

# Quizzes on Tap: Exporting a Test Generation System from one Less-Resourced Language to Another

Montse Maritxalar<sup>1</sup>, Elaine Ui Donnchadha<sup>2</sup>, Jennifer Foster<sup>3</sup>, Monica Ward<sup>3</sup>

University of the Basque Country, Spain<sup>1</sup>; Trinity College, Dublin, Ireland<sup>2</sup>; Dublin City University, Dublin, Ireland<sup>3</sup>  
{montse.maritxalar}@ehu.es; {uidhonne}@tcd.ie; {jfoster, mward}@computing.dcu.ie

## Abstract

It is difficult to develop and deploy Language Technology and applications for minority languages for many reasons. These include the lack of Natural Language Processing (NLP) resources for the language, a scarcity of NLP researchers who speak the language and the communication gap between teachers in the classroom and researchers working in universities and other centres of research. One approach to overcoming these obstacles is for researchers interested in Less-Resourced Languages (LRLs) to work together in reusing and adapting existing resources where possible. This article outlines how a multiple-choice quiz generator for Basque was adapted for Irish. The Quizzes on Tap (QOT) system uses Latent Semantic Analysis (LSA) to automatically generate multiple choice test items. Adapting the Basque application to work for Irish involved the sourcing of suitable Irish corpora and a morphological engine for Irish, as well as the compilation of a development set. Various integration issues arising from differences between Basque and Irish needed to be dealt with. The QOT system provides a useful resource that enables Irish teachers to produce both domain-specific and general-knowledge quizzes in a timely manner, for children with varying levels of exposure to the language.

**Keywords:** LRL, less-resourced languages, Irish, morphological analysis, multiple choice test item generation, educational applications

## 1. Introduction

Like many Less Resources Languages (LRLs), there are few Natural Language Processing (NLP) resources available for the Irish language, and in common with many LRLs, this is due to the limited number of researchers working in the language and the difficulty of finding the multi-skilled experts required to develop these resources. What few resources are available can be difficult to integrate into an existing application for the same reasons.

One approach to overcoming these obstacles is to adapt an existing NLP application that has been developed for another language. This involves integrating language-specific modules into the application, and is made easier if the original application has been designed in a modular fashion. This paper reports on a project to adapt a Multiple-Choice Quiz (MCQ) generation system, originally developed for Basque and subsequently adapted to English, to the Irish language, in order to be able to generate both domain-specific and general-knowledge multiple-choice quizzes for use in an Irish classroom environment.

## 2. Background

The original MCQ generation system, termed *Arikiturri*, was developed for the Basque language (Aldabe et al. 2006). The Basque term “Arikiturri” is an acronym of “Ariketen Iturria” where Iturria means “fountain” and Ariketen means “of exercises”. Therefore, the name refers to a system which is able to automatically generate tests from texts, to be included in testing tasks. In order to generate good quality items, Arikiturri makes use of two kinds of language resources: Natural Language Processing (NLP) tools and linguistic information. It is a system with an open architecture that allows the

integration of new tools and resources. In fact, the two main characteristics of the system are multilingualism and modularity. Arikiturri has been used to generate Basque MCQs (Aldabe et al., 2006) embedded in a whole text and English isolated MCQs (Aldabe et al., 2009). Almost all research on the automatic generation of MCQs based on NLP methods has focused on English and on isolated MCQs. In this article we present a system that automatically generates isolated MCQs for the Irish language.

Multiple-choice items consist of a *stem* and a set of options. The stem is the part of the item that presents the item as a problem to be solved, that is, a question or an incomplete statement. In this work, we do not generate questions but rather incomplete statements. The options are the possible answers that the students can choose from. In our approach, only one of the options will be the correct answer, i.e. the *key* and the other options of the MCQ are those that are incorrect in the given context, i.e. the *distractors*.

