Introduction to *Gramadán* and the Irish National Morphology Database

Michal Boleslav Měchura and Foras na Gaeilge
Version 2, October 2015

This is an English translation of the first two chapters of an original document in Irish titled *'Bunachar Náisiúnta Moirfeolaíochta agus Gramadán: Doiciméadúchán Teicniúil'*.

1. Introduction

This document describes two computational-linguistic resources which are available for download from www.teanglann.ie/en/gram:

Irish National Morphology Database

> This is a large collection of Irish words which records their inflected forms and linguistic properties. The database contains some 43,000 entries and covers nouns, adjectives, verbs and prepositions. It can be reused under the *Open Database Licence*.[1]

Gramadán

> This is a software library which accompanies the database. It contains functions for loading entries from the database into memory and for performing operations on them, such as composing a noun phrase from a noun and an adjective. Gramadán can be used under the *Creative Commons Attribution (4.0 International) Licence*.[2]

1.1. Feedback

As the material is being made available under an open licence, unfortunately we don't have the resources to provide continuous support to users. Feedback may be sent to aiseolas@focloir.ie.

1 http://opendatacommons.org/licenses/odbl/summary/
2 http://creativecommons.org/licenses/by/4.0/

1.2. Origin of the database

The database was created as part of the *New English-Irish Dictionary Project*, `www.focloir.ie`. Its purpose is to supply correct forms of Irish words which comply with the *Official Standard*. Every entry in the database has been proofread by the dictionary's editorial team and it is therefore believed that the database is accurate – it is not the result of statistical guesswork and it does not come with a known margin of error.

1.3. Database structure

1.3.1. The folders

Every entry in the database is stored in an XML file. The files are stored in folders according their part of speech:

Folder name	Part of speech	Number of words (at time of writing)
`noun`	nouns	~30,700
`adjective`	adjectives	~8,700
`verb`	verbs	~3,300
`preposition`	prepositions	16
`nounPhrase`	noun phrases	5

1.3.2. File names

Each file has a unique name, such as `madra_masc4.xml`, `sásta_adj3.xml`, `dúisigh_verb.xml`, `faoi_prep.xml`, `fear_poist_NP.xml`. The names follow a unified naming convention. Each name contains the following parts (with an underscore `_` between each of the two parts):

1. The first part is the word's basic form (its lemma, such as `madra`, `sásta`). If it contains a space, the space is replaced with an underscore (eg. `fear_poist`).
2. The second part is a grammatical label (eg. `masc4`, `adj3`). You are advised to ignore what the labels means because it does not contain any information which could not be derived from the XML file's contents. The label's only purpose is to distinguish between pairs of words that have the same basic form (lemma) but which constitute different words (such as `póg_fem2.xml` and `póg_verb.xml`, `aire_fem4.xml` and `aire_masc4.xml`).
3. Sometimes the grammatical labels are not enough to distinguish between a pair of words. In such cases a third part is present in the file name to make the distinction, for example `glúin_fem2_aois.xml` (*glúin* 'generation', plural *glúnta*) and `glúin_fem2_cos.xml` (*glúin* 'knee', plural *glúine*). Again, you are advised to ignore the literal meaning of these disambiguators as they may not always express fully how the two words differ in meaning – their only purpose is to make sure that each has a unique name.

The file names are unique throughout the database. Even if you placed all the files in a single folder there would be no conflict.

1.3.3. Reading the files

Each file contains a well-formed XML document, encoded in UTF-8 with Windows-style line breaks (`\r\n`). The files can be read with any software program, but Gramadán is available to make some operations easier.

1.4. Gramadán structure

Gramadán was written in the C# programming language and it is therefore a *.NET Framework Class Library*, targeting version 3.5 or newer of Microsoft's *.NET Framework.* It is being made available to you as a DLL file which you can embed in your own software.

Gramadán's source code is also being made available to you. You can open it in any text editor to examine it, and you can open the entire project in *Microsoft Visual Studio* (version 2008 or newer).

Gramadán contains a number of classes to represent objects of various types: `Noun` for nouns, `Adjective` for adjectives, `NP` for noun phrases and so on.

1.5. Documentation structure

Detailed documentation for Gramadán and the database is available in Irish only. It explains the structure of the XML files for each part of speech, how to load them in Gramadán, and what operations Gramadán can perform on them.

The documentation has three sections. **Section A** deals with nouns, adjectives and noun phrases. **Section B** deals with verbs and verbal phrases. **Section C** deals with various supplementary topics (including a library of morphological paradigms for nouns and adjectives).

The document you are reading now is an English translation of the first two chapters of the original documentation in Irish. The next chapter here will give you a preview of what the database contains and what Gramadán can do with it. For more detailed information you will need to refer to the Irish documentation.

2. Preview

This chapter will give you a preview of the kinds of data you can find in the database and the kinds of operations Gramadán can perform on them.

2.1. Nouns

You can use Gramadán to load a noun from an XML file into memory like this:

```
Noun abairtN=new Noun(@"C:\MBM\Gramadan\BuNaMo\noun\abairt_fem2.xml");
```

Now you have an object named `abairtN` which contains everything Gramadán knows about the noun *abairt* 'sentence': what gender it is (feminine), what its genitive case is (*abairte*), what its plural is (*abairtí*) and so on.

