The Focal.ie National Terminology Database for Irish: software demonstration

Michal Boleslav Měchura and Brian Ó Raghallaigh
Dublin City University, Fiontar, Ireland

In this demonstration, we will showcase the National Terminology Database for Irish focal.ie which was launched on-line in 2006 and immediately became very popular, attracting hundreds of thousands of searches every month. The Web site allows users to search a database of around 300,000 terms. It was developed to make the stock of terms of the National Terminology Committee available online, and was designed to be easy to use.

In addition to the public-facing Web site, the software comprises an editorial interface (password-protected Web site), a relational database, and a library of objects and functions that acts as an interface between the two Web sites and the database.

The public-facing Web site allows users to search the database using a 'Quick Search', a 'Complex Search', or an 'Alphabetical Listings' function. The 'Quick Search' function returns 'Similar terms', 'Exact matches', and 'Related matches' from the database. The editorial interface allows users to search and edit the data contained in the database.

While the primary requirement for the public-facing Web site is user-friendliness, the primary requirement for the database is the ability to record complex linguistic data in a logical structure. The structure adopted for the Irish lexical database is based on the conceptual model, widely considered a standard in the terminology industry (ISO 704). The database is multilingual and contains rich grammatical labelling, usage examples, definitions, as well as other information. The database, editorial tools, and public interface were developed by Fiontar in-house using Microsoft technologies. The database and Web sites are hosted by Information Systems & Services, Dublin City University.

A new version of the public site was recently launched and can be accessed at the following URL: http://www.focal.ie/.

1. Introduction

In this demonstration, we showcase the National Terminology Database for Irish focal.ie which was launched on-line in 2006 and immediately became very popular, attracting thousands of searches every month. The average number of searches per month during 2009 was 518,861, and in February 2010, the service dealt with 726,861 searches. The Web site allows users to search a database of around 300,000 terms and contains terms in Irish, English and Latin, as well as a small number of terms in other languages. It was developed to make the stock of terms of the National Terminology Committee[1] available online, and was designed to be easy to use (Měchura and Ó Raghallaigh 2009).

2. System components and functions

In addition to the public-facing Web site, the software comprises an editorial interface (password-protected Web site), a relational database, and a library of objects and function that acts as an interface between the two Web sites and the database.

3. Public features

The public-facing Web site allows users to search a database of around 300,000 terms using a *Quick Search*, an *Advanced Search*, or an *Alphabetical Listings* function. The site also contains help files, and users can fill out a form to request a term from the Terminology Committee if they cannot find it in the database. Figure 1 shows the search interface on the public Web site.

[1] The statutory body responsible for terminology development for the Irish language.

