On the compatibility of lexical resources for NooJ

Ranka Stanković, Miloš Utvić, Duško Vitas, Cvetana Krstev, Ivan Obradović

On the Compatibility of Lexical Resources for NooJ

Ranka Stanković, Miloš Utvić, Duško Vitas, Cvetana Krstev and Ivan Obradović

Abstract

Lexical resources for many languages are provided for the NooJ linguistic development environment. Meta-data descriptions of morphosyntactic and semantic properties of these languages and their resources are a mandatory part of each language module. In this paper we analyze how well the meta-data actually describe resources for a chosen subset of languages and to what extent are they compatible across languages to support multilingual processing. We show that there is place for improvement in both directions.

Introduction: Motivation, resources and task

Lexical resources for NooJ are now available in a considerable number of different languages. The compatibility of these monolingual resources, namely the extent to which they mutually correspond is thus becoming an important issue, especially given the perspective of the development of an open source version of NooJ within the framework of the CESAR project[1]. Namely, it is expected for NooJ to become an open source product running on various platforms, independent from Microsoft OS and its components, albeit compatible with them. From the point of view of Serbian, there was additional interest in this issue in view of the fact that the development of a new Serbian module for NooJ is underway.

The analysis of compatibility of NooJ resources outlined in this paper was tackled by a comparison of the annotation systems for morphological, syntactic and semantic information used in NooJ resources for Bulgarian, Croatian, English, French, Hungarian, Polish and Serbian. In addition to that, an analysis of the results obtained by application of these resources to

[1] http://www.meta-net.eu/projects/cesar/

texts of Jules Verne's novel "Around the world in eighty days" in the same languages was performed. These seven languages were selected due to the fact that both NooJ resources and aligned versions of this novel were available for them. The resources used in the analysis originated from modules available from the official NooJ resources web page[2].

The next section offers a comparison of annotation systems used in NooJ resources based on an analysis of the Dictionary Properties' Definition files. The section that follows outlines the results of lexical analysis of the application of NooJ resources to aligned texts. Finally, a section is dedicated to some related issues of compatibility and standardization. The paper ends with concluding remarks.

Comparison of annotation systems

Morphological, syntactic and semantic information in NooJ resources is represented by codes or tags, pertaining to morpho-syntactic and semantic categories, their properties and the features or values of these properties. These codes, such as N for nouns, Top for toponyms, Hyd for hydronyms, can be assigned to lexical entries and subsequently used in NooJ queries, e.g. <N+Top> for selecting nouns representing toponyms or <N+Top-Hyd>, for nouns representing toponyms that are not hydronyms.

The information on property values for specific categories can be found as metadata in the Dictionary Properties' Definition files, with the extension "def", which will be referred to as the *.def files. The general form of these metadata is "category_property=value". For example, N_Nb=s+p[3] means that the property grammatical number (Nb), in the case of category nouns (N), can have two values: s (singular) and p (plural). In addition to that, the *.def files offer the information on values of properties that inflect (see Table 4).

The content of *.def files enables NooJ to recognize properties and their values and they can also improve the visibility of representations of dictionaries in tabular form. The first step in the analysis of compatibility of resources was precisely a comparison of metadata, namely of the content of *.def files for resources of the seven selected languages.

Table 1 presents the results of the analysis of *.def files in the case of nouns. Value domains of two properties, one of them appearing in all languages and the other appearing in five out of seven, namely,

[2] NooJ linguistic modules http://www.nooj4nlp.net/pages/resources.html
[3] In the new versions of NooJ the symbol "+" was replaced by "|". We are using the old symbol for greater readability.

grammatical number and gender, are given in separate columns (2 and 3), whereas the last column shows all other properties, and their value domains.

