Dare to be different: how user needs determine termbase design

Michal Měchura[a,b], Brian Ó Raghallaigh[a], Úna Bhreathnach[a] and Gearóid Ó Cleircín[a]

[a]*Fiontar & Scoil na Gaeilge, Dublin City University, Dublin, Ireland*
[b]*Natural Language Processing Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*

Abstract
This paper describes and discusses how the design of the National Terminology Database for Irish (*téarma.ie*) has been influenced by two factors: the assumed information needs of the intended users, and the data governance needs of the publisher. In particular, we will highlight how these factors have sometimes caused our termbase design to diverge from established practices in the terminology industry and from standards such as TBX.

Keywords
online termbases, terminology in minority languages, terminology in bilingual countries

1. Introduction: who are we building a termbase for?

The National Terminology Database for Irish (NTDI) [1, 2] serves the speakers of a minority language (Irish) in a country (Ireland) where it co-exists with a majority language (English). The users are typically translators, bilingual journalists, educators and officials in public administration who are looking for translations of specialised terms from English into Irish, in fields such as public administration, sport, public health and information technology as well as various school subjects. The termbase has a public website (https://www.tearma.ie, formerly focal.ie) which handles over half a million search requests every month. The termbase is edited by a small number of terminologists through the open-source terminology management system Terminologue (https://www.terminologue.org) [3]. NTDI contains approximately 200,000 entries, each with terms in two languages.

1.1. The information needs of the end-user

When NTDI's public website was launched in 2006 it was intended as an LSP (Language for Specialised Purposes) resource as defined e.g. by [4]. However, online lexical resources for the Irish language were scarce at that time and many users have come to use the termbase as if it were an LGP (Language for General Purposes) dictionary, searching for general-purpose

vocabulary and looking for the kind of information one would normally expect to find a general-purpose dictionary. NTDI has evolved to satisfy this unique mixture of the users' *information needs* (a concept originally defined by [5]), both in its content (it contains some general-language vocabulary) and in its structure.

1.2. The data-governance needs of the terminologist

While the users' information needs are what drives the design of a termbase, the needs of the terminologists – the editors and maintainers behind the scenes – need to be taken into account as well. These are concerned mainly with *data governance*: quality control, keeping the termbase well organised and well maintained in the long run, avoiding duplicates and so on. The design of the NTDI reflects some of these needs, as we will show in the rest of this paper.

2. Some features of NTDI

We will now review some of NTDI's structural features that have been influenced by the requirements introduced above, covering both the users' information needs and the terminologist's data-governance needs.
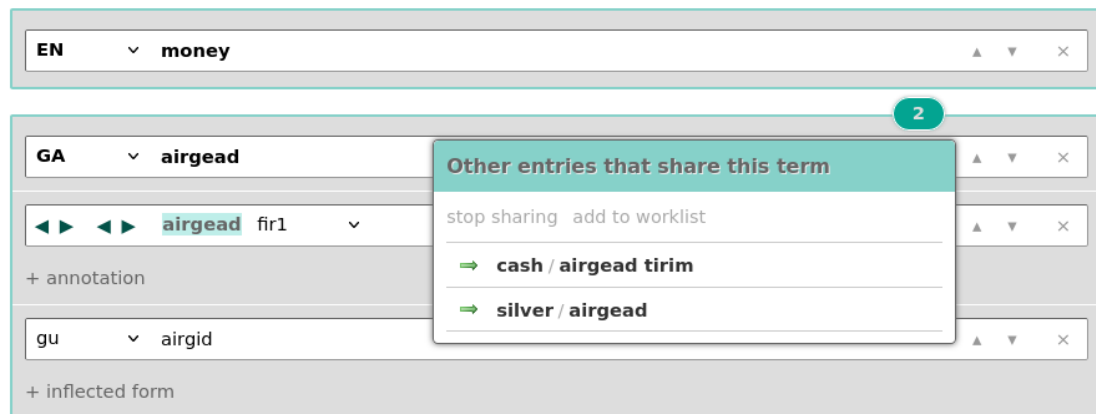
2.1. Grammatical annotations

Most termbases in the translation industry or in knowledge engineering contain only sparse grammatical information: it is expected that the user will be a (near-)native speaker and will need no help in determining the gender of nouns or the plural of noun phrases. In NTDI this assumption does not apply: NTDI is a public-service termbase, targeting the general public and serving a user community with a high percentage of learners and non-native speakers. The consequence is that terms in NTDI come with relatively rich grammatical annotations, both as labels attached to terms (part of speech, gender, inflection paradigm) and as inflected forms added to terms (plurals, genitive case). A speciality is that the termbase allows inline grammatical annotations: it is possible to attach labels not just to the entire term but also to a single word inside it, for example to the head noun of a noun phrase.

