Irish National Morphology Database: a high-accuracy open-source dataset of Irish words

Michal Boleslav Měchura
New English-Irish Dictionary Project
Foras na Gaeilge
Dublin, Ireland
mmechura@forasnagaeilge.ie

Abstract

The Irish National Morphology Database is a human-verified, Official Standard-compliant dataset containing the inflected forms and other morpho-syntactic properties of Irish nouns, adjectives, verbs and prepositions. It is being developed by Foras na Gaeilge as part of the New English-Irish Dictionary project. This paper introduces this dataset and its accompanying software library *Gramadán*.

1 Introduction

The Irish National Morphology Database is a side product of the New English-Irish Dictionary project at Foras na Gaeilge. During work on the dictionary, a requirement arose to include rich morphological information on the target (Irish) side of the dictionary. It has been decided to build a separate morphological dataset that translations in the dictionary would link to. The result can be viewed at `http://focloir.ie/` where clicking a grammatical label next to a translation opens a window listing the inflected forms and other morphological properties of the word. The same data can also be viewed separately at `http://breis.focloir.ie/en/gram/`.

2 Database design

The Irish National Morphology Database has been compiled semi-automatically from several sources available to Foras na Gaeilge, including a machine-readable version of *Foclóir Póca* and grammatical data extracted from *WinGléacht* and *focal.ie*. All data resulting from this process have been proof-read and corrected by editors working on the New English-Irish Dictionary project. Therefore, we describe the database as a high-accuracy dataset: it does not come with a known margin of error and it is meant to have normative force. The language data complies with the Official Standard for Irish (*An Caighdeán Oifigiúil* 2012).

At time of writing, the database contains 6,736 nouns, 983 adjectives, 1,239 verbs and 16 prepositions. New entries are being added continuously.

Each entry has a unique identifier consisting of the lemma followed by a grammatical label, such as `bainis_fem2`. In cases where the grammatical label is not sufficient to distinguish between homonyms, the identifier contains a "disambiguator", such as `glúin_fem2_cos` (the noun *glúin* 'knee' with plural *glúine*) versus `glúin_fem2_aois` (the noun *glúin* 'generation' with plural *glúnta*). The disambiguators (*cos* 'leg', *aois* 'age') are purely mnemotechnic: no attempt is being made to expose the semantics of the lemmas, only that two different lemmas exist with two different sets of inflected forms.

The database structure allows for variation everywhere. Every inflected form (for example, every combination of case and number) is in essence a list of variants which can contain zero, one or more

Proceedings of the First Celtic Language Technology Workshop, pages 50–54,
Dublin, Ireland, August 23 2014.

forms, each with its own grammatical properties. Thus we can accommodate cases when the Offical Standard allows for variation, such as the two genitives of *talamh* 'land' (masculine *talaimh* and feminine *talún*). On the other hand, an empty list of variants implies the form does not exist (or is not known), for example when a noun has no plural.

The entries are encoded in XML. Every entry comes in two shapes: a **minimal** format which contains the smallest necessary set of forms and properties, and an **expanded** format intended for presentation to humans. For example, in the case of nouns, the minimal entries contain only one form for each number and case (e.g. *bainis* 'wedding' in singular nominative) while, in the expanded entry, these are "expanded" to include definitiveness (*bainis* 'a wedding', *an bhainis* 'the wedding'). The expanded entries are then transformed with XSL into HTML and displayed to human users. The minimal entries are intended as a machine-readable resource that can be re-used for other purposes in language technology, such as for building spellcheckers or for query expansion in fulltext search.

Minimal entries are converted into expanded entries using *Gramadán,* a custom-built software library written in C#. *Gramadán* provides functions for performing grammatical operations such as initial mutations, constructing noun phrases from nouns and adjectives, constructing verb phrases from verbs, and so on. The process of converting a minimal entry into an expanded entry is in essence an exercise in natural language generation (where syntactic structures are serialized into strings), and *Gramadán* is in essence a software library for natural language generation in Irish.

2.1 Nouns

Listing 1 shows a typical noun entry (*abhainn* 'river')[1] in minimal format, Listing 2 shows the same entry in expanded format. Notice that each form (`sgNom` being singular nominative, `sgGen` singular genitive and so on) consists of a string (the `default` attribute) with form-specific properties: singular forms have gender while plural forms do not, the plural genitive has strength (a property which signals whether the form is weak or strong). Notice that we have decided to treat gender as a property of a word form, not a property of the whole lemma. This makes it possible to deal with cases like *talamh* 'land' which has two singular genitives, one masculine and one feminine.

2.2 Adjectives

Listing 3 shows a typical adjective entry (*bán* 'white')[2] in minimal format, Listing 4 shows the same entry in expanded format. The forms of an adjective are less evenly distributed than those of a noun: there is one singular nominative, two singular genitives (for agreement with masculine and feminine nouns) and only one plural form for all cases (the singular nominative is used for agreement with weak-plural genitive nouns). This is sufficient information for *Gramadán* to generate the forms needed for agreement with all kinds of nouns in all numbers and cases, as can be seen in the expanded format. The minimal format also contains a graded form which is used by *Gramadán* to generate comparatives and superlatives in the past and present.