Hoshino and Nakagawa (2005) report a machine-learning-based approach for the automatic generation of MCQ questions. Smith et al. (2009) present a complete architecture of a system where the distractors are generated before automatically selecting the sentences. However, in the case of Aldabe et al., (2006), Liu et al. (2005) and Sumita et al. (2005), the sentence-selection task precedes the distractor-generation task. Both approaches represent different ways of contextualising the distractors. In the present work, the selection of the sentences is also a prior task to the automatic generation of the distractors. Furthermore, the process of selecting Irish sentences is carried out manually. Indeed, the automatism of the process is limited to the generation of the distractors.

### 3. The Irish Language

We begin with a brief overview of Irish grammar, followed by a description of existing resources for Irish and a discussion of the teaching of Irish.

#### 3.1. Grammatical Properties of Irish

Irish is a member of the Celtic branch of the Indo-European family of languages. It is an inflectional language with a gender system, and it is a VSO word order language. Verbs can be inflected for tense, mood, number and person. Nouns have either masculine or feminine gender and can be inflected for case, number, definiteness and emphasis/contrast. Adjectives agree with the nouns they modify in terms of gender, case and number. The definite article also agrees with noun in terms of gender, case and number. There is no indefinite article. Prepositions can stand alone or combine with a pronoun (these forms are known as prepositional pronouns), therefore prepositions display the full paradigm of number and person.

The verbal system is quite regular, apart from eleven commonly used irregular verbs. Verb forms which are inflected for person and number are known as synthetic forms. The subject can either be a noun/pronoun or else the verb can be inflected for person and number (a synthetic verb form). However, is ungrammatical to have both at the same time.

- (1) *Chualamar* ‘We heard’ (synthetic form)
- Chuala muid* ‘We heard’ (analytic form)
- Chuala Máire* ‘Máire heard’ (analytic form)

The nominal system is more varied and irregular than the verbal system, in its inflectional system. There are a variety of ways in which the plural can be formed, and in general it is not predictable for the stem. Although a standard form will be listed in the dictionary, in practice there may be dialectal variants.

- (2) *an capall* ‘the horse’
- na capaill* ‘the horses’ (official standard)
- na caiple* ‘the horses’ (dialectal variant)

**Morphosyntax:** Inflectional forms usually require a suffix to the stem, and they frequently require modification of the stem. The modifications are of two distinct categories - initial mutations (lenition, eclipsis, prefixing of vowel-initial and s-initial words), which affect the initial sound/letter of the word, and final mutations (slenderisation, broadening, syncopation, syllable replacement), which affect the final syllable.

Irish morphology and syntax are closely inter-related. Initial mutations to a word are dependent on its syntactic relationship with a preceding word. Many functional words (determiners, prepositions, particles etc.) trigger initial mutations in the following word-form. Examples (3)-(5) show how the definite article, possessive determiners, numerals and simple prepositions trigger various initial mutations in nouns.

- (3) *cos* (fem.) ‘foot’
- an chos* ‘the foot’
- mo chos* ‘my foot’
- ar a cos* ‘on her foot’
- ar a chos* ‘on his foot’
- seacht gcós* ‘seven feet’
- faoin chos* ‘under the foot’
- (4) *arán* (masc.) ‘bread’
- an t-arán* ‘the bread’
- (5) *aerlíne* (fem.) ‘airline’
- an aerlíne* ‘the airline’

Genitive case is used in several constructions including possessive modifiers (6)-(7), after some prepositions (8), and after verbal noun constructions (9).

- (6) *an fear* (masc.) ‘the man’;
- hata an fhir* (masc.), the man’s hat
- (7) *an fhuiseog* (fem.), lark
- sciathán na fuisgeoige* ‘the lark’s wing’
- (8) *an tír* ‘the country’
- timpeall na tíre* ‘around the country’
- (9) *obair (fem.)* ‘work’
- ag déanamh oibre* ‘working’

This means that creating MCQs is not only a matter of finding a word with the relevant part of speech and semantic properties, one must also consider the local morphosyntactic environment.

