MULTEXT

LRE PROJECT 62-050

Task 1.3 Internal Standard Formats

Deliverable 1.3.1
Internal Standard Formats

Version B
March 15, 1995

Task leader :
Jean Véronis, CNRS

Contributors :
David G. Durand, Consultant
Nancy Ide, CNRS
Jacques Le Maitre, CNRS

Copyright © 1994-1995 Centre National de la Recherche Scientifique

Copyright © 1994-1995 Centre National de la Recherche Scientifique

Foreword

This document outlines the internal representation formats for the Multext corpus handling tools. It begins with an explanation of the Multext document processing model, which enables interaction between the Unix and SGML environments.

The model for data storage is based on a hyper-linked strategy instantiated in SGML and HyTime. This document outlines the overall architecture for data storage; the details of the markup are described in Deliverable 1.5.1 (Corpus Encoding Standard).

Table of contents

1. Introduction

Numerous data collection initiatives and corpus projects have recently collected large amounts of electronic texts. At the same time, the TEI and EAGLES are making substantial progress towards interchangeability and reusability of these resources. As a result, the need for generally available, flexible text analytic software tools is substantially greater than ever in the past.

Unfortunately, the software for text analysis that exists at present only begins to cover research needs. Industrial software is often expensive or unavailable and is usually hard to adapt or extend. On the other hand, academic software is often experimental, hard to install, under-documented, and sometimes unreliable. In both cases tools are typically embedded in large, non-adaptable systems which are fundamentally incompatible with each other. Although efforts to develop standards for data representation are underway, little effort has been made to develop standards for linguistic software.

As a result, there is a serious lack of generally usable tools to manipulate and analyze the text corpora and collections that are now becoming widely available. Worse, there is enormous duplication of effort: it is not at all uncommon for researchers to develop tailor-made systems that replicate much of the functionality of other systems and in turn create programs that cannot be re-used by others--thus leading to an endless ***software waste cycle***. The reusability of data is a much-discussed topic these days; similarly, ***software reusability*** is needed, to avoid the re-inventing of the wheel characteristic of much language-analytic research in the past three decades.

The LRE project Multext was established explicitly to answer this need for generally usable and reusable tools for text analysis, for use within the Language Engineering community. Therefore, rather than setting out to develop just another toolkit for text analysis, Multext is committed to the development of ***guidelines for corpus handling software***. These guidelines are intended to provide a framework both for the tools that Multext itself will develop, as well as a general framework for the further development of text handling software in the future.

2. The "Software Lego®" approach

2.1. Basic requirements

There are several known paradigms for software development. However, most of them result from experience in developing software within companies in well-structured teams or consortia, and involving professional developers. In such cases, the goal is typically obvious at the outset, and general decisions on the environment (architecture, programming languages, data types, platforms, etc.) can be taken at the beginning of the project.

In the Language Engineering community, the situation is quite different. There is no way to envision the entire functionality of a system at the outset, since there is no global project; software pieces typically grow how and when there is a demand; special and one-time use tools are common; and new unforeseen functions are likely to appear depending on research trends. The change in the scale of usable text data, from the order of a few million words to tens or hundreds of millions, is likely to affect both what language researchers do and how they do it. It is difficult to foresee the functions that text analysis will require in the future. We are certainly at the beginning of an era of exploration and experimentation, which will require flexible software that can be adapted to one-time or new applications. Given these circumstances, and given the fact that software is going to be developed by a wide range of potentially unrelated people, it is difficult to agree on strong constraints, such as the choice of a unique programming language, the choice of a class system for objects, etc.

Multext must therefore develop an entirely new paradigm which enables easy use of corpus handling tools. It must also be easy to customize the tools for special applications. Therefore this paradigm must allow for ***software evolution,*** where new tools can be developed, one tool can replace another, etc. It must allow for ***heterogeneous distributed development,*** where the developers, their skills, habits, applications, hardware, etc. are widely different and umpredictible.

2.2. Principles

These overall requirements form the basis of the "Software Lego® Approach", as outlined in the Multext Technical Annex:

2.2.1. Smaller (atomic) pieces

Existing text analytic software often comprises large, integrated systems that are nearly impossible to adapt or extend. The research community would be best served by a set of small tools that can be used alone or combined to create larger, more complex programs. Atomic modules, or "software bricks" can be bundled to perform more complex tasks, as needed. Very importantly, since each module consists of a small, easily understandable piece of code, it is easy to modify and debug. The use of small modules also allows researchers to concentrate on the underlying logic in order to develop new tools, and eliminates the need for the training and/or time to implement algorithms at lower levels. The development of increasingly complex bundles provides good, reliable software for researchers who want a ready-made high-level tool.

2.2.2. Standardization of interfaces between modules

To write compatible software, it is essential that all modules communicate effectively, that is, that their interface is standardized. This demands that internal standard formats (ISFs) for input/output and accessory files be developed, and serve as specifications for program development. It is important that the output of each function conform to the input expectations of the next. Further, it is essential that these formats be public, so that any program written anywhere by anyone can use them. The need for ISFs has been under-stressed in the past, and as a result, even a well-established system such as Unix is inconsistent in the internal formats (streams, files, command line arguments) required for even closely related functions.

2.2.3. Simple data types

Communication between tools should be by means of the simplest possible formats and data types, if possible human-readable. There is some overhead in this approach, but the speed and storage capacities of present-day machines virtually eliminate this concern. More importantly, developers gain the ability to see what is going on at each step, which enables an easy test-modify-test cycle. Finally, (for the moment) complex data types tie programs to specific languages that implement those types. The use of flat data eliminates this dependency--programs can be written in any programming language at any level: C, C++, Pascal, assembler, SNOBOL, awk, Perl, etc. There can be an exception to this principle for compiled resources (hash tables, etc.) and "binary large objects" (BLOBs--images, sound, etc.).