To do something useful with the noun you need to build a noun phrase from it. This is how you do that:

```
NP abairtNP=new NP(abairtN);
```

You now have an object that represents a noun phrase consisting of the noun *abairt* and you can write out its forms. For example, these are its singular forms in the nominative and genitive case ('the sentence' and 'of the sentence'):

```
Console.WriteLine(abairtNP.sgNomArt[0].value); //"an abairt"
Console.WriteLine(abairtNP.sgGenArt[0].value); //"na habairte"
```

You see that Gramadán automatically takes care of using the correct form of the noun, applying the correct initial mutation on it and putting the correct form of the definite article in front of it. The fields `.sgNomArt` and `.sgGenArt` mean 'singular, nominative, with article' and 'singular, genitive, with article' respectively. Similarly, you can write out the noun phrase's plural forms ('the sentences' and 'of the sentences'):

```
Console.WriteLine(abairtNP.plNomArt[0].value); //"na habairtí"
Console.WriteLine(abairtNP.plGenArt[0].value); //"na n-abairtí"
```

Nouns are discussed in **Chapter 3** and noun phrases in **Chapter 5** of the documentation in Irish.

2.2. Adjectives

You can use Gramadán to load an adjective from an XML file into memory like this:

```
Adjective morAdj=new Adjective(@"C:\MBM\Gramadan\BuNaMo\adjective\mór_adj1.xml");
```

You now have an object which contains data about the adjective *mór* 'big': what forms it has in the genitive case (*móir* for masculine nouns, *móire* for feminine nouns), what plural it has (*móra*), what its graded form looks like (*mó*) and so on.

To do something useful with the adjective, one thing you can do is build a noun phrase from it and a noun (for example *capall* 'horse'):

```
Adjective morAdj=new Adjective(@"C:\MBM\Gramadan\BuNaMo\adjective\mór_adj1.xml");
Noun capallN=new Noun(@"C:\MBM\Gramadan\BuNaMo\noun\capall_masc1.xml");
NP capallNP=new NP(capallN, morAdj);
```

If you now write out the noun phrase's forms, you will see that Gramadán has automatically taken care of gender and number agreement between the noun and the adjective, of applying the correct initial mutations in the correct places, and of using the correct form of the definite article ('the big horse' etc.):

```
Console.WriteLine(capallNP.sgNomArt[0].value); //"an capall mór"
Console.WriteLine(capallNP.sgGenArt[0].value); //"an chapaill mhóir"
Console.WriteLine(capallNP.plNomArt[0].value); //"na capaill mhóra"
Console.WriteLine(capallNP.plGenArt[0].value); //"na gcapall mór"
```

You may recall that, while most Irish adjectives are placed after the noun, some function as prefixes, such as the adjective *sean* 'old'. Facts like this are recorded in the database, Gramadán pays attention to them and attaches the adjective correctly ('the old horse' etc.):

```
Adjective seanAdj=new Adjective(@"C:\MBM\Gramadan\BuNaMo\adjective\sean_adj1.xml");
Noun capallN=new Noun(@"C:\MBM\Gramadan\BuNaMo\noun\capall_masc1.xml");
NP capallNP=new NP(capallN, seanAdj);
Console.WriteLine(capallNP.sgNomArt[0].value); //"an seanchapall"
Console.WriteLine(capallNP.sgGenArt[0].value); //"an tseanchapaill"
Console.WriteLine(capallNP.plNomArt[0].value); //"na seanchapaill"
Console.WriteLine(capallNP.plGenArt[0].value); //"na seanchapall"
```

Adjectives are discussed in **Chapter 4** and noun phrases in **Chapter 5** of the documentation in Irish.

2.3. Verbs

A verb can be loaded from an XML file into memory like this (here the verb *abair* 'say'):

```
Verb abairV=new Verb(@"C:\MBM\Gramadan\BuNaMo\verb\abair_verb.xml");
```

The object (and the XML file) contains a very large number of inflected forms for the verb in various tenses and moods, analytic and synthetic forms, the passive forms of the verb, and so on. To use the verb, you need to build a verbal phrase from it:

```
VP abairVP=new VP(abairV);
```

You can now write out the verbal phrase's forms. For example, this is how you obtain its form in the past tense, first person ('I said'):

```
Console.WriteLine(
   abairVP.tenses[VPTense.Past][VPShape.Declar][VPPerson.Sg1][VPPolarity.Pos][0].value
);
//"dúirt mé"
```

The items between square brackets tell Gramadán which form of the verbal phrase you want. For example, if you change `VPTense.Past` to `VPTense.Fut` you will get the future tense (*déarfaidh mé* 'I will say') instead of the past. If you change `VPPolarity.Pos` to `VPPolarity.Neg` you will get the negative version (*ní dúirt mé* 'I did not say') instead of the positive. And if you change `VPShape.Declar` to `VPShape.Interrog` you will get the interrogative version (*an ndúirt mé*? 'did I say?') instead of the declarative.

Sometimes, Gramadán will have more than one form for a particular tense and person. This often happens when there is a choice between a synthetic form (*dúramar* 'we said') and an analytic form (*dúirt muid* 'we said'). In cases like that you can obtain all available forms by looping through them:

```
foreach(Form f in abairVP.tenses[VPTense.Past][VPShape.Declar][VPPerson.Pl1][VPPolarity.Pos]) {
   Console.WriteLine(f.value);
}
//"dúramar"
//"dúirt muid"
```

Verbs are discussed in **Chapter 6** and verbal phrases in **Chapter 7** of the documentation in Irish.