A Two-level Morphological Analyser and Generator for Irish using Finite-State Transducers

Elaine Uí Dhonnchadha

Institiúid Teangeolaíochta Éireann
31 Plás Mhic Liam, Baile Átha Cliath 2, Éire,
and Dublin City University
Glasnevin, Dublin 11, Ireland.
elaine@ite.ie

Abstract

Computational morphology is an important part of natural language processing. Finite-state techniques have been applied successfully in computational phonology and morphology to many of the world's major languages. Celtic languages such as Modern Irish present challenging morphological features that to date have not been addressed using finite-state technology. This paper presents a finite-state two-level morphology of Irish developed using Xerox Finite-State Tools. The system encodes the inflectional morphology of all inflected parts-of-speech in Modern Irish. The morphotactics of stems and affixes are encoded in the lexicon and word mutations are implemented as a series of replace rules encoded as regular expressions. Both the lexicons and rules are compiled into finite state transducers and combined to produce a single lexical transducer for the language. A major advantage of finite-state two-level implementations of morphology is their inherent bi-directionality; the same system is used for both analysis and generation of word forms in the language. This resource can be used as a component part in many NLP applications such as spelling checkers/correctors, stemmers, and text to speech synthesisers. It can also be used in tokenising, lemmatising and part-of-speech tagging of a corpus of text. The system, which is designed for broad coverage of the language, is evaluated against the most frequently used words in a corpus of contemporary Irish texts. Finally, possible extensions to the system are suggested, such as derivational morphology and the inclusion of dialectal or historical word-forms.

1. Introduction

Computational morphology is an important step in processing natural language. Finite-state techniques have been applied successfully in computational phonology and morphology to many of the world's major languages. Celtic languages such as Irish present challenging morphological features which have not been addressed to date using finite-state technology. This paper presents a finite-state two-level morphology (Koskenniemi, 1983) of Irish developed using Xerox Finite-State Tools (Karttunen, 1994, Grefenstette *et al.* 2000).

Irish, a verb-initial, inflectional language, belongs to the Celtic branch of the Indo-European family of languages (Ruhlen, 1987). Constitutionally it is the first official language of Ireland, with English being the second (Bunreacht na hÉireann, 1937). Irish is spoken on a daily basis by ten percent of the population and over one third report an ability to speak the language (Central Statistics Office, 1998). English is however the primary language of communication in the country. Due to this relatively weak position Irish has to date suffered from a lack of basic computational language support tools and resources.

The two-level morphology for Irish presented in this paper is implemented as a finite-state lexical transducer and encodes all of the inflectional morphology of Irish. The resource can be used as a basic component in NLP applications such as spelling checkers, stemmers for Information Retrieval, text to speech synthesisers or as a component step in parsing and generation. In corpus linguistics the implementation can be used for tokenising text, lemmatising and as an input to automatic part-of-speech tagging. It could form an interface to an electronic learners' dictionary of Irish where the user could look up inflected word forms.

2. Morphological Features

In Irish morphology, the suffix is the predominant type of affix used, although there are a number of proclitics used in verbal inflection. Prefixes are mainly used derivationally (Stenson, 1981). Inflections also frequently include modification to the stem.

The system implemented covers the inflection of nouns, adjectives, verbs and prepositional pronouns. This paper focuses on some of the morphology and morphotactics of nominal inflection. In Irish, nouns are either masculine or feminine and are inflected for number and case. These inflections can have up to three components: initial mutation, final mutation and suffixation (Stenson, 1981).

2.1. Initial Mutations

Initial mutations are the changes which occur to the start of the word and the processes involved are lenition, eclipsis and the prefixing of vowel-initial and s-initial stems. These mutations started out as phonological accommodations but have over time become grammaticalised (Campbell, 2000; Ó Cuív, 1987).

Lenition is a softening of the initial consonant of a word as in the following example where a feminine noun *bean* is lenited after the definite article *an*.

