Consistent modelling of heterogeneous lexical structures

Laurent Romary, Werner Wegstein

Consistent modelling of heterogeneous lexical structures

Laurent Romary, Inria & HUB

Werner Wegstein, University of Würzburg

Pooling lexical sources – a Digital Humanities perspective

Our paper addresses the issue of interoperability between heterogeneous data sources. As a matter of fact, this issue has been the object of many debates within the Text Encoding Initiative (TEI) community and in general within many standardisation groups providing models or formats for data interchange. At the core of the problem is the usual trade-off between, on the one hand, offering a flexible platform for representing a variety of possible structures (expressivity) and, on the other hand, being able to predict under which conditions some data can be the object of a blind interchange, in particular in the context of them being processed randomly by a generic tool.

This trade-off has indeed no generic solution, but it regularly arises in defining the components of such a large-coverage modelling platform as the TEI Guidelines. The guidelines' specifications are an expression of a balance of interests between the large number and variety of cases that the user community is constantly bringing to light and the need to abstract away from such examples in order to design recommendations that a new user can easily understand and apply in the context of his own encoding endeavours.

In many of the sub-domains of the TEI Guidelines we can observe how the various constraints and the stratification of corrections or changes over time have lead to some constructs becoming hybrid data models which present the double drawback of often leaving the user in wonder as to which representation is the "optimal" one in a given context and above all of generating a wealth of heterogeneous representations in the global data space of existing TEI documents. Over the years this has become more and more an issue, in a context of better online accessibility and ever-increasing development of international collaborations between scholar.

In this paper we will focus on lexical structures, from the idea that these represent a typical case of the interoperability issue in terms of pooling data from heterogeneous sources. We have asked ourselves whether the TEI chapter ("print dictionaries") dedicated to lexical data should not be revised or at least be accompanied by further constraints on its usage so that basic operations related to the querying, displaying or merging of lexical information could be made more straightforward.

The case of lexical data as presented in a dictionary offers an interesting experimental setting for studying interoperability in the context of standardisation. It is complex enough to reflect the variability which is intrinsic to the TEI Guidelines while providing a limited observational setting for studying the granular structure of lexical entries as well as the rather high internal coherence that one specific lexical source usually has. Lexical resources also reflect the variety of analytical points of view that one may have on linguistic information ranging from quite descriptive and verbose objects in the domain of standard human-oriented dictionaries to fully structured databases like developed in the natural language processing domain.

From a digital humanities perspective, we want to understand if it is at all possible to find a good balance between expressing precise constraints on the encoding of a primary source and leaving some

freedom to the scholar who will see the encoding activity as a step in his research process. This is why we have made an attempt to identify a generic methodology for expressing encoding constraints on source texts based on the idea of local representation "crystals" (Romary, 2009). These crystals correspond to elementary constructs at a low level of granularity in a document, which, independently of the broader organisation of the document itself, may express a certain concept in an extremely regular way, thus making the further reuse of this information chunk easier. In this context, interoperability is related to the capacity of a person or a tool to process existing crystals within a document independently of its origination.

After presenting the general background for modelling and representing lexical sources we give an overview of the various crystals that form the basis of most existing types of lexical entries. For each of these crystals we make systematic recommendations with corresponding supporting arguments. In the second part of the paper we illustrate our proposals with concrete cases taken from various dictionary or lexical database projects.

Modelling tools for lexical resources

In this paper we consider only lexical resources that are encoded according to a semasiological principle, i.e. where entries are determined according to the forms of a language and are then further refined into the different senses that have been deemed relevant for this form. This *word to sense* organization is usually seen as the most appropriate for the representation of large coverage lexica, as opposed to onomasiological representations (*concept to term*), which better take into account the organisation of domain specific vocabularies (aka terminologies). The semasiological perspective is usually the underlying model for traditional print dictionaries as well as for large-scale lexica in the natural language processing domain.

There are two main international standardisation activities that are relevant for the modelling and the representation of semasiological resources. On the one hand ISO 24613, in accordance with the modelling strategy of ISO committee TC 37, provides a group of meta-models that can be combined to produce specific data models applicable to a wide range of lexical types or components (Machine Readable Lexicon, Morphology, Syntax, Semantics, Multi-word expression, etc.). Even if it provides a possible XML serialisation, LMF tends to be agnostic as to the actual implementation of the models it allows one to describe. On the other hand the TEI has been seminal in offering a reference XML vocabulary for the representation of dictionaries, which actually offers a good compliance with LMF principles[1]. However, the variety of constructions that the TEI actually allows for the representation of the same lexical phenomenon may be seen as a hindrance to the achievement of deep interoperability across heterogeneous lexical resources.

In this paper we take as our starting point the positions described by the above-mentioned ISO standard as published in 2008 and the latest release of TEI Guidelines[2] in order to provide further insights on how to build-up lexical resources or dictionaries relying on a more systematic use of standardised constructs. The work presented here is also based upon some core principles that have systematically guided our work, both theoretically but also practically in the form of examples that

[1] Some of the LMF packages such as the description of subcategorisation frames do not as yet have any equivalence in the TEI vocabulary, but the TEI extension mechanisms do indeed facilitate the description of such extensions.

[2] As the work presented in this paper also reflects the implication of the authors into the evolution of the TEI Guidelines, some of the proposed changes (in particular regarding the systematic use of <sense>) have already been integrated into the December 2012 (2.0.0, "Laurentian").

have served as experimental background for testing our proposals. Even if the present work is not about modelling XML structures at large, several of these principles are derived from a more global concept of the kind of semantics that XML constructs convey and the way to actually reflect this in the design of XML formats.

With this perspective in mind, we can already state two generic constraints that impact on the organisation and semantics of lexical structures:

- Semantic grouping, according to which features that jointly convey a given meaning in an lexical entry should be systematically grouped together, even when only one such feature occurs and even at the cost of favouring more deeply structured representations;
- Hierarchical dependency, which states that features, or groups thereof, which qualify a given level (for instance, an entry), are considered, unless stated otherwise, to be inherited further down in the sub-components (typically the senses) of the lexical entry. We refer here to Ide et al. (2000).

From these constraints we will progressively derive specific recommendation on the local organisation of lexical entries as guided by our crystal based analysis. Comparing these with real data, and in particular with legacy dictionaries, we will try to understand possible transition schemes from weakly structured data to more standardized constructs.

Core proposals – towards a systematic description of lexical crystals

Crystals as coherent sub-structures

Introducing the concept of crystals in data modelling in general and in the TEI guidelines in particular reflects the necessity to describe data structures according to their capacity to bring together, at any depth, a coherent group of components (or elements in the XML terminology). More precisely, we define a crystal as an *independent group of connected elements (clique) with semantic coherence*. A typical example of a crystal is a structured bibliographical entry as implemented by the <biblStruct> element in the TEI. It reflects a strong organization (<analytic>, <monogr> – with <imprint>, <series>), can be inserted at various places within the TEI architecture and it can be further decorated by other components or crystals (for example, <author>).

Without introducing any specific formalism here, we can outline the main aspects that have to be elicited to characterise a crystal as follows:

- The set of mandatory and optional components that may occur in the crystal
- The structural organization of the crystal, stating in particular the hierarchical relations between components
- The anchor points of the crystal where it can be further expanded
- The global semantics of the crystal

A crystal is thus a modelling tool that one can use to provide a coherent description of a subset taken from a more complex data model, as is typically the case with the TEI guidelines. To illustrate this, we can briefly present the way the TEI print dictionary chapter can serve as a basis to implement ISO 24613 and some consequences this could have on the data architecture we would recommend for the corresponding TEI elements.

As a starting point, let us consider the LMF subset depicted in Figure 1, which by essence implements the semasiological view on a lexical entry. This UML diagram states that a *Lexical Entry* is characterised by at least one *Form* component to which a hierarchically embedded series of *Sense*

components may be associated. The *Form* component if further refined by means of an optional *Form Representation* component, which can be used to represent the various concrete implementations of a form (e.g. phonetic, graphical, etc.)). Finally, each component of the meta-model (corresponding here to a UML class) can be further characterised by properties attached to each of them.

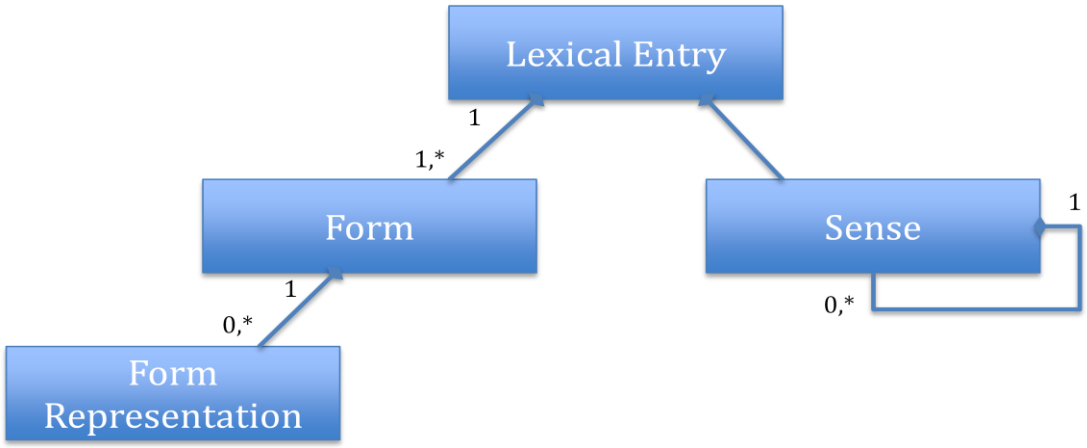

Figure 1 – The Lexical Entry sub-structure of ISO 24613 – LMF core package

Transposed within the TEI world, we can elicit the ideal counterpart as a TEI crystal rooted on the <entry> element. This crystal, depicted in Figure 2, states that the minimal lexical entry in a TEI sense is based on the <entry>, <form> and <sense> elements, with <form> being further decomposed by means of a series of elements implementing the *Form Representation* component of LMF[3]. The picture also introduces three new classes, which could gather up all further descriptive elements needed to refine <entry>, <form> and <sense>: model.entryDesc, model.formDesc, model.senseDesc.