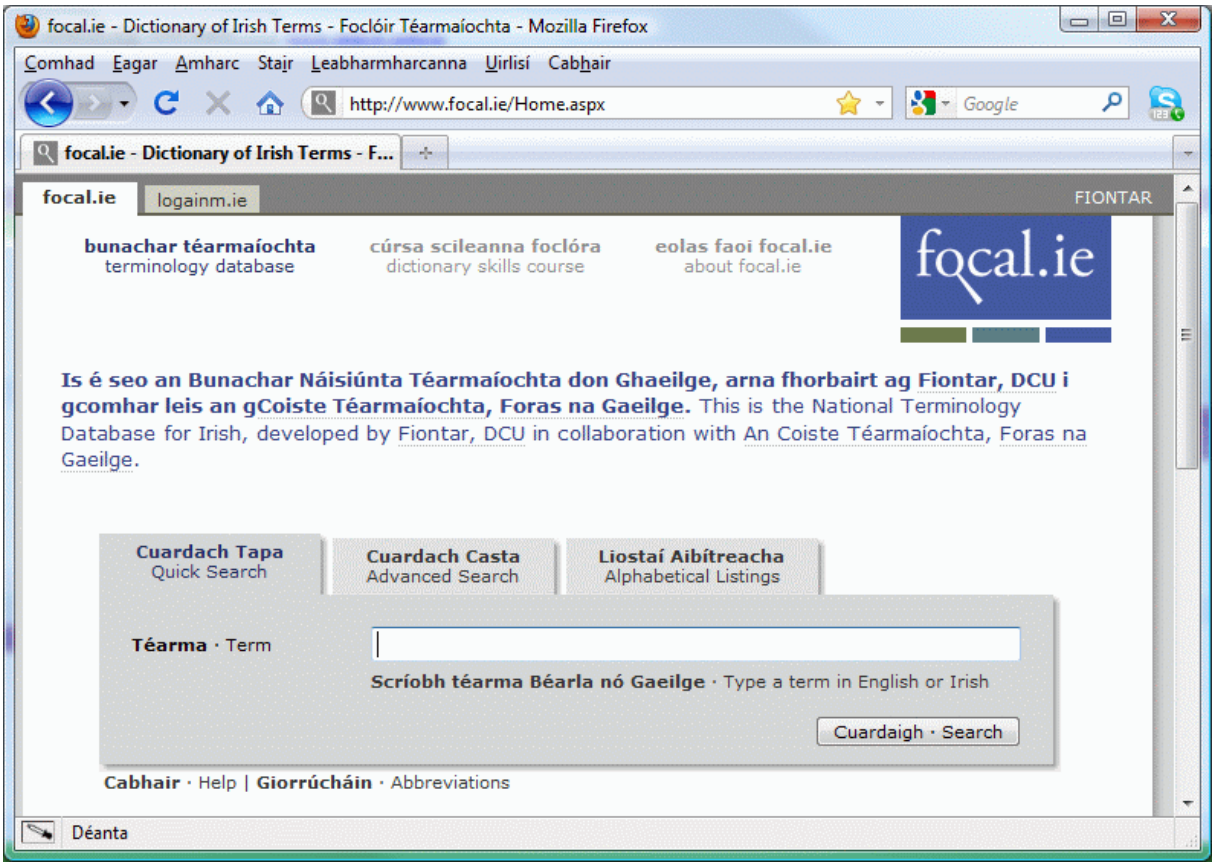

Figure 1. The search interface on focal.ie.

3.1. Quick Search

The *Quick Search* facility provides a quick way for users to search for a term in any language. Quick Search returns *Similar terms*, *Exact matches*, and *Related terms* from the database. When a search returns results in more than one language, the Exact matches and Related terms are grouped according to language, and the user can switch between them. Quick Search is designed to help users find terms even when they don't know exactly what they are looking for.

Similar terms are terms whose spelling is close to that of the text typed by the user. The function acts as a kind of spellchecker by using an implementation of the Levenshtein Algorithm to measure how far the term searched for is from every other term in the database. As well as filtering misspellings, the search for Similar terms may also offer uninflected forms when a user enters an inflected form. In the case of Irish, which is an inflected language, various inflected forms of verbs, nouns and adjectives are recorded in the database and these forms are also searched as part of the search for Similar terms.

Exact matches are terms that match exactly a normalised copy of the text entered by the user. In the case of focal.ie, normalisation involves the reduction of superfluous white space and removal of punctuation.

Related terms are terms that contain the term entered by the user. Figure 2 shows the Similar, Exact and Related results from an example Quick Search for the inflected English word *parades*.