#### 3.2. Language Technology and Irish

The development of Language Technology for Irish is quite limited. There are some basic resources such as online dictionaries, spelling/grammar checkers, a part-of-speech tagged corpus, and a text-to-speech synthesiser. A new English-Irish Dictionary is under development which will replace the current English-Irish Dictionary (Ó Dónaill, 1978). As regards NLP tools there is a finite-state morphology (Beesley and Karttunen, 2003) part-of-speech tagger (Uí Dhonnchadha and van Genabith, 2006), a partial dependency parser and a chunker (Uí Dhonnchadha and van Genabith, 2010; Uí Dhonnchadha 2010). There are very few commercial products for Irish since is a lesser-used language with little commercial potential for private developers. In order to ensure up-to-date technology for the language it would be necessary to have a government backed strategic plan and regular funding. However such a plan is not in place and funding is sporadic.

#### 3.3. Irish in the Classroom

Irish is one of the two official languages of Ireland, along with English, but only a small percentage of the population speak it as their first language on a daily basis. Irish is a compulsory subject in both primary and secondary schools, but after 13 years the level of attainment is low. There are socio-cultural, pedagogical and linguistic issues behind this discrepancy. People hold negative attitudes towards the language e.g. useless, spoken only by country people (similar to attitudes to other languages in a similar position), and forced upon them by the cultural-elite. Attitudes towards the language

are usually passed down from generation to generation to the present day, where some, but not all, parents question the need or usefulness of studying the language.

There are several pedagogical issues to consider with how Irish is taught in schools. Lack of modern pedagogical techniques (Irish is taught differently from other languages) and an emphasis on grammar and literature which most students do not understand make it hard for the students. Also, as is often the case, the best speakers do not necessarily make the best language teachers.

Furthermore, Irish is linguistically challenging for English speakers, especially as it is their first exposure to another language. Morphological issues (see Section 3.1 above), VSO word order and the orthographical issues with the language (Hickey and Stenson, 2011) make it difficult for the students to be comfortable with learning the language, let alone becoming proficient in it.

With regards to Irish schools, there are three types of schools: English-medium (the vast majority), Irish-medium in a Gaeltacht<sup>1</sup> area (very few) and Irish-medium outside a Gaeltacht area (very few, but increasing in number). As in most modern schools, there is a need for up-to-date pedagogical resources and a place for Computer-Aided Learning (CAL) resources. Outside of Irish language resources for the study of the language itself, there are very limited Irish-medium resources available for other subjects (e.g. biology), apart from textbooks and related materials. This is understandable given the market-size, lack of NLP resources and the difficulty in developing the resources (i.e. the need for a multi-disciplinary team and the difficulties e.g. financial and logistical, in assembling such a team).

However, these difficulties do not negate the need for modern Irish language technology applications in the classroom. Some schools are piloting the use of i-pads rather than textbooks and smartphone language apps. Language Technology applications could make a contribution to schools in Ireland e.g. computer-based materials for Irish or any other subject. Applications such as Quizzes-on-Tap could potentially be of benefit to both teachers and students in every classroom in the country.

## 4. Quizzes on Tap

We begin with a general description of the procedure for generating MCQs, followed by a description of the method that is used by to find distractors. The section ends with a discussion of the difficulties that arise due to morphosyntactic properties of Irish.

### 4.1. General Procedure

As explained previously, the automatization of the MCQ process is limited to the generation of the distractors of the MCQs. The generated items are tested for Irish nouns and verbs since this is the one of the type of tests teachers

---

<sup>1</sup> A Gaeltacht is a region designated as an Irish-speaking part of the country, where the majority of the population use Irish as the language of communication.

conduct in class. We have applied a corpus-based approach for the generation of the items.

The process of generating and analysing the Irish MCQs consists of the following steps:

1. **Selection of the texts:** experts on the generation of didactic resources select the texts, taking into account the level of the learners.
2. **Marking the blanks:** the terms to be considered as keys have to be relevant to the learning process of the students. The marking is carried out manually.
3. **Generation of distractors:** for each stem and key selected in the previous step, distractors are generated.
4. **Discarding incorrect distractors:** experts eliminate any automatically generated distractors that cannot fit the blank.
5. **Evaluating the distractors:** experts evaluate the quality of the remaining generated distractors.

