Finding the Right Structure for Lexicographical Data: Experiences from a Terminology Project

Michal Boleslav Měchura
Fiontar, Dublin City University
Glasnevin, Dublin 9, Ireland

Abstract
This paper deals with issues related to the design of structures for holding lexicographical and terminographical data, drawing from experiences gained during a terminology project. The issues include the structural differences between a typical dictionary entry and a typical terminographical entry, senses and concepts, semasiology and onomasiology, dictionary reversal, data conversion, polysemy and homonymy, and the grammatical labelling of multi-word items.

1 Introduction

It would appear that every dictionary and terminology database available today comes in one of two structures: either in the "lemma and senses" layout of lexicography, or the "concept and terms" layout of terminography, but no other structures seem to be common. While both these structural paradigms have been tried and tested extensively and constitute the best practices of the industry, this paper will introduce a project on which we have found that neither of these structures suits our needs completely, and consequently we have developed our own data structure in which we have combined aspects of both lexicography and terminography.

The FTU[1] project was started in the winter of 2004 by collaborating institutions in Ireland[2] and Wales.[3] The Irish half of the project has as its goal the production of an on-line English-Irish and Irish-English dictionary of specialized terminology, in many fields of human activity, which the public could access over the Internet and which the relevant authorities could use in the future as a terminology management tool. The project is substantial not only by the size of the data (there are over 200,000 dictionary entries to process) but also by the scope of uses envisaged for the end product. Our brief is to produce a software solution which is many things to many people: a terminology management system for professional terminologists, but also a publicly-accessible on-line dictionary for everyday users. This has forced us to adopt an approach which is a compromise between traditional LGP[4]-styled lexi-

[1] *Fiontar Téarmaí Unedig*, more information about the project is available online at www.focal.ie.
[2] *Fiontar*, Dublin City University; *Foras na Gaeilge*
[3] Department of Welsh, University of Wales, Lampeter
[4] Language for General Purposes

cography and traditional LSP[5]-styled terminography, and this has reflected itself in the data structure we have designed for the project.

2 The FTU data model

A (simplified) E/R diagram of the FTU terminology database is presented in Figure 1. The remaining sections of this paper will each "zoom in" on a particular aspect of the database structure and explain the factor involved in designing the database in this particular way.

