ON META-DESCRIPTIONS FOR CROSS-LINGUISTIC ELECTRONIC LINGUISTIC DATA

Pirkko Suihkonen
Max Planck Institute for Evolutionary Anthropology, Department of Linguistics
Inselstrasse 22, D-04103 Leizpig, Germany
suihkonen@eva.mpg.de

1. INTRODUCTION

The goal of this paper is to examine the typology of the meta-descriptions for machine-readable linguistic data as related to linguistic research work. The examples will be taken from the text corpora located on the University of Helsinki Language Corpus Server at the University of Helsinki, Department of General Linguistics, but the principles discussed in this paper should also be useful for the meta-descriptions of other types of electronic linguistic data. The paper is structured as follows. In section 2, we discuss the documentation and standards developed for describing electronic linguistic data. In section 3, the organisation of the meta-descriptions is presented, and in section 4, a short summary of the principles followed in this paper will be given. The paper deals with the background to the practical work that needs to be done in order to adapt electronic corpora taken from typologically diverse languages and prepared according to different kinds of principles for use as research material in cross-linguistic studies.

2. ON LINGUISTIC TOOLS AND DATABANKS: THE BACKGROUND FOR THE META-DESCRIPTION

During the last few years, the number of electronic linguistic corpora located at various universities and research centres across the globe has increased enormously. As a consequence of the history of creating electronic linguistic data, there is great variety in their systems of preparation. Various linguistic databanks have also been developed within the framework of national or international projects whose goal has been to collect and annotate electronic linguistic data. We have taken our examples from Europe, where many large corpus projects have been organised by institutions working under or connected with the European Union. The large projects MULTEXT (cf. http://www.lpl.univ-aix.fr/projects/multext/, and also http://www.ling.umu.se/asv/multext.htm/) and LE-PAROLE http://www.linglink.lu/le/projects/le-parole/) are examples of this work. Within these projects, large electronic corpora were developed in several European languages (cf. http://albion.ncl.ac.uk/esp-syn/text/5304.html; Menon & Modiano 1994). Much of the work to standardise machine-readable linguistic data has been carried out within the framework of the Expert Advisory Group on Language Engineering Standards, EAGLES that have also developed the Corpus Encoding Standard (CES) that is used in several corpora prepared under the auspices of the European Union (http://www.cs.vassar.edu/CES/CES-P.html). CES is an application of SGML "(ISO 8879: 1986, Information Processing--Text and Office Systems--Standard Generalized Markup Language) compliant with the specifications of the TEI Guidelines for Electronic Text Encoding and Interchange of the Text Encoding Initiative" (http://www.cs.vassar.edu/CES/). The recommendations for standards developed within EAGLES apply to text corpora, computational lexicons, the evaluation of natural language processing systems, computational linguistic formalisms and spoken language systems (http://www.ilc.pi.cnr.it/; http://www.ilc.pi.cnr.it/EAGLES/). In the UHLCS, there are also corpora which are prepared under the auspices of the corpus project LE-PAROLE.

The University of Helsinki Language Corpus Server located at the Department of General Linguistics in Helsinki (UHLCS; http://ling.helsinki.fi/uhlcs/) is an example of the language servers that has its origin in the activities of individual researchers interested in using the facilities offered by machine-readable electronic database. The computer corpora in the UHCLS are heterogeneous and represent different kinds of formats. There are large text collections of Finnish, Swedish, English, German and Russian, and also morphologically analyzed texts from several Uralic languages: Ingrian, Ume Saami, Erza Mordvin, Komi Zyrian, Khanty, Hill Mari, Selkup, Nenets, as well as Turkic Chuvash (Suihkonen 1998). A considerable part of corpora consists of data received from the Institute for Bible Translation. Most of these texts are written in the Cyrillic alphabet, and have been transliterated into the Latin-1 alphabet that is available in the UNIX operating system.