2.2.4. Language independence

A very important goal is to develop software for all European languages. So far, linguistic software has been developed mostly for English and is very hard (if not impossible) to adapt to other languages. Therefore, the production of language-independent tools should be encouraged. Such tools should use an engine-based approach where all language dependent materials are provided as data. Thus, extension of the tools to cover additional languages in most cases involve only providing the appropriate tables and rules.

Copyright © 1994-1995 Centre National de la Recherche Scientifique

3. Multext context

3.1. Overview of application

Multext is developing a corpus manipulation system providing all the basic functionality typically needed by academic or industrial researchers in language engineering, including tools for corpus ***annotation*** and corpus ***exploitation***. The sequence of events in corpus annotation and exploitation is typically the following:

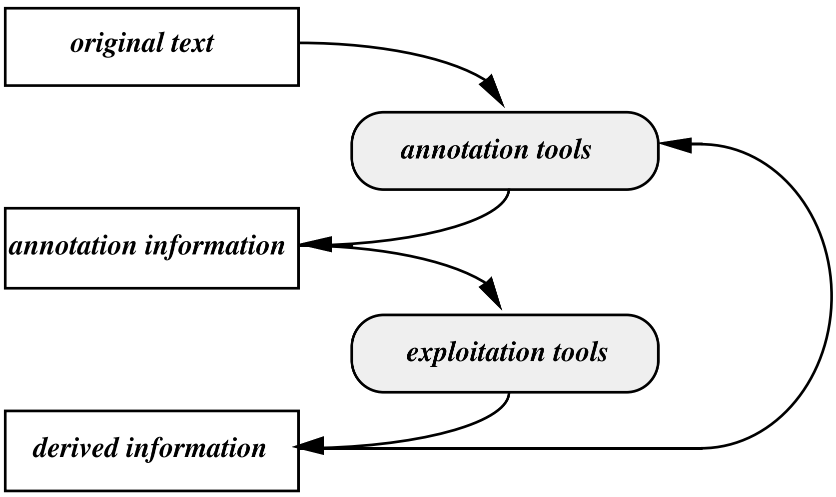

Figure 1. Overview of corpus handling system

The original text is input to the corpus annotation tools, which may accomplish tasks such as segmentation of texts into words and sentences, grammatical annotation, alignment of parallel translations, etc. The annotation tools produce ***annotation information*** (sentence and word boundaries, grammatical annotation, etc.), which may be used by other annotation tools, and/or by the exploitation tools. The exploitation tools access the information created by the annotation process in order to perform various search and retrieval tasks (browsing, concordancing, retrieval of collocations, etc.), statistical and quantitative analyses, etc. Some of the ***derived information*** created by the exploitation tools may in turn be fed back into the annotation process--for example, to augment the lexical lists used by the morphological tools, etc.

3.2. Implementation of the Software Lego® approach

The object-oriented paradigm, which is very popular at the moment in large-scale development, could offer the modularity needed to implement the Software Lego approach. However, it usually imposes:

- an object class system common to a project;
- complex data types and use of APIs;
- programming language dependence, at least for the time being (e.g. C++);

These constraints are too strong in the context of heterogeneous distributed development (that is, for the broader community, beyond Multext), especially with non-expert intervention (for assembling of modules, development of new modules, customization, etc.)

As a result, Multext has adopted the operator/stream approach, implemented in the Unix system, which has already been widely used in the language engineering community (see for example, Liberman and Marcus, 1992). In the operator/stream approach, data flows in uni-directional "streams" (or "pipelines") between tools.[1] Each of these tools is an "operator" (or "filter") that transforms the data as it passes by. Component tools are independent, and at no point are they compiled together in a single program. This approach has several advantages:

- each tool can be implemented in a different language, developed by different people, tested independently, etc.
- new tools can be plugged into the stream as needed, and tools are re-usable in other contexts.
- the choice of Unix as a platform enables exploiting the use of built-in Unix tools and utilities.
- Unix scripting facilities enable rapid and easy bundling of tools, easy modification and prototyping, etc.

This approach is readily applicable in the Multext context. For example, the text segmentation process can be accomplished with a set of micro-tools like the following (see deliverable 2.2.1):

[1] In Unix, streams are linear. We use the term "stream" here in a more general way, since non-linear streams can be emulated in Unix using temporary files.

mtsegspace	Produces a rough tokenization by segmenting the text at each space and tabulation.
mtsegpunct	Strips punctuation attached to tokens on the left or right.
mtsegfeatures	Detects and marks the typographic features of the token (uppercase letters, final period, etc.).
mtsegabbrev	Handles abbreviations.
mtsegclitics	Separates proclitics and enclitics from word itself (e.g., in French, "sais-je" = "sais" + "je").
mtsegcompound	Detects and recombines some multiword units (e.d. fixed locutions such as "parce que" in French).
mtsegnumdigit	Handles numbers written in digits, and recombines them if necessary.

etc.

These tools can be pipelined together to accomplish the full segmentation of a text. At the same time, each tool is small, self-contained, and therefore easy to understand, debug, and modify. A given tool can be easily replaced or deleted from the chain, and additional tools can be introduced anywhere in the chain as required to serve specific user needs. The only requirement to use the tools effectively is to understand the input and output format of each tool; therefore, it is necessary to provide a formal description of the I/O formats for the tools; These formats are discussed in section 5 below.

