[*Translating and the Computer 22*: Proceedings of the Twenty-second international conference…16-17 November 2000. (London: Aslib, 2000)]

Exchanging Lexical and Terminological Data with OLIF2

Susan M.McCormick
Linguistic Consultant to SAP, Walldorf, Germany
www.olif.net

1 Introduction

The OLIF2 lexicon and terminology exchange standard is currently under development within the OLIF2 Consortium, a collaborative group of industrial firms active in the field of language technology. Based on the OLIF prototype (Open Lexicon Interchange Format) that was generated as part of the OTELO and Aventinus projects, OLIF2 represents an improvement to OLIF in several important ways: First, while maintaining the simple, straightforward structure of the original OLIF, OLIF2 is now XML-compliant and will serve as the lexicographical component of the new XLT lexical/terminology exchange standard that is being developed within the framework of the SALT initiative. Second, the original OLIF language options, restricted initially to just English, German, and Danish, have been expanded to accommodate the requirements of French, Spanish, and Portuguese as well. And third, OLIF2 offers improved support for NLP systems such as machine translation, an original goal of the OTELO project, by providing coverage of a much wider and more detailed range of linguistic features.

2 Background

2.1 The OLIF2 Consortium

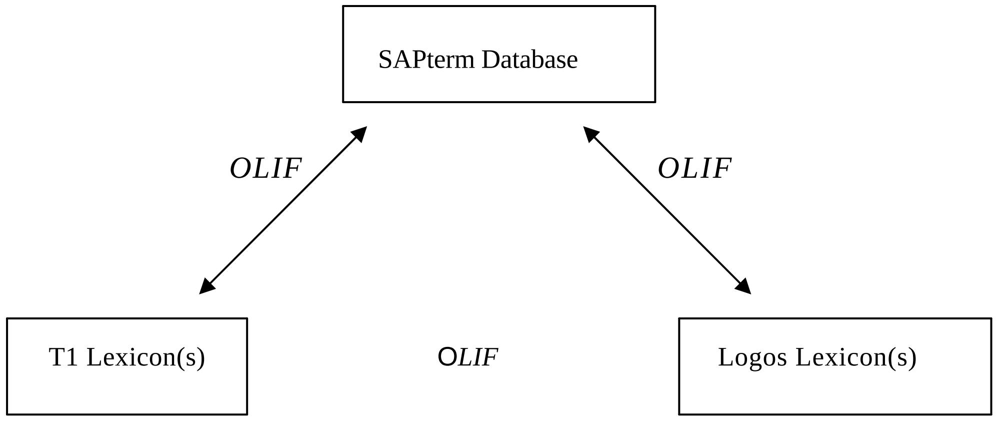
As a partner in the OTELO project, SAP of Germany was vitally interested in using the OLIF format to alleviate some of the administrative overhead generated by maintaining its large terminology set in the various language support tools it employs. These tools include the company-internal database, SAPterm, a general MultiTerm termbase, the Logos and T1 MT systems for four language pairs, and several translation memory applications. As the demand for translation at the company has grown, the task of entering and maintaining terminology, both in-house and externally, as well as the challenge of ensuring consistency among the different language tools have become increasingly onerous. Since the OLIF prototype was lacking some important features that would make it usable at SAP, the company decided to spearhead an effort to revise the standard, thus initiating the OLIF2 Consortium. As the coordinating member, it is joined in the consortium by a number of companies that develop or use language tools, including *Xerox, Sail Labs, Logos, L10NBRIDGE, Lotus, Microsoft, Trados IBM* and *Systran.* The European Commission also participates in an advisory capacity.

By the end of the year 2000, the OLIF2 consortium plans to make generally available a complete XML DTD for OLIF2 data elements. Although OLIF2 coverage for traditional dictionary handling and NLP lexicons, especially MT lexicons, is robust, it maintains the basic approach of OLIF in terms of its approach to terminology, i.e., it addresses basic terminology exchange needs, but does not duplicate well-accepted terminology exchange standards such as MARTIF in the depth and complexity of its representation. For this, users may turn to XLT, where they can avail themselves of both MARTIF and OLIF2.

