Dictionary System Shell

Diccionario Sistema

Florie Moulin,
Polytech' Montpellier / Ollscoil Luimnigh,
Montpellier / Luimneach,
France / Éire
floriemoulin@gmail.com

Laura Laluque
Polytech' Montpellier / Ollscoil Luimnigh,
Montpellier / Luimneach,
France / Éire
laura_laluque11@hotmail.fr

Gearóid Ó Néill
Ollscoil Luimnigh,
Luimneach
Éire
Gearoid.oneill@ul.ie

Resumen

En los últimos años, se ha trabajado sobre los sistemas de diccionario en la Universidad de Limerick. Actualmente se está trabajando para proporcionar un sistema de diccionario que puede ser utilizado por no-especialista para generar diccionarios de "menos recursos" idiomas.

El sistema propone distintos niveles de funcionalidad, desde el acceso a través de microfichas hasta un completo acceso de texto, pasando por el acceso a los imágenes de un diccionario gracias a los números de página. El acceso al sistema se realiza desde la web.

Palabras clave

Diccionario, sistema de generación, automatización, meta-sistema.

Summary

Over the past few years, there has been work on dictionary systems here at the University of Limerick. Currently work is in progress to provide a dictionary system which can be used by non-computer specialist to generate dictionaries for "less-resourced" languages.

The system provides for different levels of functionality, from microfiche type access through page number access to images of a book dictionary through to fairly comprehensive text access, available on the web.

Keywords

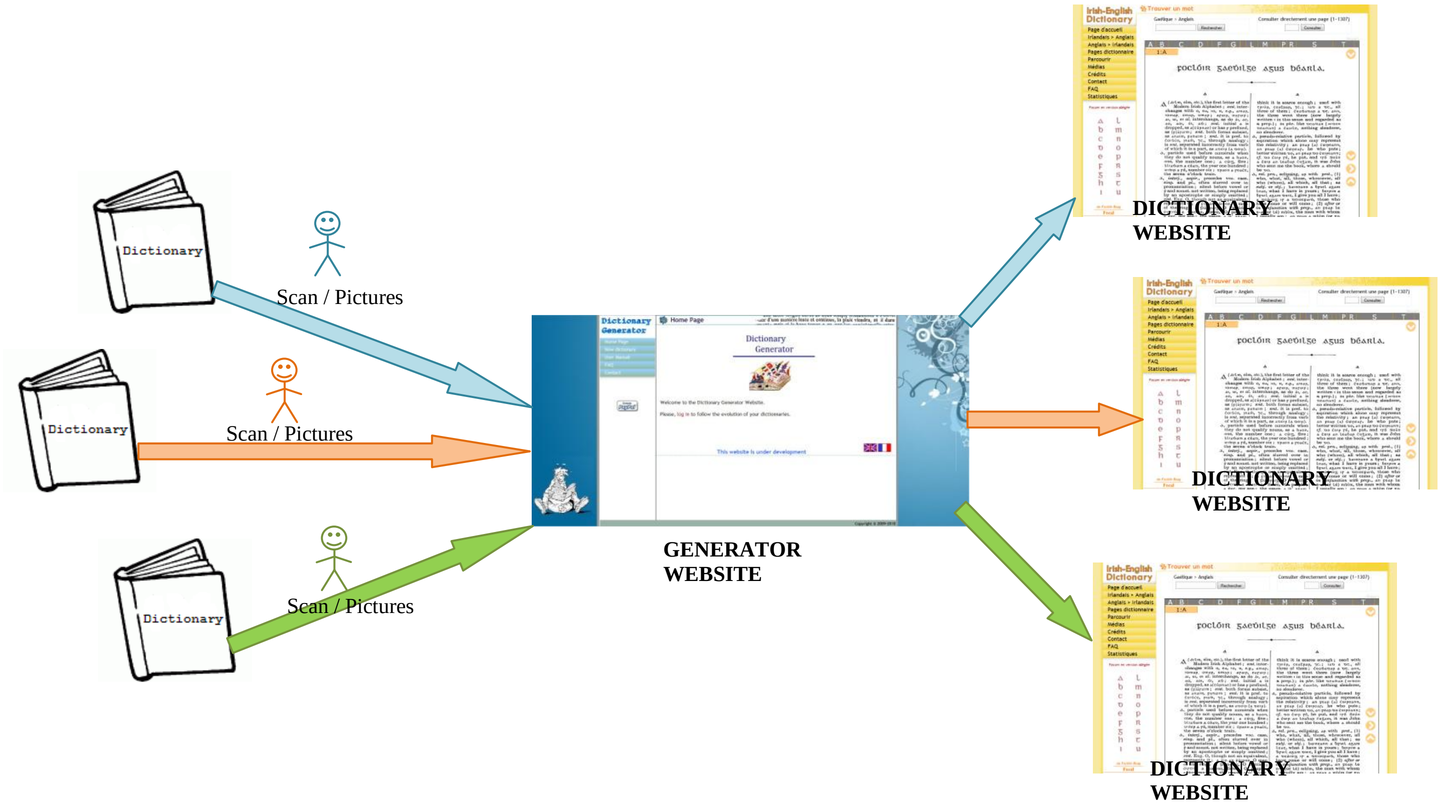
Dictionary, system generation, automation, meta-system

Background

In the early 90s a monolingual "small Irish dictionary" was published by An Gúm. McElligott et al. computerised the dictionary adding the inflected forms of the words.
It was the first mono-lingual Celtic language dictionary to be available on the world wide web. The Dinneen Irish-English dictionary, published in 1927, has been made available on the world wide web (Both are available on suitably web enabled mobile phones, for example, Samsung's Tocco).