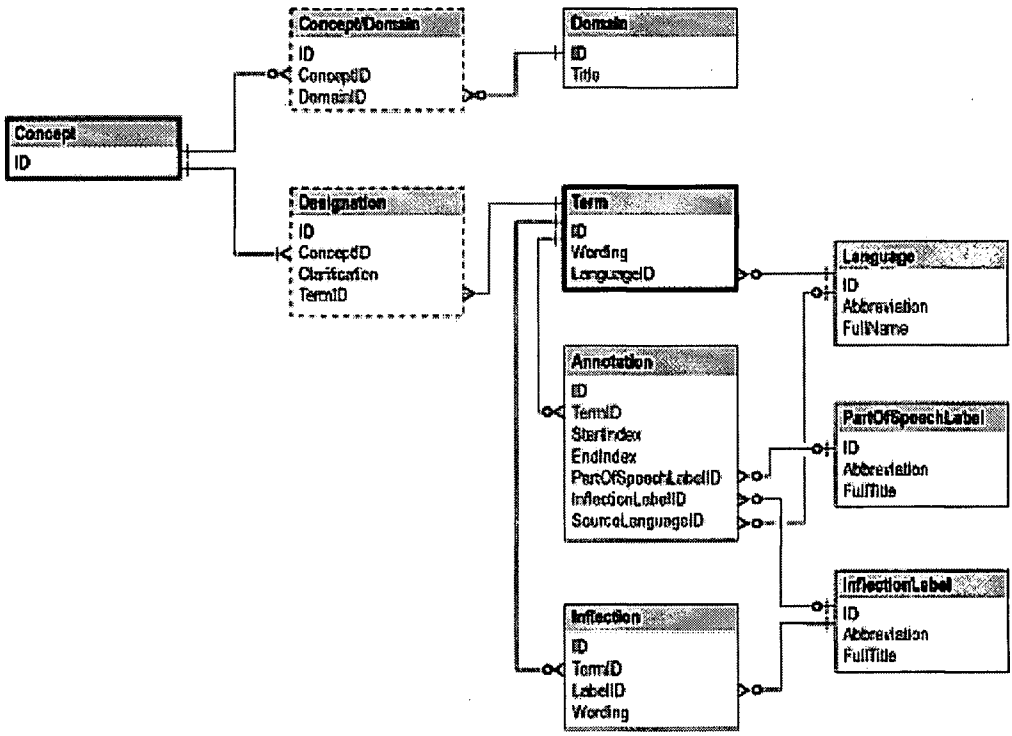


Figure 1. Simplified E/R diagram of the FTU terminology database

3 Concept-oriented approach

In traditional lexicography, the basic unit of data a lexicographer works with is a dictionary entry, organized around a lemma, and further subdivided into senses. In terminography, on the other hand, the basic unit of data is a concept. It is quite difficult to define what a con-

[5] Language for Specialized Purposes

cept is, it is an abstraction which has arisen from the need to record complex relationships between translation equivalents in more than two languages. Terminology theory dictates that a terminologist should begin by identifying the concept, and then identify all the possible terms that can be used to express the concept, in all the relevant languages.[6] In other words, the business of terminography is one of onomasiology, where the point of departure is a meaning. Lexicography on the other hand is associated with semasiology, where the point of departure is a word, not a meaning. The "concept" of terminography is roughly equivalent to the "sense" of lexicography, with the difference that a typical terminographic concept is usually realized in an LGP dictionary by multiple senses and is spread out across multiple entries. Figure 2 shows how several senses (or subsenses, in this case) of several words, which appear as separate objects in an LGP dictionary, would be treated as a single object in a concept-based terminology database.

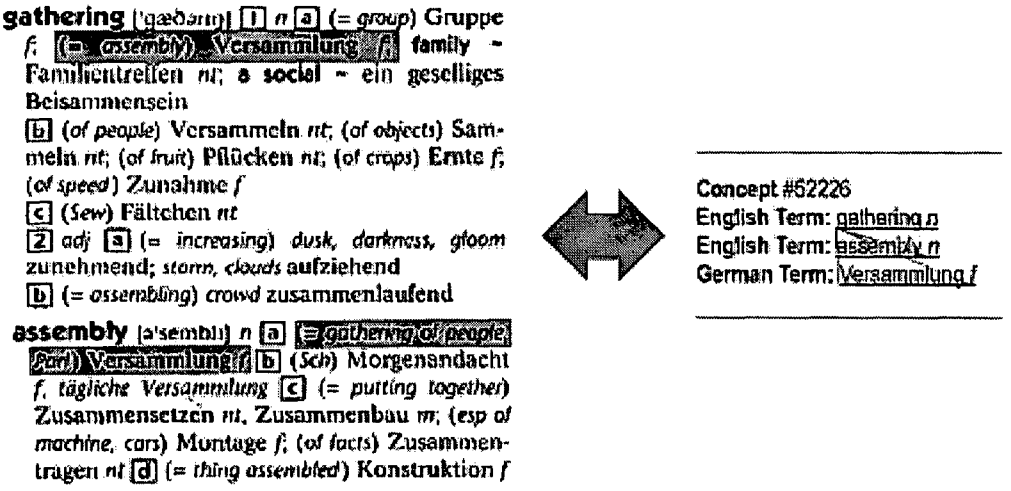


Figure 2. Senses and concepts
(reproduction from Collins German-English/English-German Dictionary)

The difference between lexicographical data structures and terminographical data structures is thus one of perspective: we work with the same kinds of data but we cluster them differently.