2.2 From OLIF to OLIF2

The original purpose of OLIF was to provide a simple, user-friendly vehicle for interfacing with multiple electronic lexical and terminological resources. While the trend in lexicon and terminology management today is generally toward standardization, electronic lexicons and termbases are still sufficiently diverse in design that users that wish to share or re-use their data are often forced to negotiate between different standards. The OTELO project members addressed this problem by developing OLIF as a common lexical resource format that would facilitate the exchange of lexical/terminological data from system to system and from user to system. For example, using the single OLIF format, SAP translators would be able to update a Logos MT lexicon with new company terms from the SAPterm database, or easily migrate terms from T1 to Logos or SAPterm:

1)



Since the lexical requirements for NLP systems like Logos or T1 are both different from one another and different from general terminology management requirements, the task of producing a central standard meant careful consideration of both system-specific requirements and general industry standards. Participating MT system lexicons were reviewed for commonality and general terminology and lexical requirements were defined.

The OLIF prototype that resulted from these efforts was a good first step in trying to bring together the disparate and often complex requirements of the electronic lexicons and terminology databases that were studied. The actual OLIF format was comparatively simple in structure and proved easy to implement. As a prototype, however, OLIF was not sufficient either to exchange data from many languages or to represent some of the grammatical information required by NLP applications that were not represented in the OTELO project. The second version of OLIF, OLIF2, is, we hope, a helpful adaptation of OLIF that addresses the shortcomings of the original format, thus making it more usable for a wider range of users.

3 The structure of an OLIF2 file

The structure of OLIF2 maintains the straightforwardness of the original OLIF, the purpose of which was to facilitate the description of a lexical/terminological entry to the extent that an NLP vendor such as Logos or Sail Labs can generate a basic, usable entry of its own from an OLIF record. Like OLIF, OLIF2 specifies a file with a *header*, which contains data that is relevant to all of the lexical/terminological entries in the file, and a *body*, which contains the entries themselves. The entry structure is relatively flat, with minimal embedding of element types.

3.1 The OLIF2 file header

The OLIF2 file header includes information on both the data in the file itself and the user. Element types and attributes that are covered include:

- **file description:** includes the filename and counts of entries, terms, concepts, and bytes.
- **public statement:** provides information on the owner and distributor of the OLIF2 document.
- **feature/value information:** contains user information on the structure of OLIF2 linguistic information, as well as information on domain hierarchy.
- **content information:** provides information on the formatting of quotation marks and typographical information such as boldfacing.

- **encoding information:** identifies the code set used; OLIF2 files are in Unicode, using eitherUCS-2, UTF-8, or ISO-646.
- **original tool:** identifies the tool that created the OLIF2 document.
- **original format:** indicates the file format of the file from which the OLIF2 document was generated.
- **creation date:** notes the creation date of the header element.
- **creator:** includes the ID of the creator of the header element.

In addition, the user may use the header to specify any siring replacements that should apply to the entire document, or to make general, informational comments on the data in the file.

3.2 The body of the OLIF2 file

The body of an OLIF2 file is a list of entries that contain data that is grouped according to the linguistic/lexical/terminological character of the information being represented. The groups are sub-lists of feature/value pairs (represented in XML as tags that reflect the element types, attributes, and values defined in the DTD), and are characterized as follows:

- **monolingual:** feature/value pairs that define monolingual data.
- **transfer:** feature/value pairs that define transfer relations between the given entry and other entries in the lexicon in different languages.
- **cross-reference:** feature/value pairs that define cross-reference relations between the given entry and other entries in the lexicon in the same language.

Transfers are represented as bilingual, unidirectional links between monolingual entities in different languages, whereas cross-referencing for relations such as synonymy, antonymy, part-whole, and orthographic variation operates within a single language.