3.3. Handling document structure

In Multext, original documents are SGML documents marked for gross structural features ("level 1", i.e. down to the level of paragraph--see deliverable 1.5.1), since SGML has become widely accepted as a text markup scheme in the language engineering community.[2] It is essential to be able to retain all the information in the SGML original--including information

---

[2] Deliverable 1.5.1 defines a preferred corpus markup style which is compliant with the guidelines developed by the Text Encoding Initiative (TEI), but any SGML document can be handled. In addition, since some users potentially have "plain text" (without explicit tags) or texts in other formats (LaTeX, etc.) that can be converted to plain text, Multext tools also allow plain text input. In the following discussion, we consider SGML input only.

about structure, rendition, etc.--throughout the annotation process, for use by the exploitation tools and possibly even the annotation tools themselves.

When the tools process documents originally encoded in SGML, it is very important for many applications that the SGML structure is fully preserved. For example :

¶ **Translation:** it might be necessary to reproduce the document in the target language with the same features (divisions, layout, etc.) as the original.

- **Navigation and display:** it would be useful to display the document in its original format and have hypertext-like capabilities based on the added annotation, e.g.:

    - show the results of searches in the actual document (e.g.: highlight all the instances of the some POS or pattern of POS in the text);
    - click on a word and get the morphosyntactic information;
    - select a fragment and have the corresponding fragment highlighted in the parallel text;

    etc.

, **Structure-sensitive operations:** it is often desirable to work with selected portions of a text, e.g.:

- count words in some element types only (e.g. excluding titles);
- use a different tagger (or different parameters) on titles;
- find mutual information between abstracts and body of text;

etc.

Also, Multext is committed to TEI-compliant SGML for the exportation of data, thus dictating that all annotation be ultimately expressible in SGML markup.

The need to retain and access information about the structure of a document throughout the annotation and exploitation processes places certain constraints on the processing model, especially for corpus annotation, which is modifying the information in the corpus. If there were no need to retain and handle the structural information about the text, it would be possible to use a simple record-field approach to represent the corpus information, as in the SUSANNE corpus (Sampson, 1992) and ET10-63/CRATER (McEnery and Daille, 1993) projects. However, document structure in SGML is hierarchical, and is therefore at odds with the record/field approach inherent in Unix stream I/O. Therefore, Multext has developed a hybrid document processing model to enable the two to work together.

Copyright © 1994-1995 Centre National de la Recherche Scientifique

4. The Multext document processing model

As mentioned above, information about document structure must be retained throughout the corpus annotation process. There are at least two ways to do this, which are discussed in the following sections.

4.1. The All-SGML model

One means to retain information about document structure is to pass the entire SGML original document, together with any markup added by previous phases of annotation, to each tool in the chain. However, this approach has several disadvantages. For processing, there are the following problems:

each tool must process the entire SGML tree, even if only certain portions of it, and certain information from it, are relevant for the analysis. This can involve potentially complex analysis of the DTD and complete parsing of the SGML tree.

portions of the text with which the tools are working (e.g. titles only) must be re-selected from the full document by each micro-tool. Such "queries" can be arbitrarily complex (e.g., select **<p>** only in **<div2>** elements, and not in **<footnote>**, etc.), demanding each tool has access to an interpreter of a language in which to express the queries.

The overhead of this kind of processing is unacceptable if it must be repeated for each of the Multext micro-tools. The only way to avoid this overhead would be to pre-compile sets of micro-tools together (for example, the suite of segmentation tools given above). However, this grossly violates the Multext atomicity principle: it precludes the possibility of skipping, replacing, or adding a given micro-tool without re-compiling. Adding micro-tools within the chain in the form of simple Unix scripts would be impossible. Adding even the most trivial tool between pre-compiled suites of micro-tools demands re-parsing the SGML document; even if an SGML API is provided to perform the parse, the range of programming language options is limited to full-blown languages such as C, and some non-trivial level of programming expertise is required. All flexibility is lost.

The all-SGML approach also locks the user into an all-SGML environment:

the tools cannot be readily used outside the SGML context, i.e., they cannot be readily used on non-SGML texts.

the ability to use and interject simple Unix tools in the chain is lost, since these tools are record/field-oriented. In addition, like most tools they necessarily operate on the entire document and are unable to select certain pieces based on a knowledge of the SGML tree. The only way to get around this problem is to re-write all Unix utilities for SGML-awareness. This is an interesting project, but it is beyond the scope of Multext.

4.2. The hybrid model

The strategy adopted in Multext to retain document structure avoids the problems with the all-SGML approach outlined above. It is based on the following observations:

most corpus annotation and exploitation tools are inherently table-to-table oriented (e.g., input a token, output a set of part of speech/lemma pairs; input part of speech/lemma pairs, output the correct one; input a list of words, output a frequency list; etc.).

most tools do not need deep knowledge of the SGML tree or, indeed, any knowledge of the SGML tree. They might need to know where a paragraph or sentence boundary (as a "milestone") occurred, but not its relation to other elements in the SGML tree (e.g., that it occurred in the fourth paragraph in the third section, etc.).

in many cases it is desirable to process only some pieces of a document at a time. For example, it is typical to process only the running prose (e.g., content of **<p>** elements) in a document when doing part of speech tagging, or, alternatively, only titles (since grammatical patterns in the two may be very different).

Therefore, Multext has adopted the following model (see Figure 2):

Select the right "chunks", using a query tool (see Deliverable 2.1.1) to enable users to precisely specify the pieces of the document in which they are interested. A cross-reference to the location of the selected chunk(s) in the original SGML tree is retained.

Convert the chunks and cross-references to tabular format.

Pass the chunk(s) and cross-references to the desired chain of table-oriented tools and/or Unix tools.

At any point, input from or export to SGML, for interchange, to use with SGML-aware tools, etc.

Note that the tabular format is not intended to be a translation of all information in the SGML document (which, as noted above, is not needed by most tools). At the same time, it is important to note that the full SGML structure is retained in the stored data and can be accessed at any point in the chain; and further that this structure is enriched by the addition of the annotation information when it is converted back to SGML format.