2.2. Term sharing

Because terms in NTDI contain a lot of grammatical annotation and because many terms are polysemous (= one term designates multiple concepts), the issue of duplication and consistency have arisen: entering the same term into several entries requires duplicate effort and can result in inconsistencies (for example, when a mistaken grammatical label is corrected in one entry but not in another). To prevent duplication and to enforce consistency, NTDI (and Terminologue) has a feature which allows for a term to be shared among several entries. Any changes made to the term in one entry (including changes to its grammatical annotation or to its infected forms) automatically become visible in the other entries too. Approximately 15% of terms in the termbase are shared like this. This is an example of a database design feature which is motivated not by the user's information needs (the end-users are probably not even aware of it)
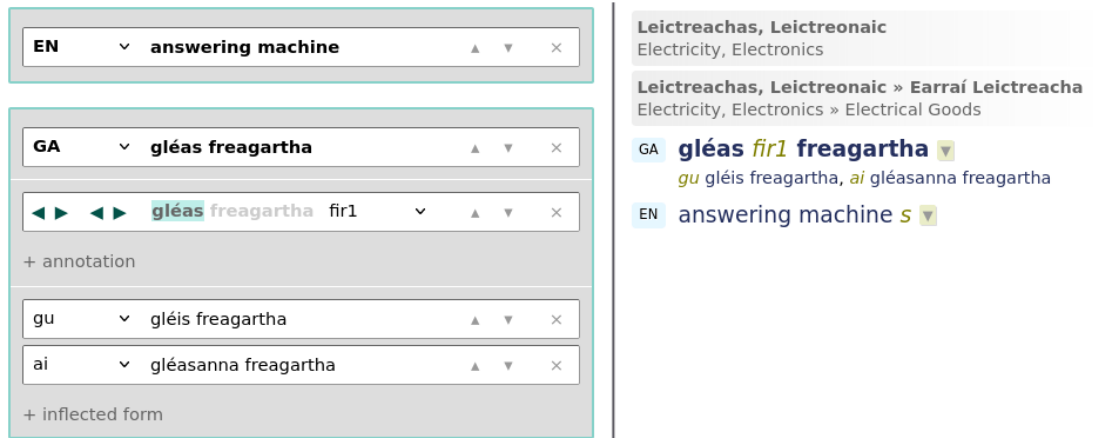
Figure 1: The Irish multiword term *gléas freagartha* with a part-of-speech label attached to the head noun and two inflected forms (genitive case and plural). Editorial interface on the left, public website on the right.

Figure 2: The Irish term *airgead*, with all its grammatical annotation, is shared by three entries.

but by the data-governance needs of the terminologists: the need to eliminate duplicate labour and inconsistency.

2.3. No ontologies

It is popular in terminology to organise entries into networks of *is-a*, *has-a* and other relations, thus building entire ontologies [6]. Ontologies are useful in a knowledge-engineering context where the goal is to enable the user to explore and understand an entire domain. In NTDI, however, this goal is almost absent. Our website traffic statistics show that most users consult NTDI not to explore an entire domain but simply to obtain translations of individual terms. NTDI users typically consult the termbase while they are doing something else: translating or writing. Because of this, the software behind NTDI (Terminologue) has no ontology-building

features. The only type of entry-to-entry relation available is a simple "see also" relation, as well as relations implicit in our relatively rich scheme of hierarchical domain labels. We find that this is sufficient for the information needs of NTDI's users.

2.4. Optional hiding of information

It is a truism that when publishing lexical resources online as opposed to on paper, one does not need to worry about space constraints, as computer memory is practically unlimited. But this does not mean that terminological entries can be arbitrarily long: we still need to take the user's cognitive capacity into account and avoid creating a situation of information overload [7]. For this reason NTDI (and Terminologue) has a feature which allows the terminologists to label certain parts of an entry as non-essential, such as protracted citations from sources, deprecated terms or certain usage examples. Such parts are hidden by default in the public user interface, while users who want to view them can reveal them by clicking a 'plus' icon.

3. Conclusion

The termbase described in this paper departs from established practice in terminology. Many of NTDI's structural features are difficult to map onto structural categories common in other terminological software and in interchange standards such as TBX (for example, TBX has no notion of term sharing). We have attempted to explain in this paper that this divergence is not arbitrary but motivated: motivated by the genre of the termbase (it is a public-service termbase), motivated by the information needs of the end-users, and last but not least, motivated by the data-governance needs of the terminologists.

Acknowledgments

References

[1] M. Měchura, B. Ó Raghallaigh, The Focal.ie National Terminology Database for Irish: software demonstration, in: A. Dykstra, T. Schoonheim (Eds.), Proceedings of the 14th EURALEX International Congress, Fryske Akademy, Leeuwarden/Ljouwert, The Netherlands, 2010, pp. 937–948.

[2] C. Nic Pháidín, G. Ó Cleircín, Ú. Bhreathnach, Building on a terminology resource – the Irish experience, in: A. Dykstra, T. Schoonheim (Eds.), Proceedings of the 14th EURALEX International Congress, Fryske Akademy, Leeuwarden/Ljouwert, The Netherlands, 2010, pp. 954–965.

[3] M. Měchura, B. Ó Raghallaigh, Introducing Terminologue: a cloud-based, open-source terminology management tool, Presented at XIX EURALEX International Congress, 2021.

[4] H. Bergenholtz, S. Tarp (Eds.), Manual of Specialised Lexicography: The preparation of specialised dictionaries, volume 12 of *Benjamins Translation Library*, John Benjamins Publishing Company, Amsterdam, 1995. doi:`10.1075/btl.12`.

[5] R. S. Taylor, The process of asking questions, American Documentation 13 (1962) 391–396. doi:`10.1002/asi.5090130405`.

[6] I. Muñoz, M. R. Zambrana, Applying ontologies to terminology: Advantages and disadvantages, Hermes: Journal of Language and Communication in Business 51 (2013) 65–77. doi:`10.7146/hjlcb.v26i51.97438`.

[7] R. Lew, G.-M. de Schryver, Dictionary users in the digital revolution, International Journal of Lexicography 27 (2014). doi:`10.1093/ijl/ecu011`.