The OLIF2 entry is itself defined as a semantic unit that is identified uniquely by a set of five obligatory monolingual features:

- **canonical form:** the entry string, represented in canonical form in accordance with OLIF2 guidelines for formulating canonical forms.
- **language:** the language represented by the entry string.
- **part of speech:** the part of speech, or word class, represented by the entry string.
- **subject field:** the knowledge domain to which the lexical/terminological entry is assigned.
- **reading number:** the number identifier used to distinguish readings for entries with identical values for *canonical form, language, part of speech,* and *subject field.*

Although the structure of an OLIF2 entry reflects a lemma-orientation and is entry-based, a concept-based structure can be easily modeled using the subject field as a conceptual identifier. The monolingual, transfer, and cross-reference feature/value groups include coverage of both linguistic and terminological information.

3.2.1 Linguistic features in OLIF2

The OLIF2 linguistic analysis includes a lexical description of morphological, syntactic, and semantic phenomena for all of the languages supported. Moreover, the new format version offers a more robust handling of selectional restrictions and lexical transformations.

The current set of linguistic features for OLIF2 entries are listed in (2). The morphology, syntax, and semantic categories relate to the monolingual block of the entry; transfer conditions, or selectional restrictions, specify conditions under which a given transfer is valid, and are listed as part of the transfer block. Also listed in the transfer block are transfer actions, or lexical transformations.

2. OLIF2 Linguistic Features

Feature	Description
Morphology:	
inflection class	Encodes the inflection pattern(s) of the entry word or head of multiword/phrasal entry.
head word	Indicates the **head word** in a multiword/phrasal entry string.
gender	Indicates grammatical **gender.**
case	Indicates grammatical **case** designation.
number	Indicates grammatical **number.**
person	Indicates **person.**
tense	Indicates verb **tense.**
mood	Indicates **mood** or mode.
aspect	Indicates verbal **aspect.**
degree type	Indicates adjectival **degree type.**
auxiliary type	Indicates the **auxiliary type** for an auxiliary verb.
Syntactic:	
syntactic type	**The syntactic type** describes the general syntactic behavior of the entry string.
syntactic position	**The syntactic position** describes the unmarked positioning of the entry string syntactically.
transitivity type	Describes the **transitivity type** of a verb.
syntactic frame	Describes the **syntactic frame** elements for the entry string (subcategorization).
preposition	Frequently-used **prepositions;** can be used to further specify syntactic frame elements.
particle	Frequently-used **verb particles;** can be used to further specify syntactic frame elements.
Semantic:	
definition	The **definition** is a prose definition of the entry string.
natural gender	The **natural gender** refers to the biological gender associated with the entry.
semantic type	The **semantic type** represents the status of the entry string with respect to a semantic type classification structure.
Transfer conditions and actions:	
context	Indicates the **context** for a given translation of a source word/phrase into a target word/phrase.
feature test	Indicates **feature** being tested in a transfer test.
string test	Indicates **string** being tested in a transfer test.
add to head	Transfer action to **add** an element to the **head** element in the target translation; type attribute is part-of-speech value.
add to context	Transfer action to **add** an element to a **context** element in the target translation; type attribute is part-of-speech value.
delete from head	Transfer action to **delete** an element from the **head** element in the target translation; type attribute is part-of-speech value.
delete from context	Transfer action to **delete** an element from a **context** element in the target translation; type attribute is part-of-speech value.

change verb form	Transfer action to **change** the **verb form** from the source to target.
change role	Transfer action to **change** the **role** of a verb argument from source to target.
translate context	Transfer action to assign a **translation to a context** element
assign case	Transfer action to **assign case** to an element in the transfer

3.2.2 Terminology features in OLIF2

The OLIF2 terminology approach offers basic handling of administrative data, as well as support for user-defined domain hierarchies. In addition, traditional dictionary categories, such as comments and examples are included in the format, as illustrated in (3):