This first presentation of the TEI lexical entry as a crystal illustrates how this concept may help describing complex structure that rely on constraints that go beyond (and deeper) than what we normally express by means of DTDs or schemas. Even if we will not systematically decompose the equivalences between ISO 24613 and the TEI in the following section, we hope that the preceding explanation will help the reader understanding the logic behind the various constraints we are now going to draw out. In a pattern quite analogous to the internal structure of the <cit> element, we see that we express the organisation of the various elements of this crystal as the actual combination of a structural description (direct dependency of an element to another) and a descriptive (further constraints applicable to the element) dimension.

[3] Ideally, this should correspond to model.formPart, but in the current version of the TEI guidelines this class is cluttered with other components which are there for purely syntactic (practical) reasons. We would limit this class to form <orth>, <pron>, <hyph>, <syll>, <stress>.

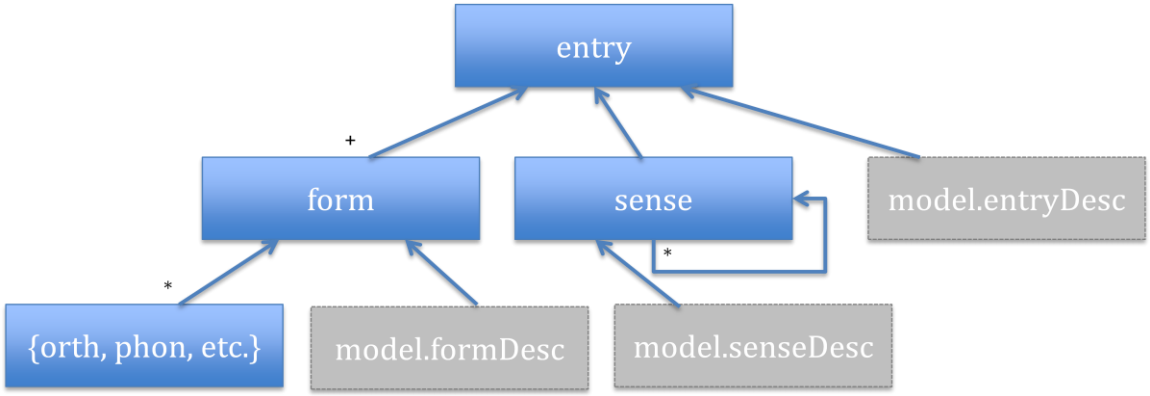

Figure 2 – the "ideal" element-class organisation of a TEI lexical entry

Morphographical descriptions

Form information in a semasiological structure corresponds to all the possible realisations of a word, whether graphical, phonetical or even iconical, which can be used as a trigger to identify the corresponding lexical unit. This may comprise abstract identifiers for the headword, namely the lemma, morphological components or categories (such as the consonantal pattern in Arabic), or any inflectional variant that can be associated with the entry.

The central issue in any morphographical description is that it should be based upon an abstract representation of *Form* as a component, which in turn groups together all the possible realisations of the corresponding form (the *Form Representation* component in LMF), as well as the associated constraints. In terms of good practices, one should thus refrain from providing a form representation (realisation) in isolation and always include it within an embedding <form> element. Besides, and unless there is only one form associated to a given lexical entry, the form type (eg. lemma or inflected form) should be marked to ensure its univocal identification.

As a consequence, the minimal structure associated with a TEI encoded lexical entry, where the only information given is that of a lemma (here, the French word "chat"; (en) "cat") should look as follows:

```
<entry>
      <form type="lemma">
            <orth>chat</orth>
      </form>
</entry>
```

On this basis, additional variants of the form (e.g. pronunciation) can be added to the same form container, together with complementary information characterising them. For instance, when more than one orthographies are used to provide the form, the appropriate type attribute should be used to qualify the corresponding orthography. In the following example, the lemma for the Korean word "치다" (*chida*; (en) "to hit") is provided in Hangul ((ko) "한글") orthography together with a Romanized form.

```
<form type="lemma">
   <orth type="한글">치다</orth>
   <orth type="romanized">chida</orth>
</form>
```

As exemplified here, we would advocate the definition of stable values for the @type attribute on <form>, adopting in particular the principle of naming orthography in their own originating language

When alternative forms are provided, indicating, for example, inflectional variations, then the variants should be entered in full in order to reflect the linguistic differences. For instance, we can take up the example ("clergyman") provided in Annex B of ISO LMF and reformulate it in TEI as follows:

```
<entry>
   <gramGrp>
      <pos>commonNoun</pos>
   </gramGrp>
   <form type="lemma">
     <orth>clergyman</orth>
   </form>
   <form type="inflected">
      <orth>clergyman</orth>
      <gramGrp>
         <number>singular</number>
      </gramGrp>
   </form>
   <form type="inflected">
      <orth>clergymen</orth>
      <gramGrp>
         <number>plural</number>
      </gramGrp>
   </form>
</entry>
```

Grammatical information

Grammatical information plays a transversal role within a dictionary entry. It is there to provide additional information to the core objects comprising the entry. In the lexicographic tradition grammatical information qualifies the "lemma", or rather — since the lemma is just a code representing the entry as a whole — syncretises the grammatical features that apply by default to all possible occurrences of the word. However, it can also occur at many other possible levels, qualifying inflected forms in a more precise way, indicating specific constraints associated to a sense, or even qualifying the occurrence within an example of phrasal expression. As a whole, grammatical features may be used at any place where the usage of a word is described.

In its usual realisation within human-oriented dictionaries, grammatical features can also be characterized as being highly informal at times, where a given grammatical constraint can for instance be represented by a prototypical morpheme (e.g. "der"/ "die"/ "das" to represent grammatical gender in German) or by means of descriptive phrase ("used in the plural form"). At best idiosyncratic codes are used (e.g. "masc.", "fém.") which are not always consistently applied within a single dictionary.

There is no doubt that such a situation prevents one from querying lexical entries with grammatical constraints in a coherent way. It is therefore a priority to establish requirements for the representation of grammatical features, which in all cases should preserve the initial editorial choices. As a basis for such recommendations we would actually recommend that TEI-based encoding of dictionary entries should be in keeping with the following elementary principles:

- Grammatical features should systematically be embedded within a <gramGrp> container element, even if only one feature is present;
- Whereas one should be flexible with the actual textual content of a grammatical descriptor, it is of utmost importance to normalize the actual intended value by mean of a @norm attribute.

For instance when a value for the grammatical gender is given by means of a determiner, the @norm attribute will provide the actual reference value (e.g. as a code from the ISOCat registry), as in for example:

```
<gramGrp>
      <gen norm="feminine">die</gen>
</gramGrp>
```

In general such grammatical descriptions should be thought of as being equivalent to the provision of feature structures and thus mappable onto an <fs> element. For instance, the preceding examples is equivalent to soothing like:

```
<fs>
      <f name="gender"><sym>feminine</sym></gen>
</fs>
```

The next stage in providing recommendation is to make sure that values for the @norm attribute are stable within a project and when possible across projects. At this stage we would recommend two complementary strategies:

- For a given project, document and publicize the values used for the norm attribute, so that the community may be aware of possible discrepancies;
- As recommended in ISO 12620, relate such values to entries in the ISO data category registry (isocat.org) so that they are mapped onto standardized conceptual references.

Let us remark that an item is on the TEI council agenda to better integrate mechanisms available in ISO 12620 within the TEI architecture to facilitate such mappings. We can thus expect that these recommendation may become standard practice within the TEI community.

Senses as systematic entry points

The representation level introduced by the Sense component in LMF and its counterpart <sense> in the TEI guidelines is an essential concept implementing the semasiological perspective of a dictionary. Still, a "lazy" encoding style for dictionary entries could lead to the idea that such a structure is superfluous when, for instance, a word can directly be described at the same level as the morphological and grammatical information by a simple definition or a translation. Indeed, it is often the case in the simplest forms of legacy lexical structures that senses are not explicitly separated out in the microstructure of the entry. We consider this as bad practice and recommend, in order to promote a better comparison of sense-related information in lexical entries, that the use of sense is systematized to enclose all descriptors that describe the *signified* (as opposed to the *signifier* in the Saussurian sense).

As can be observed from the variety of constraints that may apply to a sense element within a lexical entry, the underlying understanding of the semasiological model extends to the organisation of "senses" that do not rely on strict semantic criteria. This is not so much of a paradox when we think of the numerous ways by which semantic variation may be observed, among which we can include pure morpho-syntactic or syntactic markers. As a result, we consider that <sense> should be used to describe any subdivision reflecting the possible variations in usage for a given word. For instance, applying automatic collocation extraction tools may result in generating lexical entries automatically where senses correspond to the various collocation classes that the tool has determined.