### 4.2. Generating Distractors

Distributional similarity measures are based on the idea that the similarity of two words depends on the commonality of their contexts. Thus, two words are similar if they occur in similar contexts. In our experiments, the system uses context words and LSA (Landauer et al., 2007) a corpus-based approach, to compute the similarity measures. LSA has achieved good results in a number of NLP tasks such as Information Retrieval (Deerwester et al., 1990), evaluation of synonym test questions (Turney, 2001), and in the generation of MCQ questions (Aldabe and Maritxalar, 2010). It has also been applied in educational applications (Landauer et al., 2007). Our system makes use of Infomap software (Dorow and Widdows, 2003), a variant of LSA, to learn vectors representing the meanings of words in a vector-space known as WordSpace<sup>2</sup>. The software indexes the documents in a corpus and performs word to word semantic similarity computations based on the resulting model. As a result, the system extracts the words that best match a query according to the model. The words retrieved by the model are the lemmas of the distractors that will be generated.

The system provides InfoMap with the selected key, and in order to avoid the possibility of students guessing the correct choice by simply discarding options with the wrong morphological characteristics, the system offers as candidate distractors the first words of the output which match the PoS of the key. With the aim of generating the distractors of the MCQs, the morphological analyzer extracts the morphological information of the correct answer (i.e. the key) and attaches it to the lemmas given by Infomap. The morphological generator will produce the distractors, i.e. the inflected wordforms. In the case of Irish, the inflected wordform differs depending on the gender, therefore, the morphological information given by the correct answer (the input to Infomap) sometimes is not enough to generate the distractors corresponding to the lemma given by Infomap. For example, in the case of the word *próitéin* (feminine) one of the lemmas given by

---

<sup>2</sup> We have implemented a preprocessing submodule to enable Infomap handle accented Irish characters.

Infomap is *núicléas* (masculine). In this case the morphological generator is not able to produce the distractor because the gender of both lemmas does not match.

(10)key: *phrotéin* ‘protein’

features: próitéin+Noun+Fem+Com+Sg+Len

In order to generate an inflected form of a noun in Irish using the finite-state morphology transducer you need to supply its gender (masc/fem), type (common/proper/substantive/verbal) and number (sing./pl.) features. Only the number feature can be taken from the key as the other two features (gender and type) are specific to the distractor lemma. In order to overcome this problem we would need to look up (in a table) the gender and noun type of distractor lemmas before generating the required inflected forms.

(11)key: *núicléas* ‘nucleus’

features: núicléas +Noun+Masc+Com+Sg

Currently we use only distractors with the same gender and type as the key, which obviously limits our choice of distractor.

In addition, if the key noun is in the genitive case, the preceding definite article will vary depending on the gender of the noun. Ideally we should mark both the article and the noun and supply both article and noun for distractors in the genitive case.

(12)méid *na protéine* ‘amount of the protein’

(13)méid *an núicléas* ‘size of the nucleus’

We encounter similar problems in relation to verb features. The features tense, mood, person and number apply to all verbs, however transitivity is specific to the lemma and therefore the transitivity of the key will not apply to all verbs.

(14)key: *braitheann* ‘depends’

features: braith+Verb+VTI+PresInd

The key verb form *braitheann* ‘depends’ has the features Verb and PresInd (tense: present indicative) which are necessary for the generation of the distractor verb forms. The transitivity feature VTI (i.e. verb can be both transitive and intransitive) is specific to the key lemma. However, it makes sense to limit our distractor verbs to those with the same transitivity feature as the key, in order to fit in with the structure of the sentence.

Before creating the language models used to generate the distractors, the morphological analyser is run over the input corpus. In addition some post-processing has been implemented to detect the information necessary for the input of Infomap.