(1) *bean* 'a-woman'
*an b**h**ean* 'the woman'

Eclipsis is the prefixing of a consonant to the stem (the original consonant becomes silent) as in the following example where the noun *teach* is eclipsed following the simple preposition *i* meaning 'in'.

(2) *teach* 'a-house'
i ***d****teach* 'in a-house'

The letters *h, n* and *t* are prefixed to vowel-initial and s-initial stems in various grammatical contexts. The following are examples of some prefixed nouns.

(3) *a aois* 'her age'
a ***h****aois* 'his age'
a ***n****-aois* 'their age'

(4) *arán* 'bread'
an ***t****-arán* 'the bread'

(5) *seachtain* 'a-week'
an ***t****seachtain* 'the week'

2.2. Final Mutations

Vowel harmony within inflected forms is an important feature of Irish morphographemics. A stem with a broad ending (orthographically denoted by vowels: a, á, o, ó, u, ú) may be combined only with a broad suffix and likewise a slender stem ending (orthographically denoted by vowels: i, í, e, é) may be combined only with a slender suffix. For some words the stem itself must change to maintain this harmony (through broadening or slendering of the final syllable), and for others there are alternative suffixes.

Example (6) below shows a broad stem *cos* which is slenderised by the insertion of a slender vowel *i* before the final consonant to accommodate the slender suffix *–e*. Conversely in (7) the stem *bliain* is broadened through the removal of the final *i* in order to accommodate the broad suffix *–ta*. Finally (8) shows examples of stems which don't change since they combine with a suffix which has broad and slender allomorphs, *–anna/-eanna*.

(6) *cos* 'a-foot'
*na co****i****s****e*** 'of the foot'

(7) *bliain* 'a-year'
na blianta 'years'

(8) *carr* 'a-car'
*carr****anna*** 'cars
seit 'a-set'
*seit****eanna*** 'sets'

Some polysyllabic stems whose final syllable is unstressed undergo syncopation when a suffix is attached i.e. the vowels of the stem's final syllable are deleted.

(9) *cath****ai****r* 'a-city'
*cathr****acha*** 'cities'

The processes of broadening, slendering and syncopation of the last syllable of a word are known as final mutations (Na Bráithre Críostaí, 1999, Ó Siadhail, 1989).

3. Computational Morphology

Morphology deals with the internal structure of words. The aim of computational morphology is to encode in as efficient a manner as possible knowledge about:

a) the constituent parts of words (morphemes)
b) the rules for combining morphemes (morphotactics)
c) the effects of combining morphemes (morphographemics)

3.1. Two-level Morphology

Koskenniemi's (1983 & 1997) two-level morphology consists of three main elements - two representations and one relation. The two representations are the *surface representation* of a word – usually the orthographic (or phonemic) form – and the *lexical representation* which describes the underlying morphemic structure of the surface form. There is also a *rule* component which relates the two representations to one another. In the present implementation the lexical and surface representations are defined in the lexicon and replace rules are applied separately.

In the following examples, (10b) and (11b) are the Irish equivalents of the English words (10a) and (11a). The surface level is the fully inflected word form and the lexical level defines the stem plus a set of morphological feature tags relating to the word.

(10) a. Lexical level: walk+Verb+Past
Surface level: walked

b. Lexical level: siúil+Verb+Past+1P+Pl
Surface level: shiúileamar

(11) a. Lexical level: father+Noun+Pl
Surface level: fathers

b. Lexical level: athair+Noun+M+Com+Pl+Def
Surface level haithreacha

In (10a), mapping the lexical to the surface level only requires the addition of a suffix '*-ed*' to the stem '*walk*'. In its Irish counterpart (10b) as well as adding the suffix '*-eamar*' to the stem '*siúil*' an initial-mutation of the stem is also required – in this case lenition. This requirement is flagged in the lexicon by means of a temporary (intermediate) mark-up tag (e.g. ^Len) appended to the surface form.

All concatenative morphology is encoded in the lexicon and stem modifications are implemented using replace rules. For example an inflectional mark-up tag such as ^Len will trigger a replace rule for lenition (encoded as a regular expression) to insert '*h*' after the initial consonant in '*siúil*'.