On the FTU project we are faced with the task of computerizing a large number of manually compiled glossaries which have accumulated over many decades and have been built largely from the semasiological perspective, each entry starting with the English term and then listing Irish translation equivalents, sometimes subdivided into senses and sometimes

[6] For more information on the role of concepts in terminology theory, consult for example the first chapter of Weissenhofer (1995)

not. Some glossaries also include translation equivalents in other languages, such as Latin plant names. We needed to convert this store of eclectically structured data into a concept-based data structure to facilitate long term maintenance of the data and also to facilitate the task of dictionary reversal. While the current manually-compiled lists attended reasonably well to the needs of users looking up translation equivalents of English terms for production purposes, searches in the opposite direction usually produced results which were difficult for ordinary users to interpret. In the semasiological approach, if A is a translation of B, it does not follow automatically that B is a translation of A. We expected that if we reengineered the terminology store into a concept-oriented, onomasiological system, dictionary reversal would become an inherent feature of the system, simply a matter of displaying the same underlying data in a different way.

However, we have been forced to deviate a little from what would be considered a pure concept-oriented system. Our data sources are semasiological and converting them to an onomasiological, concept-based structure would normally require a human to analyze each entry. Since we are facing over 200,000 entries, this would be unachievable in a 30-month project. Instead, we have produced several simple, heuristic rules to follow. To start with, we presume that each dictionary entry represents a single concept. When we know beforehand that this is not the case in a particular glossary, we pre-edit it before the conversion. Then, if two entries are encountered in which exactly the same terms and words appear, we conclude that they represent the same concept, and merge them into a single concept in our database (see Figure 3). Secondly, entries in which completely different terms appear are considered to be different concepts (Figure 4). And finally, entries which have some terms in common and some different are also considered to be separate concepts, but are flagged for editorial attention (Figure 5). We have found that the division of terms into concepts achieved by this process usually makes sense, possibly because the equivalences between terms in an LSP context tend to be more straightforward than the equivalences between words in an LGP context. Editorial follow-up and clean-up is needed but the workload is significantly lower than would be required to human-analyze each individual entry.

One important fall-out, however, is that the division of terms into concepts tends to be translation-driven. If there are two concepts, one belonging to the domain of office work and one to computing and both are expressed by the same words in both English and Irish, then the system will make a single concept of them (see Figure 6).

This would be considered bad practice in terminography but it is quite common in bilingual lexicography. For example the English word *life* has 14 senses in the monolingual Oxford Advanced Learner's Dictionary but almost all of them are conflated into a single sense in the bilingual *Großes Oxford Wörterbuch für Schule und Beruf* from the same publisher because they can all be expressed by the same German word, *Leben* (Deuter 2004: p. 247). Our terminology database is organized in this way too, and we expect this to serve well the needs of our target audience, the non-specialist bilingual users, even though it departs from traditional terminographic principles.

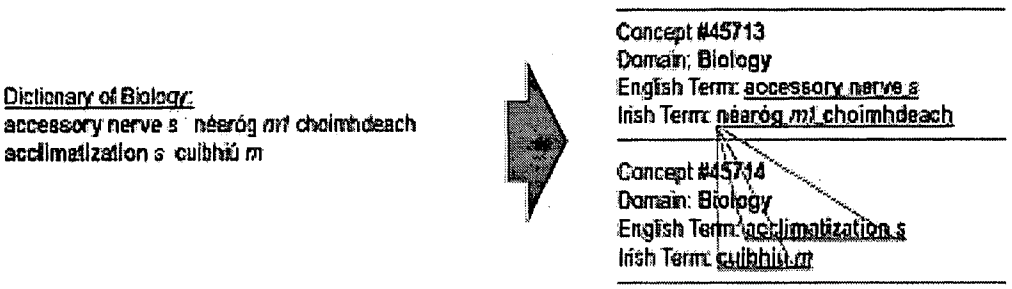


Figure 3. Dictionary entries that contain completely different terms are imported as separate concepts