The Dinneen dictionary is an Irish-English dictionary. It uses an Irish script for the Irish and a Latin script for the English. Irish and English words appear in the body of an entry.

The system is being adapted with a view to making a "shell" available such that a dictionary system can be set up requiring little computing expertise and be a monolingual or bilingual dictionary.



Dictionary Structure

For our purposes, a dictionary can be regarded as having the structure (<D>|(<N,P>)|(<N,H,P>)|(<K,N,H,M>)|(<K,A ,N,H,M>),

where <D> is an image of the whole dictionary;

where <N,P> is an ordered pair, for example, the page number and the page image;

where <N,H,P> is a triple, for example, N is the page number, H is the first headword on the page and P is the page image;

where (<K,N,H,M>) | (<K, A,N,H,M>) are ordered set of entries, described below.

K is a system generated key. N is the page number where the word occurs. H is a 'headword' structure and M is an 'information' structure.

K is generated to allow for (generous) increases in entries. Although once established, dictionaries tend to change rather slowly, the need for the generous allowance of keys is that it is expected, at least initially, there could be copious errors in an OCR phase. This, from our experience with the Dinneen dictionary, leads to entries being subsequently split or joined.

H is a minimalist 'headword' structure, namely, ((((W|S)(B1*))+)B2*)*

where each W, S is separated from another by B1 ‘a basic common’ separator, for example, a space and B2 an optional headword structure separator.

W is a string of the language (word) which can stand by itself. S is a substring of the language, possibly with a non-letter symbol of the language attached but which – ‘ordinarly’ – cannot stand by itself.

The headword can be null.

M purports to give information about the headword (if there is any). M itself has a structure, namely,

(I|F|E)*

where I is ‘metadata’, F is ‘definittion’ and E ‘examples’. Each piece of data within a structure may use separators. The information may be in any (identifiable) order. Grammar information, for example, could be included in I.

The meaning may be explicitly null.

<A>

The information can be ‘augmented’, with extra tables (see below), including multi-media data. This includes the facility to search by images, for example, there is a picture of a dog and on clicking on the dog, the entry for a dog appears.

Information Retrieval

From a dictionary can be generated a spell-checker. The spell-checker would then facilitate both the user in giving him or her more confidence and by improving the chances of getting results from a query.

The generation of the dictionary could be done in conjunction with a web browser, both to find putative headwords and examples of use and verification (probabilistic) of grammar.

Translation

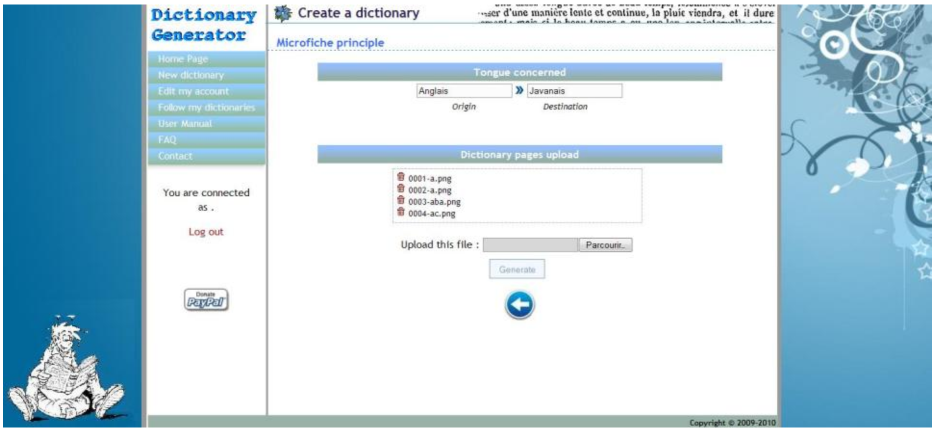
One of the authors has an interest in automatic translation (see Ó Néill) but in this talk we are mainly concerned with the dictionary structure and one particular practical application, namely the facilitation of the computerisation of dictionaries.

Page Images

The “lesser-resourced” languages vary greatly in the state of the language in relation to published material.

<D>

For those languages with printed dictionaries, one way to make the dictionary available is to scan or photograph the pages of the dictionary. The dictionary would be simply accessible through the system, rather like the microfiche systems of old.