Where replace rules are used there will be one or more intermediate levels between the lexical and the final surface form (Jurafsky, 2000). Therefore (10b) and (11b) could be re-stated as follows:

(12)b. Lexical level: siúil+Verb+Past+1P+Pl
Intermediate: siúileamar^Len
Surface level: shiúileamar

(13)b. Lexical level: athair+Noun+M+Com+Pl+Def
Intermediate1: athaireacha^Sync^Hv
Intermediate2: hathaireacha^Sync
Surface level haithreacha

In (13b) the surface level string produced by the lexicon transducer, i.e. *athaireacha^Sync^Hv,* is transformed by a h-prefix replace rule to become *hathaireacha^Sync* which in turn is transformed by a syncopation replace rule to become *haithreacha* the desired surface level form.

3.2. Finite-state technology

A two-level morphology is generally implemented as a finite-state transducer (fst). A finite-state transducer consists of a network of states and directed arcs. The arcs each have an upper level and lower level label. The fst's upper level labels correspond to the lexical representation and the fst's lower level labels correspond to the surface representation.

The network in Figure 1 encodes two Irish lexicon entries, *cat* which is a noun meaning 'cat' and *cas* a verb meaning 'turn'. By reading the upper labels of the path 0, 1, 2, 3, 5 through the network we get `cat+Noun` and by reading the lower labels we get `cat` (the ∈ symbol signifies that the `+Noun` symbol has no corresponding representation on the surface level).

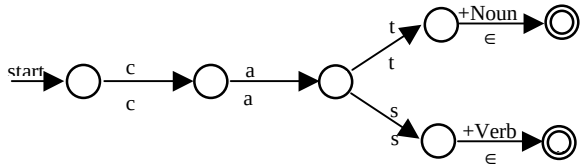

Figure 1. Finite-state network with two words.

4. Implementation

The morphological analyser and generator for Irish is implemented as a lexical transducer. Information relating to the morphotactics of stems and affixes is held in lexicons. Initial and final mutations are implemented as a series of replace rules encoded as regular expressions. Both lexicons and replace rules are compiled into finite state transducers using the Xerox Lexical Tools. Individual transducers are then composed together to produce a single transducer for the language which maps the lexical level strings to the surface level strings in one step.

4.1. Xerox Lexical Tools

Xerox has developed a set of finite-state tools which provide a means of implementing two-level morphologies. The tools are natural language independent and have been used to implement morphologies for many of the major European languages (English, Spanish, French, German etc.) as well as Arabic, Turkish, Japanese and others. The two tools used in this work are *lexc* (a lexicon compiler) and *xfst* (a general purpose finite state transducer tool). The interested reader will find full details of the tools and their uses in *Finite State Morphology: Xerox Tools and Techniques* (Beesley and Karttunen, 2002 forthcoming).

4.2. Lexicon

The lexical and surface representations of inflected word-forms are encoded in the lexicon as shown in Figure 2. A colon separates the lexical and surface representations. All lexicon entries are followed by either the name of a continuation class (a sub-lexicon) or by a '#' symbol which signifies the end of the string (this is represented in Fig 1 as a double circle). Figure 2 shows how (13b) above might be represented in a two-level lexicon.

```
Multichar_Symbols
+Noun +M +F +Com +Gen +Voc
+Sg +Pl +Def +Idf ^Len ^Sync ^Hv

Lexicon Nouns
athair+Noun+M:athair  Common-Pl;
...
Lexicon Common-Pl
+Com+Pl+Def:eacha^Sync^Hv   #;
...
```

Figure 2. Lexicon sample.

4.3. Replace Rules

Replace rules are encoded as regular expressions and take the form:

A -> B || Left-Context _ Right-Context

where A is replaced by B if it occurs between the specified left and right contexts. A and B can consist of zero or more symbols. A rule transducer specifies only the constraints necessary for that rule and all other strings pass through unchanged (Jurafsky, 2000).