We thus see the sense component/element as a generic container organizing the further description of a signifier, which may contain information related to:

- The actual syntactico-semantic restriction applicable to the sense being described: for instance by means of further grammatical constraints, a definition or some usage restriction;
- The provision of further illustrative information in particular contextualised examples or translations (see the section on the <cit> element below)
- Relational information referring to external information expressing the same meaning either within another lexical entry or an external ontological reference (e.g. Wordnet, Miller and Fellbaum 2007).

In order to actually facilitate further querying, it is important that each feature intended to be associated with a sense shall be precisely typed. This of course excludes the too general <note> element in this respect but it also requires that clearly defined typologies be associated with elements such as <usg> or, of course, <cit>. Furthermore, dictionary projects should be able to document precisely how much restrictive or illustrative information is inherited along embedded senses. For instance, a clear editorial strategy should state whether grammatical constraints replace or complete existing ones at a higher level of a sense hierarchy.

<cit> a generic linguistic quotation tool

The <cit> element in the P5 edition of the TEI Guidelines is the result of merging several previous constructs, which in the former editions of the TEI Print Dictionary chapter had been created to handle examples and translations in dictionary entries. The underlying aim of the new framework was twofold. On the one hand the objective was to provide greater coherence to the way language excerpts appear not only in dictionaries but in any textual content in general. On the other hand the TEI Council wanted to design a sound framework for dealing with additional references or constraints provided in complement to the quoted object itself, taking into account that such refinements may lead to recursive constructs. In terms of interoperability across TEI based applications the main vision behind <cit> is to provide entry points for generic searches for quoted language in texts from the point of view of both the full text content and of providing a systematized representation of complementary constraints.

Language quotations in text may indeed take many different forms. In dictionaries the most basic quotation is simply a phrase or sentence exemplifying the headword. Most of the times this quotation does not appear alone but is refined according to two main axis:

- indication of the source of the quotation, for instance (TEI P5 2.0.0): *'La valeur n'attend pas le nombre des années' (Corneille)*
- provision of further usage information, stating further constraints that the example is bound by, such as domain or pronunciation, as in (TEI P5 2.0.0): **some** … 4. (*S~* and *any* are used with *more*): Give me ~ more**/s@'mO:(r)/**

In the case of multilingual dictionaries the scope of quotations extends to all language excerpts which are used to provide equivalences for the entry (or sub-sense thereof) in the target language. In a way that is similar to the monolingual case, further refinement may indicate some source or usage information, but it may also document the target language proper. A usual case here is, for instance, the indication of the grammatical gender of a noun equivalent in the target language.

Quotation constructs are not covered in LMF but can easily be modelled as an extension to the LMF core packages. In Figure 3 one can see a simple representation for such an extension. The approach is similar to that advocated for grammatical information concerning semantic grouping and we see that the model embeds the quoted text into a quotation construct, even if no refinement is actually stated.

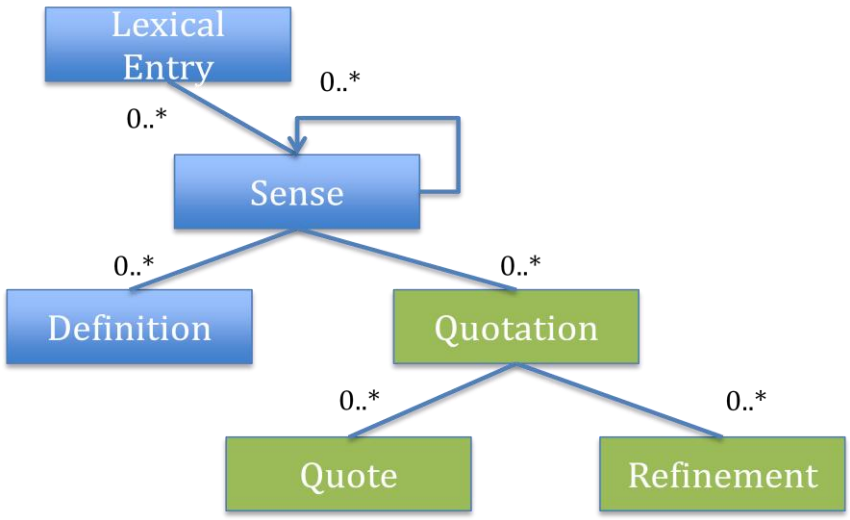

Figure 3: an LMF extension for represented quotations in dictionary

In the TEI Guidelines, the quotation construct is implemented by means of the <cit> element, which has the following characteristics:

- the quoted object may be realized not only by means of a <quote> (or <q> from the *model.qLike* class), but also as a more elaborated construct such as an XML object (<egXML>, through *model.egLike*);
- the refinement can be instantiated either as a bibliographic reference (*model.biblLike*), a pointer or external reference to a constraint (model.ptrLike), specific lexicographic features such as grammatical constraints (*model.entryPart*) or through the almost accidental inclusion of model.global in the content model of <cit>, by means of a feature structure. It should be noted that a refinement can actually be a <cit> (member of model.entryPart), thus offering a natural way to provide a translation for example ;
- the TEI Guidelines already systematize the values of the @type attribute to 'example' and 'translation' for the usages in dictionaries.

Given the variety of possible cases where <cit> may be used and the potentially infinite combinations of refinement that one could think of it may indeed be difficult to provide clear requirements for its application. Basically a proper usage of <cit> should always allow a human reader or a processor to identify one quoted object and treat all other components as refinements whose semantics is to be understood in a conjunctive way (in other words, all refinements apply en bloc to the quoted object). By default the quoted object is the first child of the <cit> element, or in general the first child that is a member of either *model.qLike* or *model.egLike*.

In the second part of this paper we will see several applications of <cit> in the context of our observational corpus, but we can quickly illustrate some basic usages of this element on the basis of examples available in the TEI Guidelines.

In the following case, which we should qualify as prototypical, a simple example for the headword is associated with a refinement giving the pronunciation of part of the quoted text:

```
<cit type="example”>
   <quote>Give me <oRef/> more</quote>
   <pron extent="part">s@'mO:(r)</pron>
</cit>
```

The next example illustrate the representation of a translation refined with grammatical feature:

```
<cit type="translation" xml:lang="fr">
   <quote>habilleur</quote>
   <gramGrp>
      <gen>m</gen>
   </gramGrp>
</cit>
```

Finally, we cannot resist presenting a recursive case where the embedded <cit> is used as an additional descriptive element for the quote text at the higher level.

```
<cit type="example”>
   <quote>she was horrified at the expense.</quote>
   <cit type="translation" xml:lang=”fr”>
      <quote>elle était horrifiée par la dépense.</quote>
   </cit>
</cit>
```

Illustrated guidelines for early printed dictionaries

Lexicographical justification

We test our encoding concepts using printed dictionaries from the second half of the 18th century for two reasons. First, in the history of English lexicography the early 18th century marks the beginning of modern dictionary practice (Landau 2001, 52). Samuel Johnson's 'Dictionary of the English Language', first published in 1755, perfectly embodies these advances in lexicography. Johnson is the first English lexicographer to include thousands of other quoted 'authorities' within his text as illustrations of word use" (Reddick 1996, 9). His dictionary also brought together "for the first time key conventions for future dictionary presentation: the folio design is a system of typography that displays the structure of each entry, though there are inconsistencies of abbreviation and ambiguities" (Luna 2005, 193). So this dictionary offers an ideal test bed to study problems in providing a consistent model for the encoding of early printed dictionaries using TEI P5.

Second, because Johann Christoph Adelung translated Samuel Johnson's dictionary into German (under the title 'Neues grammatisch-kritisches Wörterbuch der Englischen Sprache für die Deutschen; vornehmlich aus dem größern englischen Werke des Hrn. Samuel Johnson nach dessen vierten Ausgabe gezogen und mit vielen Wörtern, Bedeutungen und Beyspielen vermehrt'), starting with Johnson's dictionary also opens up additional perspectives for the study of bilingual lexicographical resources in the 18th century and research into the history of revision and the reuse of dictionaries. Adelung's translation is a rare book. Since his name does not appear on the title page nor elsewhere in the front matter, his role as a translator is little-known. It is worthwhile mentioning the publication context, because Adelung's critical examination and translation of Johnson's dictionary bridges the gap between his work on the two editions of his own German dictionaries. The first volume of his translation, containing the letters A to J, was published in 1783. This was after nearly three years of work – according to his preface p.XII – and even before he finished the fifth and last volume of the first edition of his German dictionary which he had started in 1773 ('Versuch eines vollständigen grammatisch-kritischen Wörterbuches der Hochdeutschen Mundart...1786). Thirteen years later, in 1796, he published the second volume of his translation with the letters K to Z, after having finished the first two volumes of the second and final edition of his German dictionary (Grammatisch-kritisches Wörterbuch der Hochdeutschen Mundart ... Zweyte vermehrte und verbesserte Ausgabe, Erster Theil. von A – E...1793; Zweyter Theil, von F – L...1796).

Almost at the same time Johannes Ebers used Adelung's lexicographical materials to compile a German-English counterpart in three volumes: 'New and Complete Dictionary of the German and English Languages composed chiefly after the German Dictionaries of Mr. Adelung and of Mr. Schwan. Every German Word being rendered into proper English and thoroughly enriched with

Phrases and Terms of Arts and Sciences: A Work, which will be useful and even indispensible, and therefore welcome to all such as have a Mind to translate or read the Works of either of the two Languages, elaborated by John Ebers'. Volume I, Containing the Letters A - G of the German Alphabet explained in English, Leipzig 1796'.