Meta-descriptions for distinguishing information from electronic data available on the network have been the topic of several projects over the past few years. Classifications done by libraries or projects form the basis of the meta-descriptions for documents located in various archives, libraries and museums (http://lcweb.loc.gov/ead/; http://sunsite.berkeley.edu/ead/; http://purl.org/dc/documents/rec-dces-199809.htm#; cf. also http://www.w3.org/-PICS/principles.html). One of the most important large-scale efforts in developing standards for meta-descriptions was the Text Encoding Initiative (the TEI; http://www.-tei-c.org/uic/ftp/P4beta/index.htm; Ide & Véronis 1995). The TEI project, which originated in 1987, has worked with document descriptions made with the help of the Standard Markup Language (SGML). The SGML that is an official standard developed by the Text and Office Systems Subcommittee of the International Standards Organisation (ISO; the SGML is the standard ISO 8879) contains facilities for recognising various elements in the document structure. The SGML, which was developed to transfer electronic documents within inter-machine communication, also contains the tools to be used for saving information about various textual and structural elements in the document (http://www.oasis-open.org/cover/ topics .html#SGMLDecl; cf. Bryan 1988). A new TEI Consortium has been formed to maintain and continue the work of the TEI (http://www.uic.edu/orgs/tei/; Sperberg-McQueen & Burnard 1994). The development of the facilities connected with the SGML has continued, and now it has been replaced with the Extensible Markup Language (XML) developed to be more powerful in preparing meta-descriptions. The XML is also used in the most recent works of TEI descriptions. The TEI-coding of the text itself consists of the identification of the structural elements, elements defining the text type, the elements which are different from the language itself, such as tables, figures, dates, times, abbreviations, addresses, page numbers and quotations. The coding also contains elements for identifying cross-reference relationships in the text. The TEI-coding is one of the most accurate and extensive systems for distinguishing the information of the electronic linguistic texts.

The most effective enterprise for preparing meta-descriptions and analysing the metadata available on the network is the Resource Description Framework (RDF), a foundation for processing metadata. The RDF that uses the XML contains three types of objects: resources, properties and statements. In the RDF statements, information about the data is recognised by the labelled nodes containing information on the metadata. In these nodes, the structure of the metadata can also be taken into account. The RDF is used to generate labelled directed graphs: At its core, RDF data consists of nodes and attached attribute/value pairs. Nodes can be any web resource (pages, servers, basically anything for which you can give a URI), even other instances of metadata. Attributes are named properties of the nodes, and their values are either atomic (text strings, numbers, etc.) or other resources or metadata instances Lassila 1997: http://www.w3.org/TR/NOTE-rdf-simple-intro; cf. also http://www.oclc.org/oclc/corc/ and http://www.w3.org/Metadata/Activity.html).

3. ON META-DESCRIPTIONS FOR ELECTRONIC LINGUISTIC DATA

In this section, we discuss meta-descriptions for the morphologically coded corpora located on the UHLCS. These meta-descriptions are related to the descriptions needed to give information about various types of linguistic data in general. Metadata is information about data, and meta-describtions should be made taking into account the information structure of the data. In this paper, the meta-descriptions are prepared on the basis of the information structure that can be defined from the electronic text corpora. The goal is that using the meta-descriptions, it should be possible to characterize and also generate information about the corpora. In preparing the practical meta-descriptions in this project, the morphological coding the corpora of the Uralic languages will be compared to the standard developed by the EAGLES group. In the final stage of the work, the meta-descriptions will be compatible with the same formalism used in the TEI-coding. The descriptions will be given as property-value pairs: the properties get their values from the set of values that can also receive values, etc. The databases can be given in complex relational form, which can be used as data for the programming languages. It should also be possible to add the meta-descriptions to browsable corpora. In examining information about the electronic linguistic data located in various data banks, we distinguish at least the following main groups ([] denotes the sub-classification of data):

A. Basic information to be generated from the document

1. **The type of the document** [Spoken linguistic data [Discourse, Phone conversation, Speech, Radio,…]], [Written linguistic data [Published data, Book, Newspaper article, Periodical, Comic, Network publication,…]], [Non-published linguistic data [Manuscript, [Lecture notes], [Instructions to use certain tools]],…] (on the principles of

classification and documenting the electronic corpora, cf. http://www.ilc.pi.cnr.it/EAGLES/texttyp/texttyp.html).

