

Task A Deliverable: Headword Selection Principles

Task A Deliverable: Headword Selection Principles	1
1. Task details and document plan	1
1.1 Task	1
1.2 Deliverable.....	1
1.3 Due	1
1.4 Document plan.....	1
2. User profile.....	2
2.1 Introduction	2
2.2 A user-profile for the New English-Irish Dictionary (NEID).....	2
2.2.1 Users	2
2.2.2 Uses	3
2.3 Rationale for headword selection	3
3. Standard vocabulary items	5
4. Proper Names	9

1. Task details and document plan

1.1 Task

In consultation with the Client, to agree a user-profile and identify principles for a selection of headwords which will result in a high-quality dictionary as specified.

1.2 Deliverable

Robust selection criteria, fully-tested working methodology, complete description of data sources.

1.3 Due

Month 1

1.4 Document plan

This document consists of three sections:

1. user-profile
2. standard vocabulary items
3. proper names

2. User profile

2.1 Introduction

Dictionary users have high expectations. As Samuel Johnson noted: ‘They that take a dictionary into their hands, have been accustomed to expect from it a solution of almost every difficulty’. Consequently, one of the most disappointing experiences for a user is looking up a word and finding that the dictionary does not include it. If this scenario repeats itself on any regular basis, the dictionary’s assumed reliability and claims to ‘authority’ are seriously undermined.

This presents one of the biggest challenges facing dictionary-makers: no dictionary (not even the 22-volume *OED*) can include *all* the vocabulary used within a speech community, so it follows that decisions about what to include in a dictionary (and what to exclude) will critically affect the dictionary’s reputation and perceived credibility in the marketplace.

The process of selecting vocabulary for the dictionary therefore needs to be informed by clear principles established at the outset. The task (put simply) is to include all those vocabulary items that the users of the dictionary are likely to need to look up. But how far is it possible to establish what these are? The approach we propose is to base inclusion decisions on a combination of:

- a user-profile, which identifies who will use the dictionary and what they will use it for
- corpus-derived data showing the frequency of lexical items and (just as important) the degree to which they are ‘dispersed’ across different text-types

In the current project, where corpus resources are being developed specifically for the dictionary, these two factors are interdependent, in that decisions about which texts to include in the corpus will also reflect the needs of the dictionary user – and this in turn gives the resulting statistical data greater value as a guide to inclusion. (More will be said about this in the forthcoming Corpus Design Principles document.)

2.2 A user-profile for the New English-Irish Dictionary (NEID)

The user-profile affects all aspects of dictionary content and design. Decisions about such issues as the appropriate level of grammatical information, the types of illustrative examples provided, or the metalanguage used for presenting information – all are significantly influenced by our understanding of who the user is and what he/she will use the dictionary for. But the focus here is on the selection of headwords for the dictionary, and the way that the perceived needs of users impact on these.

As for any ‘general’ dictionary, the range of users and uses will be very broad.

2.2.1 Users

language

All users are assumed to be native-speakers of English. While the dictionary may be used in many parts of the English-speaking world, its primary market is Ireland. Consequently, most users will be speakers of the variety of English spoken in Ireland (‘Irish-English’). There is a high degree of

overlap between Irish-English and British-English, but the former has its own well-documented characteristics, including words and structures transferred from Irish, survivals of English usage that are now obsolete in British English, and – especially in the north of the island – borrowings from Lallans. (Hence the importance of the planned Irish-English corpus.) It can also be assumed that there is a high level of shared cultural knowledge, at least among those users resident in Ireland, and this too has implications for the range of items that the dictionary should cover.

age

The dictionary is not aimed at children in first-level education or in the first half of second level. Apart from this, the age-range of users is very broad, from 15-16 upwards. This has implications for inclusion of vocabulary relevant to the younger end of the spectrum, in terms of both educational need and youth culture.

'occupation'

Here again, a wide range of types can be assumed, including students (at high school or university), educators, language specialists (writers, translators etc.), and general 'family' users.

education and skills

With such a varied range of users, it is impossible to target one particular subgroup. Editors must however be aware of which particular type of user needs which particular type of information, and present the data appropriately: for instance, since school students will need to be able to understand valence information, this must be couched in accessible language for them.