We test our modelling of lexicographic structures with three samples of Johnson's monolingual dictionary representing the most frequent word-classes: the adjective 'ABLE', the verb 'To APPLAUD' and all entries for the noun 'APPLE'. We further compare Johnson's apple entries with the section of apple entries in Adelung's bilingual English-German translation of Johnson's dictionary. To illustrate the differing encoding structures of bilingual German-English dictionaries we use Eber's entry 'fähig', the equivalent of 'ABLE'. As a source for this entry Ebers obviously used only the German-French dictionary of Christian Friedrich Schwan. So for reasons of reference we include Schwan's entry 'fähig' in order to illustrate dictionary reuse across languages in the 18th century. The complete sample encodings of the entries and images of the encoded pages are added as a supplement.

Typographic analysis and text encoding

Paul Luna starts his essay on the typographic design of Johnson's dictionary with some reflexions on how a typographer would analyse a dictionary: "In particular, how does a typographer look at a dictionary that is also a cultural artifact, as Samuel Johnson's *Dictionary of the English Language* undoubtedly is?" (Luna 2005, 175). Building on a more wide-ranging definition of typography as 'configuration of verbal graphic language' Luna concludes: "the main concern of this essay is not the quality of the printing, nor the nature of the paper, nor even the origin of the founts of type used to compose the *Dictionary*, but how its visual presentation reflects the structure of the text, its usability, and perhaps even its compiler's intentions" (Luna 2005, 175).

This concept comes very close to what a TEI encoding of a dictionary in an adequate granularity should achieve: reflecting the structure of the encoded text, facilitating re-usability in electronic form and – at its best – assisting in the detection of the author's intentions. In order to put our aim of a consistent modelling of even heterogeneous structures into practice we follow some basic principles.

In general we adopt a conservative editorial view for our literal transcription (cf. TEI P5 Guidelines 9.5.1 Editorial View.) and try to keep the latter close to the printed original: we do not add any character to the original text or delete it; we do not make changes in the order of the source text, we preserve the linear structures of the text with <pb>, <cb> and <lb>, and we retain the end-of-line hyphenation at this stage (cf. TEI P5 Guidelines 3.2.2 Hyphenation.). There are several good reasons for this, one of which is the scope of orthographical variation in the texts of the dictionaries. For clarity and to ensure a consistent encoding we encode only a few structurally important typographic features at the level of the lexical entry. We do not encode the indentation and alignment structure, nor do we encode italics in the contexts of part-of-speech labels (in a <pos> element), of cited forms in <etym> (if printed in italics), of the lemma used in illustrative quotations (in a <cit> element), or of the names of authors and their works in the sources for the illustrative quotations (in a <bibl> element).

Encoding practise on <entry>-level

With re-usability, interoperability and sustainability of the dictionary entries in mind, we use two attributes to customize the <entry>-element: 'xml:id' to guarantee a robust and reliable non-ambiguous identification and 'type' for the differentiation of the entries.

The 'xml:id' is composed of four parts, separated by a dot: 1. two initials of the author's name and a combination of six letters or numbers to identify the encoded edition precisely; 2. four digits for the

year of publishing; 3. six digits for the running number of the entry (set to random values in the examples); 4. the lemma, transcribed in lower case only and with any incidental spaces replaced by underlines. Thus our sample entry 'ABLE' in Samuel Johnson's dictionary is assigned the xml:id 'sjdict1f.1755.000123.able'. The first part is composed of 'sj', taken from Samuel Johnson, 'dict' reflects the title 'Dictionary of the English Language', '1' indicates the edition and 'f' the format 'folio', because edition and format are both rather important for a precise identification of the different printed editions of Johnson's dictionary. They are not necessary for Adelung (Henne 2001, 170), Ebers (Lewis 2012) and Schwan.

We use the 'type' attribute of <entry> to distinguish typographically or verbally marked types of entries and map them, as explained above, onto corresponding identifiers of the ISOcat data category registry (ISO 12620). In the majority of cases these types are 'main entry' or 'cross reference', with the only exception being attributes for the entries in Johnson's *Dictionary*.

An occasional user of Johnson's *Dictionary* may be puzzled about aspects of the typesetting of entry headwords. Thus APPLAUD and APPLE are in full caps, while APPLAUSE and APPLE TREE are in small caps. Now and then, however, entries appear typeset in italic capital letters, e.g. *ABORIGINES* and *ABRACADABRA* . In his preface Johnson explains the background for these marked differences, which for him reflect basic lexicographical distinctions: "In the investigation both of the orthography and signification of words, their ETYMOLOGY was necessarily to be considered, and they were therefore to be divided into primitives and derivatives. A primitive word, is that which can be traced no further to any *English* root; ... . Derivatives, are all those that can be referred to any word in *English* of greater simplicity." (Johnson 1755, 3f.) Thus primitives or roots are marked by 'full caps' and the derivatives by 'small caps'. The entries in italic capital letters indicate foreign words used in the English language (cf. Luna 2005, 181).

As Paul Luna notices (Luna 2005, 196 fn. 24), this distinction of entries echoes a completely different way of organizing a dictionary: word-families, represented by roots (in alphabetical order), followed by their derivatives (ordered non-alphabetically into morphological or etymological groups). Since Johnson used a single alphabetical order for all entries this organizing principle is no longer clearly visible. It is only faintly reflected in the differentiation of the lemmas. But it is still present and that is why we think it should be encoded explicitly as a significant feature of the dictionary structure. Accordingly, we map the entries in Johnson's *Dictionary* onto the ISOcat identifiers 'root', 'derivation', 'foreign', 'phrase' or 'cross-reference'. Two examples: 'ABLE' and 'APPLE *of Love*'.

```
<entry xml:id="sjdict1f.1755.000123.able" type="root">
   <form type="lemma" norm="able">
   <lb/><orth rend="ALLCAPS">A'BLE</orth><pc>.</pc>
      <gramGrp><pos norm="adjective"><abbr>adj.</abbr></pos></gramGrp>
   </form>


<entry xml:id="sjdict1f.1755.000346.apple_of_love" type="phrase">
   <form type="lemma" norm="apple of love">
      <lb/><orth><hi rend="smallcaps">APPLE</hi> <hi rend="italics">of
      Love</hi></orth><pc>.</pc>
      <gramGrp><pos norm="noun"/></gramGrp>
   </form>
   <sense>
       <cit type="encyclopedic_information">
       <quote><lb/>Apples of love are of three sorts; ...
       <bibl><author>Mortimer</author>'s <title>Husbandry</title>.</bibl>
       </cit>
   </sense>
</entry>
```

The typography of the entry — small caps for 'apple' though belonging to the 'root', italics for 'of love', and the missing word class information — indicate uncertainty about the word status of the entry. The classification as type 'phrase' may require some explanation. Valerie Adams comments in her introduction to word-formation on the distinction between 'words' and 'phrases': "Certain noun-preposition-noun phrases also show their incomplete unification by the possibility of pluralizing the first noun" (Adams 1976, 9). Since the illustrative quotation of Mortimer's book on Husbandry starts with the plural form 'apples', we regard the type 'phrase' here as justified.

The <form>-block

The <form>-element is designed to contain information on the written form — encoded using <orth> — and, if present, the spoken form — encoded using <pron> — of one headword. We use <form> with two attributes: a 'type'-attribute to distinguish the lemma from possibly listed inflected forms and – once again with re-usability, interoperability and reliable searching in mind – a 'norm'-attribute to even out orthographic variation, e.g. the use of upper or lower case, hyphenation or the use of special markers to indicate the stress position within the orthographic representation of the lemma. Johnson uses here a single prime and Adelung first adopts Johnson's stress marker in the letter range A to C. Starting within the letter range 'Co' he also extends the indication of the stress position to the distinction of long and short vowels: e. g. 'Còcoa' versus 'Cócquet', 'Cóctile'. A <stress> element, designed for stress patterns given separately, is not applicable here, apart from the fact that we did not want to split up the orthographic representation any further or change it. Ebers and Schwan do not indicate stress positions.

Within <orth> we use a 'rend' attribute depending on the context. In Johnson's *Dictionary* we use it to store his typographic differentiation of the entries. In Schwan's dictionary we apply it to distinguish two different orthographic representations of the German lemma, the first with Roman capital letters only, the second with upper and lower case, depending on the German orthography and in 'fraktur' typeface.

As the second component of form we use <gram Grp> to collect grammatical information such as part-of-speech (in a <pos> element) or gender (in a <gen> element). Quite often grammatical information precedes or follows the orthographic representation of the entry, e.g. the infinitive marker 'To' in entries for verbs in Johnson's dictionary or the determiner 'der, die, das' in German noun entries. We capture this information with a <gram> element and a 'type' attribute added for the ISOcat value. Without exception we store all elements that interpret grammatical features like <pos> or <gen> within the <gramGrp> element, once again using a 'norm' attribute to map the different grammatical descriptions given in the dictionaries on an ISOcat determiner. This way we avoid conflicts with the serial order of the printed data and can adjust inconsistencies like missing word class information, e.g. by adding an empty <pos> element with a 'norm' attribute based on information collected elsewhere in the entry. One example is Johnson's entry 'APPLAUD':

```
<pb n="148"/><cb n="APP"/>
<entry xml:id="sjdict1f.1755.000234.applaud" type="root" >
   <lb/><form type="lemma" norm="applaud">
      <gram type="infinitive-marker">To</gram>
      <orth rend="ALLCAPS">APPLA'UD</orth><pc>.</pc>
      <gramGrp><pos norm="verb"><abbr>v.a.</abbr></pos></gramGrp>
   </form>
   <etym>
   <pc>[</pc><mentioned xml:lang="la">applaudo</mentioned><pc>,</pc>
      <lang><abbr>Lat.</abbr></lang><pc>]</pc>
   </etym>
   <lb/><sense>
      <num>1.</num>
      <def>To praise by clapping the hand.</def>
   </sense>
```

```
   <lb/><sense>
      <num>2.</num>
      <def>To praise in general.</def>
   </sense>
   <cit type="example">
      <lb/><quote>I would applaud thee to the very echo,
      <lb/>That should applaud again.</quote>
      <bibl><author><abbr>Shakesp.</abbr></author><title>Macbeth</title>.</bibl>
   </cit>
   <cit type="example">
      <lb/><quote>Nations unborn your mighty names shall sound,
      <lb/>And worlds applaud that must not yet be found!</quote>
      <bibl><author>Pope</author>.</bibl>
   </cit>
</entry>
```

Our use of <pc> is governed by the principle that we avoid punctuation marks as delimiters of text in elements within <form>, such as <orth>, and within <etym>, such as <mentioned>; this is for ease of reusability and searching.

