lies in the fact that for Czech and Slovene, syncretism was not considered when encoding the wordform dictionaries. For example, in Slovene the following syncretic entries are explicitly encoded as shown:

```
Abrahama  Abraham  Npmsa
   (Noun proper masculine singular accusative)
Abrahama  Abraham  Npmsg
   (Noun proper masculine singular genitive)
Abrahama  Abraham  Npmda
   (Noun proper masculine dual accusative)
Abrahama  Abraham  Npmdn
   (Noun proper masculine dual nominative)
```

To further support the second explanation, we could add that the ratio entry/wordform for Romanian is 1.2, but if one considers expanding the case syncretism (i.e., all "direct" and "oblique" cases for nouns and adjectives are replaced by explicit N, A and G, D cases), and the expansions of the implicit “any” values, the ratio above would exceed 2.8.

Given the number of word-forms per lemma in agglutinative languages, such ambiguity might also be expected in Estonian and Hungarian, but in these languages syncretism is much less pervasive than in inflecting languages (evident in their entry/word-form ratio), which is also in the 1.2 - 1.5 range.

The table also reveals some unexpected discrepancies in lexicon construction methods and design decisions for different languages. For example, Slovene has an exceptionally high number of MSDs-far greater than Czech, which is very similar to Slovene and should be comparable. In fact, the explanation for this is that Slovene has an extremely detailed set of MSDs for Pronouns (P), which as a class encode seventeen different features for the CEE languages. Although the common tables (Erjavec & Monachini, 1997) do not show Slovene encoding more features than other languages overall, Slovene encodes more combinations. Slovene has twice as many pronouns as Czech, and inflects them 3 times as much, which accounts for a part of the difference in the number of MSDs. However, it also reflects different design decisions: where the same form is used for first, second, and third person, Czech omits the number value (implicitly adopting an “all values are possible” approach), whereas Slovene encodes it as three ways ambiguous.

The values for "PoSAmbCls" reveal that intra-category ambiguity is different in the seven lexicons. Although this might be attributable to language differences, it is also due to the omission of homographs that were not in the lexicons for Hungarian, Estonian, Czech, and English.

Table 3 provides the same statistics for the main part of speech categories (noun, verb, adjective, adverb) in the lexicons. Note that a word-form ambiguous by part of speech is counted in more than one category. "MSD AmbCls" gives the number of ambiguity classes for each part of speech; this value was computed by first isolating the entries of all word-forms which have the same part of speech and then computing the number of ambiguity classes for this set; the percentage is the ratio of the total number ambiguity classes to this number.

English, Romanian and Bulgarian have the highest proportion of verb entries and the lowest proportion of adjectives, while for Czech and Slovene the situation is exactly the reverse: adjectives have the greatest number of entries and nouns, the lowest. For the agglutinated languages, noun is the most frequent entry type. While the information in the "Entries" and "Wordform" columns is dependent on the different strategies used to handle syncretism, the "Lemmas" and "=" columns show that for all languages, noun is the most common lemma.

Lang	PoS	Entries	Wordform	=	Lemmas	MSDs	MSD AmbCls
Eng	N	32.57	47.94	42.73	48.46	14.29	55.60
	V	46.08	34.28	13.43	15.29	22.56	49.38
	A	15.40	22.22	30.50	33.73	3.01	45.64
	R	5.44	8.04	12.57	14.33	6.77	24.90
Rom	N	29.37	32.49	51.48	54.70	8.84	49.14
	V	41.45	41.02	11.92	12.68	14.57	46.49
	A	28.26	31.96	31.06	33.02	10.15	36.48
	R	0.32	0.38	3.50	3.72	1.64	9.55
Bul	N	30.59	38.20	51.80	52.87	7.69	52.71
	V	43.09	36.49	6.42	6.61	26.63	56.46
	A	17.38	22.77	10.41	10.63	2.66	34.38
	R	3.87	5.17	14.20	14.48	0.59	12.50
Cze	N	32.21	40.96	43.99	44.68	7.31	55.88
	V	10.03	26.31	18.18	18.68	12.30	23.11
	A	54.74	30.05	24.99	25.68	13.77	21.17
	R	0.87	2.73	5.86	6.02	0.21	6.73
Slo	N	22.38	30.15	44.18	44.54	4.71	48.68
	V	20.01	39.53	22.61	22.90	6.15	23.97
	A	54.81	32.02	27.49	27.84	13.41	26.53
	R	1.32	3.71	2.69	2.71	0.14	9.17
Est	N	61.16	71.02	55.29	58.54	9.41	69.17
	V	15.69	16.94	3.62	3.84	26.29	32.41
	A	19.39	23.47	26.31	27.87	13.32	35.18
	R	0.02	3.22	12.23	12.96	0.18	18.48
Hun	N	48.16	55.77	38.76	42.41	38.07	70.16
	V	25.09	26.50	8.10	7.86	9.48	20.71
	A	21.60	25.90	42.48	46.37	24.84	44.10
	R	2.55	3.15	7.52	8.17	0.65	19.04