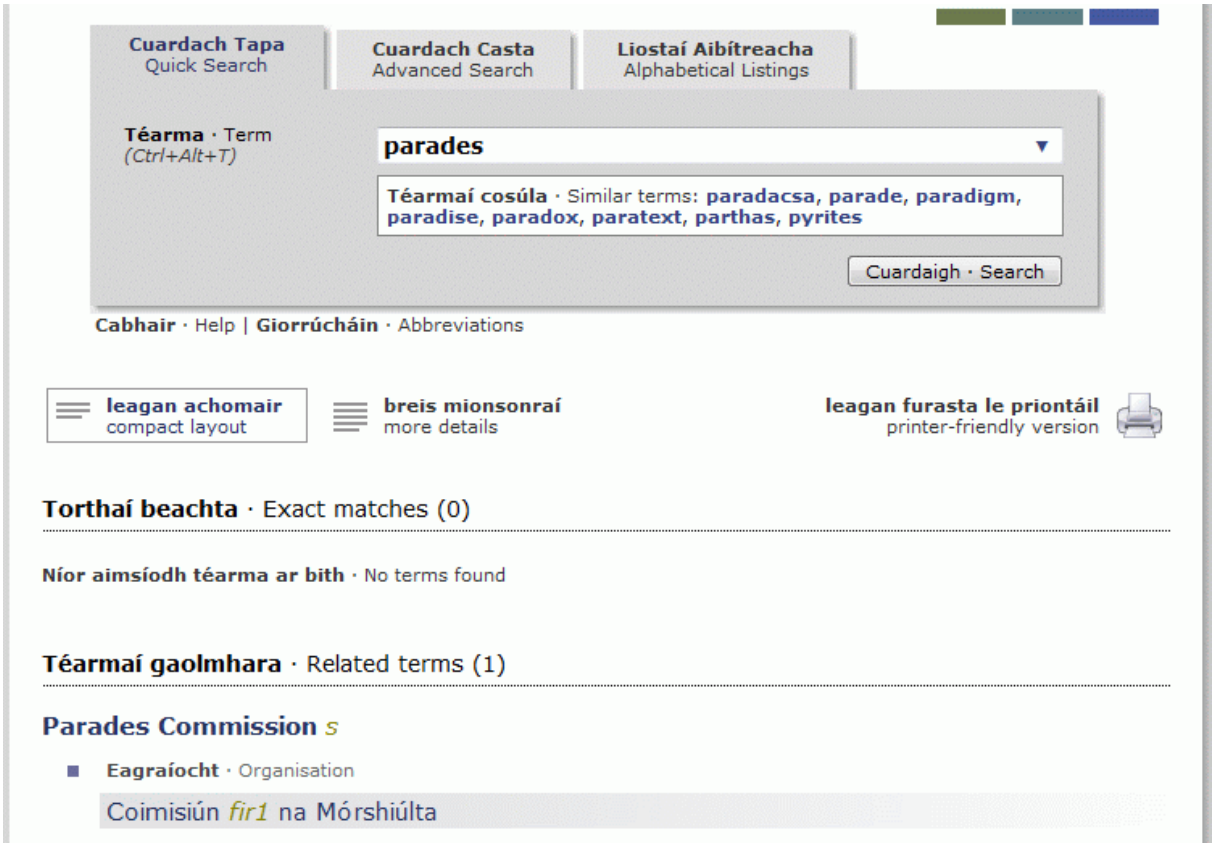

Figure 2. An example Quick Search for the word *parades*.

3.2. Advanced Search

The *Advanced Search* facility, unlike Quick Search, searches only for the text that the user enters as the search query (apart from being case-insensitive). In addition to being able to enter text to search for, the Advanced Search facility allows users to specify whether the text entered constitutes an entire term or the beginning, middle or end of a term.

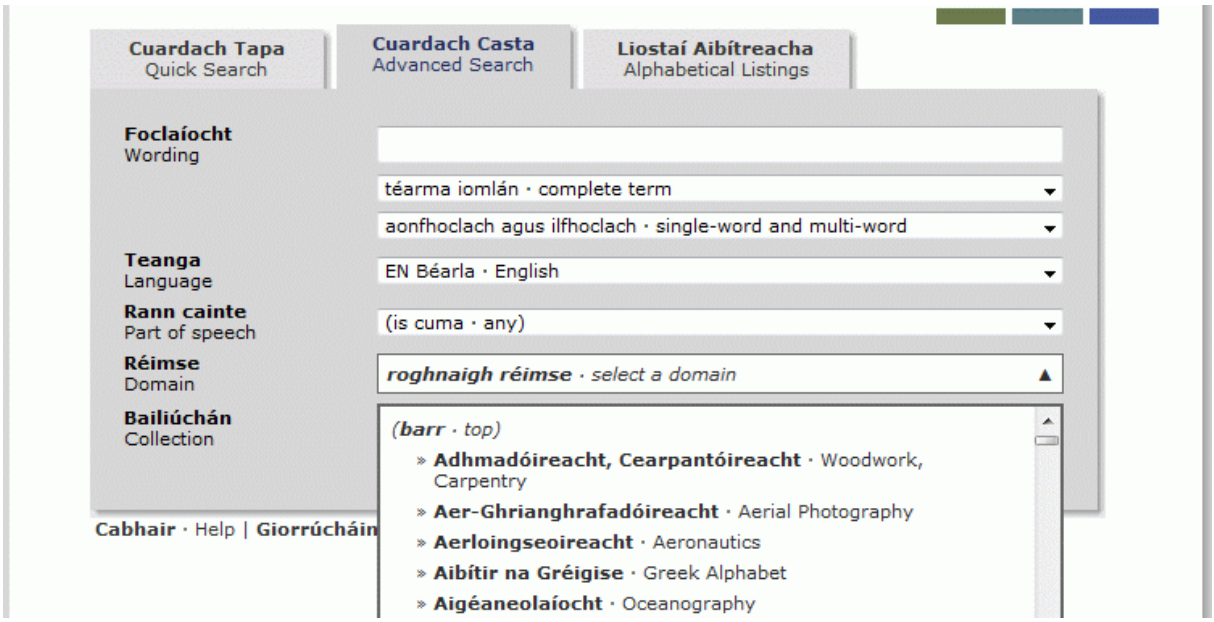

Figure 3. The Advanced Search facility on focal.ie.