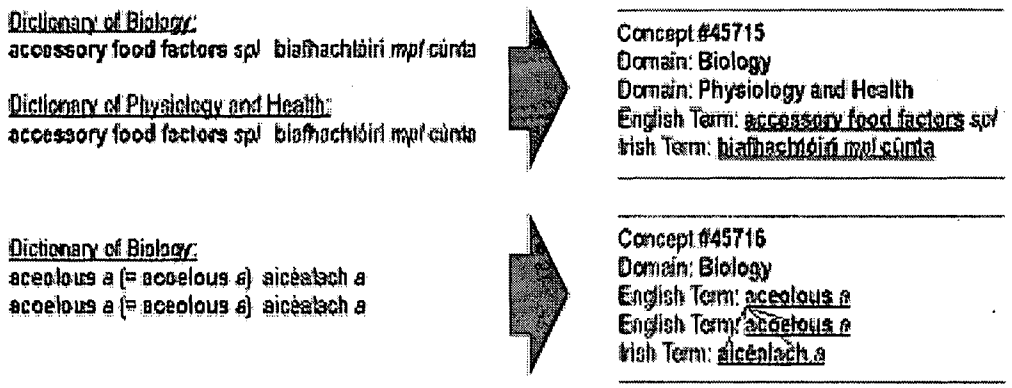


Figure 4. Dictionary entries that contain the same terms are merged into a single concept

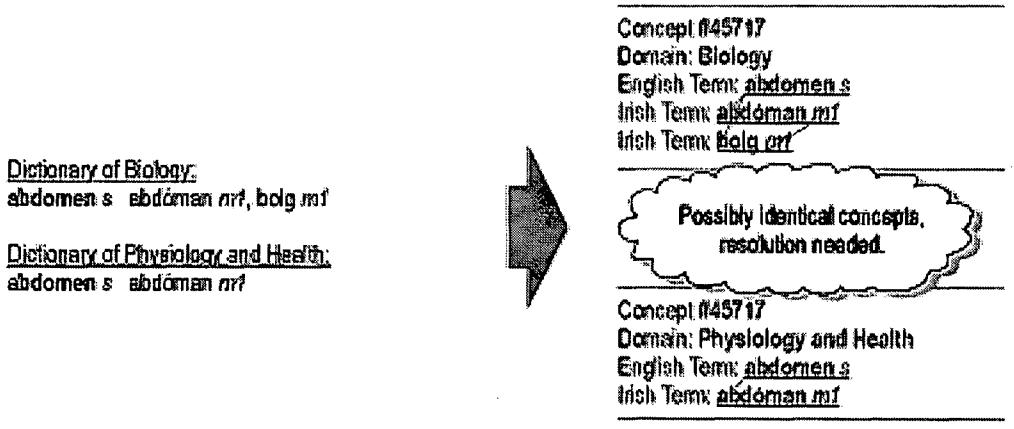


Figure 5. Overlapping dictionary entries are imported as separate concepts but flagged for editorial attention

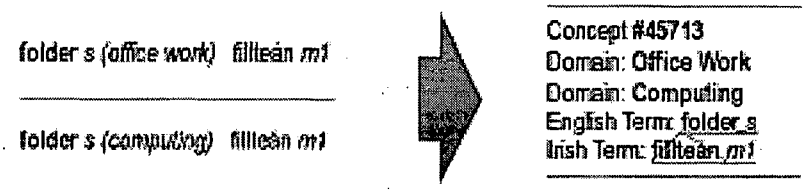


Figure 6. Two logical concepts merged into one

3 Relational data structure

Another aspect in which our database differs from a traditional terminology database and places us half-way between lexicography and terminography is that our data model is completely relational, allowing us to effortlessly resolve issues of polysemy and to display data in a user-friendly way. In a conventional terminology database, if a term designates two concepts it would be recorded twice, once in each concept. If then the terminologist updates the spelling of the term or adds grammatical information to it, they must take care to make the same changes in both concepts because there are typically no facilities to keep the two records synchronized. In the FTU database, each term is recorded only once, and instead of being *included* in the concepts it designates, it is *linked* to them. Each term can be linked to any number of concepts, and each concept can be linked to any number of terms, thus modelling polysemy (which, for the purposes of the project, we have defined as one term designating multiple concepts) and homonymy (which we define as one concept being designated by multiple terms).