2. **The language of the document** [The main language of the corpus [Indo-European languages [Germanic languages [German,…]], [Romance languages [French,…]]],… [All the languages spoken in the document [sub-classification]]. The main classification should start from the language family. The sub-classification can also be extended to apply to dialects.

3. **History of the document** [Place where the document was prepared []], [Date when the document was prepared []], [Author of the document []], [Staff or co-workers supporting the documentation []], [Information on the tools the documentation required []], [Information on the version of the document []], [Additional information []]].

4. **Special information from the fieldwork situation** [[Information on the purpose of the work []], [Name of the informant []], [Date of birth of the informant []], [All the languages the informant speaks []],[All the places the informant has lived in, and when and how long s/he has lived in each []], [Social status of the informant []], [Social class of the informant []], [Profession(s) of the informant, present and past []], [Additional information []]].

5. **Administrative information of the document** [Location of the corpus []], [Owner of the corpus []], [Steps needed before the corpus can be taken to be used as research material []], [Contact information []], [Information of the possibility to use the document in public demonstrations []], [Information of the copyright or authorization of the document []]].

B. Meta-descriptions for generating information on the data structure

Meta-descriptions of textual elements and textual structure and of documentation of the data are considered to be possible to be analyzed by using the TEI-coding as such. In this stage, information on the differences occurring in marking various structural elements must be defined. Also information about the character sets of the data and information needed for transformation of the character sets belong to this group of meta-descriptions.

C. On meta-descriptions for characterizing the data

Because the meta-descriptions should also be connected with the corpora located in various data banks, it is a plan that also differences in annotation of the corpora will be taken into account with the aid of meta-descriptions. A special problem in with this area concerns distinguishing and saving information about the typological diversity of languages. When preparing, e.g., electronic data for cross-linguistic studies, also this stage of the work should be taken into account with meta-descriptions.

In defining the meta-descriptions for distinguishing linguistic elements and combining information about the corpora annotated according to different kinds of principles, the first task concerns comparison of the tags used in annotation. In the project this paper concerns, the coding principles used in the coding of the Uralic languages will be compared with the CES standard.

The morphologically coded corpora located on the UHLCS are running texts. The coding is based on distinguishing the structural and categorial elements of language. The categorial elements are presented in the coding within the order they occur in the word form. In some corpora, information about the syntactic category or the semantic property characterizing the categorial tags is given after each category. The morphologically tagged word forms are translated into English, Swedish, German or Russian (the sample of Ume Saami is from the corpus prepared by Olavi Korhonen, the sample of Hill Mari is from the corpus prepared by André Hesslebäck, the sample of Selkup is from the corpus made by Jarmo Alatalo, and the sample of Finnish is done by using the automatic morphological analyzer of Finnish (Koskenniemi 1983; the sample of Udmurt is prepared for demonstrating the of annotating the corpora of the Uralic languages (Suihkonen 1998), http://www.ling.helsinki.fi/uhlcs/samples/; in the Finnish example, the capital letters are marked with an asterisk).

Udmurt

*Dzhog*_ADV_MAN	*fast, soon*
*ortts'+i+z*_V_-CONT_-TRA_+FIN_IND_PAST_SG	*to pass (away)*
*zarn'i*_N_-COUNT_SG_NOM|A_REL_SG_NOM	*golden*
*kuaro*_A_REL_SG_NOM	*with leaves*
*dyr*_N_+/-COUNT_SG_NOM	*time*
,	
*dzhog+en*_ADV_MAN_INSM	*fast*
*vu+i+zy*_V_-CONT_-TRA_+FIN_IND_PAST_PL3	*to come*
*zhob*_A_REL_SG_NOM	*nasty, unpleasant*
*siz'yl*_N_+COUNT_SG_NOM|A_REL_SG_NOM|ADV_TIME	*autumn*
*nunal+jos*_N_+COUNT_PL_NOM	*day*
.	

Selkup

Castrén: 1. Die gestohlene Frau.

mee uørkøsøwøt tagøn.
Wir wohnten im Sommer.

mee	P	SBJ
uørkøsøwøt	V PRT 1P	VER
tagøn	N INS	SAD

.