	Nb	Gen	Specific
Fr	s+p	m+f	N_Struct=AAN+AN+E01+NA+NAA+NAN+NCN+NDN+NN+NPN+NPNPN+NPV+NX+NX4+PN+VN+VV+VX+XA+XN+XV+XX N_Sem=Anl+Hum+HumColl+Conc+Alim+Anim+Abst+Pol+PR
En	s+p	m+f +n	N_Struct=2XN+NCN+NPN+NsN+NX+PMCO+XN N_Distribution=Abst+Anl+AnlColl+Coll+Conc+Hum+HumColl+PR+Ntime+Ntps+Unit
Hu	PL		N_Case=NOM+ACC+DAT+ILL+INE+ELA+ALL+ADE+ABL+SUB+SUP+DEL+INS+FAC+CAU+FOR+TER+SOC+ESS+TEM N_Posessive=PS N_OwnerNumber=ps_sg+ps_pl N_OwnerPerson=ps_1+ps_2+ps_3 N_Possessee=ANAP
Pl	s+p	mo+ mr+ mz+f +n	N_Sem=Hum+Conc+Abstr+Org+Text+ConcColl+Cpmc+Immeub+Qual+Anim+Loc+Pdc+Sent+Quant+Mat+Liq+Alim+Vehicl+Pr+Tmp+Atm+Geom+CollHum+CollImmeub+Mach
Hr	s+p	m+f +n	N_Case= Nom+G+D+Acc+Voc+L+I N_Type=c+vl+r N_Sem=geo+etn+ust+ime+vr+kr+inc
Bg	s+p		N_Case=v N_Definiteness=0+d+h+l+o N_Counting_Form=c
Sr	s+p +w	m+f +n	N_Animacy=v+q N_Distribution=Adj+Aug+Bot+Col+Coll+Dem+Der*+Ek+El+FG+Hip+Hum+HumN+Ijk+Ik+HumColl+MG+Myth+Neg+NG+Num+NumN+Pej+Pl+PT+Relig +Zool

Table 1. Analysis of *.def metadata: NOUNS

In Table 2 we present the corresponding results for properties of adjectives shared by the majority of languages. In this case, grammatical number and gender are also the two predominant categories.

As for the verbs, represented in Table 3, it can be observed that there are three verb properties that *.def files for all seven languages share: number, person and tense or mood, whereas the gender property is this time shared by four out of seven languages – in Serbian the values of this property were omitted due to oversight.

	Nb	Gender
Fr	S+p	m+f
En		
Hu	PL	
Pl	S+p	mo+mr+mz+f+n
Hr	S+p	m+f+n
Bg	S+p	m+f+n
Sr	S+p+w	m+f+n

Table 2. Analysis of *.def metadata: ADJECTIVES

	Nb	Gender	Pers	Tense/ Mood
Fr	S+p	m+f	1+2+3	INF+G+PP+PR+IP+S+C+I+PS+SI+F
En	S+p		1+2+3	G+INF+PP+PR+PT
Hu	E+t		1+2+3	K+M+F+P
Pl	S+p	mo+mr+ mz+f+n	1+2+3	F+G+H+J+K+P+Q+S+T+U+Z
Hr	S+p	m+f+n	1+2+3	INF+INFk+PR+IMP+AO+PDR+PDT+ IMT+GPS+GPP+PER+FUT1+FUT2
Bg	S+p	f+m+n	1+2+3	R+E+D+I
Sr	S+p+w		x+y+z	W+P+A+I+Y+G+T+F+S+X

Table 3. Analysis of *.def metadata: VERBS

The final table pertaining to results of the analysis of *.def files, namely Table 4, shows how inflection properties of languages, that is, the features that inflect are represented in *.def files for the seven observed languages.

