Lexicographic Profiling: an Aid to Consistency in Dictionary Entry Design

Sue Atkins
Lexicography MasterClass Ltd.
River House
Church Lane
Lewes
East Sussex BN7 2JA
UK

Valerie Grundy
Lexicographic Consultant
Bellevue
07440 Boffres
France

Abstract
The first phase of any new dictionary project includes the detailed design of the dictionary entries. This involves writing sample entries. Deciding which lemmas to write entries for tends to be random, for lack of theory-based criteria for selection. This paper describes a database of the lexicographic properties of English lemmas, reflecting the work of theorists such as Apresjan, Cruse, Fillmore, Lakoff, Levin and Mel'Ëuk. It was used successfully in the selection process during the planning phase of a major new English bilingual dictionary, ensuring that all major policy decisions on entry structure were made early in the project, and resulting in a DTD well able to cope with 50,000 entries and a Style Guide already comprehensive at the start of the main project. The method described here ("lexicographic profiling") is applicable to any language, although of course the actual properties of lemmas are to some extent language specific.

1 The background

The paper describes a systematic approach to the microstructure design for a totally new corpus-based bilingual dictionary, tried and tested in Phase 1 of the *New English Irish Dictionary* (NEID) project.[1] The following packages were completed on time and within the rather modest budget in the course of the one-year first phase of this project:

[1] This project was launched in the summer of 2003. The contract for Phase 1 of the project was awarded to the Lexicography MasterClass Ltd. The project was directed by Sue Atkins, Adam Kilgarriff, and Michael Rundell, with Valerie Grundy as Managing Editor.

1. a design for a customised lexicographical corpus;
2. a 225 million word corpus of English (Hiberno-, British and American), and a 30 million word corpus of Irish, all linguistically annotated to a high level;
3. software: a corpus query system with the corpora loaded;
4. a user profile, a set of headword selection principles and a headword list;
5. a list of linguistic labels for marking register, style, domain, variety etc.;
6. 100 sample bilingual entries (English-Irish) covering the full range of entry types;
7. 50 'template' (model) entries, for certain lexical sets;
8. a detailed description of the proposed entry structures needed for the dictionary;
9. a document type definition (DTD);[2]
10. software: a dictionary production system with the DTD loaded;
11. a comprehensive style guide (lexicographers' manual);
12. a functioning web-based 'reading-and-marking' programme;
13. software: project-management tools (scheduling, textflow, budgets etc.);
14. a Linguistic Advisory Board consisting of 30 leading linguists, to advise on dictionary policy and comment on dictionary sample text;
15. a business plan for Phase 2 (dictionary production), with detailed schedules and budgets.

The work on lexicographic profiling described in this paper contributed substantially to tasks 6-11 inclusive of the above list.

2 How we set about it

For the dictionary writing software to serve the project correctly, the major policy decisions on entry structure, content and layout must be taken before the DTD is finalized. Similarly, the Style Guide must be comprehensively drafted, although of course this document is inevitably fine-tuned during the first year or so of a dictionary project. Our aim therefore in selecting the headwords for the 100 sample entries which would inform the DTD and Style Guide was to choose words which would bring out, as far as possible, the whole spectrum of problems, situations, and combinations likely to be encountered by the lexicographers and translators working on Phase 2 of the project.

2.1 Identifying lexicographic problems

In the course of our work as senior dictionary editors we had already compiled a database of words that had proved problematic in other dictionaries, together with the specific problems they raised. Lemmas were grouped according to part of speech, and within these groups there were a number of sub-categories relevant to that part of speech. Each of the problems in our original list had been analysed in detail, resulting in a relational database in which were recorded, for each lemma, the lexicographic properties that would require policy deci-

[2] The commercial dictionary production software, supplied by IDM Paris (www.idm.fr), required customization for the project: in particular of course the DTD had to be written, defining the structure of the various types of entry.

sions for the dictionary macro- and microstructure. Table 1 shows some examples of the kinds of issue we recorded at this point.