In testing our encoding concept we encountered some phenomena — 'word class' in grammar and 'hyphenation' in orthography — which prompted us to reinforce our aim of consistently modelling heterogenous lexicographical data through normalization. The first case has to do with an old problem of word classes: 'adjective' and 'adverb' in German. Ebers defines the part-of-speech information in his entry 'fähig' with the abridged terms in Latin 'adj. et adv.', 'Adjectivum et Adverbium'. This concept – one word, two word classes – is not compatible with the present-day understanding of word classes in German: since 'adverbs' in German generally are never inflected and 'fähig' is capable of inflection, this word is generally regarded as an 'adjective' in any present-day dictionary of German. Of course, we do not alter Ebers' word class definition, but we suggest resolving the word class conflict in this and in comparable cases by standardizing the 'norm' attribute in <pos> by using the ISOcat value 'adjective' only. Ebers' example entry 'fähig 'in abridged form:

```
<entry xml:id="jedictge.1788.000999.fähig" type="main_entry">
   <form xml:lang="de" type="lemma" norm="fähig">
      <lb/><orth>Fähig</orth><pc>,</pc>
      <gramGrp>
         <pos norm="adjective">
         <abbr xml:lang="la">adj.</abbr> et <abbr xml:lang="la">adv.</abbr>
         </pos>
      </gramGrp>
   </form>
   <sense> ... </sense>
</entry>
```

The second phenomenon has to do with hyphenation, an old problem primarily but not only in the English Language. First, consider Johnson's noun compounds with 'apple' in abridged form:

```
<entry xml:id="sjdict1f.1755.000347.apple-graft" type="derivation">
   <form type="lemma" norm="apple graft">
      <lb/><orth rend="smallcaps">APPLE-GRAFT</orth><pc>.</pc>
      <gramGrp><pos norm="noun"><abbr>n.s.</abbr></pos></gramGrp>
   </form>
   <etym><pc>[</pc>from
      <mentioned corresp="xml:id=sjdict1f.1755.000345.apple">apple</mentioned
      <lbl>and</lbl>
      <mentioned corresp="xml:id=sjdict1f.1755.009999.graft">graft</mentioned>
      <pc>.]</pc>
   </etym>
<sense> ... </sense>
```

```
</entry>

<entry xml:id="sjdict1f.1755.000348.apple-tart" type="derivation">
   <form type="lemma" norm="apple tart">
      <lb/><orth rend="smallcaps">APPLE-TART</orth><pc>.</pc>
      <gramGrp><pos norm="noun"/></gramGrp>
   </form>
   <etym><pc>[</pc>from
      <mentioned corresp="xml:id=sjdict1f.1755.000345.apple">apple</mentioned>
      <lbl>and</lbl>
      <mentioned corresp="xml:id=sjdict1f.1755.029999.tart">tart</mentioned>
      <pc>.]</pc>
   </etym>
<sense> ... </sense>
</entry>

<entry xml:id="jdict1f.1755.000349.apple_tree" type="derivation">
   <form type="lemma" norm="apple tree">
      <lb/><orth rend="smallcaps">APPLE TREE</orth><pc>.</pc>
      <gramGrp><pos norm="noun"><abbr>n.s.</abbr></pos></gramGrp>
   </form>
   <etym><pc>[</pc>from
      <mentioned corresp="xml:id=sjdict1f.1755.000345.apple">apple</mentioned>
      <lbl>and</lbl>
      <mentioned corresp="xml:id=sjdict1f.1755.039999.tree">tree</mentioned>
      <pc>.]</pc>
   </etym>
<sense> ... </sense>
</entry

<entry xml:id="jdict1f.1755.000350.apple_woman" type="derivation">
   <form type="lemma" norm="apple woman">
      <lb/><orth rend="smallcaps">APPLE WOMAN</orth><pc>.</pc>
      <gramGrp><pos norm="noun"><abbr>n.s.</abbr></pos></gramGrp>
   </form>
   <etym><pc>[</pc>from
      <mentioned corresp="xml:id=sjdict1f.1755.000345.apple">apple</mentioned>
      <lbl>and</lbl>
      <mentioned corresp="xml:id=sjdict1f.1755.049999.woman">woman</mentioned>
      <pc>.]</pc>
   </etym>
   <sense> ... </sense>
</entry>
```

Apart from the special case 'APPLE *of love*', both APPLE-GRAFT and APPLE-TART are hyphenated, whereas APPLE TREE and APPLE WOMAN are spelled as two separate words. There is no consistent distinction here between open (word-spaced) and hyphenated compounds. Noel Osselton gives a compact résumé of the puzzling phenomenon 'variation of hyphenated compounds' as entries and their steady downgrading in the second half of the dictionary from the letter M onwards (Osselton 2005). Against this background we have modelled the 'norm'-attribute of <form> in order to provide the best support for search procedures: we have retained the original hyphenated and open compound spellings (as in Johnson's text), but have encoded the open or word-spaced form on the 'norm'-attribute as the standardized form.

In his translation of Johnson's apple entries Adelung takes a different view. He unifies the hyphenated spelling for all the apple compounds, downgrades the hybrid entry 'Apple of love' to appear as a form mentioned within the base entry 'apple', and adds more compounds, taken from other sources mentioned in the introduction (abridged examples, full transcription in appendix):

```
<entry xml:id="jagkwbed.1783.000999.apple" type="main_entry">
   <form  xml:lang="en" type="lemma" norm="apple">
      <lb/><orth>'Apple</orth><pc>,</pc>
      <gramGrp xml:lang="de">
      <pos norm="noun"><abbr xml:lang="la">subst.</abbr> </pos>
      </gramGrp>
      <pc>(</pc><pron>äpp'l</pron><pc>,</pc>
   </form>
   <etym><mentioned><lang xml:lang="ang"><abbr>angels.</abbr></lang>
       <lang xml:lang="nds"><abbr>niederd.</abbr></lang>aep- <lb/>pel</mentioned>
       <pc>,</pc> <mentioned><lang xml:lang="de">deutsch</lang> Apfel</mentioned>
       <pc>.</pc><pc>)</pc>
   </etym>
   <sense xml:lang="de">
      <num>1)</num>
      <def>Die Frucht des <lb/>Apfelbaumes,</def>
      <cit type="translation"><quote>der Apfel.</quote></cit>
   </sense>
   <sense xml:lang="de">
      <num>2)</num>
      <cit type="encyclopedic_information">
      <quote>Wegen eini-<lb/>ger Ähnlichkeit in der Gestalt ...</quote>
      </cit>
      <cit type="encyclopedic_information">
      <quote><mentioned xml:lang="en">The Apple of love, Love-apple</mentioned>
         o-<lb/>der <mentioned xml:lang="en">Wolf's Peach</mentioned>,
      <cit type="translation" xml:lang="de"><quote>Liebesapfel</quote>
      </cit>
      <term xml:lang="la">Lycoper-<lb/>sicon<name nymRef="Linné">Linn.</name>
      </term>auch wohl eine Art des <term xml:lang="la">Sola-<lb/>num</term>;
      <mentioned xml:lang="en">the Mad-apple</mentioned>,
   </sense>
   <sense xml:lang="de">
      <num>3)</num>
      <usg>Figürlich,</usg><def>die Pupille in dem Auge,</def>
      <cit type="translation"><quote>der <lb/>Augapfel,</quote></cit>
      <xr type="synonym "><lbl>welcher wohl auch
         <ref xml:lang="en" target="xml:id=adwbeng1.1783.009999.eye-ball">
         Eye-ball</ref> ge-<lb/>nannt wird.</lbl>
      </xr>
   </sense>
</entry>

<entry xml:id="jagkwbed.1783.001000.apple-coar" type="main_entry">
   <form xml:lang="en" type="lemma" norm="apple coar">
      <lb/><orth>'Apple-coar</orth><pc>,</pc>
      <gramGrp xml:lang="de"><pos norm="noun">subst.</pos></gramGrp>
   </form>
   <etym><lbl>von</lbl>
      <mentioned xml:lang="en" corresp="#adwbeng1.1783.000999.apple">
      apple 1)</mentioned>
   </etym>
   <sense>
      <def>der Griebs oder Gröbs in dem Apfel.</def>
   </sense>
</entry>

<entry xml:id="jagkwbed.1783.001001.apple-graft" type="main_entry">
   <form xml:lang="en" type="lemma" norm="apple graft">
      <lb/><orth>'Apple-graft</orth><pc>,</pc>
      <gramGrp><pos norm="noun">subst.</pos></gramGrp>
   </form>
```

```
   <sense>...</sense>
</entry>

<entry xml:id="jagkwbed.1783.001002.apple-loft" type="main_entry">
   <form xml:lang="en" type="lemma" norm="apple loft">
      <lb/><orth>'Apple-loft</orth><pc>,</pc>
      <gramGrp><pos norm="noun">subst.</pos></gramGrp>
   </form>
   <sense>...</sense>
</entry>

<entry xml:id="jagkwbed.1783.001003.apple-monger" type="main_entry">
   <form xml:lang="en" type="lemma" norm="apple monger">
      <lb/><orth>'Apple-monger</orth><pc>,</pc>
      <gramGrp><pos norm="noun">subst.</pos></gramGrp>
   </form>
   <sense>...</sense>
</entry>

<entry xml:id="jagkwbed.1783.001004.apple-paring" type="main_entry">
   <form xml:lang="en" type="lemma" norm="apple paring">
      <lb/><orth>'Apple-paring</orth><pc>,</pc>
      <gramGrp><pos norm="noun">subst.</pos></gramGrp>
   </form>
   <sense>...</sense>
</entry>

<entry xml:id="jagkwbed.1783.001005.apple-roaster" type="main_entry">
   <form xml:lang="en" type="lemma" norm="apple roaster">
      <lb/><orth>'Apple-roaster</orth><pc>,</pc>
      <gramGrp><pos norm="noun">subst.</pos></gramGrp>
   </form>
   <sense>...</sense>
</entry>

<entry xml:id="jagkwbed1.1783.001006.apple-squire" type="main_entry">
   <form xml:lang="en" type="lemma" norm="apple squire">
      <lb/><orth>'Apple-squire</orth><pc>,</pc>
      <gramGrp><pos norm="noun">subst.</pos></gramGrp>
   </form>
   <sense>...</sense>
</entry>

<entry xml:id="jagkwbed.1783.001007.apple-tart" type="main_entry">
   <form xml:lang="en" type="lemma" norm="apple tart">
      <lb/><orth>'Apple-tart</orth><pc>,</pc>
      <gramGrp><pos norm="noun">subst.</pos></gramGrp>
   </form>
   <sense>...</sense>
</entry>

<entry xml:id="jagkwbed.1783.001008.apple-thorn" type="main_entry">
   <form xml:lang="en" type="lemma" norm="apple thorn">
      <lb/><orth>'Apple-thorn</orth><pc>,</pc>
      <gramGrp><pos norm="noun">subst.</pos></gramGrp>
   </form>
   <sense>...</sense>
</entry>

<entry xml:id="jagkwbed.1783.001009.apple-tree" type="main_entry">
   <form xml:lang="en" type="lemma" norm="apple tree">
      <lb/><orth>'Apple-tree</orth><pc>,</pc>
```

```
      <gramGrp><pos norm="noun">subst.</pos></gramGrp>
   </form>
   <sense>...</sense>
</entry>

<entry xml:id="jagkwbed.1783.001010.apple-woman" type="main_entry">
   <form xml:lang="en" type="lemma" norm="apple woman">
      <lb/><orth>'Apple-woman</orth><pc>,</pc>
      <gramGrp><pos norm="noun">subst.</pos></gramGrp>
   </form>
   <sense>...</sense>
</entry>
```