	Inflection
Fr	M+f+s+p+1+2+3+Inf+G+PP+PR+Ip+S+C+I+PS+IS+F
En	1+2+3+m+f+s+p+G+PR+PP+INF+PT
Hu	NOM+ACC+DAT+ILL+INE+ELA+ALL+ADE+ABL+SUB+SUP+DE L+INS+FAC+CAU+FOR+TER+SOC+ESS+PL+PS+ps_sg+ps_pl+ps_1 +ps_2+ps_3+ANAP+SUPER+COMP+T+I+KIJ+MULT+FELT+FELSZ +e+t+1+2+3+FF+FOK+VMIB+VMIF+VMIA+VHIN+VHINN+Os+sA g+HAT+UN
Pl	M+f+s+p+Bi+Ce+Do+In+Lo+Mi+Wo+F+G+H+J+K+P+Q+S+T+U+Z
Hr	INF+INFk+PR+PRk+PRd+IMP+AO+AO2+PDR+PDT+IMT+GPS+GP P+s+p+1+2+3+m+f+n+np+cp+sp
Bg	s+p+0+d+h+l+o+v+c+f+m+n+y+z+i+a+t+1+2+3+b+w+x+A+B+C+K+ P+F+M+N+R+E+D+I+Y+Z+X+W+Q
Sr	m+f+n+1+2+3+4+5+6+7+s+p+w+v+q+g+i+a+b+c+d+k+W+P+A+I+Y +G+T+F+S+X+x+y+z+h

Table 4. Analysis of *.def metadata: INFLECTION

It is obvious from these tables that only a few properties are shared by all seven languages, and that the number of shared features is also very low. In addition to that, the same features are often differently coded, as for example the three values for person in the case of Serbian verbs, which are coded with x, y and z, as opposed to all other languages which use 1, 2 and 3. While the latter can be considered a mainly technical issue, there are some other, much more serious differences pertaining to the representation of specific phenomena. They are often a consequence of established practices of representation of these phenomena in grammars. For example, the number category does not exist in English and French, which use the determiner category instead, whereas other languages do not recognize the determiner category at all. As to properties, there are, for example, three types of masculine gender in Polish (animate person, animate animal, and inanimate), while Serbian and Croatian do not recognize this difference, although a similar differentiation for the masculine gender exists in both languages. Instead, there is a property in Serbian – animacy – with two possible values, animate (v) and inanimate (q), while the Croatian *.def file does not envisage such a possibility. A standardization effort aimed at defining at least common feature codes would certainly be a step forward.

Linguistic analysis results

In the next step, NooJ resources were applied to texts of Verne's novel and a linguistic analysis of the results obtained was performed. The analyzed texts were in XML format in compliance with TEI, and their alignment was performed at the sentence level using the ACIDE system (Obradović et al 2008), which can handle aligned texts in various formats (TEI, TMX, html, Vanilla). During the alignment process, the texts were segmented in such a way as to establish a one-to-one correspondence between the aligned segments and the original text in French.

An example follows, showing the introductory chapter title and its first sentence in each of the seven languages:

<tu><text lng="fr"**> **
<head>I Dans lequel Phileas Fogg et Passepartout s'acceptent réciproquement l'un comme maître, l'autre comme domestique**</head>**
** <seg>**En l'année 1872, la maison portant le numéro 7 de Saville-row, Burlington Gardens -- maison dans laquelle Sheridan mourut en 1814 --, était habitée par Phileas Fogg, esq., l'un des membres les plus singuliers et les plus remarqués du Reform-Club de Londres, ...**</seg> ...**
<text lng="en"**> **

<head>Chapter I in which Phileas Fogg and Passepartout accept each other, the one as master, the other as man**</head>**
** <seg>**Mr. Phileas Fogg lived, in 1872, at No. 7, Saville Row, Burlington Gardens, the house in which Sheridan died in 1814. He was one of the most noticeable members of the Reform Club, …**</seg> ...**

<text lng="hu"**> **
<head>ELSÕ FEJEZET Egy megállapodásról, melynek értelmében Phileas Fogg és Passepartout kölcsönösen gazdává, illetve inassá fogadja egymást**</head>**
** <seg>**1872-ben a Saville Row 7. szám alatt (Burlington Gardens) - e házban halt meg 1814-ben Sheridan - Phileas Fogg esq., a londoni Reform Club tagja lakott, aki, bár nyilvánvalóan minden igyekezetével azon fáradozott, …**</seg> ...**