As is shown in Figure 3, Advanced Search also allows users to narrow their search according to a number of other parameters. Users can specify if they want to restrict results to single-word or multi-word terms. Additionally, they can specify what language they want results in, or what part of speech, domain or collection (source) they are interested in. Any combination of these criteria can be selected by the user. Wildcards may also be used when searching using Advanced Search.

3.3. Alphabetical Listings

The *Alphabetical Listings* facility allows users to get terms of a particular domain or collection in a particular language in the form of an alphabetical list, such as you would find in a printed dictionary.

4. Database structure

While the primary requirement for the Web site is user-friendliness, the primary requirement for the database is the ability to record complex linguistic data in a logical structure. The structure adopted for the Irish lexical database is based on the conceptual model, widely considered a standard in the terminology industry (ISO 704).
The database is multilingual and contains rich grammatical labelling, usage examples, definitions, as well as other information.

5. Editorial features

This section will demonstrate some of the editorial features available to the terminologists who work on focal.ie. The focal.ie Web site has a password-protected editorial section which allows the editors to carry out various activities, including:

- editing the terminology collection
- editing metadata lists such as language names and domain labels
- reviewing terminology enquiries received from the public
- reviewing and analyzing the public Web site's usage statistics
- controlling aspects of the public Web site's content such as announcements and 'terms of the day' that appear on the home page
- changing one's own password, adding and removing users
- accessing the editorial history of any entry and any user

In this section, we will concentrate on the first bullet point, that is, editing data in the terminology collection.

5.1. Editing the concepts

In the editorial interface, the focus of the terminologist's attention is a concept (as is common in terminology management and as defined, for example, in ISO 704). A concept includes data of several types, including the terms that designate that concept.

Figure 4. Terms that designate a concept, with one term highlighted.

Figure 4 shows the editing window for one of the concepts designated by the English term *protector*. Notice that the terms that designate this concept are listed in three language groups: English terms, Irish terms and terms in other languages. It is possible to add a new term by clicking *nua* 'new' in the appropriate language group. An existing term can be edited by clicking on it and removed by clicking the *bain* ('remove') option that appears next to it when the mouse pointer is positioned over it.

A term can be accompanied with labels that indicate its acceptability (colloquial, deprecated etc.) and clarify its meaning (roughly corresponding to transfer comments in ISO 12620). These can be added by clicking the appropriate options that appear next to the term when the term is highlighted (*soiléiriú* 'clarification', *inghlacthacht* 'acceptability') and edited by simply clicking on them as Figure 5 demonstrates.

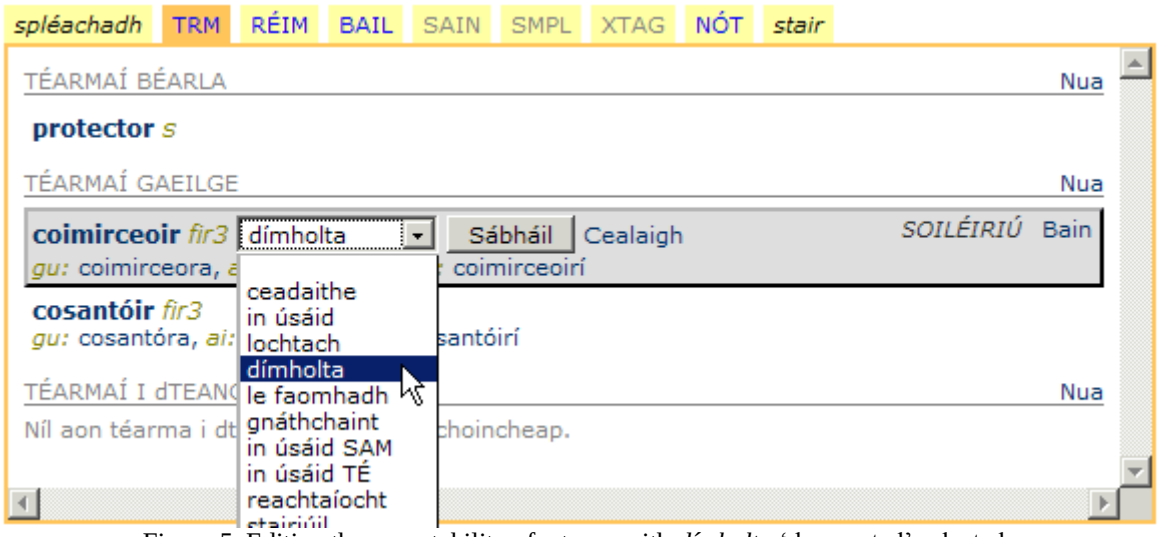

Figure 5. Editing the acceptability of a term, with *dímholta* 'deprecated' selected.