These examples illustrate that, despite differences in detail, the <entry> and <form>-information can be encoded using the same pattern. Missing standard information (like wordclass) can be supplied without modification of the printed text. Even if the encoding cuts into typographical structures (e.g. <pron> in Adelung's dictionary), it does not corrupt the text.

<etym> between etymology and word-formation

As noted above, Johnson emphasized in his preface the importance of etymology. Accordingly, he started his dictionary with a grammar, which is defined as follows: "Grammar, which is the art of using words properly, comprises four parts; Orthography, Etymology, Syntax, and Prosody" (Johnson 1755, 39). The part on etymology opens with the definition: "ETYMOLOGY teaches the deduction of one word from another" (Johnson 1755, 44) and the introduction to the chapter 'Of DERIVATION' finally points to Johnson's core concept: "That the English language may be more easily made understood, it is necessary to enquire how its derivative words are deduced from their primitives, and how the primitives are borrowed from other languages" (Johnson 1755, 47). In compound word entries he uses square brackets following the part-of-speech information to mark the root components of the compound – his derivatives (eg. in APPLE-GRAFT: [from *apple* and *graft*]); for root entries he provides information about related words in the Indo-European, Romance or Germanic languages, if necessary with an English translation (e. g. in ABLE: [*habile*, Fr. *habilis*, Lat. Skilful, ready.]). In accordance with Johnson's concept we suggest using the <etym>-element for both cases. The <etym> element requires no additional attribute, since its content structure clearly indicates to what type of entry <etym> belongs and how it has to be interpreted:

```
<entry xml:id="sjdict1f.1755.000123.able" type="root">
   <form>...</form>
   <etym><pc>[</pc>
      <mentioned xml:lang="fr" >habile</mentioned><pc>,</pc>
         <lang><abbr>Fr.</abbr> </lang>
      <mentioned xml:lang="la">habilis</mentioned><pc>,</pc>
         <lang><abbr>Lat.</abbr></lang>
      <lb/><gloss xml:lang="en">Skilful<pc>,</pc> ready<pc>.</pc>
      </gloss><pc>]</pc>
   </etym>
```

In the entry 'ABLE' the content of <etym> consists of <mentioned> plus <lang>, possibly combined with <gloss>, so it must be a root entry.

```
<entry xml:id="sjdict1f.1755.000347.apple-graft" type="derivation">
   <form>...</form>
   <etym><pc>[</pc>from
      <mentioned corresp="xml:id=sjdict1f.1755.000345.apple">apple</mentioned
      <lbl>and</lbl>
      <mentioned corresp="xml:id=sjdict1f.1755.009999.graft">graft</mentioned>
      <pc>.]</pc>
```

```
   </etym>
```

In the entry 'APPLE-GRAFT' the content of <etym> consists of two <mentioned> elements, both with 'corresp' attributes that point to other entries within the same source, so it must be a derivation. The price to pay for this concept is the additional effort of identifying and inserting the corresponding 'xml:id's. But from our point of view the resulting network of linked entries is worth the effort.

Stepwise refinement of <sense>: <num>,<def> and <gramGrp> with <gram>

The function of <sense> as a container for the semasiological information of dictionary entries has been outlined above. Some sections of the encoding of 'ABLE' can illustrate the flexibility of the concept of "crystals" (cf. p. 2f.) for the encoding of complex semantic structures. The first step of refinement adds <num>-elements to label the different numbered <sense>-values which are typical for the semantic analysis of Johnson's dictionary.

```
<entry xml:id="johdel_1755_1_a00000_able">
   <form> ... </form>
   <etym> ... </etym>
   <sense>
      <lb/><num>1.</num>
      <def>...</def><cit>...</cit><cit>...</cit>
   </sense>
   <sense>
      <lb/><num>2.</num>    <def>Having    power    sufficient;    enabled.</def>
         <cit                                                    type="example">
            <lb/><quote>All  mankind  acknowledge  themselves  <hi  rend="italics">
            able</hi>  and  sufficient  to  <lb/>  do  many  things,  which
            actually they never do.
            </quote>
            <bibl><author>South</author>'s <title>Serm.</title></bibl>
        </cit>
   </sense>
   <sense>
      <lb/><num>3.</num>
      <gramGrp>
         <gram>Before a verb, with the participle<hi rend="italics">to</hi></gram>
      </gramGrp>,
      <def>it signifies generally hav-<lb/>ing the power</def>;
      <gramGrp>
         <gram> before a noun, with <hi rend="italics">for</hi></gram>
      </gramGrp>,
      <def> it means <hi>qualified</hi></def>.
```

In a second step — <num>3.</num> — one <sense>-element is used to combine the morpho-syntactic features '*able* + *to* before a verb' in the <gramGrp>-container with <gram> and the semasiological definition 'signifies generally having the power' contained in the <def>-element. In a comparable construction with *able*, the morphosyntactic feature 'before a noun, with *for*' in <gramGrp> and <gram> is connected with the definition 'it means qualified' in <def>.

```
<cit                                                    type="example">
   <lb/><quote>Wrath is cruel, and anger is outrageous; but who is
   <hi   rend="italics">able   <lb/>   to</hi>   stand   before   envy?</quote>
   <bibl><title>Prov.</title>
      <biblScope type="part">xxvii.</biblScope>
      <biblScope type="ll">4.</biblScope>
   </bibl>
</cit>
```

```
<cit                                                                type="example">
   <lb/><quote>There   have   been   some   inventions   also,   which   have   been
   <lb/><hi rend="italics">able for</hi> the utterance of articulate sounds,
   as        the        speaking        of        <lb/>certain        words.</quote>
   <bibl><author> Wilkin</author>’s <title>Mathematical Magic</title>.
   </bibl>
</cit>
```

The <cit>-examples that follow in <sense><num>3.</num> repeat the structures and illustrate both usages. Typographically the respective phrases ‘*able to*’ and ‘*able for*’ are marked by italics in the quoted text. Of course, the refinement of the encoding could be extended to word level and features of syntactical analysis, building on the encoding reached so far. But this is beyond what we wanted to illustrate in this paper.