2.2.2 Uses

at school and university

for translations, exercises, assignments, dissertations, exam preparation etc.

general 'lay' consultation

spelling, pronunciation, crosswords, dispute-resolution etc.

by writers

this encompasses functions such as technical, pedagogical, informative, official or literary writing and translating

in professional contexts

for use in the workplace, in business, law, medicine, the media, and other fields

2.3 Rationale for headword selection

Our starting point is a basic core of general vocabulary which is 'register-neutral': that is, words and phrases that can be shown to be current in a wide range of text-types. Corpus-derived wordlists and frequency data now make it possible to identify this part of the lexicon with increasingly high levels of reliability. Thus far, the task is relatively straightforward.

Beyond this central core, there is a vast amount of vocabulary that is characteristic of specific environments and unlikely to appear much (or at all) outside of them. These environments reflect factors such as regional variety, diachronicity, style and register, professional and technical domains, and the language of subcultures, and we will refer to them here as *sublanguages*. The user-profile will help us determine the relative importance – for users of the NEID – of each of these sublanguages, while empirical data from the corpora will indicate the relative importance of individual items within the sublanguages.

A few examples will help to illustrate how the user-profile and frequency data interact to inform inclusion decisions

medical vocabulary: bones, organs, surgical procedures

word	BNC frequency	in how many texts	include?
sternum	40	19	Y
triquetral	0	0	N
pancreas	182	40	Y
islets of Langerhans	2	2	N
Caesarian (section)	59	36	Y
thoracoplasty	2	1	N

In addition to the frequency data shown here, we can assume that – even though some of our user-group will expect to be able to find Irish equivalents for medical vocabulary – the most specialized, least frequent terms (which appear only in texts written by specialists for specialists) will be needed only by a tiny minority; and it is reasonable to assume that this category of user will have their own professional reference resources to draw on.

In addition to these general criteria (which would apply to any dictionary with English as its source language), it is vital that we also take account of the specifically Irish base of the user-group. This argues for including items which would not necessarily appear in dictionaries published in other parts of the English-speaking world, such as: Irish mythical figures (*Finn Mac Cool*, *Queen Maeve*), terms from Irish sports such as hurling (*full-forward*), Irish expressions and colloquialisms (*culchie*, *blow-in*), and names or abbreviations of Irish places and institutions (*Lansdowne Road*, *ICTU*).

The electronic version of the dictionary will accommodate an expanded version of the print dictionary wordlist, and this will allow, if desired, a broader coverage of the various types of vocabulary described in this document. The exact nature of the additional material for the electronic dictionary wordlist should be specified in the tenders offered in Phase 2 for the task of developing the electronic dictionary.

Building headword lists is not an exact science but we are confident that the methodology outlined here will enable us to select appropriate non-core items with as much rigour and systematicity as the idiosyncrasies of natural languages allow.

3. Standard vocabulary items

It is possible that the dictionary entry design eventually chosen will handle one or two of these items, such as phrasal verbs, as secondary rather than primary (main-entry) headwords, but for the moment they are shown in the Include list.

Type	Property	Include	Exclude
Word class	part of speech : noun, verb, adjective, adverb, preposition, determiner, particle, interjection	all major parts of speech e.g. <i>table, give, splendid, badly, in, the, up, ouch!</i>	N/A
Lexical form	variant forms:	all corpus-attested, e.g. <i>aluminium</i> and <i>aluminum, judgement</i> and <i>judgment</i> etc.	N/A
	variant spellings	all corpus-attested, e.g. <i>harbour</i> and <i>harbor, analogue</i> and <i>analog</i> etc	N/A
	inflected forms: morphological	irregular plurals of nouns, irregular comparatives and superlatives of adjectives e.g. <i>children, better, best</i> Note that this refers only to inflections to be shown as <i>headwords</i> . The separate question of which inflections will be shown <i>within</i> a dictionary entry is not addressed in this document.	all regular morphological inflections and strong verb inflections, e.g. <i>giving, gives, gave, given, cats, leaves, harder, softest</i>
	inflected forms: derivational	all e.g. <i>highhandedness, blissful, nakedly</i>	N/A
Lexical Structure	single word	all e.g. <i>table, give, splendid, badly, enough</i>	N/A
	partial word: bound prefix	N/A	all, e.g. <i>im-</i> as in <i>impossible</i>