A concept can contain data of many other types beside terms, including domain labels, example sentences, definitions and cross-references. These are accessed by clicking on tabs as indicated in Figure 6. The interface that appears under each tab follows the same pattern as has just been demonstrated: items are edited by either clicking on them or by positioning the mouse over them and choosing from the options that appear.

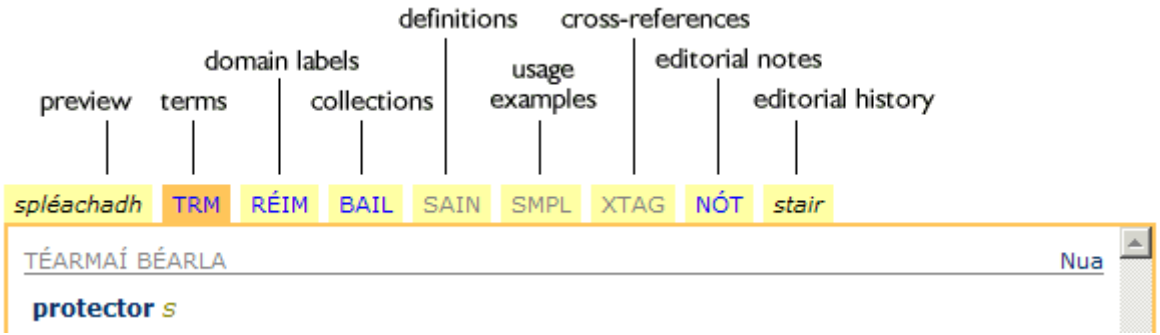

Figure 6. Tabs are used to access the various data types contained in a concept.

The data types that a concept can contain (beside terms) are:

- Domain labels: each concept can have one or more domain labels. Note that in focal.ie, domain labels are hierarchical; for example the domain *Biology* contains the subdomains *Anatomy*, *Physiology* and *Microbiology*. A concept can be assigned to a domain at any level in the hierarchy, and the hierarchy is exploited on the public Web site. If a user wishes to obtain a list of terms from *Biology*, terms from its subdomains will be included in that list also.
- Collections: each concept can be a member of one or more collections. Collections are mainly used to organize concepts into groups for editorial purposes, for example to group concepts that have been published, are to be published, in a particular printed dictionary.
- Definitions: each concept can have one or more definitions. A definition can be bilingual, in Irish only or in English only. A definition can have one or more domain labels, just like a concept can.
- Usage examples: each concept can have one or more usage examples. A usage example is not just a single sentence, it is in fact a complex object that can contain any number of sentences in Irish and English that are mutually equivalent. A usage example can also have directionality: some usage examples are only applicable for the English-to-Irish direction, and are thus only displayed on the public Web site when a search from English to Irish has been preformed.
- Cross-references: each concept can participate in a cross-reference. A cross-reference is not just a clickable hyperlink: like a usage example, it is in fact an object with its own internal structure. A cross-reference is an object that groups two or more concepts together. Thus, if concept A contains a cross-reference to concept B, the database structure automatically guarantees that concept B will contain a corresponding cross-reference to concept A. The type of the cross-reference (hypernym, meronym, etc.) is not specified, the cross-references are simply presented as ‘see also’ hyperlinks on the public Web site.

5.2. Sharing terms between concepts

An important aspect of the database structure that every editorial user needs to understand before they start working in the editorial interface is that in focal.ie, terms can be shared among concepts. If a polysemous term like *award* or *drive* or *folder* designates more than one concept, then the term is not entered in the database multiple times: instead, it is only entered once and then *linked* to all the concepts it designates.

This affects the way editorial users add new terms to concepts. If an editorial user wishes to add a new term, and if that term already exists in a different concept in the database, then the

system offers the user to *link* to the existing term instead of creating a new one. Figure 7 demonstrates this. There is also an option to override this recommendation and to create a new term anyway but this is not recommended.