headword	issues
box	homograph issues: whether or not to make different entries with homograph numbers for the different core meanings of the lemma ('container', 'sport', 'botanical' *etc.*)
box	whether to organize the senses according to part of speech or according to meaning; e.g. should *box* the noun, meaning 'container', be located close to *box* the verb meaning 'put into a container' and separate from the other noun *box* (the plant), or should all the noun senses be grouped, then all the verb senses?
speck	an instance of a type of word that always raises problems in bilingual dictionaries: when used as itemizers,[3] such words often give rise to a plethora of translations according to the noun itemized (*speck of dust, speck of mud, specks of yellow etc.*)
echo	organization of multiple parts of speech (noun, modifier, verb), including representatives of main subtypes of verb (intransitive, transitive, report verb) and wide range of directional PPs.
echo	participation in multiword expressions (*cheer to the echo, echo chamber, echo back* etc.): where these should be shown (as headwords, nested headwords, within the entry in a special section etc.)
little	how to deal with not just the adjective but parts of speech that are notoriously awkward to pin down in a bilingual entry (broad negative quantifier/determiner)
might	as a modal verb, frequently raises problems of target-language equivalence: where no direct equivalent exists, particular lexicographical strategies need to be developed (specially adapted entry layout, discursive notes etc.)

Table 1. Some of the issues noted in our survey

2.2 Classifying the problems

The structured data constituted a database of ***lexicographic properties of lemmas*** (henceforth the lemma properties database, or LPD), i.e. those properties that have an immediate impact on lexicographic decisions. We included in this database concepts from linguists such as Apresjan, Cruse, Fillmore, Lakoff, Levin and Mel'Ëuk, whose work in lexical semantics gave us a clearer view of what we had to do. An example is the use we made of Apresjan's seminal study of regular polysemy (see 4.2.2).

This allowed us to produce lexicographic profiles of words which we were considering as candidates for the 100 sample entries, and to select the final 100 headwords on a systematic basis, ensuring that most of the problems the dictionary team would come across in Phase 2 of the project had been tackled and solved in this preliminary stage.

3 The problems: principal microstructure decisions

The main microstructure issues on which decisions had to be taken are summarised in

[3] This is Mel'Ëuk's 'Sing' lexical function; see Mel'Ëuk *et al* (1984/1988/1992)

this section: although our particular work was establishing the infrastructure for an English-Irish dictionary with a detailed user profile (and therefore focussing only on English), the decisions discussed here are relevant to every bilingual dictionary. We provided the publisher with seven batches of sample entries. Each batch focussed on one of the key issues listed below, and consisted of 12–15 entries and a discussion document offering alternative ways of handling the data and proposing our preferred method.[4]

3.1 High-level structural division of lexical entries

This is the first decision regarding the overall microstructure organization: should the top-level organizing principle for grouping the various *lexical units*[5] of a lemma be (1) meaning or (2) part of speech? This is a decision which must be made very early in the planning phase because of its implications for the DTD.

3.2 Treatment of multi-word expressions and derived forms

How should multiword expressions (MWEs: essentially, compounds, phrasal verbs, idioms) and derived forms be dealt with within the microstructure and what principles could be established for deciding which types of MWE (if any) should be given headword status. An early decision is needed on this point too, in order to establish a basic DTD.

3.3 Treatment of linguistic labels in dictionary and database

The actual types of label (domain, register, style, variety etc.) and the specific labels to be made available to the lexicographers were established separately. This batch of sample entries focussed on the circumstances in which a label should be used – in the initial source-language analysis, in the translation material, and in the final version of the dictionary entry – and those where some other type of indicating material would be more appropriate. The scope of the label, both in SL and in TL text, was also discussed.

3.4 Treatment of Irish regional variants in dictionary and database

The issue here was to determine which regional vocabulary and syntactic variants should be included in the target-language material, how these should be labelled (if at all), and – in the absence of an accepted 'standard' Irish – whether there should be any prioritizing distinctions made among the various regional equivalents.

3.5 Function word entries: content, style and layout

The main decisions here related to the words to be considered as 'function words' (in particular, the grey area of English adverbial and prepositional particles), and how the *functional* equivalents in the target language should be set out and exemplified. The methodology de-

[4] Final decisions on all of these points were the responsibility of the publishing institution, Foras na Gaeilge.