Type	Property	Include	Exclude
	partial word: bound suffix	N/A	all e.g. <i>-ness</i> as in <i>happiness</i>
	partial word: productive prefix	<i>ex-</i> (<i>wife, mayor, teacher...</i>), <i>anti-</i> (<i>war, abortion...</i>)	N/A
	partial word: productive suffix	all current productive suffixes, e.g. <i>-gate</i> (as in <i>Watergate, Parkgate</i> , although the most frequent of these forms will be headwords in their own right); <i>-free</i> (as in <i>additive-free</i>)	N/A
	abbreviation: alphabetism	all current e.g. <i>RTÉ, WMD, MEP, EEA</i> (<i>European Environment Agency</i>), i.e., e.g.	N/A
	abbreviation: acronym	all current e.g. <i>Unesco, EFSA</i> (<i>European Food Safety Authority</i>), <i>LUAS</i>	N/A
	multiword: compound : noun, verb, adjective, adverb	all current e.g. <i>house agent, centre of excellence, blind drunk, dead centre</i>	N/A
	multiword: hyphenated compound: noun, verb, adjective, adverb	all current e.g. <i>decision-making, deep-fry, holier-than-thou, high-handedly</i>	N/A
	multiword: phrasal verb : 2-part and 3-part	all current e.g. <i>pass out, smack of, come up with</i>	N/A
	phrases: idioms, proverbs, quotations, other fixed and semi-fixed phrases	N/A	exclude all (they will be in body of entry) e.g. <i>to be too big for one's boots, kith and kin, all's well that ends well,</i>
Sublanguage	region: Irish-English	current and dated items from Irish-English e.g. <i>graip, craic, skite</i> Note however that much of what is generally identified as Irish-English vocabulary consists of specifically Irish	rare and obsolete terms

Type	Property	Include	Exclude
		<p>uses of common polysemous words: e.g. <i>guards</i> ('police' in Irish English), <i>hook</i> (in hurling), or <i>give out</i> (=criticise, complain). In these cases, specific senses will be labelled, but this is not an issue for headword selection.</p>	
	<p>region: other varieties of English: include</p>	<p>common items from:</p> <ul style="list-style-type: none"> ▪ British English: e.g. <i>dialling code</i>, <i>lawyer</i>, <i>courgette</i>, <i>aubergine</i> ▪ American English e.g. <i>dime-store</i>, <i>attorney</i>, <i>zucchini</i>, <i>eggplant</i> <p>and highly frequent current items from other Englishes (such as Canadian, Indian, and Australian English), especially when these have some currency in Ireland.</p>	<p>very infrequent items from British and American English; most items from other Englishes, e.g (Canadian) <i>beater</i>, <i>allophone</i>; (Australian) <i>award wage</i>;</p>
	<p>domain: 'school-curriculum-oriented' vocabulary (associated with mathematics, the sciences, economics, history, geography, environmental science, music, art, languages and literature, religious studies etc.)</p>	<p>all words appearing regularly in Irish educational texts, and in general, non-specialist texts, including words with specifically Irish connections or referents, e.g. <i>gamma rays</i>, <i>transubstantiation</i>, <i>biosphere</i>, <i>greenhouse gas</i>, <i>cubism</i>, <i>aquifer</i>, <i>fractal</i>, <i>gamelan</i>, <i>hexameter</i>, <i>assonance</i>, <i>gouache</i></p>	<p>terms which occur only in texts written by specialists for specialists, e.g. scholarly journals such as <i>Journal of Biochemistry</i>, <i>Molecular Biology and Evolution</i>, <i>Progress in Physical Geography</i>.</p>
	<p>domain: vocabulary from professional disciplines (medicine, the law, computing, engineering, etc.)</p>	<p>all words commonly found in general, non-specialist texts, e.g. <i>tibia</i>, <i>lien</i>, <i>quark</i>, <i>screensaver</i>; in particular, words appearing regularly in the Irish media, Irish educational texts, and other vocabulary items with specifically Irish connections or referents.</p>	<p>terms which occur only in texts written by specialists for specialists e.g. professional or scholarly journals such as <i>Irish Medical Journal</i>, <i>Computational Linguistics</i>, <i>Irish Criminal Law Journal</i></p>
	<p>domain: sports</p>	<p>all words appearing regularly in the Irish media and other general, non-specialist texts e.g. <i>striker</i>, <i>line-up</i>; most terms</p>	<p>rare and esoteric terms used only in texts written by specialists for specialists, e.g. sport magazines such</p>