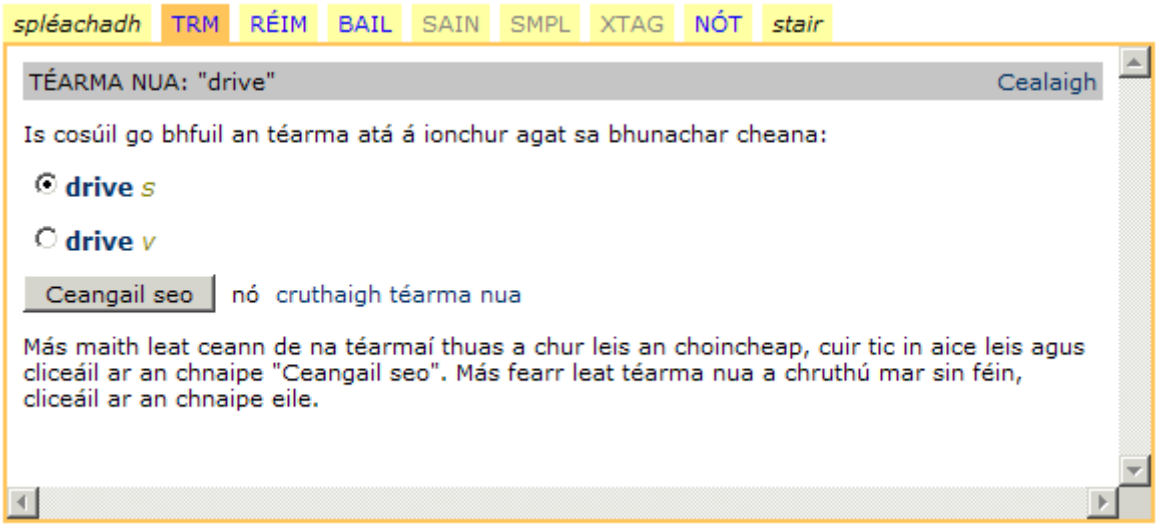

Figure 7. A user has just attempted to add a term that already exists in another concept.

This 'sharing and linking' architecture allows us to achieve consistency of grammatical annotation across concepts. In the focal.ie database, terms (especially Irish terms) are annotated with large amounts of grammatical annotation such as part of speech and inflected forms. Without 'sharing and linking', a polysemous Irish term like *fillteán* 'folder' or *tiomántán* 'drive' would need to be entered in the database as many times as there are concepts it designates, and its grammatical annotations would also need to be replicated the same number of times. It would be difficult to achieve consistent annotation in such a situation. Therefore, a decision was made to employ a relational data model in which terms and concepts are linked in a many-to-many fashion (for more details on the data model used in focal.ie see Měchura (2006)).

An added benefit of the 'sharing and linking' architecture is that the system can easily switch between a concept-oriented view and a term-oriented view of the data. In the editorial interface, users work mainly with concepts. On the public Web site, however, users search for terms and the system presents data to them in a term-oriented layout: for every term that matches the user's query, the system simply looks up the concepts the term is linked to and presents them in a bulleted list that looks like a conventional dictionary entry.

Figure 8a. The editorial view of *tiomántán*.

Figures 8a and 8b demonstrates this duality of views in practice. There are two concepts in the database linked to the Irish word *tiomántán*. From the editorial user's point of view, these are two separate entities. For the public user who has just performed a search for *tiomántán*, however, the two concepts are combined in real time into a single dictionary-like entry.

Figure 8b. The public view of *tiomántán*.

5.3. Editing the terms

Terms are edited in a separate window in the editorial interface. Because the same term can be shared among several concepts, editorial users need to understand that the changes they make to a term may affect more than one concept. To make users aware of this, the system displays shared terms with a different background colour, as Figure 9 shows.

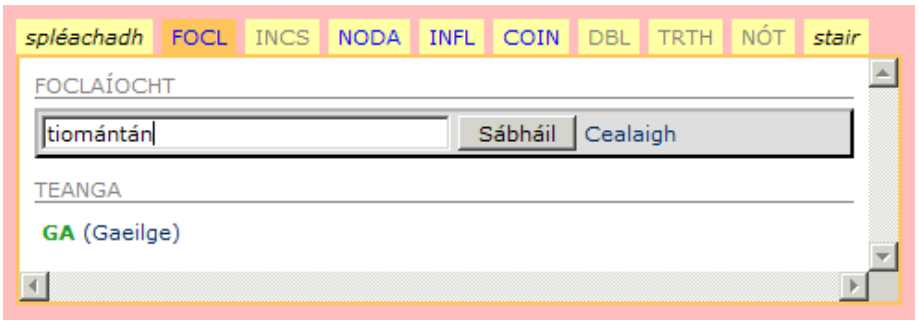

Figure 9. The editing window for the polysemous Irish term *tiomántán*.
The red background serves as a warning that the term is linked to more than concept.

