4. Primary linguistic data

4.1. File structure

4.1.1. Database structure

The units of a database are records. A record consists of fields. As to field structure, there are two kinds of databases:

- database with rigid field structure: the field structure is a property of the database and therefore the same for all records;
- database with flexible field structure: the field structure is a property of each record and may therefore differ from record to record.

Database programs usually allow for databases of only one kind. In what follows, a database with flexible field structure will be assumed. However, the flexibility requirements are such that they can be met by a database with rigid field structure simply by providing the maximum number of fields and not filling in some fields in some records.

4.1.2. The text database

The following guidelines apply to a sample text which is used as a self-sufficient set of data. They apply analogously to a set of example sentences that are stored separately and may be retrieved for use in a metalinguistic context.

A database may contain several sample texts (or sets of example sentences). From the title of the text, an abbreviation is derived that is repeated in all of its records (in the record id).

The text is broken down into units whose length does not exceed the size of a print line. Preferably, such a database unit should be a syntactic unit: maximally, a sentence, minimally, a phrase. Each such unit founds a record.

4.1.3. The null record

The first record of a text, called the null record, has the following field structure:

1. **Text id**: abbreviation identifying the text, followed by record number 000.
2. **Title**.
3. **Name of author**.
4. **Date of production**.
5. **Publication**: bibliographical data, if the text has been published.
6. **Type of text**, according to some classification of genres.
7. **Name of analyst**.
8. **Comment**: explanation of any non-standard features, e.g. special symbols.
9. **Date of last modification**.

4.1.4. Body of text database

Each record except the null record has the following maximum field structure:

1. **Record id**: Abbreviation identifying the text, followed by record number (normally three digits, with leading zeros).
2. **Orthographic representation**: Original orthographic form of the text, transliterated if necessary.
3. **Phonetic representation**: Broad phonetic transcription.
4. **Phonemic representation**: Text words represented as sequences of phonemes.
5. **Prosodic representation**: Prosodic structure according to Du Bois et al. 1992.
6. **Morphological representation**: Text broken up into morphemes (each given in morphophonemic representation), with suitable boundary symbols according to section 3.1.
7. **Morphological gloss**: Representation of each item in field 6 by its meaning or grammatical function (grammatical category label), according to section 3.1.
8. **Grammatical tagging**: Grammatical categories of items in field 6, structural information (e.g. bracketing). Details in section 4.2.
9. **Translation**: Idiomatic translation into one of Eurotyp's official languages.
10. **Descriptors**: Keywords identifying interesting linguistic features of the record. They are taken from the set in section 2.2.
11. **Comments**: Free format remarks, including problems.

4.1.5. Subset of fields in a record

The field structure of a given record is an appropriate subset of the above field structure. In the selection of this subset, the following considerations apply:

- #1 is necessary, or else no reference to the record will be possible.
- #2 or #4 or both are necessary.
- The necessity of #3 depends on the kind of text; if the text was never recorded, it will be superfluous.
- The necessity of #4 depends on the language. It will be necessary if it differs substantially from #6 minus morphological boundary symbols, as is the case in languages with a heavily obliterative phonology.
- #6 is necessary for a language with some morphology and if the data are to be used for grammatical analysis.
- #7 is necessary if the data are to be used for grammatical analysis.
- #9 is necessary if the data are to be used in arbitrary scientific contexts. It will generally be in English.
- The records of a text should be uniform in terms of fields #1 - 9. #10 and 11 are optional from record to record.

Du Bois, John W. & Schuetze-Coburn, Stephan & Paolino, Danae & Cumming, Susanna 1992, *Discourse transcription.* Santa Barbara: UCSB, Department of Linguistics (Santa Barbara Papers in Linguistics, 4).

4.2. Tagging: Coding the linguistic structure of a text

[to be filled in]

Back to index