Create a dictionary

Microfiche principle

Tongue concerned

Anglais » Javanais

Origin Destination

Dictionary pages upload

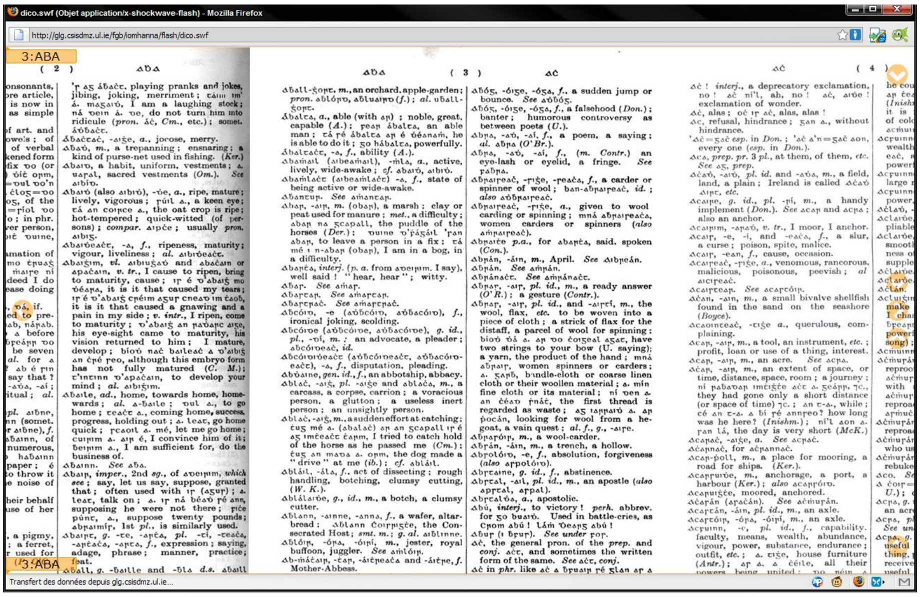
0001-a.png
0002-a.png
0003-aba.png
0004-ac.png

Upload this file :

You are connected as .

Copyright © 2009-2010

Figure 1 : User upload of dictionary pages


dico.swf (Objet application/x-shockwave-flash) - Mozilla Firefox

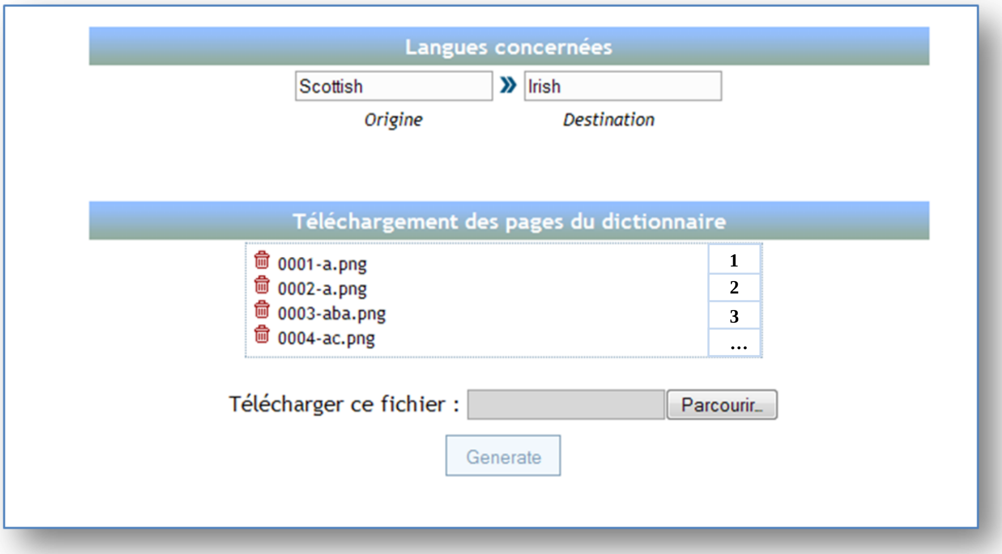
http://glg.csisdmz.ul.ie/fgb/iomhanna/flash/dico.swf

Transfert des données depuis glg.csisdmz.ul.ie...

Figure 2 : Example of microfiche system (Dineen Dictionary)

<N,P>

The next stage up is to provide page numbers for the scanned pages. As well as providing the "microfiche" style of access, it also allows for access by page.