[5] This term is used as defined in Cruse (1986): essentially, a lemma in one of its meanings.

vised for dealing lexicographically with *semantic* equivalence is not very helpful in such cases.

3.6 Miscellaneous style issues

This batch of entries dealt with problems not addressed elsewhere. Each of them impacted less critically on DTD design than those discussed in 3.1 – 3.5, but nonetheless had to be accounted for in the Style Guide and DTD. They included dealing with abbreviations and acronyms, productive prefixes and suffixes, and contracted forms, as well as showing verbs with obligatory adjuncts, and resolving the various problems raised by nouns functioning as itemizers.

3.7 Issues relating to Irish equivalence

The issues raised in this batch of entries are those which cause problems in every bilingual dictionary: source-language (SL) items with no direct target-language (TL) translation; those with partial semantic congruence, where either the SL item or the TL item is more specific than the other; instances where the denotational meaning of SL and TL items matched but they diverged along the axes of register, style, or pragmatic force etc.; the approach to finding equivalents of culture-bound encyclopedic items such as *Downing Street*; and in what circumstances a simple TL paraphrase of the SL item should be acceptable.

4 Analysing the problems: the lemma properties database

The database is designed to be stored in Microsoft Access or a similar database package and to be queried on the basis of property type, actual property, or instantiating lemma. Still incomplete, it contains 27 classes of lexicographic properties, which account for 581 actual properties.

4.1 Classes of lexicographic properties

While some of these properties relate to the lemma (headword) itself, the great majority relate to specific parts of speech and therefore to the lexical unit (LU) rather than the lemma. The principal properties recorded about the lemma itself were:

- the various wordclasses of its LUs;
- its lexical form (simple word, prefix, abbreviation, hyphenated, multiword etc.);
- its corpus profile (corpus frequency, preferred text type, regional variety etc.);
- the labels needed for its lexicographic description (domain, register etc.);
- miscellaneous properties (monosemous, polysemous, with non-homophone homograph etc.);
- its participation in multiword expressions;
- its participation in regular polysemy types (see 4.2.2);
- its participation in a lexical set for which a model ('template') entry was envisaged.

The remainder of the properties in the database were recorded individually for nouns, verbs, adjectives, adverbs and function words and focussed on:

• noun, verb etc. types (common/proper nouns; lexical, auxiliary, copular verbs and subtypes of these; gradable, ungradable etc. adjectives and adverbs);

• semantic classes (human, artifact etc. nouns; motion, sound etc. verbs; colour, sound, existential etc. adjectives; manner, degree, time etc. adverbs);

• morphological properties, both inflectional and derivational;

• syntactic properties (nouns that modify other nouns, nouns postmodified by an adverbial phrase, as in *the journey home*; the wide range of syntactic behaviour, especially complementation, of verbs, adjectives and adverbs);

• participation in Levin-type alternations (for verbs and nouns);

• some other miscellaneous properties for each part of speech.

4.2 Some examples of lexicographic properties

Examples of two of these classes follow: *lexical form of the lemma (headword)* and *regular polysemy participant.*

4.2.1 Properties relating to the lexical form of the headword

The lexical form – or variations on this – of the lemma raises issues of content and layout which impact on both macrostructure and microstructure. Some examples are shown in Table 2.

property	e.g.	at headword	issue
simple word	*book, give, France*		(standard headword)
abbreviation (initial letters, pronounced as letters)	*the EC, the BBC, WMD*	**European, British, weapon**	Where will the full information be given? Will there be cross-references? If the target language has also an abbreviated form, will the full form of the target language be included and if so where?
acronym (initial letters, pronounced as a word)	*Unesco, Nato*	**united, north**	(as for abbreviation above)
prefix (productive)	*anti-war, ex-wife*	**anti-, ex- ...**	How to handle semantically rich prefixes which may be attached to a wide range of words.
suffix (productive)	*talkfest, additive-free, work-shy*	**-fest, -free , -shy**	How to handle similarly productive suffixes.
multiword lemmas	*all right, in spite of*	**all, right, spite ...**	Should there be multiword headwords? If so, define criteria for headword status. Decide how (if at all) to show links with other headwords.
hyphenated word	*broad-leafed, flop-eared*	**broad, leaf, flop, ear**	How to handle compounds (hyphenated or non-hyphenated) – as headwords? Secondary headwords? Defined? Not defined? etc.
has variant spelling	*aluminium /aluminum*	**aluminium**	Should both versions have headword status or should one variant be given within the entry for the other? If two headwords, where should the information be given?