<text lng="pl"**> **
<head>Rozdział pierwszy: W KTÓRYM FILEAS FOGG I PASSEPARTOUT AKCEPTUJĄ SIĘ NAWZAJEM **</head>**
** <seg>** W roku 1872 dom opatrzony numerem siódmym przy Saville Row w Burlington Gardens - dom, w którym w 1814 zmarł Sheridan * - zamieszkiwał szanowny pan Fileas Fogg, jeden z najwybitniejszych i najbardziej oryginalnych członków klubu "Reforma" w Londynie, ...**</seg> ...**

<text lng="hr"**> **
<head>PRVO POGLAVLjE U kojem se Phileas Fogg i Passepartout uzajamno priznaju, prvi kao gospodar, a drugi kao sluga**</head>**
** <seg>**U ulici Saville row kod Burlington-Gardena, u kući broj 7, u kojoj je 1816. umro Sheridan, stanovao je 1872. Phileas Fogge sq. najčudnovatiji i najistaknutiji član kluba Reform u Londonu, ...**</seg> ...**

<text lng="bg"**> **
<head> I глава. Филиас Фог и Паспарту се харесват като господар и прислужник **</head>**
** <seg>**През 1872 година в къщата на “Савил роу” № 7, Бърлингтън Гардънс – същата, в която през 1814 година почина Шеридан, – сега живееше Филиас Фог. Той беше един от най-странните и видни членове на Реформаторския лондонски клуб, ...**</seg> ...**

<text lng="sr"**> **
<head>I GLAVA: u kojoj Fileas Fog i Paspartu stupaju u međusobni odnos gospodara i sluge**.</head>**
** <seg>**Godine 1872, u kući broj 7 u Ulici Sevil-rou Bar-lington Gardenz, u kojoj je 1816.godine umro Šeridan, stanovao je gospodin Fileas Fog, jedan od najčudnovatijih i najzapaženijih članova londonskog Reform-kluba, …**</seg> …</tu>**

It should be noted that the resources used in this analysis, namely those that were available at the NooJ resource web page, often do not represent “full” and updated versions of resources. Results are also influenced by the choice of specific resources and the order they were applied for

languages with more than one dictionary and morphology graph. Hence, they should be taken with a grain of salt.

An interesting result of the analysis was that categories exist in dictionaries that cannot be found in corresponding *.def files, which should not be the case. For example, when Polish resources are concerned, NUM, ADV, PREP are not found in the *.def file, but they nevertheless appear in the text. The Bulgarian *.def file does not define the gender category, while it appears in Bulgarian dictionaries: masculine gender - господар,N+M, feminine gender - глава,N+F and neutral gender - лице,N+NE. Conversely, which is though not unexpected, there were *.def categories which did not appear in text dictionaries.

In addition to that, further review of the analyzed texts after the application of resources revealed that important differences often exist between semantic codes in metadata (in the *.def file) and those in the text dictionary. For example, in Croatian, the semantic code *cons* appears in the text dictionary (hraniti,V+FLX=BRANITI+Sem=cons+Prelaz=pov), although it does not exist in the *.def file. Conversely, semantic codes *geo* (place), *etn* (ethnic), *ust* (institution) etc. appear in the *.def file but they cannot be found in the text dictionary despite the fact that the text contains nouns that belong to these semantic categories. In French, N+PRENOM extracts both given names and family names, although the code used suggests family names only. In addition to that, PRENOM as a category does not appear in the *.def file, whereas the category PR, which appears in the *.def file, cannot be found in the text dictionary.

The results of the analysis point to the need of making a stronger connection between metadata from *.def files and resources – NooJ dictionaries and grammars. We suggest the replacement of the textual form of metadata by a more formal form in XML format, with a defined schema which would enable a stricter control and tighter relation between the *.def file and the dictionary. Given that Data Categories are standardized (ISO 2009) it would be good if some form of establishing a relation between them and NooJ existed.

Along the lines of the observation that it would be particularly important if the concept of local grammar could be generalized in such a way that one grammar extracts (approximately) the same concepts from aligned texts in different languages, using the information stored in the system of electronic dictionaries (Vitas et al. 2008), in the next step of the analysis, an attempt was made to construct NooJ expressions that could extract the same concepts in all languages.