Langues concernées

Scottish » Irish

Origine Destination

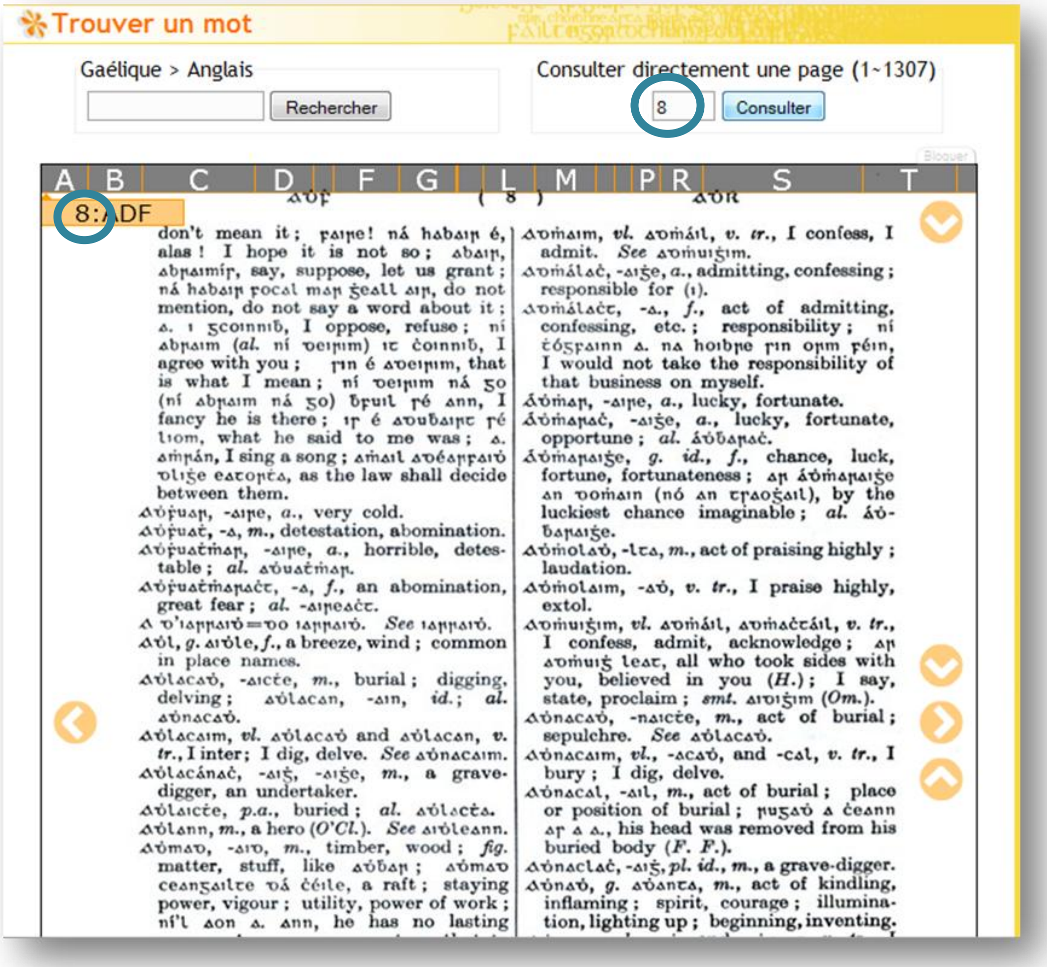
Téléchargement des pages du dictionnaire

0001-a.png	1
0002-a.png	2
0003-aba.png	3
0004-ac.png	…

Télécharger ce fichier : Parcourir...

Generate

Figure 3 : User upload of pages and page numbers association



aḋf (8) aḋn

don't mean it; faire! ná habair é, alas! I hope it is not so; abair, abraimís, say, suppose, let us grant; ná habair focal mar ġeall air, do not mention, do not say a word about it; a. i gcoinniḃ, I oppose, refuse; ní abraim (al. ní deirim) it ċoinniḃ, I agree with you; sin é adeirim, that is what I mean; ní deirim ná go (ní abraim ná go) ḃfuil sé ann, I fancy he is there; is é aduḃairt sé liom, what he said to me was; a. aṁrán, I sing a song; aṁail adéarfaiḋ dliġe eatorṫa, as the law shall decide between them.

aḋḟuar, -aire, *a.*, very cold.

aḋḟuaṫ, -a, *m.*, detestation, abomination.

aḋḟuaṫṁar, -aire, *a.*, horrible, detestable; *al.* aḋuaṫṁar.

aḋḟuaṫṁaraċt, -a, *f.*, an abomination, great fear; *al.* -aireaċt.

a d'iarraiḋ=do iarraiḋ. *See* iarraiḋ.

aḋl, *g.* aiḋle, *f.*, a breeze, wind; common in place names.

aḋlacaḋ, -aicṫe, *m.*, burial; digging, delving; aḋlacan, -ain, *id.*; *al.* aḋnacaḋ.

aḋlacaim, *vl.* aḋlacaḋ and aḋlacan, *v. tr.*, I inter; I dig, delve. *See* aḋnacaim.

aḋlacánaċ, -aiġ, -aiġe, *m.*, a grave-digger, an undertaker.

aḋlaicṫe, *p.a.*, buried; *al.* aḋlacṫa.

aḋlann, *m.*, a hero (*O'Cl.*). *See* aiḋleann.

aḋmad, -aid, *m.*, timber, wood; *fig.* matter, stuff, like aḋḃar; aḋmad ceangailte dá ċéile, a raft; staying power, vigour; utility, power of work; ní'l aon a. ann, he has no lasting

adṁaim, *vl.* adṁáil, *v. tr.*, I confess, I admit. *See* adṁuiġim.

adṁálaċ, -aiġe, *a.*, admitting, confessing; responsible for (i).

adṁálaċt, -a., *f.*, act of admitting, confessing, etc.; responsibility; ní ṫógfainn a. na hoibre sin orm féin, I would not take the responsibility of that business on myself.