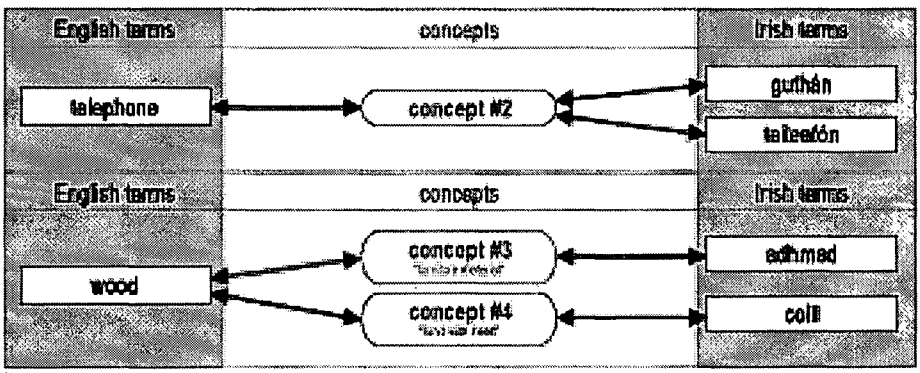


Figure 7. Terms and concepts linked to each other in different ways.
The Irish words guthán and teileafón are homonyms,
and the English word wood is polysemous ('material' and 'vegetation').

The advantages of this approach are manifold. In addition to recording polysemy effortlessly it also relieves the editorial staff of having to re-enter duplicate information. The terms on the Irish side of the dictionary usually have a lot of grammatical information associated

with them, and it would be a bad use of human resources to have to enter this information several times for duplicate records of the same term – never mind the danger of inconsistency. At the user interface level, when a user searches for a term, the system navigates the relational structure to quickly look up all the concepts the term is associated with, and arranges them in a bulleted list underneath the searched term, thus effectively compiling a conventional dictionary entry "on the fly" in which concepts are represented as senses.

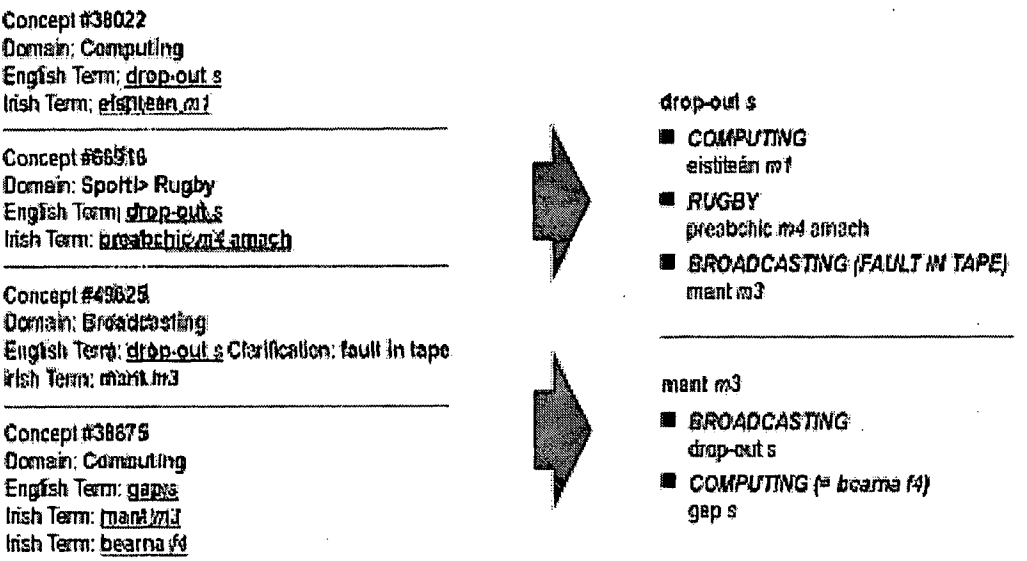


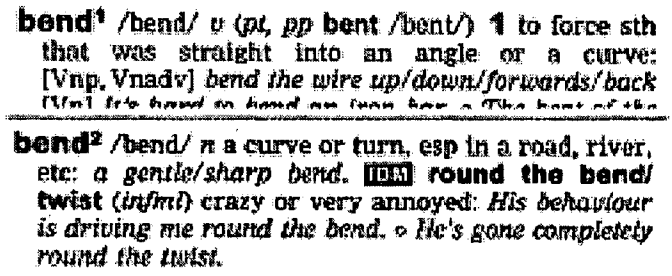
Figure 8. Concepts, and their on-screen representations as dictionary entries.