3. OLIF2 Terminology Features

Feature	**Description**
geographical usage	Refers to the **geographical usage,** or dialect, represented by entry string.
entry type	The **entry type** indicates the shape/structure of the entry string.
entry status	Indicates the **entry status** of an entry within a given lexicon/termbase.
entry source	Refers to the **entry source,** or the lexicon/termbase that the entry originated from.
entry ID	The **entry ID is a** user-defined numeric identifier associated with the entry.
originator	The **originator** is the individual who originated the entry.
updater	The **updater** is the individual who last modified the entry.
modification date	The **modification date** indicates the date that the entry was last modified.
example	The **example** is a sample text or portion of text that contains the entry string as an illustration of usage.
usage note	Indicates a **usage note** for entry siring
note	Refers to **note,** or commentary, on entry by lexicographer/terminologist.
administrative status	Indicates the **administrative status** of an entry relative to a given work environment
company	Indicates the **company**/organization for whom entry is valid.
abbreviation	Indicates an **abbreviated** form of the entry string.
deprecated synonym	Indicates a rejected or **deprecated synonym** for the entry string.
time restriction	Refers to **time restriction,** or the period of time during or since which usage of the entry is valid.
product	Indicates the **product** for which entry is valid.
project	Indicates the **project** for which entry is valid

4 OLIF2 Entries in XML

As noted in section (1) above, OLIF2 is XML-compliant. The sample entry in (4) shows both the basic structure of an OLIF2 entry and its representation in the revised format. (4) is taken from SAPterm and encodes the German noun *Briefkurs* in the subject field *general accounting/financial* with its English transfer *bank selling rate:*

```
<entry>
   <mono>
      <canForm>Briefkurs</canForm>
      <language>de</language>
      <ptOfSpeech>noun</ptOfSpeech>
      <subjField>gac-fi</subjField>
      <readingNo>1</readingNo>
      <entryType>cmp</entryType>
      <entryStatus>term</entryStatus>
      <entrySource>sterm</entrySource>
      <company>sap</company>
      <originator>fischerf</originator>
      <updater>hansenpou</updater>
      <modDate>1999-28-01</modDate>
      <adminStatus>ver</adminStatus>
      <usage>online</online>
      <note>online-a</note>
      <gender>(m)</gender>
      <inflection>n-15</inflection>
      <synType>cnt</synType>
      <semType>meas</semType>
   </mono>
   <transfer>
      <canForm>bank selling rate</canForm>
      <language>en</language>
      <ptOfSpeech>noun</ptOfSpeech>
      <subjField>gac-fi</subjField>
   <equival>full</equival>
</transfer>
</entry>
```

5 Conclusion

OLIF2 should offer users a respite from the repetitive task of coding and re-coding lexical or terminology entries for systems and databases with incompatible standards. Since OLIF2 Consortium members are committed to supporting the new format, users of Logos, for example, will be able to easily migrate their Logos entries either to other Logos systems or to another MT system, such as Comprendium. The inclusion of OLIF2 in XLT, the new lexical-terminology exchange standard being developed by SALT, means as well that terminological data that is compliant with the MARTIF standard can be integrated into Logos or Comprendium lexicons via the new format. In addition, OLIF2 will make it much easier for users to compare data in different lexicons and termbases, a task that is often necessary in order to ensure that the data are consistent with one another and up-to-date. Maintaining lexical and terminology sets in different lexicons and termbases should therefore be substantially simplified with the new format.

The attraction of OLIF2 is clearly not restricted to lexicons and terminology databases, but extends to other NLP tools that connect in important ways with terminology and lexicon maintenance. For example, spell and grammar checkers, term management software, tools for Controlled Language, taggers, and tools for information classification and retrieval could all benefit from a standard format that allows data exchange from tool to tool. OLIF2 offers a means of bringing all of these applications together to improve efficiency and productivity for users.

References

Lieske, Christian (2000) *OLIF2 DTD Proposal.* in Documents www.olif.net.

McCormick, Susan et al. (2000) *Proposal for the Structure and Content of the Body of an OLIF2 File.* in Documents www.olif.net

Spaeth, Mark, G. Thurmair, and J. Ritzke (1998) *Final Specification: The Open Lexicon Interchange Format - OLIF.* OTELO Project report LE 1 2703-LR1.1.

Thurmair, Gregor, J. Ritzke, and S. McCormick (1999) *The Open Lexicon Interchange Format - OLIF.* In TAMA '98 - Terminology in Advanced Microcomputer Applications. Proceedings of the 4th TermNet Symposium, Vienna. 237-262.