Figure 9 shows the editing window for the Irish term *tiomántán*. In this windows, the user can change the term's wording and the term's language by clicking on them. But a term is a

complex object that consists of more than just a string of characters: it also contains inline grammatical annotations and a list of inflected forms. These can accessed by clicking the respective tabs at the top of Figure 10.

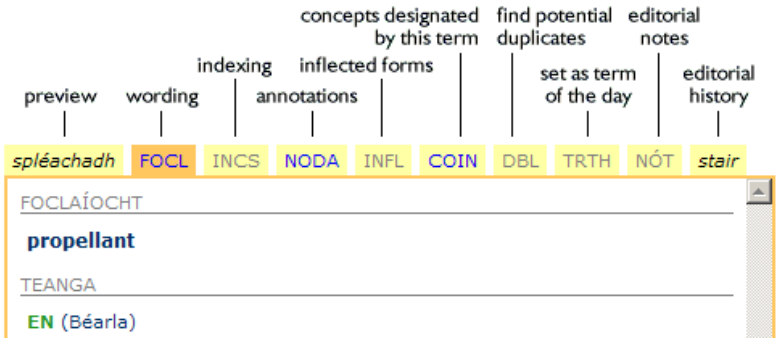

Figure 10. Tabs are used to access the various data types contained in a term.

It is worthwhile to focus on how the focal.ie database handles grammatical annotations. In focal.ie, it is possible to attach a grammatical annotation (such as a part-of-speech label) not only to the whole term but also to a substring within the term. This allows the editorial users to annotate multi-word terms extensively. Figure 11a shows how the Irish term *an Afraic Theas* 'South Africa' is annotated.

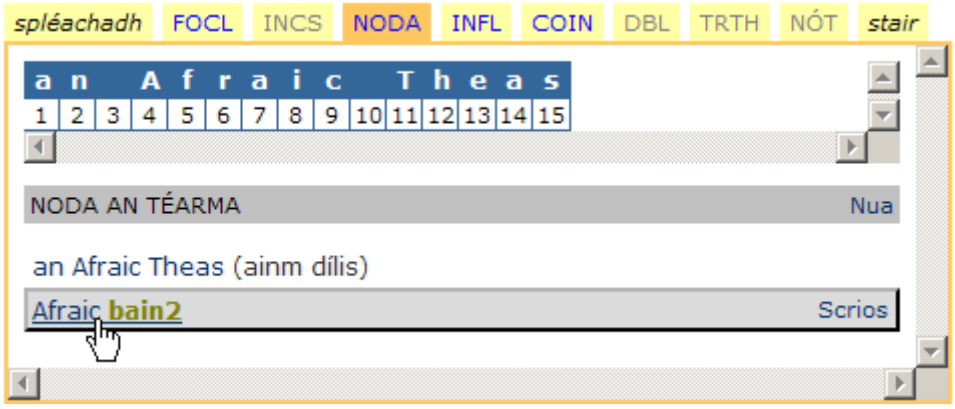

Figure 11a. Grammatical annotation of *an Afraic Theas* in the editorial interface.

The term consists of a definite article, followed by a noun, followed by an adjective. There are two annotations attached to the term: one to the noun *Afraic* that identifies it as a feminine noun of the second declension and another to the whole term that identifies it as a proper name. On the public Web site, these annotations are inserted into the term and the appropriate section of the term is highlighted when the user positions the mouse over it, as Figure 11b shows.

Figure 11b. Grammatical annotation of *an Afraic Theas* on the public Web site.

5.4. Other editorial features

The focal.ie editorial interface contains many other features that cannot be dealt with here, including a section for editing metadata lists (language names, part-of-speech labels, etc.) and a section for analyzing traffic and usage statistics. For example, Figure 12 shows a listing of the most frequent 'quick' searches made during the month of February, 2010. Any searches that returned no results would be highlighted in red in the list. Lists like these are used to develop the content in the focal.ie database and to improve the performance of the search algorithm, as discussed in Měchura (2008).