However, that turned out to be a rather difficult task, due to the fact that semantic codes were different in all languages, and mappings of

corresponding codes did not exist and were difficult to establish. That was in compliance with the observation that the solution of such problems depends on at least two components (Vitas et al. 2008). One of them is how "faithful" is the translation compared to the original, and the other – the manner in which specific information is marked in the system of electronic dictionaries. The first condition was satisfied to a great extent, but the second, as it has already been demonstrated, was not.

Instead of universal NooJ expressions, which turned out to be unfeasible, mutually corresponding NooJ expressions were produced on basis of semantic codes discovered in the previous step through the analysis of annotations (text dictionary). These expressions were subsequently applied to appropriate texts, and the results are shown in Table 5. There are considerable differences, as for example in the case of proper names, which range from 122 in Croatian to 1364 in Serbian. However, these data are also not fully comparable, as semantic categories in Serbian are hierarchically organized – for example, +NProp (proper name) is hierarchically superior to categories +Dr (state) and +Gr (settlement), whereas this is not the case, for example with the French dictionary. Moreover, in French, the category +PAYS plays a double role – to mark that a lemma represents a country (e.g., Chine,N+PAYS) and to mark the country to which a city belongs (e.g. Yokohama,N+VILLE +PAYS="Japon"). This explains, to some extent, the difference in the number of recognized "states" in the French and Serbian texts (Table 5). On the other hand, inhabitants of states and cities are not marked as proper names in the French dictionary (e.g. Anglais,N+z1+m+s) as opposed to the Serbian dictionary (e.g. Englez,N+m+s+1+v+Hum+NProp+Top+Inh).

Another interesting illustration is the proper name *Londres*, which appears in the French original in the first sentence of the text. In Polish (*Londynie*) and Croatian (*Londonu*) translations it is also a proper name, but it does not appear in the English translation at all. In Hungarian (*londoni*), Bulgarian (*лондонски*) and Serbian (*londonskog*) translations it appears in the form of a relational adjective. In dictionaries of respective languages, these words are represented as follows:

FR: Londres,N+VILLE+PAYS="Angleterre"

PL: -

HR: Londonu,London,N+Case=D+Nb=s+Type=vl+Gender=m

HU: - (London,N+ne)

BG: -

SR: londonskog,londonski,A+a+d+m+s+2+g+PosQ+Top+IsoGB

During the analysis, concordances, annotated text and semantic codes were manually inspected and compared, in an attempt to determine what

precisely each specific semantic code stands for, as in general they represent abbreviations, and a description of the meaning of specific codes is nowhere available.

	Fr	En	Hu	Pl	Hr	Sr
Proper	+PRENOM 475	+PR 449	+ne 895		+vl 122	+Nprop 1364
Pays	+PAYS 371					+Dr 194
Towns	+VILLE 291					+Gr 411
Hum	+Hum 250	+Hum 2706		+Hum 414		+Hum 3661
HumColl	+HumColl 20	+HumColl 139		+CollHum 94		

Table 5. NooJ expressions used with noun patterns

Another experiment along the same lines was performed with NooJ graphs. In order to analyze the results of the application of same/similar graphs in the text, a simplified graph aimed at extracting measures was produced. The graph does not contain all measures, but rather represents a version simplified for this occasion. Figure 1 depicts the main graphs for English, Serbian and Croatian texts.