Ume Saami

Umesamisk text från Malå, berättad av LARS SJULSSON, Malå. Version med morfologisk kodning 981027. OKn.

Die PART då/så
dahta PRON DEM +COUNT SG NOM den/det
bálgies N +COUNT SG NOM stig
guhtjuotuvvij V PASS -CONT -TRA +FIN IND PRET SG3 kalla
gïrkuobálgies N +COUNT SG NOM kyrkostig
.

Hill Mari

Ramstedt, G.J. (1902). Bergstscheremissische Sprachstudien. Pp. 169-173. Memoires de la Société Finno-Ougrienne 17. Finno-Ugrian Society. Helsinki. 169.

perºi_[Russian] once
iktö_N_CARD a
komi_N_PROP komi
lömèn_ADJ named
örºezö_N_NOM_SG boy
roºotnjÉkesh_N_LAT_SG to work
pÉrash_V_INF to come
ºèröm_N_ACK_SG place
köchèl_GERUND seeking
ken_V_2PRET_3SG he went
.

Finnish

* talvi o+n * suome+ssa kylmä .
winter_N_SG_NOM be_V_PRES_SG3
Finland_N_PRP_SG_INE cold_A_SG_NOM
`The winter is cold in Finland.'

```
("<*>")
("" ("talvi" N NOM SG)) `winter'
(""
("olla" COP V PRES ACT SG3)) `to be'
("<*>")
(""
("suomi" PROP N INE SG)) `Finland'
(""
("kylmä" A POS NOM SG)) `cold'
("<.>")
("<*>")
(""
("muutto_lintu" N NOM PL)) `migratory bird'
(""
("saapua" V PRES ACT PL3) `arrive'
("saapuva" V PCP1 ACT A POS NOM PL)) `arriving'
(""
("kevät" N ADE SG)) `spring'
("<.>")
("<*>")
```

The computer corpora located on the UHLCS are heterogeneous, and when preparing the morphologically coded Uralic languages to be transformed to CES standard, the first step concerns unifying the morphological tags. In this stage, also the tags needed to be added to the tag vocabulary of CES should be evaluated. In spite of that it concerns transforming individual tag sets, the principles can also be used in unifying the corresponding coding of the other morphologically annotated linguistic corpora.

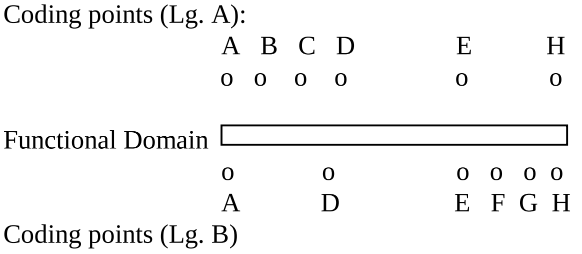
Unifying the coding principles in the way that the specific properties of a language distinguished in the coding will be saved is more complex and time consuming. Within the framework of the EAGLES, the general structural framework for defining the linguistic categories and properties concerning the analysis of the lexicon is called **linguistic architecture**. Linguistic architecture includes three kinds of schematic levels: **meta-schemata**, **schemata** and **instances of schemata**. In meta-schemata, the conditions for the well-formedness of schemata are defined, and in schemata, the logical format of language specific and level-wise linguistic information are defined in the linguistic architecture (Menon & Modiano 1994: 3-4; the standards include the TEI-coding). Some projects of the European Union have concentrated in description of lexicon. Within the framework of the GENELEX, linguistic architecture consists of the articulation between the morphological, syntactic and semantic levels. Each lexical item is seen through these three layers or a set of information concerning one layer. **Morphological units** at the morphological level include information on the morphological and grammatical properties connected with the word forms, etymological information, and possible information of abridged forms. The **syntactic layer** includes, e.g., information about the structures and positions of constructions and the order of the elements in the construction, the syntactic and transformational operations, and the grammatical and thematic functions of the syntagma. The properties of various syntactic structures are also defined in the model (Menon & Modiano 1994: 12 - 17). The principles developed within the framework of the EAGLES are exploited in the MULTEXT program that has worked efficiently for standardization for the lexical encoding initiatives in Europe. In the standard created by the MULTEXT program, the lexica are of two types:

(a) word forms, containing a word form, morpho-syntactic information, lemma and TAG; (b) lemmas, containing lemma, morpho-syntactic and inflectional information

Linguistic descriptions that are given with TAGs are based on attribute|value formalism. (http://www.ilc.pi.cnr.it/EAGLES96/lexarch/node15.html; http://www.ilc.pi.cnr.it/EAGLES-96/morphsyn/morphsyn.html; cf. also Bel, Calzolari & Monachini 1994):

(a) attributes are marked by their positions in the string of codes;
(b) values are represented by a single character and
(c) a special marker reflects the non-applicability of a given attribute.

In the morphologically coded corpora on the UHLCS, the functional categories of the word forms are distinguished. When planning the meta-description for this stage of the work, the functional domain of linguistic categories and the linguistic structure should be distinguished. In spite of the fact that the coding of the structural categories in languages vary, the typological variety within the particular functional domain is limited. (Givón 1984: 36-39).



In the morphological coding of the Uralic languages, the positions of constructions are recognised as soon as they occur in the text investigated. Location of the element in the hierarchical linguistic structure forms another distinctive parameter. Linearity of the expressions and the elements in the expressions is distinguished in coding the running text, which is the topic of the coding. That support defining information on the order of the structural elements located on the functional domain.

4. SUMMARY

The meta-descriptions collecting information of the extralinguistic data, i.e., the history of the document and the information space connected with the document. This kind of information can be defined on the basis of information structure available from the electronic linguistic data. The TEI-coding forms the basis for defining the meta-descriptions of the linguistic structure and characterizing elements in the linguistic documents. These meta-descriptions form an interface between the extralinguistic and linguistic data. The meta-descriptios of linguistic data should be defined in the way that the differences occurring in linguistic annotation can be neutralized, and the information of the typological differences about linguistic data can be taken into consideration in the cross-linguistic studies. The variety in linguistic analysis supports getting different kinds of information of language. The meta-descriptions can also form a level that can be used in distinguishing structural properties of languages. The meta-descriptions collected from electronic linguistic data should be taken into account in the stage of the work they concern.

REFERENCES

Bel, Nuria, Calzolari, Nicoletta & Monachini,Monica (eds.): LRE Project 62-050 MULTEXT.Workpackage 1. Milestone B Deliverable D1.6.1B. Common Specifications and Notation for Lexicon Encoding and Preliminary Proposal for the Tagset. Work in Progress. Report.

Bryan, Martin. 1995 [1988]. SGML. *An Author’s Guide to the Standard Generalized Markup Language*. Addison-Wesley Publishing Company. Wokingham, England, & al.

Givón, Talmy. 1984. *Syntax. A Functional-Typological Introduction* I. John Benjamins Publishing Company. Amsterdam/Philadelphia.

Ide, Nancy & Véronis, Jean. 1995. *Text Encoding Initiative. Background and Context.* Kluwer Academic Publishers.

Dordrecht.

Koskenniemi, Kimmo. 1984. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. Publications* 11. Department of General Linguistics. University of Helsinki.

Lassila, Ora. 1997. Introduction to RDF Metadata. http://www.w3.org/TR/NOTE-rdf-simple-intro.html.

Menon, B. & Modiano, N. 1994. EAGLES, WG-Lexicon. Task Group on Lexicon Architecture. Draft - Report.

Sperberg-McQueen, C. M. & Burnard, Lou. 1994. *Guidelines for Electronic Text Encoding and Interchange* (TEI P3). Vol. 1 and 2. The Association for Computers and Humanities (ACH), The Association for Computational Linguistics (ACL), The Association for Literary and Linguistic Computing (ALLC). Chicago.

Suihkonen, Pirkko. 1998. *Documentation of the Computer Corpora of Uralic Languages at the University of Helsinki*. *Technical Reports* TR-2. Department of General Linguistics, University of Helsinki.