ISO TC 37/SC4 N047
2003-05-06

Language Resource Management

Descriptors and Mechanisms for Language Resources

File ID SC4N047.doc (37 ko)
SC4N047.pdf (46 ko)

Title:	Language resource management – Morphosyntactic annotation framework; Call for contribution.
Editor(s):	Eric de la Clergerie, Lionel Clément
Source:	AFNOR
Project number:	N/A for submission
Status:	Call for contribution
Date:	2003-04-22
Agenda / Action:	Deadline for collecting replies from Member bodies: the 6 June 2003. The received material will be collected, and will be examined during the forthcoming Sapporo meeting.
References:	TC 37/SC 4 N035, N036

Mr. Key-Sun Choi - SC4 Secretary – KORTERM - 373-1 Guseong-dong Yuseong-gu - Daejeon 305-701 - Korea
82 42 869 35 65 – fax: 82 42 869 87 90 - kschoi@cs.kaist.ac.kr – http://korterm.kaist.ac.kr

Objectives

In order to illustrate the Morphosyntactic Annotation Work Item Proposal, the various members are invited to submit representative samples of such annotations for many languages and several linguistic phenomena. Annotation guidelines and annotation tagsets are also welcome. The objective of this call is to collect enough material about morpho-syntactic annotations to cover the current practices in the international community. The collected material will be used to help in the definition of a standard meta-model for morpho-syntactic annotation.

Each Member body is invited to synthesize all collected material into one contribution to be sent to Lionel Clément (lionel.clement@inria.fr), who is in charge of gathering the complete material.

Types of materials

Samples

We consider two main levels related to morpho-syntactic annotation to be covered by short but representative samples in various languages:

Each sample may come with comments about the illustrated phenomenon, a translation in English and an ASCII transliteration for readability.

Segmentation — This level is in charge of delimiting elementary segments in text or speech (or in multimedia documents). For textual document, we would like to illustrate phenomena such as the tagging of abbreviations, morphological amalgams, segmentation alternation, compound words…

Tagging — This level identifies morpho-syntactic properties of above-mentioned segments, such as part-of-speech tags and grammatical (sub) categories. These tags are generally organized in tagsets, eventually with complex structures. The diversity of languages (morphology) raises different views or approaches for annotations, as illustrated by the following non-exhaustive list:

- Agglutinative languages (Finnish, Magyar, Turkish, Nippon, Swahili…);
- Inflectional languages (French, German, Spanish, English,…);
- Isolating languages (Chinese);
- Amalgams phenomena (clitics in Romance languages, compounds in Germanic languages), inflexionnal and derivational morphology, …

The annotation process may also depend of the kind of writing and more generally by the kind of document being annotated. We would like therefore to collect samples for different kinds of documents:

- Textual documents;
    - Phonetic writing;
    - Consonant writing (Arabic);
    - Syllabic languages (Kana);
    - Ideogram languages (Chinese);
- Oral transcription;
    - Orthographic or phonetic transcription;
    - Specific tags and marks used in speech transcription (intonation curves, speech overlaps, repetitions, alternative transcriptions for a sequence of spoken data,…):
- Electronic documents such as hypertexts and multimodal documents;
- Annotated documents such as editorial meta-data and annotations used in literary genre.

Guidelines

Guidelines have been proposed by different organizations or companies to help or standardize the annotation process of corpora. They may provide definitions of tagset elements; include a glossary of linguistic terms and phenomena (definition of a compound, the restriction of a definition of a word, combinatory definition of a tag). They may also provide hints to solve some problems (for instance in case of ambiguity). More generally, they provide information about annotation practices that we think useful to collect.

Tagset

An analysis of a large sample of the various morpho-syntactic tagsets is needed to identify the possibilities of convergence between them, at least on some basic core. We therefore wish to collect various tagsets for part of speech, grammatical categorization, morphology… Besides the collection of tags, we are also concerned by the structure of tagsets (list, simple or multiple hierarchy…) and the structure of tags (simple name, feature structure,…)