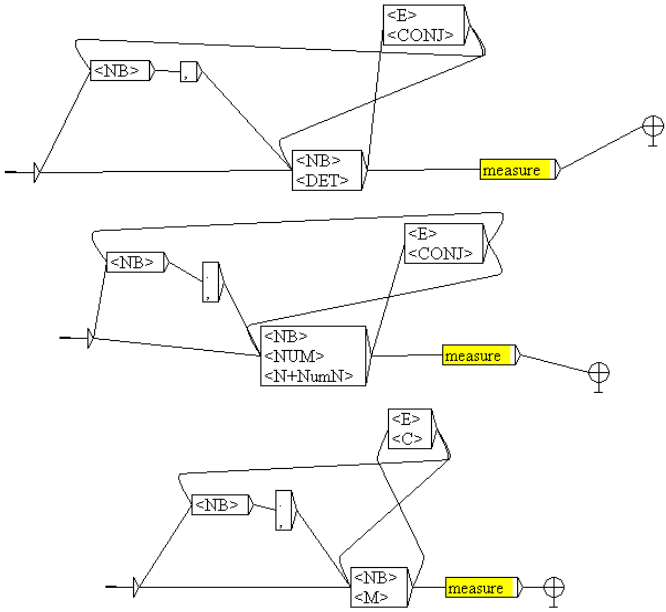

Fig 1. Graph examples for measure expressions (En-top, Sr-middle, Hr-bottom)

The differences in the graphs result from different ways categories and properties are determined and annotated. All three graphs recognize numbers written in digits by the <NB> pattern, whereas for numbers written in words the English graph uses the <DET> pattern, which also recognizes other determiners, whereas Croatian uses <M> and Serbian uses <NUM> and <N+NumN> for numerical nouns. All these graphs, after they recognize one or more numbers connected by a comma or a conjunction (<CONJ> or <C>), invoke their specific „measure" sub-graphs which recognize different measure units, either as character strings: kilometer, meter, foot, feet, pond, mile, or as patterns <N+Unit>, <N+Mes>.

An example of aligned concordances obtained by the application of these three graphs on the text is:

> En: … embraces ***fourteen hundred thousand square miles***, upon which is spread unequally a population of one hundred and eighty millions of souls.
>
> Hr: …zaprema površinu od ***četrdeset tisuća četvornih milja***, na kojem je nejednako rasprostranjeno sto i osamdeset milijuna stanovnika.
>
> Sr: …obuhvata površinu od ***četiri stotine hiljada kvadratnih milja***, na kojoj je nejednako raspoređeno stanovništvo od sto osamdeset miliona ljudi.

However, a more detailed analysis of aligned text revealed examples where a measure expression was recognized in some but not in all languages. For instance, a measure expression "four hundred horse power" was recognized in English and Croatian, but not in Serbian, because the colloquial "horses" has been used instead of "horse-power". The same example shows that the expression "seventeen hundred and seventy tons" was not recognized in Croatian because of the insertion *brutto registarskih* 'gross register'.

> En: …built of iron, weighing about ***seventeen hundred and seventy tons***, and with engines of ***four hundred horse-power***.
>
> Hr: …brod od čelika i na vijak; istiskivao je tisuću sedam stotina i šezdeset brutto registarskih tona, a strojevi su imali ***četiri stotine konjskih sila***.
>
> Sr: …bio je gvozdeni brod sa propelerom i imao je zapreminu od bruto ***hiljadu sedam stotina tona***, sa snagom od četiri stotine konja.

Another example is the following sequence, which was not recognized in the Croatian text, because *stotine*, a form of the number *stotina*

'hundred' was marked in the NooJ dictionary only as a noun (without any further explanation) and not as a number:

Hr: …čelični parobrod na vijak sa dvije palube, koji istiskuje dvije tisuće i osam stotina tona,…

En: … built of iron, of ***two thousand eight hundred tons*** burden,…

However, the majority of omissions were not made because various encodings of numerals were not captured, but were rather due to peculiarities of translations and inadequacies of simplified graphs.

Finally, there was some incorrect recognition as well. Some of it results from the usual ambiguity, as illustrated by the following Serbian example (*vrsta* is a measurement unit, but also has another meaning, namely 'type'):

Sr: Imao je sve ***3 vrste*** jedara i mogao je opremiti...

En: she carried ***brigantine, foresail, storm-jib, and standing-jib***, and was well rigged for running before the wind...

The next source of false recognition is more serious. Namely, in the English e-dictionary numbers are marked as determiners, which represent a wider category, as illustrated by the following example.

En: Squarely in his armchair, ***his feet*** close together like those...

Standardization

A starting point for the standardization of grammatical attributes, their values and codes can be find in MULTEX-East, a valuable resource whose fourth edition, published in 2010, includes (to a certain extent) as much as 17 languages (Erjavec 2010). One of its especially convenient features is that languages encompassed by the CESAR project, and analyzed in previous sections (except French) are all to be found in MULTEXT-East.