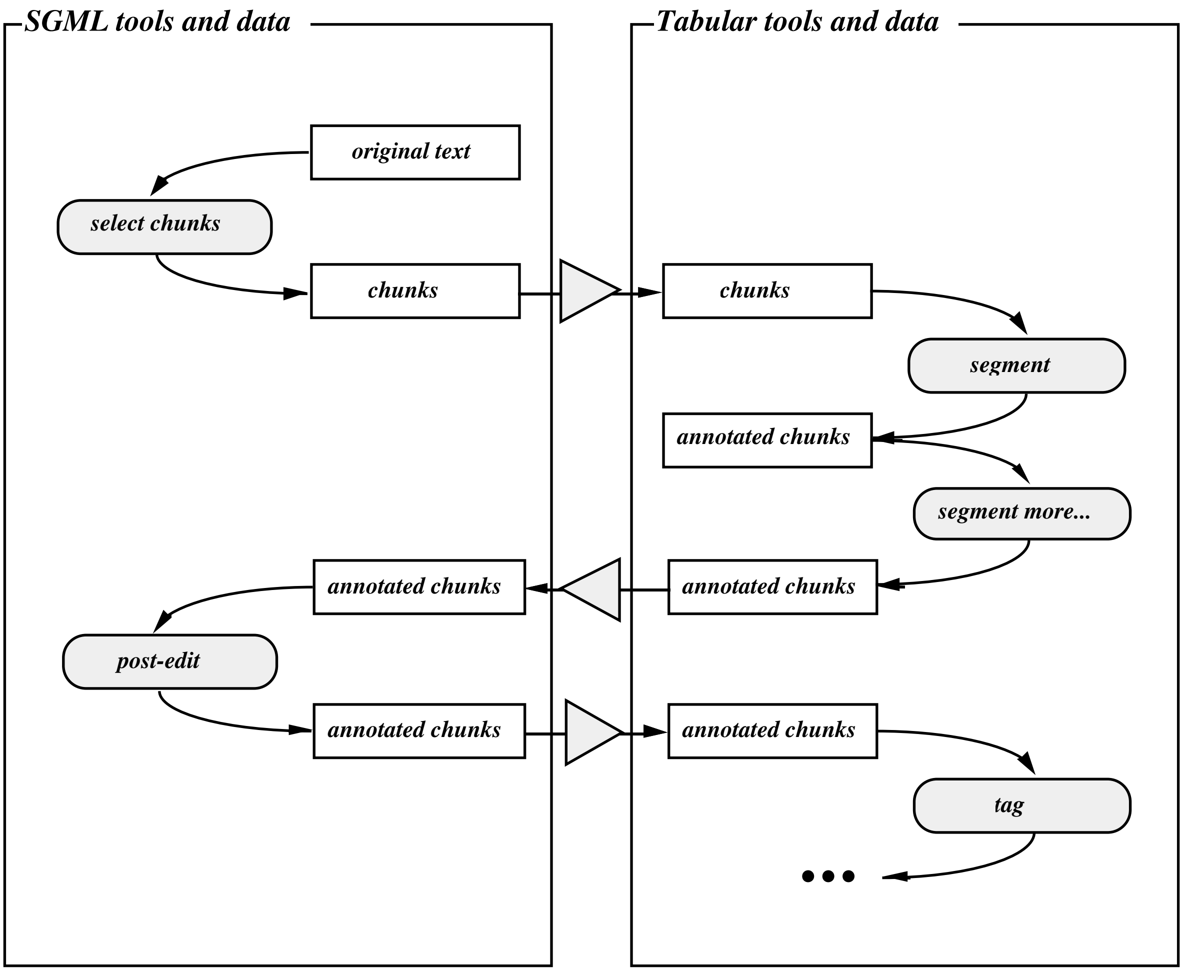

Figure 2. Hybrid model

This model has the following advantages:

simplicity of tool input and output, making it easy to read, debug, etc, as specified in the Multext software principles;

ability to use Unix functions, utlitites etc. anywhere desired in the chain;

input to and output from the tools can be stored in SGML format, which is rapidly becoming an interchange standard for corpora;

option to use SGML tools on the data after storage in SGML format;

non-SGML input and output is possible for users without data in SGML form, or who are not interested in retaining annotation results in SGML format.

Let us assume the following normalized SGML document (nSGML)[3] :

```
...
<TEXT LANG=FR>
<BODY>
<DIV0 TYPE=DOCUMENT>
<DIV1 TYPE=article N=4>
<TITLE>
Implantation à Eloyes
</TITLE>
<P>
L'usine, qui devrait être implantée à Eloyes (Vosges) représente un investissement
d'environ 3,7 milliards de yens. Elle fabriquera des pièces détachées pour la filiale
de Minolta en RFA.
</P>
</DIV1>
<DIV1 TYPE=article n=5>
<TITLE>
Projet Port-Mariane
</TITLE>
<P>
A Montpellier, après que le maire J. Besse ait annoncé la création du port fluvial,
se pourrait-il qu'aujourd'hui le projet Port-Mariane soit définitif ?
</P>
</DIV1>
</DIV0>
</BODY>
</TEXT>
```

[3] Note that this example is drawn from the current Multext corpus. This corpus utilizes an interim encoding scheme which does not conform to the specifications in deliverable 1.5.1. This example conforms to the current format of the corpus, which has not yet been made consistent with the final markup specifications.