## 5. Preliminary Evaluation

In order to carry out a proof-of-concept evaluation of the QOT system, we created two LSA models using a general Irish language corpus of 12 million words and a smaller scientific corpus of almost 300,000 words. The smaller scientific corpus was used as a development corpus

during the initial system development. The larger corpus was used to create the LSA model used in our evaluation. Nouns, adjectives, verbs and adverbs were used to create the models.

Our test set consists of 21 MCQ stem/key pairs. 10 of these are on the topic of biology and are suitable for teenagers attending an Irish-speaking secondary school. The remaining 11 are general-knowledge items, aimed at teenagers attending both Irish-speaking and English-speaking secondary schools. We examine the distractors produced by the QOT system for each of these 21 items.

The QOT system produced distractors for 15 of the 21 input items. 7 of these 15 required no editing whatsoever, i.e. all of the generated distractors were suitable. An example is the following (the key is underlined and the distractors are listed below):

*Bionn teocht áirithe agus pH áirithe ag teastáil le go n-oibreodh na heinsimí ar fad i gceart.*

‘For the enzymes to work properly particular temperature and pH levels are required’.

1)*frasaíocht* ‘precipitation’

2)*ocsaíd* ‘oxide’

3)*radaíocht* ‘radiation’

4)*báisteach* ‘rainfall’

5)*ocsaigin* ‘oxygen’

6)*aeráid* ‘climate’

7)*nítrigin* ‘nitrogen’

The remaining 8 items were associated with distractors of varying quality. The unsuitable distractors fell into two categories:

### A distractor that is semantically too similar to the key.

An example is the following:

*Is deacair sútha talún na hÉireann a fháil i gcaitheamh an gheimhridh.*

‘Irish strawberries are hard to find in winter’.

1)*shamhraidh* ‘summer’

2)*shneachta* ‘snow’

3)*sheaca* ‘frost’

4)*gearraigh* ‘spring’

5)*ghriain* ‘sun’

6)*theasa* ‘heat’

Here, distractors 2) and 3) (*snow* and *frost*) are unsuitable as distractors because they make sense in the context of the stem.

### A distractor that is grammatically incorrect.

An example is the following:

*Chuir a mháthair pláta bia ar an mbord.*

‘His mother put a plate of food on the table’.

1)*glasra* ‘vegetable’

2)*ithe* ‘eat’

3)*arán* ‘bread’

4)*béile* ‘meal’

5)*ocras* ‘hunger’

6)*bainne* ‘milk’

7) *éasc* ‘a flaw in timber’  
8) *tobac* ‘tobacco’

In this example, distractors 1) and 3) (*vegetable* and *bread*) are unsuitable not only because they are too similar in meaning to the key (as in the previous example) but also because they are not correctly inflected for the genitive case. Distractor 7) is an interesting example since it is similar in spelling to the Irish word for *fish* (*éisc*), but semantically quite unrelated. As such, it makes a good distractor.

The first problematic case illustrates the difficult nature of the task itself – it is a non-trivial problem to generate good distractors which can be reliably used to test a person’s knowledge of a particular subject area. The second problematic case is related to the integration of the Irish morphological analyser/generator into the system – there remains quite a bit of work to be done here to guarantee morphologically well-formed distractors. The generation of ill-formed distractors does however have the potential to be of use in applications which aim to test grammatical knowledge.

In general, the distractors produced for the scientific sentences were of higher quality than those produced for the general sentences.

## 6. Conclusion

We have described the process of adapting a multiple-choice-question generation system, developed initially for Basque, to the Irish language. Although the underlying system is, in theory, language-independent, in practice, much work still needs to be done to port it from one language to another. The resulting system, Quizzes on Tap, relies on a morphological analyser/generator for Irish, and aims to take into account the complexities of Irish morphosyntax. Although the system development is still in its early stages, the results of our initial evaluation are encouraging enough to suggest that the system has the potential to be useful to teachers.