Table 2. Issues raised by the lexical form of the lemma

4.2.2 Properties relating to regular polysemy relationships

This class of properties is very large (163 items): it was compiled initially on the basis of Apresjan's work (see Apresjan 1973), and expanded as other instances of the phenomenon surfaced in our analysis. This property does not generate lemma-specific problems. However, the identification of types of regular polysemy allows the Style Guide to include instructions on how to deal with headwords which participate in each particular type of relationship, making the overall approach to this aspect of the lexicography much more consistent. Table 3 contains a very short extract showing some types of regular polysemy.

property	instantiating lemmas	e.g.
container *n*->contents *n*	*box, tin, can, pocket, bag, phial, case*	• *six boxes on the floor* • *he ate the whole box himself*
container *n*->amount *n*	*box, glass, phial, case*	• *six sherry glasses* • *add half a glass of wine*
container *n*->to place in container *vt*	*pocket, box, can, tin*	• *she had some money in her pocket* • *he pocketed the change*
dance *n*->dance-music *n*	*tango, quickstep, reel*	• *I'm learning the tango* • *they played several tangos*
dance *n*->do that dance *vi*	*tango, quickstep, foxtrot ..*	• *I'm learning the tango* • *they tangoed round the room*
fruit of plant *n*->plant *n*	*raspberry, apple, orange, strawberry*	• *a bowl of raspberries* • *a field of raspberries*
tree *n*->its wood *n*	*oak, cedar, elm, pine, mahogany ..*	• *three old pines behind the house* • *table made of pine*
tree / wood *n*->of the colour of that wood *n mod*	*mahogany, walnut, oak, yew*	• *made of mahogany* • *the rich mahogany of her hair*
animal *n*->its fur or skin *n*	*mink, squirrel, fox, leopard, crocodile*	• *a lake full of crocodiles* • *a crocodile handbag*

Table 3. Some types of regular polysemy recorded in the database

5 Solving the problems: lexicographic profiles

A list was drawn up of over 400 lemmas known to be problematic, each lemma was checked against the appropriate property classes and its lexicographic profile was extracted. Our aim was to collect words as disparate as possible with regard to all aspects of inherent properties and corpus use, thus ensuring policy decisions on as many as possible points of macrostructure and microstructure. One example suffices, that of *echo*.

5.1 Case study: lexicographic profile of echo

Table 4 contains the lexicographic profile of the lemma *echo*, drawn from our database. The order in which the properties are listed has no significance. These profiles help to ensure that the words chosen as headwords of sample entries cover as many issues as possible. The implications for dictionary entry structure of most of these properties are for the most part transparent, but an explanatory note has been included where there could be some confusion.

