Modelling frequency data – Methodological considerations on the relationship between dictionaries and corpora

Gerhard Budin, Karlheinz Mörth, Laurent Romary

Modelling frequency data

Methodological considerations on the relationship between dictionaries and corpora

Gerhard Budin, Karlheinz Moerth, Laurent Romary

The research questions addressed in our paper stem from a bundle of linguistically focused projects which –among other activities– also create glossaries and dictionaries which are intended to be usable both for human readers and particular NLP applications. The paper will comprise two parts: in the first section, the authors will give a concise overview of the projects and their goals. The second part will concentrate on encoding issues involved in the related dictionary production. Particular focus will be put on the modelling of an encoding scheme for statistical information on lexicographic data gleaned from digital corpora.

The mentioned projects are tightly interlinked, are all joint endeavours of the Austrian Academy of Sciences and the University of Vienna and conduct research in the field of variational Arabic linguistics. The first project, the *Vienna Corpus of Arabic Varieties* (VICAV), was already started two years ago on the basis of a low budget scheme and was intended as an attempt at setting up a comprehensive research environment for scholars pursuing comparative interests in the study of Arabic dialects. The evolving VICAV platform aims at pooling linguistic research data, various language resources such as language profiles, dictionaries, glossaries, corpora, bibliographies etc. The second project goes by the name of *Linguistic Dynamics in the Greater Tunis Area: A Corpus-based Approach*. This three-year project which is financed by the Austrian Science Fund aims at the creation of a corpus of spoken youth language and the compilation of a diachronic dictionary of Tunisian Arabic. The third project which has grown out of a master's thesis deals with the lexicographic analysis of the Wikipedia in Egyptian vernacular Arabic. In all these projects, digital data production relies on the Guidelines of the TEI (P5), both for the corpora and the dictionaries. The dictionaries compiled in the framework of these projects are to serve research as well as didactic purposes.

Using the TEI dictionary module to encode digitized print dictionaries has become a fairly common standard procedure in digital humanities. Our paper will not resume the TEI vs. LMF vs. LexML vs. Lift vs. ... discussion (cf. Budin et al. 2012) and assumes that the TEI dictionary module is sufficiently well-developed to cope with all requirements needed for the purposes of our projects. The basic schema used has been tested in several projects for various languages so far and will furnish the foundation for the intended customisations.

Lexicostatistical data and methods are used in many fields of modern linguistics, lexicography is only one of them. Modern-time dictionary production relies on corpora, and statistics–beyond any doubt– play an important role in lexicographers' decisions when selecting lemmas to be included in dictionaries, when selecting senses to be incorporated into dictionary entries and so forth. However, lexicostatistical data is not only of interest for the lexicographer, it might also be useful to the users of lexicographic resources, in particular digital lexicographic resources. The question as to how to make such information available takes us to the issue of how to encode such information.

Reflecting on the dictionary–corpus–interface and on the issue of how to bind corpus-based statistical data into the lexicographic workflow, two prototypical approaches are conceivable: either statistical information can statically be embedded in the dictionary entries or the dictionary provides links to services capable of providing the required data. One group of people working on methodologies to implement functionalities of the second type is the *Federated Content Search* working group, an initiative of the CLARIN infrastructure which strives to move towards enhanced search-capabilities in locally distributed data stores (Stehouwer et al. 2012). FCS is aiming at heterogeneous data, dictionaries are only one type of language resources to be taken into consideration. In view of more and more dynamic digital environments, the second approach appears to be more appealing. Practically, the digital workbench will remain in need of methods to store frequencies obtained from corpus queries, as human intervention will not be superfluous any time soon. Resolving polysemy, grouping of instances into senses remain tasks that cannot be achieved automatically.

Which parts of a dictionary entry can be considered as relevant? What is needed is a system to register quantifications of particular items represented in dictionary entries. The first thing that comes to mind are of course headwords, lemmata. However, there are other constituents of dictionary entries that might be furnished with frequency data: inflected wordforms, collocations, multi word units and particular senses are relevant items in this respect.

The encoding system should not only provide elements to encode these, but also allow to indicate the source from which the data were gleaned and how the statistical information was created. Ideally, persistent identifiers should be used to identify not only the corpora but also the services involved to create the statistical data.