Table 3. Lexicon data by part of speech

A comparison of the "Wordform" and "Lemmas" columns shows that for all languages except for Czech and Slovene, verbs exhibit a high lemma/wordform ratio. In these languages, therefore, verbal wordforms are strongly marked and easily recognizable. This observation was confirmed by several tagging and guessing experiments (see Tufis & Mason, this volume). For Czech and Slovene, verb identification is the same as for other parts of speech.

Table 3 also shows that recognition of adjectives is easier in the Slavic languages than the others, due to distinct graphemic marking. Nouns are somewhat easier to differentiate from the other parts of speech for the two agglutinative languages. In general, the largest number of MDSs[3] are defined for verbs; However, Czech and Slovene allocate higher proportion of tags to Adjectives, while for Hungarian, almost two-thirds of the total number of tags are for nouns and adjectives.

The last column in Table 3 shows that nouns are included in half of the ambiguity classes, with Hungarian at the extreme (70.16% of the total number of MSD ambiguity classes include at least a nominal MSD).

Table 4 shows the degree of ambiguity for each language lexicon by MSD; for example, for English, 75.01% of the wordforms in the lexicon have only one MSD; 15.27% have 2 MSDs, etc.

Lang	1	2	3	4	5	6	7	8	9
Eng	75.01	15.27	1.11	4.22	3.82	0.41	0.13	0.02	0.00
Rom	85.18	10.75	2.59	0.99	0.07	0.10	0.00	0.00	0.00
Bul	75.26	4.87	1.85	0.61	0.10	0.01	0.00	0.00	0.00
Cze	40.10	21.64	15.05	5.61	3.69	3.89	1.56	0.28	8.17
Slo	34.89	29.00	14.32	4.27	6.69	3.75	0.80	0.58	5.70
Est	72.65	16.16	6.66	3.17	0.25	0.70	0.32	0.06	0.02
Hun	78.02	18.07	3.21	0.58	0.08	0.04	0.00	0.01	0.00

Table 4: MSDs per wordform

Table 5 gives the non-ambiguous wordforms and lemmas.

Language	Wordforms %MSD non-amb	Lemmas %MSD non-amb
English	75.01	76.43
Romanian	85.18	88.24
Bulgarian	75.26	93.41
Czech	40.10	94.80
Slovene	34.89	97.61
Estonian	72.65	88.87
Hungarian	78.02	81.73

Table 5: Non-ambiguous wordforms and lemmas

Czech and Slovene include several words with an exceptionally large number of MSDs (48 and 49 for Czech, and 54, 55, 56 and 57 for Slovene) and their lexicons therefore contain fewer unambiguous wordforms. However, in terms of **lemmas**, the lexicons for Czech, Slovene, and Bulgarian exhibit the lowest ambiguity, indicating that intra-category (inflectional) ambiguity is greatest for these languages.

The Romanian lexicon exhibits the lowest wordform ambiguity (fewer than 15% of the words have more than one MSD). As noted above, the non-ambiguity values in Table 5 result directly from the strategies used to handle syncretism. It would seem that for statistical tagging, encoding syncretism and using "any value" attributes (removable from a tagset encoding) likely leads to greater tagging accuracy, and certainly increases the efficiency of training and tagging (see Tufis & Mason, 1998).

[3] This is valid for the parts of speech shown in Table 3. When considering all the parts of speech, the highest number of MSDs are used for encoding the pronouns.