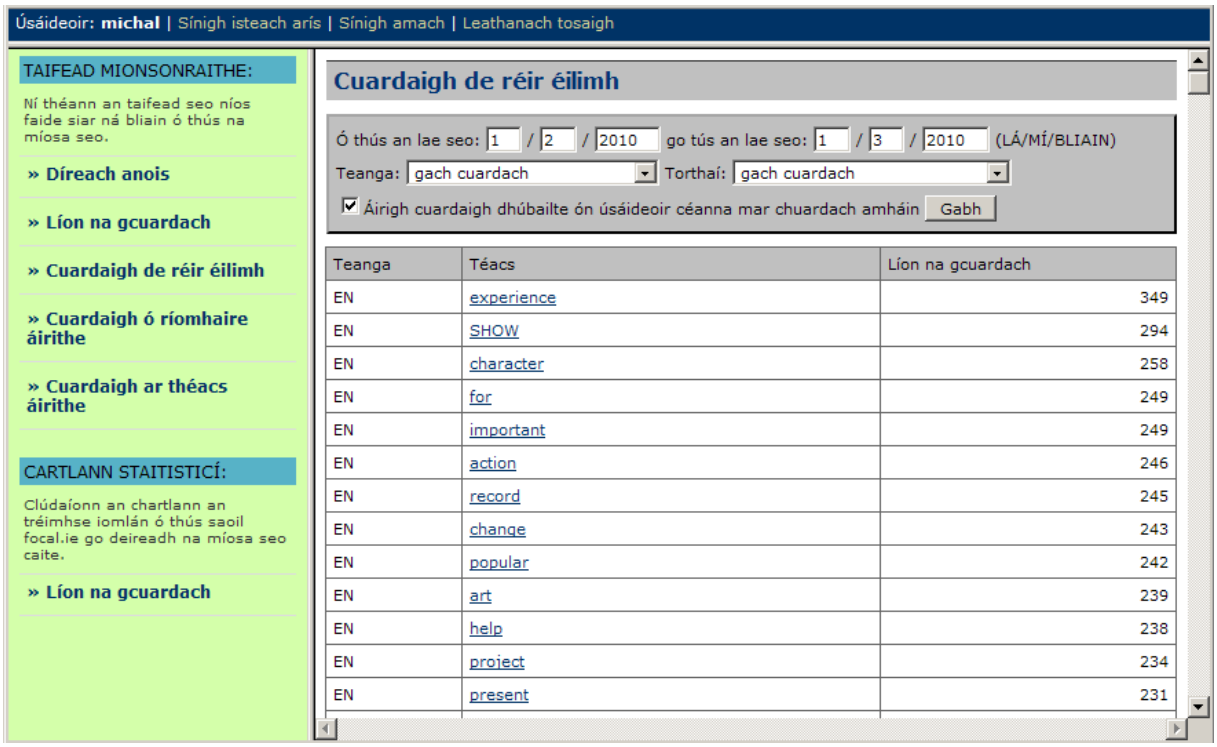

Figure 12. Most frequently made searches in February 2010.

6. Technical infrastructure

The service is hosted by Information Systems & Services, Dublin City University (DCU ISS). The database and Web sites are hosted on two separate Windows servers, one data, one Web, both of which are managed by DCU ISS.

The data server is running Microsoft SQL Server 2005 Standard Edition. The Web sites use Microsoft ASP.NET 3.5 which is installed on the Web server.

DCU ISS provide a backup service for both the data and Web servers.

7. Conclusion

A new version of the public site was recently launched and can be accessed at the following URL: http://www.focal.ie/.

8. Standards

ISO 704:2009 *Terminology work – Principles and methods.*
ISO 12620:2009 *Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources.*

References

focal.ie Irish National Terminology Database [online]. http://www.focal.ie [access date: February 2010].

Měchura, M. B. (2006) 'Finding the right structure for lexicographical data: experiences from a terminology project'. In *Proceedings of the 12th Euralex International Congress*. Turin: Edizioni dell'Orso. 189-198.

Měchura, M. B. (2008). 'Giving Them What They Want: Search Strategies for Electronic Dictionaries'. In *Proceedings of the 13th Euralex International Congress*. Barcelona: Universitat Pompeu Fabra. 1295-1299.

Měchura, M. B.; Ó Raghallaigh, B. (2009). 'User-Friendliness: The Key to Promoting a Minority Language on the Internet'. In *International Conference on Minority Languages (ICML) XII*. Tartu: University of Tartu.