áḋṁar, -aire, *a.*, lucky, fortunate.

áḋṁaraċ, -aiġe, *a.*, lucky, fortunate, opportune; *al.* áḋḃaraċ.

áḋṁaraiġe, *g. id.*, *f.*, chance, luck, fortune, fortunateness; ar áḋṁaraiġe an doṁain (nó an tsaoġail), by the luckiest chance imaginable; *al.* áḋḃaraiġe.

aḋṁolaḋ, -lta, *m.*, act of praising highly; laudation.

aḋṁolaim, -aḋ, *v. tr.*, I praise highly, extol.

adṁuiġim, *vl.* adṁáil, adṁaċtáil, *v. tr.*, I confess, admit, acknowledge; ar adṁuiġ leat, all who took sides with you, believed in you (*H.*); I say, state, proclaim; *smt.* aidiġim (*Om.*).

aḋnacaḋ, -naicṫe, *m.*, act of burial; sepulchre. *See* aḋlacaḋ.

aḋnacaim, *vl.*, -acaḋ, and -cal, *v. tr.*, I bury; I dig, delve.

aḋnacal, -ail, *m.*, act of burial; place or position of burial; rugaḋ a ċeann ar a a., his head was removed from his buried body (*F. F.*).

aḋnaclaċ, -aiġ, *pl. id.*, *m.*, a grave-digger.

aḋnaḋ, *g.* aḋanta, *m.*, act of kindling, inflaming; spirit, courage; illumination, lighting up; beginning, inventing.

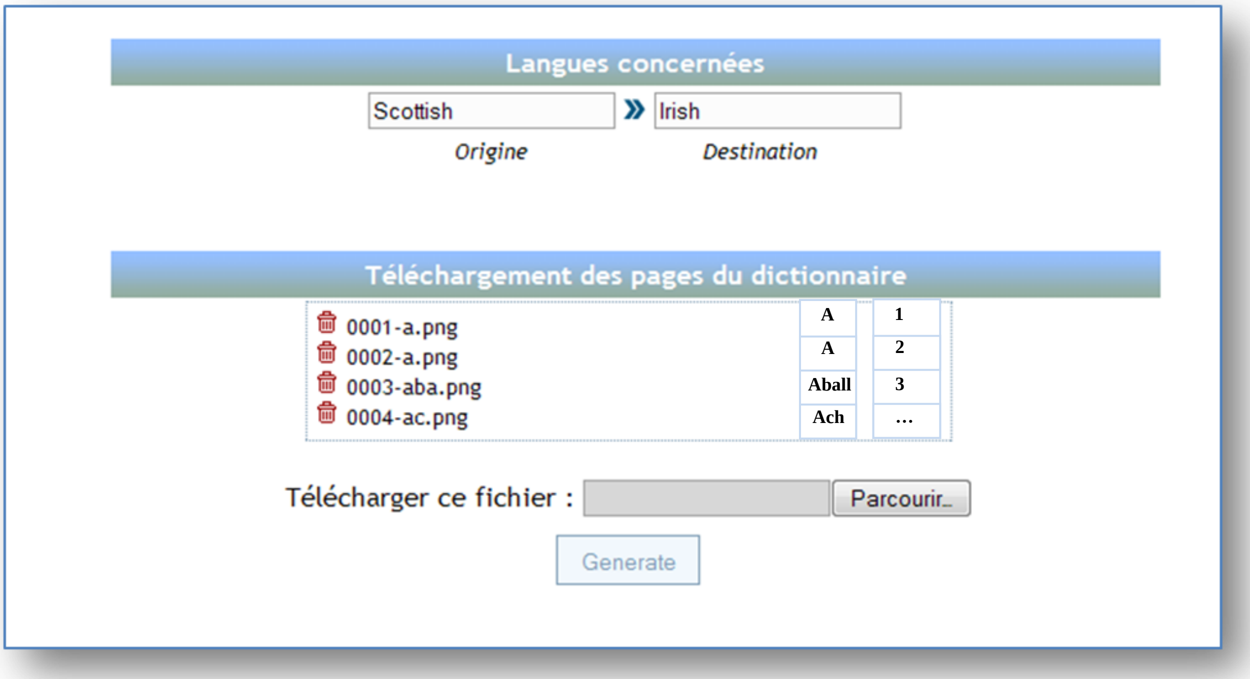
Figure 4 : Example of page access (Dineen Dictionary)

<N,H,P>

The third level of access for the scanned image is by word. This is achieved by specifying the first headword on each page, along with the page number. It requires a lexicographical (alphabetic type) ordering. When presented with a word, it finds the page within which the word would occur. It is then up to the reader to search the page visually to locate the word or determine that the word is not in the dictionary.

The above levels of dictionary usage is relatively easily and cheaply achieved – it does not require optical character recognition. It can be achieved by someone with no expertise in computing. It does require the user to be able to scan or photograph the pages and to enter them into a computer.

The user of the dictionary can search by substrings, so exact spelling is not a requirement.