It is important to note that these figures say relatively little about ambiguity rates in running text; rather, they provide an index of ambiguity according to the encoding schema as well as an index of the degree of homography and syncretism that has been considered by the lexicon designers. Ambiguity in running text is considered in section 4, below.

4. Comparison and Distribution in Corpora

The statistics in the previous section describe distribution and use of MSDs in the lexicon. It is expected that at least some of these statistics will differ for running texts, which represent actual usage of the lexical stock. In this section we present the results of a preliminary investigation of the distribution of MSDs and PoS categories in running text.

The MULTEXT-East corpus includes a parallel component comprised of translations of Orwell's *Nineteen Eighty-Four*. The remainder is a comparable corpus in terms of its size and type, covering two domains: fiction and newspaper articles. Sixteen of the nineteen sub-corpora have been encoded to level 1 according to the specifications of the Corpus Encoding Standard (Ide, 1998), and therefore include encoding for general elements such as paragraphs, titles, notes, etc., and some encoding of dates, proper names, abbreviations, numbers, etc.

The procedure for generating the statistics in this section involved as a first step the segmentation of the encoded texts, followed by the assignment of all applicable MSDs to each lexical token. The resulting MSD-annotated text was then manually disambiguated. To date, we have considered two of the MULTEXT-East languages (Romanian and Slovene) in this analysis, together with English.

Figure 6 summarizes ambiguity information from the corpus of *Nineteen Eighty-Four*.

A "token" is considered to be a lexical unit as identified by an automatic segmentation tool developed in the MULTEXT project -i.e., a word-form, whereas "types" refers to unique tokens. Note that a token is not necessarily an orthographic word: orthographic words may be split into several tokens by the segmenter. For example, the Romanian "da-mi-l" (give it to me) is split into three tokens. Similarly, several orthographic words may be combined into a single token: for example, the Romanian words "de la" are combined into one token "de_la". Within the project, lexicon entries and segmentation rules have been developed hand-in-hand for all the languages, to ensure consistency in the definition of word-forms, etc. (see Ide, 1996).

In the table, %MSD/Unamb is the percentage of tokens in the text which have only one possible MSD based on the information in the lexicon; %PoS/Unamb is the percentage of tokens which, although possibly having several MSDs, have only one PoS assigned to them.

Language	English	Romanian	Slovene
Tokens	5532	101449	5472
Types	1729	14040	2128
%MSD/Unamb	61.4	66.44	26.9
%MSD/Amb	38.6	32.56	73.1
%PoS/Unamb	72.2	70.2	69.8

Table 6. Summary of corpus statistics (English, Romanian[4], Slovene)

The ambiguity percentages here are very different from the same percentages computed for the lexicons (Table 5); this is expected since in the lexicons, a lexical item (word-form MSD) appears only once, while in the running texts the number of occurrences of a given token may be quite large.

Language	%MSD/Unamb	
	Lexicon	Corpus
English	75.01	61.4
Slovene	34.89	26.9
Romanian	85.18	66.44

Table 7: Lexicon and Corpus ambiguity

This indicates some MSD-ambiguous items appear quite frequently in the corpus, while a substantial number of MSD-unambiguous items in the lexicon do not appear there or are not very frequent. Similarly, all three languages show higher PoS ambiguity (or equivalently lower PoS unambiguity) in the corpus than in the lexicon.

5. Conclusion

The paper provides an overview of the morpho-syntactic descriptions, lexicons, and lexical items in the corpus of the MULTEXT-East project, comprising six Central and Easter European languages from three language families together with English as the hub.

A primary contribution of this work is, of course, the provision of widely available lexical and corpus resources for the languages of the project. The complete documentation of the MULTEXT-East project together with HTML corpus samplers is available on the WWW at http://nl.ijs.si/ME/. The entire corpus is available on CD-ROM through the TELRI concerted action (see Erjavec *et al.*, 1998), together with four new translations of *Nineteen Eighty-Four* in Latvian, Lithuanian, Serbian, and Russian. These translations are encoded in the same way as the MULTEXT-East corpus, using the CES specifications, and the Latvian, Lithuanian, and Serbian translations are sentence-segmented and aligned with the English. The CD-ROM is available for research purposes only, on a per-cost basis.