	property	of lemma *echo*	note on policy decision
1	lexical form: simple word		
2	corpus profile[6]: raw frequency: lemma	30.65 per million words	Frequency and ranking of lemmas and LUs can affect the depth of treatment, particularly in learners' dictionaries.
3	corpus profile: frequency rank: noun	5089	*echo* is the 5089th most frequent noun in the corpus (see note above)
4	corpus profile: frequency rank: verb	3581	*echo* is the 3581st most frequent verb in the corpus (see note above)
5	belongs to a 'word family'	*echoing* adjective (*echoing footsteps*)	Morphologically related headwords can influence treatment of a lemma.
6	polysemous lemma	several dictionary senses e.g. 1) *his voice echoed* 2) *I echo these sentiments*	Relates to sense ordering, also hierarchical or non-hierarchical sense numbering etc.
7	p-o-s = noun	*Can you hear the echo?*	
8	p-o-s = verb	*His voice echoed down the hall.*	
9	noun subtype: count noun	*... sending out sounds and listening to the echoes*	Subtypes of noun relate to decisions on p-o-s markers in dictionary. e.g. *nc, nu* ...
10	noun subtype: uncount noun	*... using string sections and lots of echo to make records that appealed to older listeners.*	See note at 9.
11	noun syntax: (head of NP) subject of VP	*The unexpected echo startled us all.*	Making sure all syntactic possibilities are covered: some nouns rarely or never occur in this position, and this must be noted where relevant.
12	noun syntax: (head of NP) object of VP	*We heard a loud echo.*	See note above at 11.
13	noun: semantic frame[7]: sound	*We heard a loud echo.*	This is a concept from frame semantics; see Atkins *et al* (2003) and Fillmore *et al* (2003). Noting the semantics helps to ensure that the sample headwords are as disparate as possible, semantically.
14	noun: semantic frame: becoming aware	*It was an echo of former happiness.*	See note at 13.
15	noun: morphology: plural in *–es*	*echoes*	How to show source language plurals in the dictionary.
16	noun: morphology: verb-derived zero nominal	1) *it will echo* 2) *hear an echo*	See Levin (1993). Check if relationship holds across semantically distinct LUs. Policy on how to handle this if so and if not.

[6] Statistics taken from the British National Corpus (www.natcorp.ox.ac.uk/).

[7] "Semantic frames are schematic representations of situation types (eating, spying, removing, classifying, etc.) together with lists of the kinds of participants, props, and other conceptual roles that are seen as components of such situations." FAQs on FrameNet website, www.icsi.berkeley.edu/~framenet/. See also Fontenelle (2003).

	property	of lemma *echo*	note on policy decision
17	noun is element in MWE	*echo chamber* *to cheer to the echo*	
18	in MWE type: idiom: morpho-syntactic flexible	*cheered/cheering to the echo*	How to show flexibility (or not) of MWEs.
19	in MWE type: compound noun	*echo chamber*, plural *echo chambers*	How to handle plural of compound nouns: here the plural marked on 2nd component, but some compound nouns (e.g. *court martial*) mark plural on 1st component.
20	noun: metaphorical extension	*... uncanny echoes of past events.*	How to handle productive use of metaphor in language, yielding uses which are not 'set' enough to be treated as dictionary senses. See Lakoff & Johnson (1980).
21	verb subtype: intransitive	*The shot echoed across the valley.*	Subtypes of verb relate to decisions on p-o-s markers in dictionary: e.g. *vt, vi, vti, ...* How to deal with transitivity/ intransitivity in entry layout.
22	verb subtype: transitive	*He echoes this view.*	See note above at 21.
23	verb subtype: reporting verb	*"Because of the play?" he echoed mockingly.*	See note above at 21.
24	verb semantic frame: sound	*The sound echoed round the room.*	See note at 13.
25	verb: semantic frame: communication	*I am happy to echo those sentiments.*	See note at 13.
26	verb: morphology: weak vb *-es, -ing, -ed*	*echoes, echoed, echoing*	How to show verb inflections, which type to include, and which to consider default.
27	verb : metaphorical extension	*We painted the beams green to echo the colour of the furniture.*	How to handle metaphorical meaning extensions which are not 'set' enough to be treated as dictionary senses. See Lakoff & Johnsen (1980).
28	verb: in alternation: swarm: X verbs in Y -> Y verbs with X	1) *their voices echoed in the hall* 2) *the hall echoed with their voices*	See Levin (1993).
29	verb is element in MWE type: phrasal verb - V+ADV intransitive	*The sound ricocheted off the walls and echoed back.*	How to handle intransitive verb + particle phrasal units.
30	verb is element in MWE type: phrasal verb - V+ADV transitive	*The stage but echoes back the public voice.*	How to handle transitive verb + particle phrasal units.
31	label type: domain: telecommunications	*There's an echo on the line.*	A labelling policy is essential, together with a list of agreed labels and how they will be used.