Langues concernées

Scottish » Irish

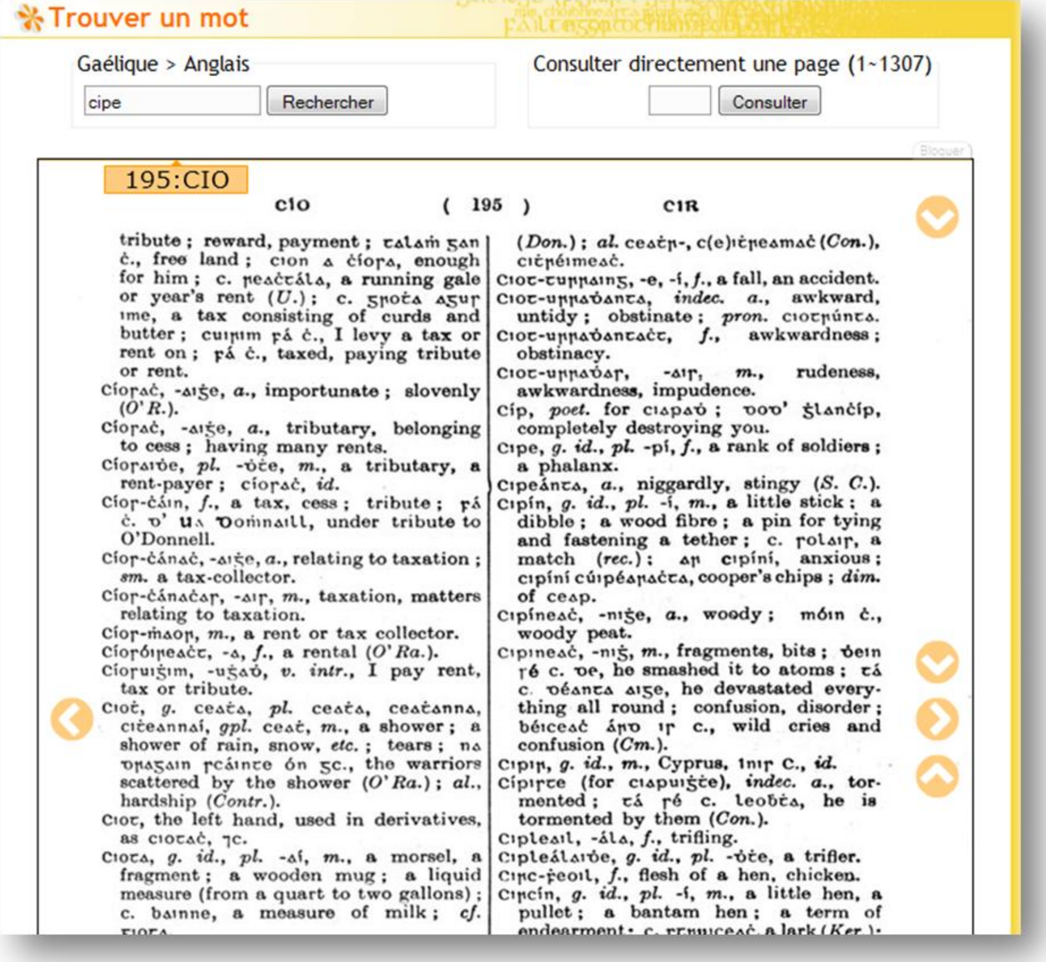
Origine Destination

Téléchargement des pages du dictionnaire

0001-a.png	A	1
0002-a.png	A	2
0003-aba.png	Aball	3
0004-ac.png	Ach	…

Télécharger ce fichier :

Figure 5 : User upload of pages, page numbers and headwords association



Trouver un mot

Gaélique > Anglais

cipe

Consulter directement une page (1~1307)

195:CIO

cio (195) cir

tribute; reward, payment; talaṁ gan ċ., free land; cion a ċíosa, enough for him; c. reaċtála, a running gale or year's rent (*U.*); c. groṫa agus ime, a tax consisting of curds and butter; cuirim fá ċ., I levy a tax or rent on; fá ċ., taxed, paying tribute or rent.

Cíosaċ, -aiġe, *a.*, importunate; slovenly (*O'R.*).

Cíosaċ, -aiġe, *a.*, tributary, belonging to cess; having many rents.

Cíosaiḋe, *pl.* -ḋṫe, *m.*, a tributary, a rent-payer; cíosaċ, *id.*

Cíos-ċáin, *f.*, a tax, cess; tribute; fá ċ. d' Ua Doṁnaill, under tribute to O'Donnell.

Cíos-ċánaċ, -aiġe, *a.*, relating to taxation; *sm.* a tax-collector.

Cíos-ċánaċas, -ais, *m.*, taxation, matters relating to taxation.

Cíos-ṁaor, *m.*, a rent or tax collector.

Cíosóireaċt, -a, *f.*, a rental (*O'Ra.*).

Cíosuiġim, -uġaḋ, *v. intr.*, I pay rent, tax or tribute.