Bilingual dictionaries: a shift of perspective

The consistent modelling of heterogeneous lexical structures, we suggest, can be extended to the more complex structures we find in the two bilingual dictonaries, Adelung’s English-German translation of Johnson’s dictionary and Ebers’ ‘New and Complete Dictionary of the German and English Languages’ compiled using Adelung’s and Schwan’s lexicographical materials. At the same a comparable precision in the encoding can be achieved. Let us first compare the entry ‘Apple-tart’ in Johnson’s dictionary and Adelung’s translation:

```
<entry xml:id="sjdict1f.1755.000348.apple-tart" type="derivation">
   <form type="lemma" norm="apple tart">
      <lb/><orth rend="smallcaps">APPLE-TART</orth><pc>.</pc>
      <gramGrp><pos norm="noun"/></gramGrp>
   </form>
   <etym><pc>[</pc>from
      <mentioned corresp="xml:id=sjdict1f.1755.000345.apple">apple</mentioned>
      <lbl>and</lbl>
      <mentioned corresp="xml:id=sjdict1f.1755.029999.tart">tart</mentioned>
      <pc>.]</pc>
   </etym>
   <sense>
      <def>A tart made of apples.</def>
      <cit type="example">
         <lb/><quote>What, up and down carv’d like an apple-tart.</quote>
         <lb/><bibl><author>Shakespeare</author>'s
         <title>Taming of the Shrew</title>.
         </bibl>
      </cit>
   </sense>
</entry>

<entry xml:id="jagkwbed.1783.001007.apple-tart" type="main_entry">
   <form xml:lang="en" type="lemma" norm="apple tart">
      <lb/><orth>'Apple-tart</orth><pc>,</pc>
      <gramGrp xml:lang="de" ><pos norm="noun">subst.</pos></gramGrp>
   </form>
   <sense xml:lang="de" >
      <def>eine Torte von Ä-<lb/>pfeln,</def>
      <cit type="translation"><quote>eine Äpfeltorte.</quote></cit>
   </sense>
</entry>
```

In contrast to Johnson, Adelung, meeting the requirements of an English-German dictionary, left out the <etym>-element on word-formation and the Shakespeare quotation. He translated Johnson’s

definition of ‘apple-tart’ almost literally into German and then added the slightly strange German compound ‘Aepfeltorte’.

The encoding of the translation becomes more complex because of the mix of two languages; at the same time we need to keep an eye on the extension and inheritance of the 'xml:lang' attribute. The use of the German plural form ‘Äpfel’ may have been inspired by Johnson’s plural definition and the fact that a decent apple-tart requires more than one apple. Ten years later, in his monolingual German dictionary ‘Grammatisch-kritisches Wörterbuch der Hochdeutschen Mundart, Band 1. Leipzig 1793, S. 412, the entry shows no umlaut and the definition is derived from a recipe that puts the sliced apples on top.

In a final look at Ebers' German-English dictionary, the randomly chosen sample entry ‘fähig’ shows the problems in encoding bilingual dictionaries when translation from mother-tongue into a foreign language is involved.

```
<entry xml:id="jedictge.1788.000999.fähig" type="main_entry">
   <form xml:lang="de" type="lemma" norm="fähig">
      <lb/><orth>Fähig</orth><pc>,</pc>
      <gramGrp><pos norm="adjective">
      <abbr xml:lang="la">adj.</abbr> et <abbr xml:lang="la">adv.</abbr></pos>
      </gramGrp>
   </form>
   <sense>
      <def xml:lang="de">tüchtig, geschickt</def>
      <cit type="translation" xml:lang="en">
         <quote>capable, able, apr, fit, proper.</quote>
      </cit>
      <cit type="example" xml:lang="de">
         <quote>zu etwas fähig seyn,</quote></cit>
      <cit type="translation" xml:lang="en">
         <quote>to be capable or <lb/>fit for a Thing.</quote></cit>

      <lb/><cit type="example" xml:lang="de">
         <quote>sie ist des Erbrechts nicht fähig</quote></cit>
      <cit type="translation" xml:lang="en">
         <quote>she is <lb/>incapable for Succession.</quote></cit>
   </sense>
   <sense>
      <def xml:lang="de">fähig, lehrsam, gelehrig,</def>
      <cit type="translation" xml:lang="en">
      <quote>docile, teach- <lb/>able.</quote></cit>

      <lb/><cit type="example" xml:lang="de">
         <quote>fähig etwas zu erfinden</quote></cit>
      <cit type="translation" xml:lang="en">
         <quote>inventive.</quote></cit>
      <cit type="example" xml:lang="de">
         <quote>der Unterweisung fähig</quote></cit>
      <cit type="translation" xml:lang="en">
         <quote>susceptible of <lb/>Discipline, of Instruction</quote></cit>
      <lb/><cit type="example" xml:lang="de">
         <quote>er ist fähig alles zu unternehmen</quote></cit>
      <cit type="translation" xml:lang="en">
          quote>he <lb/>is a Man that will undertake any <lb/>Thing</quote></cit>
   </sense>
   <sense>
      <def xml:lang="de">fähig machen,</def>
      <cit type="translation" xml:lang="rn">
        <quote>to enable or fit, to in- <lb/>capacitate, to habilitate.</quote>
```

```
      </cit>
      <lb/><cit type="example" xml:lang="de">
        <quote>der Hunger macht einen zu allem fähig,</quote></cit>
      <lb/><cit type="translation" xml:lang="en">
         <quote>Hunger breaks through Stone-<lb/>Walls, or Hunger drives
         the Wolf <lb/>out of the Forest.</quote></cit>
      <lb/><cit type="example" xml:lang="de">
         <quote>einen wieder fähig machen,</quote></cit>
      <cit type="translation" xml:lang="en">
         <quote>to rehabi-<lb/>litate, re-enable, re-instate, re- <lb/>store,
          or re-establish one</quote></cit>
   </sense>
</entry>
```

At first glance the main lexicographical problem here is to specify the different senses of ‘fähig’, first in German (in <sense> + <def>), then in translating the German adjectives into the English equivalents (using <cit type=translation>), and finally in adding English translations (in <cit type=”translation”> of German example phrases (in <cit type=”example”>) containing the adjective. Compared with Johnson’s dictionary the senses are not numbered and the principle of their order is not quite clear.

Recalling the longish title of Ebers’ dictionary, ‘New and Complete Dictionary of the German and English Languages composed chiefly after the German Dictionaries of Mr. Adelung and of Mr. Schwan’, it is worthwhile taking a closer look at Ebers’ possible sources. Adelung’s entry ‘fähig’ in his ‘Grammatisch-kritisches Wörterbuch der Hochdeutschen Mundart, Vol. 2, Leipzig 1796, p. 11 is built around two numbered senses and looks completely different (google link: [to be supplied later]). But checking the dictionary of Christian Friedrich Schwan, Nouveau dictionnaire de la langue allemande et françoise : Composé sur les dictionnaires de M. Adelung et de l'Acad. Françoise ; enrichi des termes propres ... 1. Qui Contient Les Lettres A - G De L'Alphabet Allemand, Expliqué Par Le François, Mannheim: Schwan et Fontaine, 1782, p. 519, shows clearly how Ebers had compiled this entry of his dictionary:

```
<entry xml:id="csdictaf.1782.000999.fähig" type="main_entry">
   <form xml:lang="de" type="lemma" norm="fähig">
      <lb/><orth rend="ALLCAPS">FÆHIG</orth><pc>,</pc>
      <pc>(</pc><orth rend="fraktur">fähig</orth><pc>)</pc>
      <gramGrp>
      <pos norm="adjective"><abbr xml:lang="fr">adj. &amp; adv.</abbr></pos>
     </gramGrp>
   </form>
   <sense>
      <def xml:lang="de">tüchtig, geschikt;</def>
      <cit type="translation" xml:lang="fr">
         <quote>Capable, habile, propre.</quote></cit>
      <cit type="example" xml:lang="de"><quote>Zu etwas fähig seyn;</quote></cit>
      <lb/><cit type="translation" xml:lang="fr">
         <quote>être capable de qq. ch. être propre à une chose.</quote></cit>
     <lb/><cit type="example" xml:lang="de">
         <quote>Sie ist des Erbrechts nicht fähig;</quote></cit>
     <cit type="translation" xml:lang="fr">
         <quote>elle n'est pas <lb/>habile à succéder.</quote></cit>
   </sense>
   <sense>
      <abbr>It.</abbr><def xml:lang="de">Fähig, lehrsam, geleh-<lb/>rig</def>
      <cit type="translation" xml:lang="fr"><quote>docile.</quote></cit>
      <cit type="example" xml:lang="de">
         <quote>Fähig etwas zu erfinden;</quote></cit>
     <cit type="translation" xml:lang="fr"><quote>inven-<lb/>tif.</quote></cit>
     <cit type="example" xml:lang="de">
```

```
         <quote>Der Unterweisung fähig;</quote></cit>
      <cit type="translation" xml:lang="fr">
         <quote>susceptible de di-<lb/>scipline.</quote></cit>
      <cit type="example" xml:lang="de">
         <quote>Er ist fähig alles zu unternèhmen;</quote></cit>
      <lb/><cit type="translation" xml:lang="fr">
         <quote>il est homme à tout entreprendre.</quote></cit>
      <cit type="example" xml:lang="fr">
         <quote>Dinge, die<lb/>nicht jedermann zu verstehen fähig ist;</quote>
      </cit>
      <cit type="translation" xml:lang="fr">
         <quote>des <lb/>choses qui ne sont pas à la portée de tout
         le mon-<lb/>de</quote></cit>
      <cit type="example" xml:lang="de">
        <quote>Er ist nicht fähig, euch in geringsten zu<lb/>schaden</quote></cit>
      <cit type="translation" xml:lang="fr">
         <quote>il est incapable de vour nuire aucunement.</quote></cit>
      <lb/><cit type="example" xml:lang="de"><quote>Fähig machen</quote></cit>
      <cit type="translation" xml:lang="fr"><quote>habiliter.</quote></cit>
      <cit type="example" xml:lang="de">
         <quote>Der Hunger macht <lb/>einen zu allem fähig;</quote></cit>
      <cit type="translation" xml:lang="fr">
         <quote>la faim chasse le loup hors<lb/>du bois.</quote></cit>
      <cit type="example" xml:lang="de">
         <quote>Einen wieder fähig machen;</quote></cit>
      <cit type="translation" xml:lang="fr">
         <quote>réhabi-<lb/>liter qq. un.</quote></cit>
   </sense>
</entry>
```

With the exception of two phrases — ‘Dinge, die nicht jedermann zu verstehen fähig ist’ and ‘Er ist nicht fähig euch in geringsten zu schaden‘ — he has copied the German text of Schwan’s dictionary and replaced the French translation equivalents by English ones. The encoding problems remain the same and we think that the solution we propose is adequate.