Table 4. Lexicographic profile of *echo*

6 Applying lexicographic profiles

The lexicographic profiles of the 100 headwords selected for the sample entries generated principled decisions on the type of information to be included in the dictionary entry, how that information was to be shown, how different types of entry were to be structured and how the information types fitted together to make up that structure. From there, we were able to go on and establish the DTD and write the Style Guide.

6.1 The DTD

The DTD defines the structure of the different types of entry in the dictionary, and the use of the lemma properties database in the selection of sample entries enabled us to write the DTD with a high level of confidence that no major issue had been overlooked.

Every aspect of the lexicography of each headword was thoroughly studied: for instance, our analysis of different multiword expressions led us to conclude that – in this one-volume print dictionary – we needed to retain a good deal of flexibility with regard to the status of compounds. Thus, in the DTD, a Compound Container can contain essentially the same structure as a full lexical Dictionary Entry, allowing lexicographers at the final stage of editing[8] to decide (on the basis of the Style Guide) whether a specific compound should be given headword status or shown within the body of an entry, as this DTD extract[9] shows:

```
<!ELEMENT DEnt HwdGp, (SenBlk | FwkMWEBlk | PhrVBlk | UsgNoteCnt |
XRefCnt | FwkSenCnt | RegVarNoteCnt | COMMENT)*) >
<!ELEMENT CpdCnt (CpdGp, (SenBlk | FwkSenCnt | UsgNoteCnt | XRefCnt |
COMMENT)*) >
```

6.2 The Style Guide

On the basis of a comprehensive sampling of headwords participating in multiword constructions, the Style Guide contained quite explicit guidelines on how to handle these, as is shown in the following extract from the manual used by editors compiling the bilingual entry:

> **14.8 Compounds - nouns**
> In the framework, all compounds associated with a headword are contained in the CPDBlk within the MWEBlk. You should consider each carefully. All those that merit full compound-headword status according to the principles discussed above should be promoted to full headword entries and treated separately as such (CpdEnt). Any that do not should be subsumed in the modifier sense within the main entry.

[8] The three stages of editing were (1) initial source-language analysis, recorded in a database ("the framework"); (2) target-language translations added to database; (3) extraction from this material of the polished bilingual entry.

[9] The abbreviations CPDBlk etc. are labels used to define elements of entry structure in the DTD and to signpost them in the dictionary editing software.

7 Cùonclusion

The lemma properties database provided a systematic approach to policy decision-making in a new corpus-based dictionary project. By applying the ideas of theoretical linguists, and using our many years of lexicographic experience, we ensured that most difficulties which would be encountered during the course of dictionary editing had been carefully considered and instructions on how to handle them encoded into the style guide, and that the software was running on a comprehensive and flexible DTD.

References

Apresjan, J. D. (1973), 'Regular Polysemy', *Linguistics* 124, pp. 5-39.

Atkins, B. T. S., C. Fillmore, J., Johnson, C. R. (2003), 'Lexicographic relevance: selecting information from corpus evidence', in *International Journal of Lexicography*, Oxford, OUP: 16:3, pp. 251-280.

Cruse, D. A. (1986), *Lexical Semantics*, Cambridge University Press, Cambridge, UK.

Fillmore, C. J., Johnson, C.R., Petruck, M. (2003), 'Background to FrameNet', in *International Journal of Lexicography*, Oxford, OUP: 16:3, pp. 235-250.

Fontenelle, Th. (2003) (guest editor), *International Journal of Lexicography*, Oxford, OUP: 16:3 (issue devoted to FrameNet Project).

Lakoff, G., Johnson, M. (1980), *Metaphors We Live By* University of Chicago Press, Chicago, USA.

Levin, B. (1993), *English Verb Classes and Alternations: A Preliminary Investigation*, University of Chicago Press, Chicago, USA.

Mel'čuk, I., Arbatchewsky, N., Dagenais, L., Elnitsky, L., Iordanskaja, L., Lefebvre, M. N., Mantha, S., Lessard, A. (1984/1988/1992), *Dictionnaire Explicatif et Combinatoire du Français Contemporain: Recherches Lexico-Sémantiques I, II, III*, Montréal, Les Presses de l'Université de Montréal.