Cioṫ, *g.* ceaṫa, *pl.* ceaṫa, ceaṫanna, ciṫeannaí, *gpl.* ceaṫ, *m.*, a shower; a shower of rain, snow, *etc.*; tears; na dragain scáinte ón gc., the warriors scattered by the shower (*O'Ra.*); *al.*, hardship (*Contr.*).

Ciot, the left hand, used in derivatives, as ciotaċ, ⁊c.

Ciota, *g. id.*, *pl.* -aí, *m.*, a morsel, a fragment; a wooden mug; a liquid measure (from a quart to two gallons); c. bainne, a measure of milk; *cf.* ciota.

(*Don.*); *al.* ceaṫr-, c(e)iṫreamaċ (*Con.*), ciṫréimeaċ.

Ciot-tuirling, -e, -í, *f.*, a fall, an accident.

Ciot-urradanta, *indec. a.*, awkward, untidy; obstinate; *pron.* ciotrúnta.

Ciot-urradantaċt, *f.*, awkwardness; obstinacy.

Ciot-urradas, -ais, *m.*, rudeness, awkwardness, impudence.

Cíp, *poet.* for ciapaḋ; dod' ġlanċíp, completely destroying you.

Cipe, *g. id.*, *pl.* -pí, *f.*, a rank of soldiers; a phalanx.

Cipeánta, *a.*, niggardly, stingy (*S. C.*).

Cipín, *g. id.*, *pl.* -í, *m.*, a little stick; a dibble; a wood fibre; a pin for tying and fastening a tether; c. solais, a match (*rec.*); ar cipíní, anxious; cipíní cúipéaraċta, cooper's chips; *dim.* of ceap.

Cipíneaċ, -niġe, *a.*, woody; móin ċ., woody peat.

Cipineaċ, -niġ, *m.*, fragments, bits; ḋein sé c. de, he smashed it to atoms; tá c. déanta aige, he devastated everything all round; confusion, disorder; béiceaċ árd is c., wild cries and confusion (*Cm.*).

Cipir, *g. id.*, *m.*, Cyprus, inis C., *id.*

Cípiste (for ciapuiġṫe), *indec. a.*, tormented; tá sé c. leoḃṫa, he is tormented by them (*Con.*).

Cipleail, -ála, *f.*, trifling.

Cipleálaiḋe, *g. id.*, *pl.* -ḋṫe, a trifler.

Circ-ḟeoil, *f.*, flesh of a hen, chicken.

Circín, *g. id.*, *pl.* -í, *m.*, a little hen, a pullet; a bantam hen; a term of endearment; c. [illegible] a lark (*Ker.*);

Figure 6 : Example of word access (Dineen Dictionary)

Text Entries

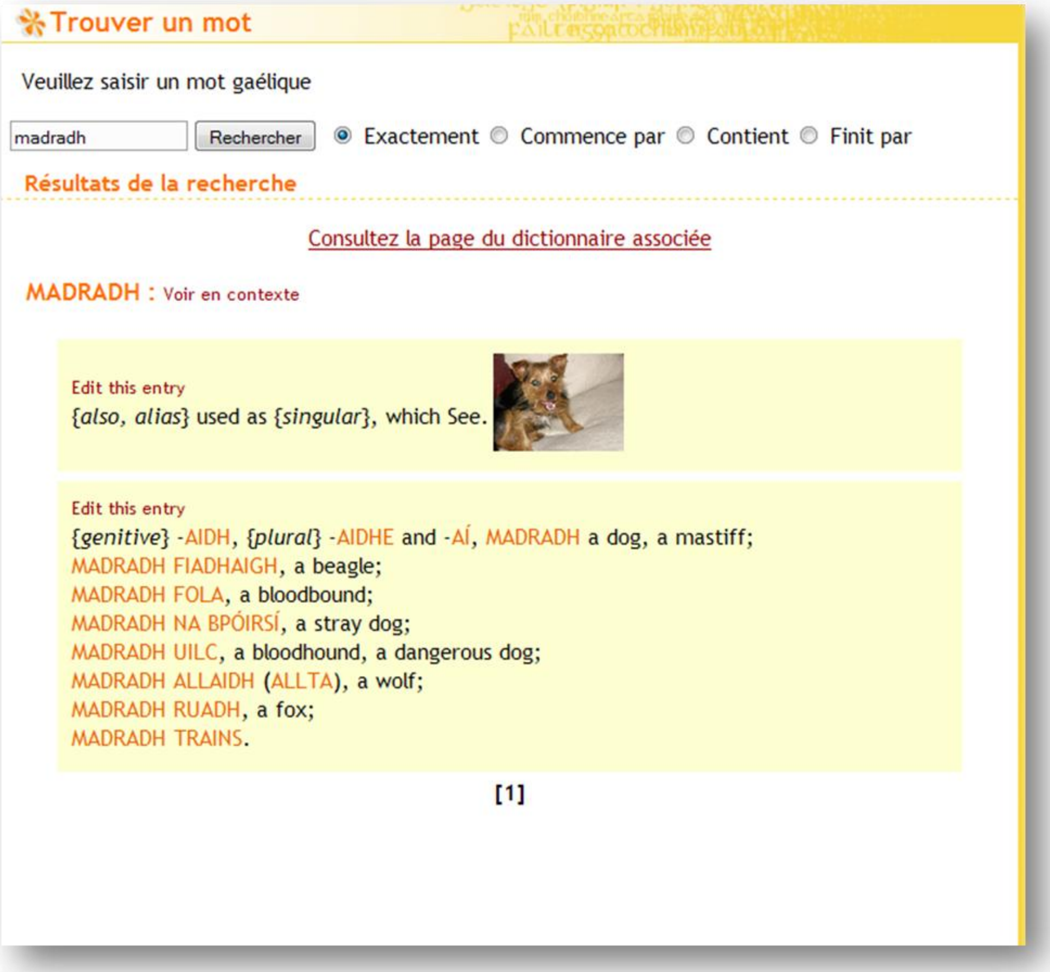
<K,N,H,M>
The next processing level up – perhaps ironically - is searching by text. To search by text requires either the dictionary to be typed in, collected electronically or to use optical character recognition (OCR). OCR works fairly well, especially on "well-known" alphabets and, of course, "well-resourced" languages. Although we are concerned here with "less-resourced" languages, the dictionary might well be a bilingual dictionary into or from a "well-resourced" language. If the lesser-resourced language has a distinctive script, wich is not available directly in the OCR software, then there is a likelihood of increased errors.