The SGML extractor (*mtsgmlql*--see Deliverable 2.1.1), given an appropriate query, extracts the elements to be processed through the chain of annotation tools; e.g., to get all the paragraphs of our example, the query is

```
select
   <chunk>
      every PCDATA descendant_from p
   </chunk>
from p in every P descendant_from number( file( example ) )
```

The output of the extractor is also a text in nSGML format. For the example above, the result has the following form[4]:

```
<chunk>
<pcdata loc="2.1.1.1.2.1">
L'usine, qui devrait être implantée à Eloyes (Vosges) représente un investissement
d'environ 3,7 milliards de yens. Elle fabriquera des pièces détachées pour la filiale
de Minolta en RFA.
</pcdata>
</chunk>
<chunk>
<pcdata loc="2.1.1.2.2.1">
A Montpellier, après que le maire J. Besse ait annoncé la création du port fluvial,
se pourrait-il qu'aujourd'hui le projet Port-Mariane soit définitif ?
</pcdata>
</chunk>
```

The results of applying the segmentation tools can be translated back into SGML, yielding something like the following:

```
<tok from="2.1.1.1.2.1\1">L'</tok>
<tok from="2.1.1.1.2.1\3">usine</tok>
<tok from="2.1.1.1.2.1\8">,</tok>
...
```

---

[4] Note that #PCDATA are considered here as pseudo-elements in the SGML tree, which is consistent with the HyTime philosophy. Therefore the extractor allows #PCDATA pseudo-elements to be wrapped in a **<pcdata>** tag and treated as regular elements.

5. Tabular formats

Multext is committed to defining internal formats for communication between its tools, and for making these formats publicly available. This is essential if the tools are to be easily usable and reusable, as the Technical Annex stresses.

5.1. Logical formats

Tables (or "relations") are familiar in the database world, consisting of rows and columns corresponding to specific fields of data. Relational databases map readily into Unix sequential files. Rows in the tables correspond to Unix lines, columns correspond to fields. A simple example:

ID	TOKEN	POS	LEMMA
1	La	DMF	le
2	petite	AFS	petit
3	porte	NCFS	porte
...			

In a non-disambiguated corpus, you may have redundancies in the rows of the tables, as in the following:

ID	TOKEN	POS	LEMMA
1	La	DMF	le
1	La	PPJ	le
2	petite	AFS	petit
2	petite	NCFS	petit
3	porte	NCFS	porte
3	porte	VM1S	porter
3	porte	VM3S	porter
...			

Copyright © 1994-1995 Centre National de la Recherche Scientifique

To avoid this redundancy, values can be "packed" in a non-first normal form relation (NF^2), that is, a relation in which values can be either atomic values or sub-tables (Abiteboul and Bidoit, 1986; Roth et al. 1988). A new table containing sub-tables can be constructed on this basis:

ID	**TOKEN**	**(POS LEMMA)***	
1	La	DMF	le
		PPJ	le
2	petite	AFS	petit
		NCFS	petit
3	porte	NCFS	porte
		VM1S	porter
		VM3S	porter
...			

The notation (ID TOKEN (POS LEMMA)*)* denotes the type of the table in the example above, which consists of three attributes. The notation (POS LEMMA)* denotes a single (relation-valued) attribute. This type of notation will be used to describe the input and output format of tools (see for example Deliverable 2.2.1).

Sub-tables such as those in the example above can be considered (and processed) as a whole. We provide below a Unix representation for embedded tables that packs sub-tables in a single field. This enables skipping a sub-table altogether, performing a **sort**, a **join**, a **uniq**, etc. on an entire sub-relation, building an index on a sub-table, etc. We limit the nesting of relations to one level, which is a reasonable compromise between added expressiveness and added notational complexity. In addition, we have not yet found examples in the chain of annotation tools where a deeper embedding would be useful.

5.2. Physical formats

There is a susbtantial inconsistency within Unix commands and tools as to what constitutes a field or (conversely) what constitutes a separator. The space vs. tab problem is well known. In addition, tools do not always recognize 8-bit characters, and lines have various length limitations.

We therefore define the following conventions for (possibly nested) tables :

Each row is a line.
(Top level) separator is a tab.
Subtables are coded using '\' (backslash) as field separator and '|' (pipe) as row separator.
Fields cannot contain end-of-line, tab, backslash, pipe or space.
Fields can use any other printable ISO 8859-1 character.
When needed in a field, space must me recoded as underscore (e.g. New_York).
If end-of-line, tab, backslash, pipe or a non ISO 8859-1 character is needed in a field it should be coded using an appropriate SGML entity.
Line length is not limited.

Example:

```
1   La        DMF\le|PPJ\le
2   petite    AFS\petit|NCFS\petit
3   porte     NCFS\porte|VM1S\porter|VM3S\porter
...
```

Implications for the tools:

they should process 8 bit characters;
they should accept long lines;
separators should be re-definable;
character counts, etc. must be adapted to SGML entities.

Many public tools (e.g. GNU) statisfy the first two or three contraints. The use of SGML entities in the stream is very unfortunate, but seems unavoidable--there will always be some character that does not belong to the chosen character set.

5.3. Example

At the moment, the segmentation and lexical lookup tools use the ftabular formats defined in figure 3 for input and output. The atomic-valued fields and the sub-tables are as follows:

Copyright © 1994-1995 Centre National de la Recherche Scientifique

ACT	action associated with a token
CLASS	class
CNAME	class name
COM	comment
CORPTAG	tags specific for corpus processing
CPOS	vector of possible positions of a token in a compound
FEAT	typographical features of tokens
LEM	lemma
LEXDESCR	lexical description
POS	position of a token within an SGML element
TREELOC	location of element in SGML tree
TOK	token