Type	Property	Include	Exclude
		relating to Irish sports and sports popular in Ireland, such as hurling e.g. <i>hurley, corner-back, half-forward, foot pass, square ball.</i>	as <i>Irish Lady Golfer</i> , books such as <i>Flyfishing in Ireland, The Irish Sports Almanac</i> , and their British and US equivalents.
	domain: politics and current affairs	all words appearing regularly in the Irish media and other general, non-specialist texts; cover principal institutions, titles, posts, entities and processes (national and local) in Ireland (including Northern Ireland), and – with a somewhat higher threshold of inclusion – in the US and the EU: e.g. <i>Dáil, Taoiseach, T.D., Labour Relations Commission, Northern Ireland Assembly, House of Commons, Defense Department</i>	rare and esoteric terms used only in texts written by specialists for specialists, e.g. journals such as <i>Parliamentary Affairs, Strategic Survey</i> , and books such as <i>Hegemony or Survival? America's Quest for Global Dominance.</i>
	domain: banking, business, management, the media	all words appearing regularly in the Irish media and other general, non-specialist texts; e.g. <i>benchmarking, asset-stripping, digital divide, narrowcasting, CEO, stakeholder</i> ; in particular relating to principal institutions, titles, posts, entities and processes in Irish, British, American and European economies e.g. <i>Bank of Ireland, Irish Stock Exchange, FTSE index</i>	rare and esoteric terms used only in texts written by specialists for specialists, e.g. <i>Journal of Financial Econometrics</i> , sections of the <i>Financial Times</i> etc.
	slang & jargon	terms appearing regularly in the Irish media and in general, non-specialist texts e.g. (prison slang) <i>stir</i> ; (computing jargon) <i>hacker, nerd</i> ; (email jargon) <i>btw, wrt</i>	obscure or rare terms e.g. (prison slang) <i>ass betting</i> ; (computing jargon) <i>bear paw</i> ; (military slang) <i>click, hooch, PX</i> ; (Cockney rhyming slang) <i>Ruby (Murray = curry)</i>
	dialect	a wide range of Irish dialect expressions; only the most frequent of English regional, Scottish and Welsh	infrequent English, Welsh or Scottish dialectal terms; very obscure or rare Irish dialectal terms

Type	Property	Include	Exclude
		dialect words and phrases, especially where there is evidence of them appearing in novels by popular writers such as Catherine Cookson and Irvine Welsh e.g. <i>lough, loch</i>	Irish dialectal terms
	style: literary, poetic, bureaucratic, journalese writing:	items necessary for understanding of literary texts or current general reading or listening e.g. <i>casement, revels</i>	obscure or rare terms e.g. <i>incarnadine, bosky</i>
	register: formal, informal, very informal	all current, unless offensive e.g. <i>cool, muppet, rasta</i>	obscure or rare terms
	time: obsolete (archaic), obsolescent (old-fashioned), ephemeral etc.	items necessary for understanding of literary texts or current general reading or listening e.g. <i>helpmeet, fie!, mayhap</i>	obscure or rare terms e.g. <i>gadzooks, anight, afeard, yclept</i>
	restrictive usage: offensive terms – racist, sexist, etc. :	current and recent terms with adequate warnings as to use e.g. <i>teague, paki, jock, nigger, fairy</i>	(inclusion – or otherwise – of the most extreme members of this set will need to be agreed between the Contractor and Client)
	restrictive usage: offensive terms	current and recent terms with adequate warnings as to use e.g. <i>shit, bloody</i>	(inclusion – or otherwise – of the most extreme members of this set will need to be agreed between the Contractor and Client)

4. Proper Names

The User Profile – and specifically the anticipated use of the dictionary by schoolchildren, students, academics and writers/journalists – argues for fairly wide coverage of this part of the lexicon. We propose an inclusion algorithm that combines:

- systematic inclusion of certain ‘basic’ closed sets (e.g. continents, planets, zodiac signs)
- a hierarchy of ‘relevance’ with Ireland and Irish-related items having most favoured status, then ‘rest of Europe’, then ‘rest of the world’
- an approach to selecting from open-ended sets that reflects:
 - corpus frequency

- 'profile' (how well-known and widely-known the item is)
- the existence (or not) of additional meanings or connotations (e.g. where *Stormont* has a meaning beyond the physical building, and *Orwellian* means more than 'relating to Orwell')
- whether or not there is an Irish name for the place, person etc.