Conclusion

Above we applied our encoding suggestions for the <form>-block to Johnson's entry '*To* APPLAUD' but did not comment on the unusual structure of the elements <sense> and <cit>: two numbered senses, followed by two quotations. A look at the last edition, which was considerably revised and prepared for publication by Johnson himself, the fourth folio edition of 1773, can make the author's original intentions clearer. Thanks to Anne McDermott's excellent CD-ROM edition, published in 1996, we have access to an SGML encoding of the texts of both the first and fourth folio edition and can not only compare the texts themselves but also the change over the years from TEI P3 SGML of 1994 to the current TEI Guidelines P5 using XML Schema.

First folio edition [TEI P5]:

```
<entry xml:id="sjdict1f.1755.000234.applaud" type="root" >
   <lb/><form type="lemma" norm="applaud">
      <gram type="infinitive-marker">To</gram>
      <orth rend="ALLCAPS">APPLA'UD</orth><pc>.</pc>
      <gramGrp><pos norm="verb"><abbr>v.a.</abbr></pos></gramGrp>
   </form>
   <etym>
   <pc>[</pc><mentioned xml:lang="la">applaudo</mentioned><pc>,</pc>
      <lang><abbr>Lat.</abbr></lang><pc>]</pc>
   </etym>
   <pc>[</pc><mentioned xml:lang="la">applaudo</mentioned><pc>,</pc>
```

```
      <lang><abbr>Lat.</abbr></lang><pc>]</pc>
   </etym>
   <lb/><sense>
      <num>1.</num>
      <def>To praise by clapping the hand.</def>
   </sense>
   <lb/><sense>
      <num>2.</num>
      <def>To praise in general.</def>
   </sense>
   <cit type="example">
      <lb/><quote>I would applaud thee to the very echo,
      <lb/>That should applaud again.</quote>
      <bibl><author><abbr>Shakesp.</abbr></author><title>Macbeth</title>.</bibl>
   </cit>
   <cit type="example">
      <lb/><quote>Nations unborn your mighty names shall sound,
      <lb/>And worlds applaud that must not yet be found!</quote>
      <bibl><author>Pope</author>.</bibl>
   </cit>
</entry>
```

Ann McDermott Fourth folio edition [TEI P3 SGML]:

```
<ENTRYFREE ID="J4APPLAUD-1" N="1999" TYPE="4">IV
   <FORM>
      <HI REND="ital">To</HI> <HI REND="acp">APPLA'UD.</HI>
   </FORM>
<PB SIG="Bb2r" MACFILE=":4:100:148.CAL" PCFILE="4\100\148.CAL">
   <POS><HI REND="ital">v.a.</HI></POS>
   <ETYM>[<HI REND="ital">applaudo,</HI> Lat.]</ETYM>
   <SENSE N="1">
      <DEF>
      <NUM>1.</NUM> To praise by clapping the hand.
      </DEF>
   <EG TYPE="verse">
      <QUOTE>
      <L>I would <HI REND="ital">applaud</HI> thee to the very echo,</L>
      <L>That should <HI REND="ital">applaud</HI> again.</L>
      </QUOTE>
      <AUTHOR><HI REND="ital">Shakesp.</HI></AUTHOR>
      <TITLE><HI REND="ital">Macbeth.</HI></TITLE>
   </EG>
   </SENSE>
   <SENSE N="2">
      <DEF>
      <NUM>2.</NUM> To praise in general.
      </DEF>
   <EG TYPE="verse">
      <QUOTE>
      <L>Nations unborn your mighty names shall sound,</L>
      <L>And worlds <HI REND="ital">applaud</HI> that must
         not yet be sound!</L>
      </QUOTE>
      <AUTHOR><HI REND="ital">Pope.</HI>
      </AUTHOR>
   </EG>
   </SENSE>
</ENTRYFREE>
```

As result we can conclude:

1. The transcription of the entry APPLAUD in the SGML version of the fourth folio edition shows clearly that Johnson had intended to illustrate each definition with an illustrative quotation, as elsewhere in the dictionary, and that the unusual structure of the first folio text — two numbered senses, followed by two quotations — is simply a typesetting error.

2. Both encodings have many structural features in common: with the exception of <cit> and <pc> all elements used in our encoding were available in TEI P3, whereas the mechanisms usable at the attribute level are not comparable. But the main difference is the style of the encoding: although the SGML version is very close to the typography of the text, our suggestions aim more at interpreting typographical detail in order to capture lexicographic and linguistic data and to constrain encoding options in favour of robust interoperability and reusabilty of resources.

References

Adams, Valerie. (1976). *An Introduction to Modern English Word-Formation*. London: Longman.

Henne, Helmut ed. 2001. *Deutsche Wörterbücher des 17. und 18. Jahrhunderts. Einführung und Bibliographie.* Hildesheim/Zürich/New York: Olms.

ISO 12620:2009 Terminology and other language and content resources -- Specification of data categories and management of a Data Category Registry for language resources

ISO 24613:2008 Language resource management - Lexical Markup Framework (LMF)

Ide, N., Kilgarriff, A., Romary, L. (2000). A Formal Model of Dictionary Structure and Content. *Proceedings of Euralex 2000,* Stuttgart, 113-126 - http://hal.archives-ouvertes.fr/hal-00164625

Johnson, Samuel. 1755. *A Dictionary of the English Language: in which the Words are deduced from their Originals, and Illustrated in their Different Significations by Examples from the best Writers. To which are prefixed, a History of the Language, and an English Grammar*.In Two Volumes. London: W. Strahan.

Lewis, Derek. 2012. *Die Wörterbücher von Johannes Ebers. Studien zur frühen englisch-deutschen Lexikographie.* PhD diss. University of Würzburg (with a catalog of Ebers' publications, in print).

Luna, Paul. (2005). "The typographic design of Johnson's Dictionary." In Anniversary Essays on Johnson's Dictionary, edited by Jack Lynch and Anne McDermott, 175–197. Cambridge: Cambridge University Press.

McDermott, Ann, ed. (1996). *Samuel Johnson, A Dictionary of the English Language, on CD-ROM.The First an Fourth Editions.* Cambridge: Cambridge University Press.

Miller George A. and Christiane Fellbaum (2007) “WordNet then and now”, Language Resources and Evaluation, Volume 41, Number 2, 209-214, DOI: 10.1007/s10579-007-9044-6

Osselton, Noel E. (2005). "Hyphenated compounds in Johnson's Dictionary." In *Anniversary Essays on Johnson's Dictionary*, edited by Jack Lynch and Anne McDermott, 160–174. Cambridge: Cambridge University Press.

Reddick, Allen. 2006. *The Making of Johnson's Dictionary 1746 – 1773*. Revised Edition. Cambridge: Cambridge University Press.

Romary, Laurent. 2009. “ODD as a generic specification platform”. Text encoding in the era of mass digitization - Conference and Members' Meeting of the TEI Consortium – http://hal.inria.fr/inria-00433433

TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.0.2. Last updated on 2nd February 2012. Accessed May 7, 2012. TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Keywords

Dictionary encoding;

semasiological dictionary;

entry; form; sense;

Samuel Johnson, Dictionary of the English Language

Abstract

Our paper outlines a proposal for the consistent modelling of heterogeneous lexical structures in semasiological dictionaries, based on the element structures described in detail in chapter 9 (Dictionaries) of the TEI Guidelines. The core of our proposal describes a system of relatively autonomous lexical 'crystals' that can — within the constraints of the element definition — be combined to form complex structures for the description of morphological form, grammatical information, etymology and word-formation and meaning. A core element for the semasiological structure of an entry is <sense> used generically in combination with the <cit>-element.

The encoding structures we suggest guarantee sustainability and support re-usability and interoperability. This paper presents case studies of dictionary entries which illustrate our concepts and test their usability. The dictionaries themselves were all published in the second half of the 18th century and are chosen because of the interesting relationships which they exhibit with one another.

Our starting point is Samuel Johnson's Dictionary of the English Language of 1755, the basis of modern English lexicography. We use an adjective (able) and noun samples to demonstrate a baseline encoding with lexical 'crystals'. We then apply the model to a translation of this dictionary into a bilingual English to German dictionary by Johann Christoph Adelung, the most important German lexicographer of the time, and to the German-English dictionary of Johannes Ebers dating from 1796, which re-used Adelung's (and Johnson's) material.

In the above context we comment on encoding issues involving <entry>, <form>, the element <etym>, and on refinements to the internal content of <sense>.