CORPINFO	= (LEM CORPTAG)*
LEXINFO	= (LEM LEXDESCR)*
REF	= (TREELOC POS)*

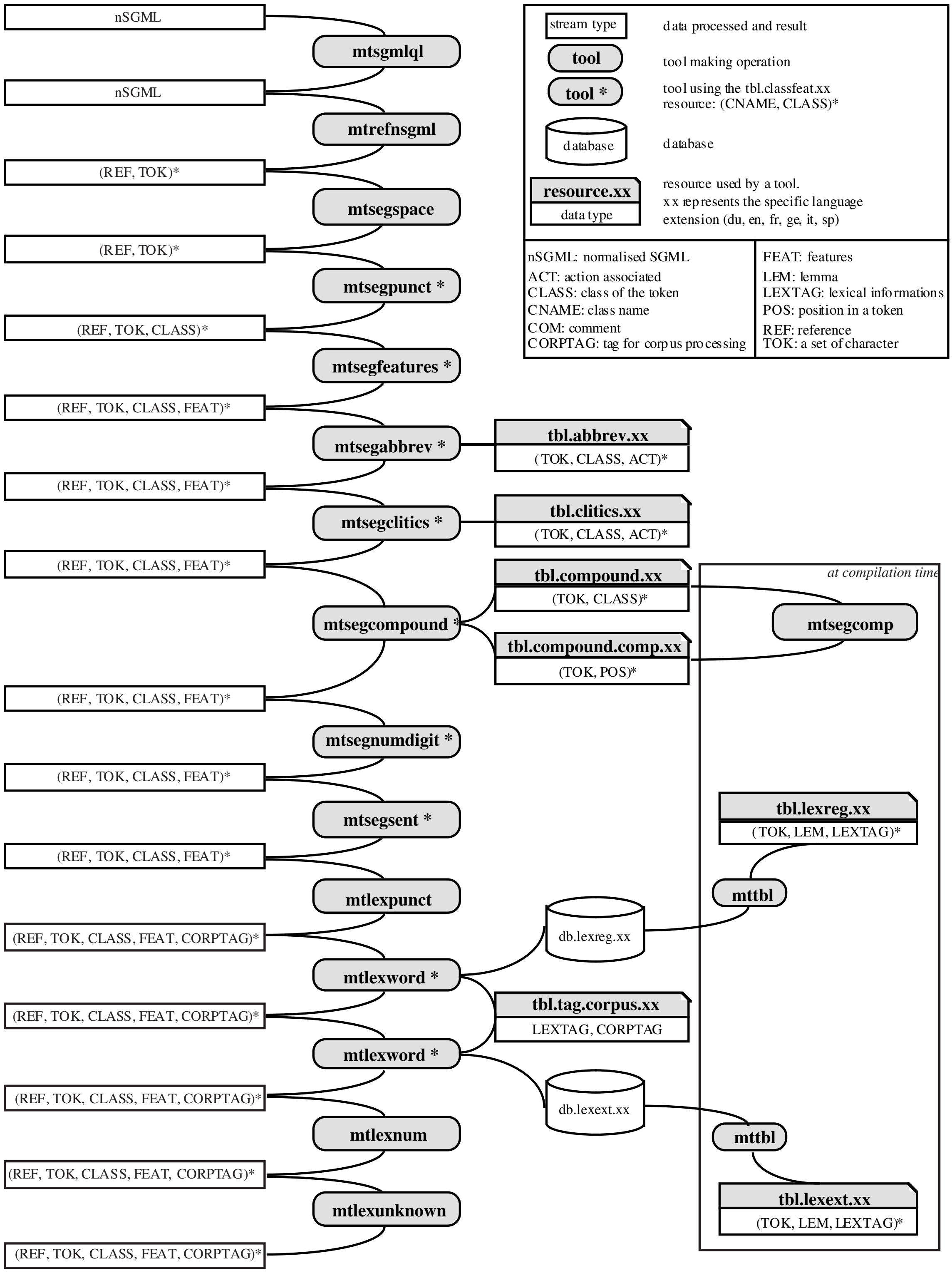

Figure 3. Tabular formats for a subset of tools

6. Storage and interchange formats

6.1. Hypertext approach

As described above, data can be exported to SGML at any point during the annotation chain. Annotation information could be merged into the original document, or maintained separately. In Multext, we have adopted a strategy whereby annotation information is not merged with the original, but rather retained in separate SGML documents and linked to the original or other annotation documents using the HyTime/SGML standard for hypertextual linkage (DeRose and Durand, 1994). The original or "hub" document remains stable and is not modifed in any way by any process. This strategy has the following advantages:

it avoids the creation of potentially unwieldy documents--envision, in a worst case, a single document containing segmentation and part of speech markup, plus markup for alignment with translations in the eleven EU languages, plus alignment with the speech recording, plus variant part of speech taggings from several taggers, etc.!

it avoids problems with markup containing overlapping hierarchies, which are not allowable in SGML. For example, annotation added by a given tool (e.g., sentence boundaries markup) could conflict with markup in the original (e.g., markup for emphasis). Or, conflicting markup could be produced by different versions of the same tool (e.g., two different word segmenters could produce two different sets of word boundaries).

different versions of the same kind of annotation (e.g., different POS annotation) can be associated with the text.

it simplifies the input to the tools, since very few tools need access to all of the original information plus all of the annotation information, and in some cases would not be able to use all of this information (e.g., variant encodings produced by different versions of the same tool).

it is trivial to map the annotation information in tabular format, which contains information about the location of elements in the input SGML document, into a HyTime/SGML file containing this information and links to the input document. In this way, annotation information is added to the SGML document at its appropriate location.

it is very much in line with what is evolving in the SGML world, and it is likely that SGML/HyTime software which handles complex linkages will be available in the near future.

annotation can be accomplished by associating the SGML original or other annotation documents with other, pre-existing documents--e.g., instead of generating a document containing POS markup and linking it to the original, links could be made directly with lexicon entries.

it gives easy access to the original SGML document (or, as mentioned above, any among several versions annotated for certain features) for use by other applications.

6.2. Document linkage architecture

6.2.1. Overview

The hyper-document comprising each text in the Multext corpus and its annotations will consist of several documents, each described below.