We basically see three options to go about the encoding problem as such: (a) to make use of some TEI elements with very stretchable semantics such as *note*, *ab* or *seg* and to provide them with *type* attributes, (b) to make use of TEI feature structures or (c) to develop a new customisation. We will discuss why we have discarded the first option, will present a provisional solution on the basis of feature structures and discuss pros-and-cons of this approach. As is well known, feature structures are a very versatile, sufficiently well-explored tool for formalising all kinds of linguistic phenomena. One of the advantages of the *fs* element is that it can be placed inside most elements used to encode dictionaries.

```
<entry xml:id="mashcal_001" >
   <form type="lemma">
      <orth xml:lang="ar-arz-x-cairo-vicavTrans">mašʕal</orth>
      <orth xml:lang="ar-arz-x-cairo-arabic">مشعل</orth>

      <fs type="corpFreq">
         <f name="corpus" fVal="#wikiMasri"/>
         <f name="frequency"><numeric value="6"/></f>
      </fs>
   </form>

   <gramGrp>
      <gram type="pos">noun</gram>
      <gram type="root" xml:lang="ar-arz-x-cairo-vicavTrans">šʕl</gram>
   </gramGrp>

   <form type="inflected" ana="#n_pl">
      <orth xml:lang="ar-arz-x-cairo-vicavTrans">mašāʕil</orth>
      <orth xml:lang="ar-arz-x-cairo-arabic">مشا عل</orth>

      <fs type="corpFreq">
```

```
        <f name="corpus" fVal="#wikiMasri"/>
        <f name="frequency"><numeric value="2"/></f>
      </fs>
    </form>
```

The paper will be concluded by first considerations considering a more encompassing ODD based solution. We hope the work could lead to the introduction of a comprehensive set of descriptive objects (attributes and element) to describe frequencies in context, encompassing: reference corpus, size of reference corpus, extracted corpus, size of extracted corpus and various associated scores (standard deviation, t-score, etc.).

Selected references

Banski, Piotr, and Beata Wójtowicz. 2009. FreeDict: an Open Source repository of TEI-encoded bilingual dictionaries. In *TEI-MM*, Ann Arbor. (http://www.tei-c.org/Vault/MembersMeetings/2009/files/Banski+Wojtowicz-TEIMM-presentation.pdf)

Bel, Nuria, Nicoletta Calzolari, and Monica Monachini (eds). 1995. Common Specifications and notation for lexicon encoding and preliminary proposal for the tagsets. MULTEXT Deliverable D1.6.1B. Pisa.

Budin, Gerhard, Stefan Majewski, and Karlheinz Mörth. 2012. Creating Lexical Resources in TEI P5. In *jTEI* 3.

Hass, Ulrike (ed). 2005. *Grundfragen der elektronischen Lexikographie: Elexiko, das Online-Informationssystem zum deutschen Wortschatz*. Berlin; New York: W. de Gruyter.

Romary, Laurent, Susanne Salmon-Alt, and Gil Francopoulol. 2004. Standards going concrete : from LMF to Morphalou. In *Workshop on enhancing and using electronic dictionaries.* Coling 2004, Geneva.

Romary, Laurent, and Werner Wegstein. 2012. Consistent Modeling of Heterogeneous Lexical Structures. In *jTEI* 3.

Sperberg-McQueen, C.M., Lou Burnard, and Syd Bauman (eds). 2010. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Oxford, Providence, Charlotteville, Nancy. (http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf)

Stehouwer, Herman, Matej Durco, Eric Auer, and Daan Broeder. 2012. Federated Search: Towards a Common Search Infrastructure. In: *Calzolari, Nicoletta; Choukri, Khalid; Declerck, Thierry; Mariani, Joseph (eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012).* Istanbul.

Werner Wegstein, Werner, Mirjam Blümm, Dietmar Seipel, and Christian Schneiker. 2009. Digitalisierung von Primärquellen für die TextGrid-Umgebung: Modellfall Campe-Wörterbuch. (http://www.textgrid.de/fileadmin/TextGrid/reports/TextGrid_R4_1.pdf)