This is the layout that the dictionary will offer to non-specialist users, while linguists and terminologists will edit the data in a concept-oriented layout.

The relational model which we have adopted has repercussions for the homonymy/polysemy debate. Essentially, when a word with two meanings is encountered, the lexicographer has the option to either treat it as a single word designating two separate concepts (polysemy) or as two words, each designating its own concept, which just happen to have the same spelling (homography) or pronunciation (homophony).[7] In our system, since we cannot research each word and term individually, we have adopted a simple principle: if two terms have exactly the same spelling and if they have the same grammatical information attached to them (for example if they belong to the same word class), then they are a single polysemous term. In all other cases we are dealing with different terms. This principle is very easy for a computer to follow while converting data from manual glossaries to the new format,

[7] For a debate of the homonymy/polysemy divide and an example of how different dictionaries resolve it differently, see section 3.5.2 of Saeed (1997: 64).

and we have found that in most cases it resolves cases of duplicity between all the different glossaries successfully. It does result in some unusual behaviour, though. For example, verbs and nouns which have the same spelling in English are treated by the system as separate words, while some lexicographers prefer to treat them as a single word, as for example in Figure 9.


bend[1] /bend/ *v* (*pt, pp* **bent** /bent/) **1** to force sth that was straight into an angle or a curve: [Vnp, Vnadv] *bend the wire up/down/forwards/back*

bend[2] /bend/ *n* a curve or turn, esp in a road, river, etc: *a gentle/sharp bend.* **round the bend/twist** (*infml*) crazy or very annoyed: *His behaviour is driving me round the bend.* ◦ *He's gone completely round the twist.*

Figure 9. A noun and a verb treated as homonyms
(reproduction from the Oxford Advanced Learner's Dictionary)

Also, in some isolated cases, the identity of spelling is just a coincidence, for example the system treats as a single polysemous word "adder" when it designates a viper and "adder" when it designates a device for adding numbers, although most English speakers would probably intuitively feel that these are two separate words which just happen to look the same but have different morphological histories ("adder" as a device for adding numbers is very obviously a composite of "add" + "-er" while "adder" the viper is not).

4 Multi-word items

Most terms in the FTU database are multi-word items rather than single lemmas, as is common in LSP terminology. There is a tradition in Irish lexicography to annotate headwords and terms with extensive grammatical information (word class, gender, declension) and the FTU database had to accommodate that. In many conventional dictionary-writing systems[8] and terminology management systems,[9] a grammatical label can only be attached to the whole term but not to an individual word inside it. We have designed our system to overcome this obstacle. When attaching a label to a term, the user can choose a substring of the term and declare that the label only pertains to that substring. At presentation time, the system inserts the label into the term but inside the database, the term is stored unbroken.

[8] Examples include TshwaneLex and Lexique Pro.
[9] Examples include Trados MultiTerm and Star WebTerm.

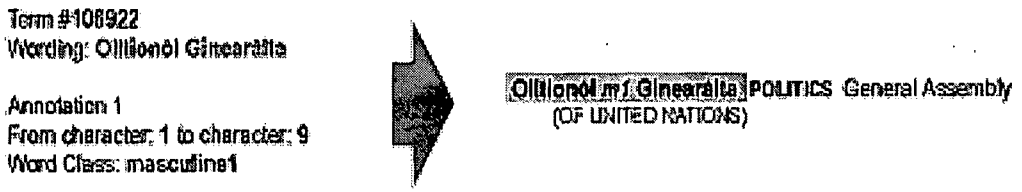


Figure 10. An in-line grammatical annotation and its on-screen representation

This solution not only facilitates searching (a search for *Olltionól Ginearálta* will match *Olltionól (m1) Ginearálta* without any additional programming) but also allows us to store many kinds of information about individual words, not only the part of speech but also the inflected form in which the word occurs in the current term (genitive, plural, etc.), whether it is a borrowed word (borrowed words are displayed in italic type at presentation time), and even which language the word has been borrowed from, if known. For example in Figure 11, both words are annotated. The first is a noun of the word class *masculine4* and is a borrowed word from Latin. The second is also a noun, it is of the word class *feminine3*, and it appears here in the genitive case.