The base or "hub" document is the unannotated document which serves as input to the segmenter, containing only markup of types 0 and 1 (defined in Deliverable 1.5.1). The hub document is in a normalized SGML format (nSGML), as defined in the appendix. The hub document is "read only" and is not modified in the annotation process.

As mentioned above, the output of any tool in the chain can be exported to SGML format. Each exported document is a proper SGML document with a DTD, containing annotation information linked to its appropriate location in the hub document. Thus the annotation document contains what can be seen as "remote markup", virtually added to the hub. Some annotation documents will be permanent and retained for distribution, while others will be retained only temporarily (possibly, to use as input later on to additional annoation tools), at the user's option. Below we discuss the important annotion documents that can be created from the output of the tools.

Links between documents can be one-way or two-way. A one-way link points ***from*** a given location--i.e., the location at which the pointer appears in an SGML document--***to*** another location in the same or a different SGML document.[5] Two-way links associate two (or more) locations by pointing to both; the location of the two-way link itself is not dependent upon the location of either of its ends. In Multext, all annotation documents will be linked to the hub or other annotation documents using one way links. The exception is output of the aligner for parallel texts, which will consist of an SGML document containing ***only*** two-way links associating locations in two documents in different languages. The two linked documents are two annotation documents containing segmentation information, such as sentence boundaries. The overall architecture is described by figure 4:

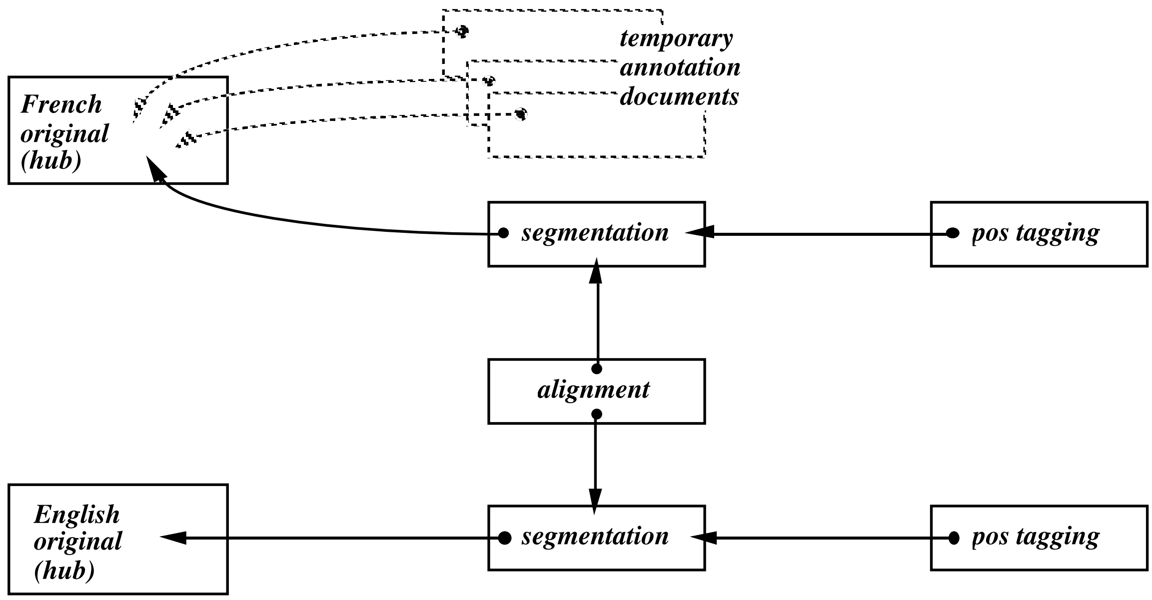

Figure 4. Multext document linkage architecture

[5] Note the distinction between files, which are storage units, SGML documents, which may be composed of a number of files by means of external entity references, and HyTime hyper-documents, which consist of either one SGML document that uses HyTime links to connect parts of that document, or several SGML documents connected by HyTime links.

6.2.2. Segmenter output

The output of the segmenter consists of markup signalling sentence and word boundaries. Retaining segmentation markup in a separate SGML document with links to the original avoids problems of overlapping hierarchies--that is, invalid nesting of SGML elements--that are likely to arise when markup for sentence and word boundaries, which in itself forms a well-formed hierarchy, is superimposed on markup recording typographic phenomena recorded in the hub document, such as emphasis. For example, if the hub document contains the following markup:

```
It certainly wasn't <hi r=it>him. He</hi> would never do such a ...
```

segmentation could produce

```
<s>It certainly wasn't <hi r=it>him.</s> <s>He</hi> would never do such a ...
```

which is invalid SGML since the **<hi>** and **<s>** elements are not properly nested. See the discussion in Deliverable 1.5.1.

The information in the segmentation document will be associated with the relevant locations in the hub via HyTime and/or TEI links. Both HyTime and the TEI provide several different means to specify the location of the endpoints of a link. One of the most common is to point to a unique identifier, specified in the ID attribute on the target tag. However, since segmentation documents are pointing into the hub document, which is read-only, it may be necessary to modify the original by adding ID attributes to each tag, as a prior step. Another lmethod is to use locations in the SGML tree to point to specific elements (e.g., the third child of the second child of the first child of the root). Using this method solves the read-only problem, eliminates the indexing needed for IDs, and provides a semantically meaningful location that might be useful for the tools.

For segmentation purposes, we also need to point to locations that are not the entire content of an SGML element--for example, to a specific character that is the start of a sentence. Both HyTime and TEI allow for mixing pointing mechanisms in the same location indicator. Therefore, segmentation markup will be linked to the hub document via one-way links, specified in two steps:

¶ tree addressing to point to the nearest enclosing tag for the location in question;

• reference to the position of a specific character from the location given in step ¶.