2.3 Verbs

Listing 5 shows an extract from a typical verb entry (*bagair* 'threaten')[3] in minimal format, Listing 6 shows a corresponding extract from the same entry in expanded format. Verbs are more complicated than nouns and adjectives in the sense that they contain many more forms. In the Irish National Morphology Database, a verb has forms for up to six **tenses** (past, past continuous, present, present continuous, future, conditional) and two **moods** (imperative, subjunctive). Note that we treat the conditional as a tense because it has the properties of a tense, even though grammar books traditionally categorize it as a mood.

The difference between a tense and a mood is that a tense can generate forms that are either declarative or interrogative, while a mood can only generate declarative forms (*bagair!* 'threaten!', *ná bagair!* 'don't threaten!'). Consequently, every tense form in the minimal format is labelled as being either **dependent** or **independent**, while mood forms have no such distinction. The dependent and independent forms are identical for many verbs, but different for some irregular ones (e.g. *déan* 'make'

[1] For a user-friendly presentation of the noun, see `http://breis.focloir.ie/en/gram/abhainn`

[2] For a user-friendly presentation of the adjective, see `http://breis.focloir.ie/en/gram/bán`

[3] For a user-friendly presentation of the verb, see `http://breis.focloir.ie/en/gram/bagair`

in the past tense: independent *rinne*, dependent *dearna*). The independent forms generate positive declarative forms (*rinne mé* 'I made'), the dependent forms generate all others (*ní dhearna mé* 'I didn't make', *an ndearna mé?* 'did I make?', *nach ndearna mé?* 'didn't I make?')

Additionally, every tense and mood form is assigned to a **person,** which in our analysis is a conflation of person, number and other features: there is a "base" person from which analytic forms are generated (*rinne* 'made' → *rinne muid* 'we made'), there are singular/plural first/second/third persons for synthetic forms (*rinneamar* 'we made'), and there is an "autonomous" person for passive forms of the verb (*rinneadh* 'was made').

A typical verb has, in its minimal format, about 60 individual forms. This is the set from which *Gramadán* can generate a verb phrase in any tense or mood, person, number, polarity (positive or negative) and shape (declarative or interrogative). Unlike other parts of speech where the rules for generating an expanded entry from a minimal one are completely regular, the verbal component in *Gramadán* has some hard-coded exceptions for a small number of irregular verbs. Also, the verb *bí* 'be' is quite exceptional as it is the only verb that has both a present tense (*tá* 'is') and a continuous present tense (*bíonn* 'habitually is'); other verbs only have a continuous present tense (their non-continuous present tense is built analytically from the verbal noun). Finally, the Irish National Morphology Database does not include the copula *is*, as we do not think it is as a verb.

3 More about *Gramadán*

The tool used for processing data in the Irish National Morphology Database, *Gramadán*, deserves separate mention. Besides converting entries from minimal to expanded format, *Gramadán* has additional features both below and above the level of words.

Below the level of words, for nouns and adjectives that have not been included in the Irish National Morphology Database yet, *Gramadán* is able to derive their forms and properties from knowing which inflection class they belong to. Unlike the traditional inflection classes found in Irish dictionaries, *Gramadán* uses a radically different system, inspired by Carnie (2008), where singular and plural classes are separate.

Above the level of words, *Gramadán* can be used as a realisation engine in an NLG (natural language generation) setting. *Gramadán* is able to use data from the Irish National Morphology Database to construct noun phrases, prepositional phrases and rudimentary clauses while respecting the rules of gender and number agreement, initial mutations, case inflections and so on. This aspect of *Gramadán* is in development and the goal is, eventually, to cover all the basic syntactical phenomena of Irish including the construction of clauses containing the copula and the construction of numbered noun phrases (noun phrases with cardinal and ordinal numerals).

While many of *Gramadán's* features are used for processing the Irish National Morphology Database, it is an independent software tool which has potential applications beyond it.

4 Future plans

The Irish National Morphology Database is work in progress and will continue to be developed by Foras na Gaeilge along with other outputs from the New English-Irish Dictionary project. Once the database structure has been finalized and detailed documentation has been produced, the whole dataset (along with its accompanying tool, *Gramadán*) will be released under an open-source licence and made available for download on the Internet. In the longer term, we plan to develop the natural language generation aspect of *Gramadán* and to use it as a basis for assistive language technology, as well as to inform applied research into Irish morphosyntax.