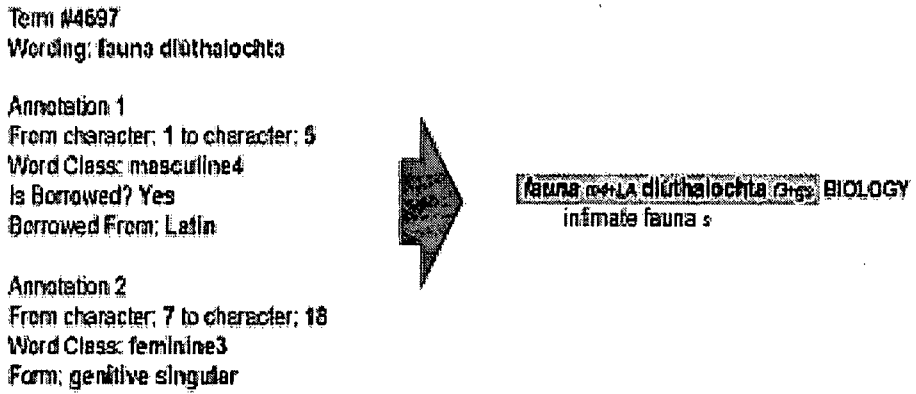


Figure 11. A term with extensive in-line annotation

4 Conclusion

The data structures employed by lexicography and terminography have traditionally been very different. However, modern-day developments have introduced the need to make terminology databases user-friendly to a wide audience, as specialized terminology becomes a part of everyday life for the general public. This has made it necessary for terminology projects such as FTU to revise this age-old separation between terminography and lexicography and to devise a new data structure to satisfy these requirements.

On the one hand, FTU is an LSP project and has employed the concept-oriented approach because its vocabulary comes from specialized areas of human activity and the correspondence between translation equivalents is usually more straightforward than in LGP. On the

other hand, FTU is an LGP project, and the concept-oriented data model has been extended to accommodate the efficient handling of polysemy and to facilitate the on-screen display of data as conventional dictionary entries. Providing useful information to the non-specialist user is a priority, reflected for example in the in-line grammatical annotations attached to terms.

A valuable lesson learned from the project so far has been that it pays to reflect on the database structure in which we store our data. We are grateful to our project partners in Wales and to numerous other international experts whom we have consulted for helping us craft a data structure which, unconventional as it may be, serves our needs more efficiently than either of the two conventional data models in their pure form would.

References

A. Terminology Management Systems and Dictionary Writing Systems

LexiquePro, information on-line at http://www.lexiquepro.com/.

Star WebTerm, http://www.star-solutions.net/html/eng/produkte/webterm55.html.

Trados MultiTerm, http://www.trados.com/multiterm.

TshwaneLex, http://www.tshwanedje.com/tshwanelex/.

B. Dictionaries

Terrell, P. (1999), *Collins German-English/English-German Dictionary Unabridged* (Fourth Edition), Glasgow & New York, HarperCollins.

Hornby, A. S., Crowther, J [ed] (1995), *Oxford Advanced Learner's Dictionary of Current English* (Fifth Edition). Oxford, Oxford University Press.

C. Other Literature

Deuter, M. (2004), 'A Tale of Two Halves: Writing a Bilingual Dictionary for Students of English', in Williams, G., Vessier, S. [ed] *EURALEX 2004: Proceedings of the Eleventh EURALEX International Congress*, Lorient, Université de Bretagne-Sud

Saeed, J. I. (1997), *Semantics*, Oxford, Blackwell.

Weissenhofer, P. (1995), *Conceptology in terminology theory, semantics and word-formation*, Vienna, TermNet.