Type	Referents	Include	Exclude
Place names	closed sets of 'basic' items: oceans, continents, planets, countries, counties in Ireland, the four Irish provinces	all – there is no obvious or non-invidious way of selecting subsets of these: <i>Atlantic, Asia, Saturn, Switzerland, Meath, Leinster</i> options: <ul style="list-style-type: none"> ▪ capital cities: include all major capitals, and minor capitals in current news ▪ U.S. states: include all, provided there is an Irish translation ▪ British counties: include all, provided there is an Irish translation 	N/A
	open sets of major geographic features: seas, lakes, rivers, mountains/ranges, regions, non-capital cities, islands, topographic features	include on basis of frequency and well-knownness (which often correlates with size), with a lower threshold for anything Irish; in the case of towns and cities in Ireland - set threshold based on population Ireland: <i>Irish Sea, Bantry Bay; Lough Neagh; River Blackwater; Mourne Mountains, Macgillycuddy's Reeks; Connemara, Dingle;</i>	Ireland: precise thresholds to be agreed during compilation of the dictionary. rest of world: <i>Tasman Sea, Sea of Okhotsk; Lake Balkhash; Paraná, Mackenzie; Mt Nanga Devi, Sierra de Guadarrama; Transcaucasia; Leicester, Chennai; Andaman Islands; Ngorogoro Crater</i>

Type	Referents	Include	Exclude
		<p><i>Cork, Limerick; Achill Island; the Burren, Giant's Causeway</i> rest of world: <i>Mediterranean, Sea of Galilee; Lake Superior; Amazon, Thames; Everest, Snowdon, the Andes; Middle East, Balkans, the Lake District; New York, Barcelona, Shanghai; Sardinia, Isle of Man; Grand Canyon, Sahara Desert, Bermuda Triangle</i></p>	
	<p>famous places and buildings: major battlefields, important buildings, major airports, sites of religious significance</p>	<p>include on basis of frequency and high profile. Many place names have additional uses, developed either through metonymy (<i>The White House</i>=the US administration), or connotation ('like Fort Knox' connotes 'high levels of security'), and this is a further argument for inclusion:</p> <p>Ireland: <i>Ballinamuck, Temple Bar, Stormont, Leinster House, Phoenix Park, Knock</i></p> <p>rest of the world: <i>Gettysburg, Ten Downing Street, Scotland Yard, the Pentagon, Cape Kennedy, Heathrow, Lourdes, Mecca, Fort Knox, the Parthenon, Harley Street</i></p>	<p>exclude obscure, rarely referred to items with no special cultural significance</p>
	<p>extra-terrestrial places/objects (apart from planets): stars, galaxies, moons/satellites, constellations, comets etc.</p>	<p>include on basis of frequency and high profile: <i>asteroids, Milky Way, Halley's comet, Sea of Tranquillity, the Big Dipper</i></p>	<p>exclude obscure, rarely referred to items with no special cultural significance: <i>Sea of Ingenuity</i> (on the moon), <i>Lysithea</i> (moon of</p>