References

An Caighdeán Oifigiúil [the Official Standard]. 2012. Houses of the Oireachtas, Dublin. http://tinyurl.com/coif2012 (accessed 8 May 2014)

breis.foclóir.ie: Dictionary and Language Library. http://beis.focloir.ie/

Andrew Carnie. 2008. *Irish Nouns: A Reference Guide.* Oxford University Press, Oxford.

focal.ie: National Terminology Database for Irish. http://www.focal.ie/

foclóir.ie: New English-Irish Dictionary. http://www.focloir.ie/

Foclóir Póca, Irish-English/English-Irish dictionary. 1986. An Gúm and Department of Education, Dublin.

WinGléacht: CD-ROM. 2007. An Gúm, Dublin.

Appendix A. Code listings

Listing 1. The noun 'abhainn' in minimal format

```
<noun default="abhainn" declension="5" disambig="" isProper="0" isDefinite="0"
allowArticledGenitive="0">
 <sgNom default="abhainn" gender="fem"/>
 <sgGen default="abhann" gender="fem"/>
 <plNom default="aibhneacha"/>
 <plGen default="aibhneacha" strength="strong"/>
</noun>
```

Listing 2. The noun 'abhainn' in expanded format

```
<Lemma lemma="abhainn" uid="abhainn_fem5">
 <noun gender="fem" declension="5">
  <sgNom><articleNo>abhainn</articleNo><articleYes>an abhainn</articleYes></sgNom>
  <sgGen><articleNo>abhann</articleNo><articleYes>na habhann</articleYes></sgGen>
  <plNom><articleNo>aibhneacha</articleNo><articleYes>na haibhneacha</articleYes></plNom>
  <plGen><articleNo>aibhneacha</articleNo><articleYes>na n-aibhneacha</articleYes></plGen>
 </noun>
</Lemma>
```

Listing 3. The adjective 'bán' in minimal format

```
<adjective default="bán" declension="1" disambig="">
 <sgNom default="bán"/>
 <sgGenMasc default="báin"/><sgGenFem default="báine"/>
 <plNom default="bána"/>
 <graded default="báine"/>
</adjective>
```

Listing 4. The adjective 'bán' in expanded format

```
<Lemma lemma="bán" uid="bán_adj1">
 <adjective declension="1">
  <sgNomMasc>bán</sgNomMasc><sgNomFem>bhán</sgNomFem>
  <sgGenMasc>bháin</sgGenMasc><sgGenFem>báine</sgGenFem>
  <plNom>bána</plNom><plNomSlen>bhána</plNomSlen>
  <plGenStrong>bána</plGenStrong><plGenWeak>bán</plGenWeak>
  <comparPres>níos báine</comparPres><comparPast>ní ba bháine</comparPast>
  <superPres>is báine</superPres><superPast>ba bháine</superPast>
 </adjective>
</Lemma>
```

Listing 5. Extract from the verb 'bagair' in minimal format

```
<?xml version='1.0' encoding='utf-8'?>
<verb default="bagair" disambig="">
 <verbalNoun default="bagairt"/>
 <verbalAdjective default="bagartha"/>
 <tenseForm default="bagair" tense="Past" dependency="Indep" person="Base"/>
 <tenseForm default="bagraíomar" tense="Past" dependency="Indep" person="Pl1"/>
 <tenseForm default="bagraíodar" tense="Past" dependency="Indep" person="Pl3"/>
 <tenseForm default="bagraíodh" tense="Past" dependency="Indep" person="Auto"/>
 ...
</verb>
```

Listing 6. Extract from the verb 'bagair' in expanded format

```
<Lemma lemma="bagair" uid="bagair_verb">
 <verb>
  <vn>bagairt</vn>
  <va>bagartha</va>
  <past>
```

```
   <sg1><pos>bhagair mé</pos><quest>ar bhagair mé?</quest><neg>níor bhagair mé</neg></sg1>
   <sg2><pos>bhagair tú</pos><quest>ar bhagair tú?</quest><neg>níor bhagair tú</neg></sg2>
   <sg3Masc><pos>bhagair sé</pos><quest>ar bhagair sé?</quest><neg>níor bhagair sé</neg></sg3Masc>
   <sg3Fem><pos>bhagair sí</pos><quest>ar bhagair sí?</quest><neg>níor bhagair sí</neg></sg3Fem>
   <pl1>
    <pos>bhagraíomar</pos><pos>bhagair muid</pos>
    <quest>ar bhagraíomar?</quest><quest>ar bhagair muid?</quest>
    <neg>níor bhagraíomar</neg><neg>níor bhagair muid</neg>
   </pl1>
   <pl2><pos>bhagair sibh</pos><quest>ar bhagair sibh?</quest><neg>níor bhagair sibh</neg></pl2>
   <pl3>
    <pos>bhagair siad</pos><pos>bhagraíodar</pos>
    <quest>ar bhagair siad?</quest><quest>ar bhagraíodar?</quest>
    <neg>níor bhagair siad</neg><neg>níor bhagraíodar</neg>
   </pl3>
   <auto><pos>bagraíodh</pos><quest>ar bagraíodh?</quest><neg>níor bagraíodh</neg></auto>
  </past>
  ...
 </verb>
</Lemma>
```