The presentation of the observed languages in MULTEXT-East and their relation with the description in NooJ resources cannot be tackled here in detail - for more details see (Erjavec et al. 2003). We will thus only offer a couple of illustrative examples pertaining to nouns, and three of their attributes in particular: category, gender and number. The category attribute has three possible values in MULTEXT-East: common, proper, and gerund. NooJ resources for all observed languages treat all nouns by

default as common, and hence do not envisage a specific tag for them. When proper names are concerned, it has been observed that the annotation system differs from one language to another. According to MULTEXT-East the attribute gerund is only recognized in Polish, whereas for other languages it is not applicable. In the Polish e-dictionary available under NooJ it is not specifically coded. As for the number, MULTEXT-East envisages five values: singular, plural, dual, count, and collective. For English, Polish, Croatian and Hungarian, only singular and plural are applicable, while Serbian and Bulgarian also use count. In Serbian this number corresponds to paucal (a small number – 2, 3 and 4). However, the collective category is not totally inapplicable, and is thus coded with the tags +Coll and +HumColl in Serbian and +ConcColl and +CollHum in Polish (see Table 2). According to the same table the Bulgarian NooJ resources do not recognize values other than singular and plural. Finally all observed languages, except Hungarian, according to MULTEXT-East, recognize the gender values masculine, feminine and neutral, whereas none of them uses the value common. This, as we have seen, differs from the representation of this attribute for Polish in NooJ resources – the traditional Polish representation has probably been abandoned for the sake of fitting into the description of the MULTEXT-East schema.

This brief and non-exhaustive analysis shows that special attention should be paid to morphosyntactic representation in NooJ resources if useful and comparable resources are to be realized.

Concluding remarks

The analysis of compatibility of NooJ resources outlined in this paper revealed a considerable heterogeneity at various levels. Not many properties and features are shared by seven observed languages, and features are also often differently coded. The analysis also revealed categories that appeared in dictionaries but not in corresponding *.def files, which was quite unexpected. Important differences were also found between semantic codes in the *.def files and those in the text dictionaries.

All results point to the need of establishing a better relation between metadata from *.def files and NooJ resources. A more formal form, preferably in XML format with a corresponding schema would certainly be very helpful. This would contribute to the desired, more structured relation between *.def file and dictionaries. A starting point in what would be a standardization effort pertaining to grammatical attributes and their corresponding values and codes can be found in MULTEXT-East project.

References

Erjavec, T., 2010, MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC'10*, Paris: ELRA. (http://nl.ijs.si/ME/V4/)

Erjavec, T., Krstev, C., Petkevič, V., Simov, K., Tadić, M., Vitas, D., 2003, "The MULTEXT -East Morphosyntactic Specifications for Slavic Languages", in *Proc. of the Workshop on Morphological Processing of Slavic Languages: EACL 2003*, Budapest, Hungary, eds. T. Erjavec and D. Vitas, pp. 25-32

Obradović, I. Stanković, R., Utvić, M. 2008, "Integrisano okruženje za pripremu paralelizovanog korpusa" (An integrated environment for the preparation of aligned corpora), *Zbornik radova međunarodnog simpozijuma Razlike između bosanskog/bošnjačkog, hrvatskog i srpskog jezika, Graz, Austria, April 2007, B. Tošović (ur.), pp.563-578, Münster, Germany, LITVerlag*.

Vitas, D., Koeva, S., Krstev, C., Obradović, I., 2008, "Tour du monde through the dictionaries", *Actes du 27eme Colloque International sur le Lexique et la Gammaire, L'Aquila*, 10-13 septembre 2008, eds. M. Constant, T, Nakamura, M. De Gioia, S. Vecchiato, Universite Paris-Est, Institut Gaspard-Monge.

ISO. (2009) ISO 12620 Terminology and other language and content resources – Data Categories – Specification of data categories and management of a data category registry for language resources.