The system can take text entries, ranging from untagged to extensively tagged. It can also be related to the scanned images, allowing for results from comprehensive text searches to be in the form of the original scanned image.

This latter feature pre-empts, to some extent, disputes about an entry - relative to the source - being correct or not. It can also facilitate the use of a cherished script.

The system also allows for "reverse" look-up. The entry can be scanned for words or substrings and the context, as well as the headword, shown.

The user is prompted for the type of information to be presented. At the level of the scanned pages, the file name of the images. The default is the text used to name the images, a suitable extension and then the system will expect the pages to be in lexicographic order. The user is free to specify otherwise.



Trouver un mot

Veuillez saisir un mot gaélique

madradh [Rechercher] ● Exactement ○ Commence par ○ Contient ○ Finit par

Résultats de la recherche

Consultez la page du dictionnaire associée

MADRADH : Voir en contexte

Edit this entry
{*also, alias*} used as {*singular*}, which See.

Edit this entry
{*genitive*} -AIDH, {*plural*} -AIDHE and -AÍ, MADRADH a dog, a mastiff;
MADRADH FIADHAIGH, a beagle;
MADRADH FOLA, a bloodbound;
MADRADH NA BPÓIRSÍ, a stray dog;
MADRADH UILC, a bloodhound, a dangerous dog;
MADRADH ALLAIDH (ALLTA), a wolf;
MADRADH RUADH, a fox;
MADRADH TRAINS.

[1]

Figure 7 : Example of OCR use (Dinneen Dictionary)

Meta-tables

The arrangement of information for storing in a database is often done by arranging data into tables. The system allows for the user to specify a particular feature of a language to be incorporated into the actual dictionary, explicitly. This facility is also used in the Irish-English dictionary to add other facilities. It is currently being used to add an option to search by phonetic representation.

As mentioned earlier, the dictionary may be searched using pictures.

An Aid

A tool which will be supplied along with dictionary is a tool for correcting spelling, using dynamic programming and n-grams. To be used effectively, this tool would require some knowledge of the language or languages in the dictionary.

A quasi-static system, with possible errors cannot be used for automatic self-correction but some degree of error correction can be achieved. If the user has access to (correct) word lists, then quite a lot can be done by way of auto-correction.

In the Dinneen case, many errors were corrected for the English, using a source external to the Dinneen dictionary. The spelling corrector used a dynamic programming approach (see, for example, Levenshtein_distance).

Tagging

The text, if supplied untagged, will be treated as consisting of a headword followed by the body of the entry. Since reverse look-up is possible, no information is lost but information retrieval might be more difficult and less meaningful.

At the other extreme, the text may be fully tagged.

Once created, the system can be edited.

Rules for Distinguishing Languages

If the dictionary is a bilingual dictionary, with languages mixed in the body of the entry, then rules can be provided for the languages or alternatively ‘carefully’ selected samples can be provided. The rules would be of the type single, double or triple letter combinations that occur in one language but not in the other language.

Some System Details

A system prototype was developed in Prolog and then converted to MySQL because variants of SQL are more commonly taught than Prolog.

Test

Our first test will be to set up a system for a monolingual Scottish Gaelic dictionary. (It is hoped to have the first tests completed by the middle of July.)

Free

Most of the functionality described is currently available for the Dinneen Irish-English dictionary and work is in progress to make the system available as a “shell”, from which one can build a particular web-based language dictionary. It will be free to non-profit organisations.

In the first instance the ‘new’ dictionaries, will be made available, free of charge, on a Department of Computer Science server, in the University of Limerick.

Future Work

It would be interesting to develop a facility to allow for entries to be searched for by sound, graphically or by gesture.