The following replace rule implements the prefixing of *h* to a vowel-initial stem which contains the *^Hv* mark-up tag.

```
0 -> h || .#. Vowel ?* ^Hv
```

The letter *h* is inserted, (literally zero is replaced by *h*) before a word-initial vowel (`.#.` means the start of the word). The vowel must be followed by zero or more symbols followed by the `^Hv` mark-up tag.

4.4. Noun Lexical Transducer

The noun lexical transducer developed in this implementation contains a number of distinct parts. It consist of lexicons followed by a number of replace rules. Word stems are listed in the lexicons, which are divided into various sub-lexicons (continuation classes) depending on how the stems are inflected for case and number.

Morphological tags are appended to the lexical representation. Suffixes and inflectional mark-up tags relating to stem modifications are appended to the lexicon transducer surface representation. The lexicon is then compiled into a finite-state transducer using *lexc*.

Replace rules relating to particular processes (e.g. initial mutation, harmony checking etc.) are grouped together and compiled into transducers using *xfst*.

Both the lexicon transducers and the replace rule transducers are then composed together in the order shown in Figure 3 below to produce a single transducer for nouns.

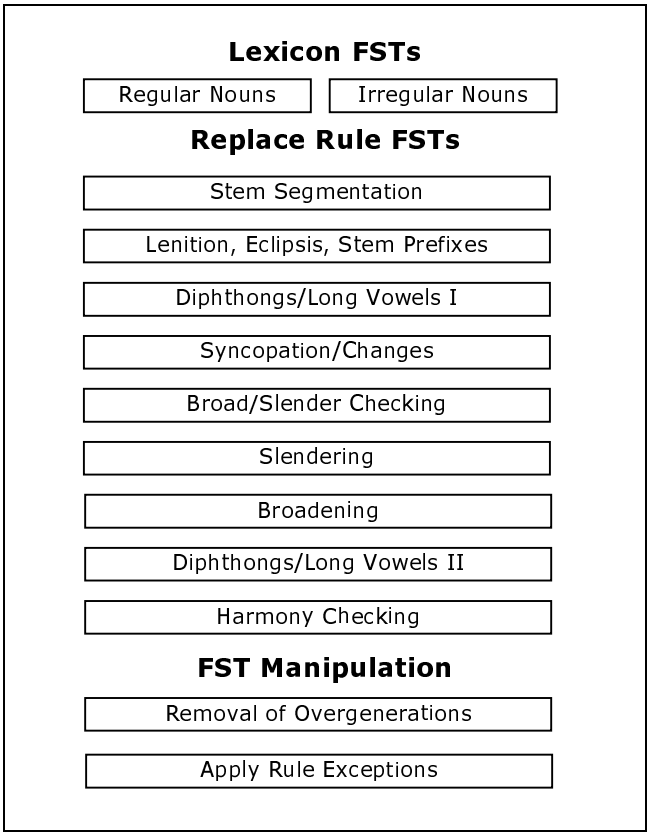

Figure 1. Components of the noun lexical transducer

5. Testing and Evaluation

The *xfst* tool has several useful features that enable rigorous and consistent testing. During development, an old network (surface level) can be subtracted from a new network (surface level) to see which words have been added and alternatively by subtracting a new network from an old network one can see if words have been lost. If this check is performed after each change to the system, any unintentional effects can be quickly spotted and problems can be rectified before continuing (Beesley and Karttunen, 2002 forthcoming).

The upper-level lexicon network can be compared against a lexical tag grammar to ensure that all strings are well-formed. The same type of test can be carried out on the surface network to check that it conforms to the required inflectional mark-up grammar.

The system is designed for broad coverage of the language. This is evaluated by comparing the output against a list of the most frequently used words in a corpus of contemporary Irish texts (Ó Cróinín and Uí Dhonnchadha, 1998). A list of unique words (240,000 approx.) was extracted from a corpus of over 14 million tokens. The list was arranged in order of frequency of use and the 1000 most frequently used types were extracted.

At present the system generates over 11,000 inflected forms from approximately 1000 roots. Initial results show that the system recognises over 80% of the one thousand most frequently used types in the corpus.