Type	Referents	Include	Exclude
			Jupiter), <i>Schwassmann-Wachmann 1</i> (comet)
	imaginary, biblical or mythological places	include on basis of frequency and high profile: <i>Shangri-La, Ruritania, Garden of Eden, Lilliput, Hades, Armageddon</i>	exclude obscure, rarely referred to items with no special cultural significance
	nicknames for places	include on basis of frequency and high profile: <i>Tinseltown, The Big Apple, The Square Mile</i>	exclude obscure, rarely referred to items with no global cultural significance: <i>The Big Easy, Beantown, Foggy Bottom, the Granite City</i>
People			
	: first names, surnames	<ul style="list-style-type: none"> ▪ first names where there is an Irish equivalent: <i>Joseph, Mary, Peter</i>, etc ▪ Irish surnames – this is not without its problems, but we believe surnames could be a popular feature with many users. The CIE proposes that we select 100 common surnames and include these in the print edition. 	exclude obscure, rarely referred to items
	famous (real) people: historical figures, rulers, writers, artists, musicians, religious leaders etc	include on basis of frequency and high profile, where there is an Irish translation: <i>Wolfe Tone, Daniel O'Connell, J.M. Synge, Beethoven, Shakespeare, Julius Caesar, Cromwell, the Buddha</i>	exclude obscure, rarely referred to items
	mythological, semi-historical, biblical, or fictional characters	include on basis of frequency and high profile, with special consideration for the	exclude obscure, rarely referred to items

Type	Referents	Include	Exclude
		particular issues in Irish regarding Biblical names. Many of these names are used allusively, as stereotypical embodiments of some particular quality (e.g. <i>Lothario</i>), and this is a further argument for inclusion <i>Anna Livia, Queen Maeve, Solomon, Jezebel, Canute, Colonel Blimp, Robin Hood, Jekyll and Hyde, Mr Pooter, Cinderella, Lady Macbeth, Walter Mitty</i>	
	derived adjectives from famous people	as headword where frequent and where the meaning is more specialised/connotative than simply 'by or in the manner of X': <i>Orwellian, Kafkaesque, Paisleyite, Dickensian</i>	exclude obscure, rarely referred to items. Where the adjective means only 'by or in the manner of X' (e.g. <i>Shakespearian</i>), show as derived forms at main headword
	nationalities (nouns and adjectives), and names for natives of Irish cities, counties, or regions	include on basis of frequency and high profile: <i>Kerryman, Galwegian, Liverpoolian; French, American, Chinese,</i>	for less frequent/central items, show as derived form at country headword (e.g. <i>Venezuelan</i> at <i>Venezuela</i>)
	peoples, e.g Native-American people, ancient and medieval people, ethnic groups (nouns and adjectives)	include on basis of frequency and high profile <i>Minoans, Hittites, Celts, Vikings, Kurdish, Roma, Walloons, African-American, Apache, Aztec, Picts, Saami</i>	exclude obscure, rarely referred to items
Other names			
	festivals, ceremonies	include on basis of frequency and high profile: <i>Samhain, Lughnasa; the Assumption, Fourth of July, Ramadan, Christmas, Thanksgiving,</i>	exclude obscure, rarely referred to items

Type	Referents	Include	Exclude
	organizations, clubs, political parties, institutions, government departments	<i>Bar-mitzvah, Valentine's Day</i> include on basis of frequency and high profile: <i>Alliance Party, ICTU, Court of Criminal Appeal, Office of the Revenue Commissioners, European Central Bank, Parades Commission (Northern Ireland) Shinners; Freemasons, Republican Party, Liberal Democrats, National Guard, Ivy League, Al-Qaeda</i>	exclude obscure, rarely referred to items
	languages: national and major regional languages, major language groups/families	<i>Dutch, Mandarin, Flemish, Ulster-Scots, Arabic, Hindi, Sanskrit, Basque, Dravidian, Indo-European</i>	exclude obscure, rarely referred to items
	trademarks for products and services	include on basis of frequency and high profile: <i>Babygro, Band Aid, Bluetooth, Frisbee, Yellow Pages</i>	exclude obscure, rarely referred to items.
	beliefs and religions, and their adherents (nouns and adjectives)	include on basis of frequency and high profile: <i>Church of Ireland, Taoism, Hindu, Wee Frees, Baptist, Marxism, Freudian, Jain, Amish, Scientology, Salvation Army, Albigensian</i>	exclude obscure, rarely referred to items.
	miscellaneous	include on basis of frequency and high profile: <i>IMPAC Prize, Holy Grail, Heath Robinson, Hang Sen Index, Book of Kells, Wade vs. Roe, Academy Award, the Holocaust, Good Friday Agreement</i>	exclude obscure, rarely referred to items. Titles of books, operas, songs etc will rarely merit inclusion as headwords, but will sometimes be covered within other headwords