## References

Aldabe I., Lopez de Lacalle M., Maritxalar M., Martinez E., Uria L. (2006). ArikIturri: An Automatic Question Generator Based on Corpora and NLP Techniques In: M. Ikeda, K. Ashley, and T.-W. Chan (Eds.), *The 8th International Conference on Intelligent Tutoring Systems* (Book Series: Lecture Notes in Computer Science), Volume 4053, Springer-Verlag, pp. 584-594.

Aldabe I., Maritxalar M., and Mitkov R., (2009). A Study on the Automatic Selection of Candidate Sentences and Distractors. In: Dimitrova, V., Mizoguchi, R., Boulay, B., Graesser, A. (Eds.). *Artificial Intelligence in Education. Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*. (Book Series: Frontiers in

Artificial Intelligence and Applications), Volume 200, IOS press, pp. 656-658

Aldabe I., Maritxalar M. (2010). Automatic Distractor Generation for Domain Specific Texts. In: Loftsson, H., Gñvaldsson, E. R, Helgadóttir S. (Eds.). *The 7th International Conference on NLP, IceTAL* (Book Series: Advances in Natural Language Processing: Lecture Notes in Computer Science), Volume 6233, Springer-Verlag, pp. 27-38.

Chao-Lin L., Chun-Hung W., Zhao-Ming G and Shang-Ming H. (2005): Applications of Lexical Information for Algorithmically Composing Multiple choice Cloze Items. In *Proceedings of the 2nd Workshop on Innovative Uses of NLP for Building Educational Applications*, Ann Arbor, ACL, pp. 1-8.

Deerwester S., Dumais S., Furnas G., Landauer T. and Harshman R. (1990): Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, Volume 41(6), pp. 391-407.

Dorow B. and Widdows D. (2003): Discovering corpus-specific word senses. In Proceedings of EACL.

Hickey, T., and Stenson, N., 2011. Irish Orthography: what teachers and learners need to know about it, any why? *Language, Learning and Technology*, Vol., 24, Issue 1, 2011.

Hoshino, A. and Nakagawa, H. (2005). A real-time multiple-choice question generation for language testing. In *Proceedings of the 2nd Workshop on Innovative Uses of NLP for Building Educational Applications*, Ann Arbor. ACL, pp. 17-20

Landauer T.K., McNamara D.S., Dennis S. and Kintsch W. (2007): *Handbook of Latent Semantic Analysis*. Mahwah NJ: Lawrence Erlbaum Associates.

Ó Dónaill, N. (1978). *Foclóir Gaeilge-Béarla*, Dublin, An Gúm

Smith, S., Kilgarriff A., Sommers S., Wen-liang G., and Guang-zhong W. (2009). Automatic cloze generation for English proficiency testing. *Proceedings of LTTC conference*.

Sumita, E., Sugaya, F. and Yamamoto, S. (2005). Measuring Non-native Speakers’ Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions. *Proceedings of the 2nd Workshop on Innovative Uses of NLP for Building Educational Applications*, Ann Arbor, ACL, pp. 61-68.

Turney P. (2001): Mining the web for synonyms: Pmi-ir versus lsa on toefl. *Proceedings of the Twelfth European Conference on Machine Learning*, pp. 491-502.

Uí Dhonnchadha, E. (2010) *Natural Language Processing Tools: Developing a Part-of-Speech Tagger and Partial Dependency Parser for Irish*, Saarbrücken, LAP LAMBERT Academic Publishing.

Uí Dhonnchadha, E. and van Genabith, J (2006). Scaling an Irish FST morphology engine for use on unrestricted text, In: Yli-Jyrä, A., Karttunen, L., Karhumäki, J. (Eds.). *Finite-State Methods in Natural Language Processing* (Book Series: Lecture Notes in Artificial Intelligence), Springer-Verlag, pp. 247 – 258.

Uí Dhonnchadha, E and Van Genabith, J. (2010). Partial Dependency Parsing for Irish, *LREC2010: Language Resources and Evaluation Conference, Malta*.