Also of interest is the fact that only 53% of the most frequently used words in the corpus of Irish are found among the fifteen thousand (approx.) Irish headwords of An Foclóir Póca [The pocket dictionary] (An Roinn Oideachais, 1986).

6. Further work

Although all of the linguistic phenomena relating to the inflectional morphology of Irish are implemented in the transducer the lexicon itself is currently quite small. An investigation into the semi-automatic extraction of stems from a machine-readable dictionary could prove to be a very efficient means of increasing the coverage of the system.

Issues relating to standard and dialectal forms can also be addressed at the morphological level by including all dialectal variations of word-forms along with the appropriate dialect tag. These tags can then be used to extract only the forms relating to a particular dialect plus all common forms etc. Similar techniques could be used to encode historical forms.

As this system currently deals only with inflectional morphology, work remains to be carried out in the area of derivational morphology.

7. Conclusion

Two-level morphology is proving to be very well suited to Irish morphology. A major advantage of finite-state two-level implementations of morphology is their inherent bi-directionality; the same system is used for both analysis and generation of word forms in the language.

A very useful set of lexical tools have been developed by Xerox which reduce the time involved in developing a lexical transducer by allowing the user to concentrate on stating morphological rules rather than programming finite-state compilers and manipulation programs.

8. References

An Roinn Oideachais, 1986. *Foclóir Póca* [Pocket Dictionary] *English-Irish Irish-English*. Baile Átha Cliath: An Gúm.

Beesley, K. and Karttunen, L., 2002 (forthcoming). *Finite State Morphology: Xerox Tools and Techniques.* Cambridge University Press.

Bunreacht na hÉireann (Constitution of Ireland), 1937. Baile Átha Cliath: Oifig an tSoláthar.

Campbell, G.L., 2000. Irish. *In: Compendium of the Worlds Languages.* vol 1. 2nd ed. Routledge.

Central Statistics Office, 1998. *Census 1996, Volume 9, Irish language*. Dublin: Stationary Office.

Grefenstette, G., Schiller, A. and Aït-Mokhtar, S., 2000. Recognising lexical patterns in text. *In*: F. Van Eynde and D. Gibbon (eds.), *Lexicon Development for Speech and Language Processing* Kluwer Academic Publishers. 141-168.

Jurafsky, D. and Martin, J.H., 2000. *Speech and Language Processing.* Upper Saddle River, N.J.: Prentice Hall.
Koskenniemi, K., 1983. Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. Ph.D. thesis, University of Helsinki.
Karttunen, L., 1994. Constructing lexical transducers. *In: COLING-94*, Kyoto, Japan.
Koskenniemi, K 1997. Representations and Finite-State Components in Natural Language. *In:* E. Roche and Y. Schabes (eds.), *Finite-State Language Processing*. MIT Press. 99-116.
Na Bráithre Críostaí, 1999. *Graiméar Gaeilge na mBráithre Críostaí.* [The Christian Brothers' Irish Grammar]. (2nd ed). Baile Átha Cliath: An Gúm.
Ó Cróinín, D. and Uí Dhonnchadha, E., 1998. The LE-PAROLE project and The National Corpus of Irish. *In Proceedings: Workshop on Language Resources for Minority European Languages (LREC)*. Granada, Spain.
Ó Cuív, B., 1987. Sandhi phenomena in Irish. *In:* H. Andersen (ed.). *Sandhi Phenomena in the Languages of Europe*. Mouton de Gruyter, 395-414.
Ó Siadhail, M., 1989. *Modern Irish.* Cambridge University Press.
Ruhlen, M., 1987. *A Guide to the World's Languages.* Volume 1: Classification. Stanford University Press.
Stenson, N., 1981. *Studies in Irish Syntax*. Tübingen: Gunter Narr.

9. Acknowledgements

The author wishes to thank Foras na Gaeilge for its financial support and Kenneth Beesley, Xerox Research Centre Europe whose generous advice and assistance is greatly appreciated.