One motivation for the quantitative studies presented in this paper is the need to develop automatic tagging mechanisms for the languages of the project. The first decision that needs to be made here is, of course, choosing the appropriate tagset for each language. While several tagsets exist for the English language, as well as some harmonized tagsets for Western European languages, these tagsets are of limited use for MULTEXT-East due to the

[4] The Romanian data refers to the whole book, while English and Slovene just to the part 1 of the book.

considerable differences between the MULTEXT-East languages (except for Romanian) and Western European languages.

There is little experience in probabilistic tagging of Central and Eastern European languages. The few known results on large tagsets show poor results (Hajic and Hladka, 1998). This is because highly inflected and free word order languages require extremely large tagsets (approximately 1500 for Czech, and potentially millions for Estonian or Hungarian). The corpus size necessary to train a probabilistic tagger to reasonable accuracy (e.g., 95%) with a tagset of this size is on the order of tens of millions of words, which is well beyond the scope of the project.

MULTEXT-East included a phase where the lexical MSDs for each language were mapped to a significantly smaller "corpus tagset", chosen to enable probabilistic tagging. It is well known that collocational stochastic tagging methods (digram, trigram, n-gram) cannot discriminate all the fine-grained distinctions made in the MSDs. Therefore the corpus tagsets comprise broader categories that collapse or eliminate MSD values or (in some cases) features which a stochastic tagger cannot reliably disambiguate. By separating the morpho-syntactic specifications and the corpus tagset, the latter can be developed and fine-tuned, based on experimentation with a stochastic tagger, using the morpho-syntactic descriptions as the starting point.

Tufis (1998) and Tufis & Mason (1998) propose a methodology for tagset design and probabilistic tagging, called "tiered tagging", which attempts to find a middle ground between (large) fine-grained morpho-syntactic tagsets and the resources needed in statistical disambiguation. This approach has been extremely encouraging for the first experiments on Romanian (accuracy higher than 97% with 611 tags and a hand-disambiguated training corpus of less than 100,000 words). We plan to extend the tiered-tagging experiments to all the languages in the project.

Acknowledgements

The authors would like to acknowledge the contribution of the following people to the development of the lexical specifications: M. Monachini, R.Pavlov, L.Dimitrova, L.Sinapova and K.Simov, V.Petkevic, H.J.Kaalep, L.Tihanyi, A.M.Barbu, and P.Holozan. We would also like to acknowledge V. Patrascu and G. Priest-Dorman for their work on the preparation of the corpus and generation of some of the statistics.

Bibliographical References

Bel, N., Calzolari, N. and Monachini, M., eds. (1995). Common Specifications and Notation for Lexicon Encoding. Deliverable 1.6.1. MULTEXT Project LRE 62-050. http://www.lpl.univ.aix.fr/projects/multext/LEX/LEX1.html

Carpenter, B. (1992). The Logic of Typed Feature Structures. *Tracts in Theoretical Computer Science,* Cambridge University Press.

Erjavec, T., Ide, N., Petkevic, V., & Véronis, J. (1996). MULTEXT-East: Multilingual Text, Tools and Corpora for Central and Eastern European Languages. *Proceedings of the First TELRI European Seminar,* 87-98. Project description, documentation and resources available at http://nl.ijs.si/ME/.

Erjavec, T. & Ide, N. (1998). The MULTEXT-East Corpus. *First International Language Resources and Evaluation Conference,* Granada, Spain (this volume).

Erjavec,T., Lawson, A., & Romary, L.(1998). East meets West: Producing Multilingual Resources in a European Context. *First International Language Resources and Evaluation Conference,* Granada, Spain (this volume).

Erjavec, T., Monachini, M., eds.(1997). Specifications and Notation for Lexicon Encoding of Eastern Languages. Deliverable 1.1F. MULTEXT-East Project COP-106. http://nl.ijs.si/ME/

Hajic, J., Hladka, B. (1998). Czech Language Processing / POS Tagging. *First International Language Resources and Evaluation Conference,* Granada, Spain (this volume).