The segmentation document could consist of only the remote markup identifying word and sentence boundaries, punctuation, abbreviations, etc. with no tag content. For example, consider the following text:

```
...
<P>
L'usine, qui devrait être implantée à Eloyes (Vosges) représente un investissement
d'environ 3,7 milliards de yens. Elle fabriquera des pièces détachées pour la filiale
de Minolta en RFA.
</P>
...
```

The following two tags point to the first two words inside the **<p>** element:

```
<tok from="2.1.1.1.2.1\1" to="2.1.1.1.2.1\2">
<tok from="2.1.1.1.2.1\3" to="2.1.1.1.2.1\6">
...
```

We use here a short notation to specify the target location in the hub document. The notation "2.1.1.1.2.1\1" is equivalent to the TEI notation "CHILD (2) (1) (1) (1) (2) (1) STRLOC (1)". See Deliverable 1.5.1 for specification of addressing mechanisms.

The above markup is straightforward, but for the time being we recommend retaining the actual token string as the tag content, as follows:

```
<tok from="2.1.1.1.2.1\1">L'</tok>
<tok from="2.1.1.1.2.1\3">usine</tok>
...
```

This method duplicates some of the content of the hub document in the segmentation document; however, the segmentation document becomes self-standing. Since many tools will use this document as input, this has processing advantages. This method also allows eliminating the location of the end of the token, since this information is implicit in the string itself.

Another option would be to retain sentence markup in a document separate from the document containing token markup; however, we recommend retaining them in the same document to avoid the multiplication of documents and to provide a self-standing document containing both kinds of information, since sentence and token boundaries are very often used together. Since markup for sentences and tokens forms a well-formed tree, this presents no problem for a straightforward DTD.

```
<s from="2.1.1.1.2.1\1">
<tok from="2.1.1.1.2.1\1">L'</tok>
<tok from="2.1.1.1.2.1\3">usine</tok>
...
</s>
...
```

6.2.3. Tagger output

The output of the Tagger consists of a separate SGML document containing the annotation information generated by the tool--i.e., lemma and part of speech tags for the tokens in the text. We see two main options for retaining the annotation information generated by the tagger:

¶ the output of the tagger can be stored in a separate document and linked to the segmentation document with one-way links, as follows :

Segmentation document :

```
<s from="2.1.1.1.2.1\1">
<tok from="2.1.1.1.2.1\1">L'</tok>
<tok from="2.1.1.1.2.1\3">usine</tok>
...
</s>
...
```

Tagger document :

```
<gram from="1.1"><lem>le</lem><pos>DFS</pos></gram>
<gram from="1.2"><lem>usine</lem><pos>NCFS</pos></gram>
...
```

- since the tagger has all segmentation information available to it during processing, another option is to create a new annotation document containing the segmentation information as well as the POS category information. This new document is linked to the hub with one-way links. If a segmentation document already exists, this document could be discarded if desired.

```
<s from="2.1.1.1.2.1\1">
<tok from="2.1.1.1.2.1\1"><form>L'</form><lem>le</lem><pos>DFS</pos></tok>
<tok from="2.1.1.1.2.1\3"><form>usine</form><lem>usine</lem><pos>NCFS</pos></tok>
...
</s>
...
```

This second method avoids problems that may arise due to updating of documents. We can imagine, for example, that segmentation boundaries may be changed after part of speech tagging (e.g., at the post-editing stage).

6.2.4. Aligner output

The output of the aligner consists of a series of links between two documents, identifying corresponding segments (sentence, word) in two texts. Sentences, and words are easily adressable by their location in the SGML tree in the segmentation documents under analysis. The links between the two segmentation documents are stored in a third document :

```
<link targets="1.1;1.1">
<link targets="1.2,1.3;1.2"
<link targets="1.4;">
```

The first link shows a one-to-one mapping of two sentences; the second shows the mapping of two sentences in the first document on to one sentence in the second; and the third shows a sentence in the first document with no corresponding translation in the second. The exact mechanisms for specifying links are discussed in Deliverable 1.5.1.

References

Abiteboul, S. and Bidoit, N. (1986) Nonfirst normal form relations: an algebra allowing data restructuring. *J. Comput. Sys. Sci.* 33, 3, 361-393.

Liberman, M., Marcus, M. (1992). *Tutorial on Text Corpora,* presented at the Association for Computational Linguistics Annual Conference.

McEnery, T., Daille, B. (1993) *Database design for corpus storage: the ET10-63 data model.* ET10-63 report.

DeRose, S.J., Durand, D.G. (1994) *Making HyperMedia work: A users's guide to HyTime.* Kluwer Academic. Boston.

Roth, M. A., Korth, H. F., Silberschatz, A. (1988). Extended algebra and calculus for nested relational databases. *ACM Transactions on Database Systems,* 13, 4, 389-417.

Sampson, G. (1992). *The SUZANNE corpus.* University of Sussex, Bringhton.

Appendix

Normalized SGML format (nSGML)

SGML is a very complex language that enables the encoding of the same facts in many synonymous ways. Some SGML constructs are extremely difficult to process (entity replacement, minimization, etc.). Multext has therefore adopted the approach that all SGML documents be rendered into a normalized SGML format (nSGML), which is valid SGML with the following additional constraints:

the defined character set is ISO-8859-1;

tags and corresponding attributes appear on the same line;

each line starts at the first position (i.e., no leading blanks or tabs);

no carriage returns should appear in #PCDATA (note that carriage returns in legacy documents should have been interpreted as paragraph breaks, etc.);

no markup minimization is used (in particular, all closing tags are present);

all entities are expanded except for character entities;

no character present in character set should be represented by an entity (e.g. "é" must not be represented by "é").

Note that this specification for nSGML is preliminary and is subject to further refinement.