Ide, N., ed. (1996). MULTEXT-East Language-Specific Resources. Deliverable D1.2. MULTEXT-East Project COP 106. http://www.1pl.univ-aix.fr/projects/MULTEXT-East/MTE2.html

Ide, N. (1998). The Corpus Encoding Specification: SGML Guidelines for Encoding Linguistic Corpora. *First International Language Resources and Evaluation Conference,* Granada, Spain (this volume). See also http://www.cs.vassar.edu/CES/

Ide, N., & Véronis, J. (1994). MULTEXT (Multilingual Tools and Corpora). *Proceedings of the 14th International Conference on Computational Linguistics,* COLING'94, Kyoto, Japan, 90-96.

Mason, O., & Tufis, D. (1997). Probabilistic Tagging in a Multi-lingual Environment: Making an English Tagger Understand Romanian. *Proceeding of the Third International TELRI Seminar*, Montecatini, October 1997, 165-68.

Monachini M., & Calzolari, N., Eds. (1996). Synopsis and Comparison of Morpho-syntactic Phenomena Encoded in Lexicons and in Corpora: A Common Proposal and Application to European Languages. EAGLES document EAG-CLWG-MORPHSYN/R, Pisa.

Tufis, D., Barbu, A., M., Patrascu, V., Rotariu, G., & Popescu C. (1997). Corpora and Corpus-Based Morpho-Lexical Processing. In Tufis D., Andersen P. (eds.): *Recent Advances in Romanian Language Technology*, Editura Academiei, Bucharest, 35-56.

Tufis, D. (1998). Tiered Tagging. *International Journal on Information Science and Technology*, 1:2, Editura Academiei, Bucharest, 1998

Tufis, D. & Mason, O. (1998). Tagging Romanian texts: A Case Study for QTAG, A Language Independent Probabilistic Tagger. *First International Language Resources and Evaluation Conference,* Granada, Spain (this volume).

Tzoukermann, E., Radev, D. (1997). Tagging French Without Lexical Probabilities: Combining Linguistic Knowledge and Statistical Learning. *cmp-lg/9/10002*, 10 October, 1997.

Appendix: Tables for NOUN

11 Positions

```
**** **** **** **** **** ---- ---- ---- ---- ---- ----
PoS  Type Gend Numb Case Def  Cltc Anim OwnN OwnP OwdN
**** **** **** **** **** ---- ---- ---- ---- ---- ----
```

P	ATT		VAL	C	EN	RO	SL	CS	BG	ET	HU
					x	x	x	x	x	x	x
1	Type		common	c	x	x	x	x	x	x	x
			proper	p	x	x	x	x	x	x	x
2	Gender		masculine	m	x	x	x	x	x		
			feminine	f	x	x	x	x	x		
			neuter	n	x	x	x	x	x		
3	Number		singular	s	x	x	x	x	x	x	x
			plural	p	x	x	x	x	x	x	x
			dual	d			x	x			
		l.s.	count	t					x		
4	Case		nominative	n			x	x	x	x	x
			genitive	g			x	x		x	x
			dative	d			x	x			x
			accusative	a			x	x			x
			vocative	v		x		x	x		
			locative	l			x	x			
			instrumental	i			x	x			x
		l.s.	direct	r		x					
		l.s.	oblique	o		x					
		l.s.	partitive	1						x	
			illative	x						x	x
			inessive	2						x	x
			elative	e						x	x
			allative	t						x	x
			adessive	3						x	x
			ablative	b						x	x
		l.s.	translative	4						x	
			terminative	9						x	x
			essive	w						x	x
		l.s.	abessive	5						x	
		l.s.	komitative	k						x	
		l.s.	aditive	7						x	
		l.s.	temporalis	m							x
		l.s.	causalis	c							x
		l.s.	sublative	s							x
		l.s.	delative	h							x
		l.s.	sociative	q							x
		l.s.	factive	y							x
		l.s.	superessive	p							x
		l.s.	distributive	u							x
5	Definiteness		no	n		x			x		
			yes	y		x			x		
		l.s.	short_art	s					x		
		l.s.	full_art	f					x		
6	Clitic		no	n		x					
			yes	y		x					
7	Animate		no	n			x	x			x
			yes	y			x	x			x
8	Owner_Number		singular	s							x
			plural	p							x
8	Owner_Number		singular	s							x
			plural	p							x
9	Owner_Person		first	1							x
			second	2							x
			third	3							x
10	Owned_Number		singular	s							x
			plural	p							x
					EN	RO	SL	CS	BG	ET	HU