The DANTE database: a User Guide

Michael Rundell, Sue Atkins
Lexicography MasterClass
E-mail: michael.rundell@lexmasterclass.com, sue.atkins@lexmasterclass.com

Abstract

DANTE – the Database of ANalysed Texts of English – is a lexical database which provides a corpus-based description of the core vocabulary of English. It records the semantic, grammatical, combinatorial, and text-type characteristics of over 42,000 single-word lemmas and 23,000 compounds and phrasal verbs, and it also includes over 27,000 idioms and phrases. Every fact recorded in the database is derived from a systematic analysis of a 1.7 billion-word corpus and supported by corpus examples. The complete text of DANTE from M to R is freely available online (at www.webdante.com), and the full database is available through research or commercial licences (http://dante.sketchengine.co.uk/). The website provides basic information about DANTE and a Help function to assist users who wish to search the database. This User Guide is intended to complement the information on the website and explain the rationale of the various components of DANTE's microstructure.

Keywords: lexical database; query-builder; inherent grammar; syntactic context; label; domain; multiword; chunk; collocation; support verb; support preposition; itemiser; pragmatics

1. Introduction

DANTE is a lexical database which provides a fine-grained, corpus-based description of the core vocabulary of English. Every fact recorded in the database is derived from, and explicitly supported by, evidence from a 1.7 billion-word corpus of current English. Almost all of these facts are machine-retrievable.

DANTE – the Database of ANalysed Texts of English – was designed and created for Foras na Gaeilge by the Lexicography MasterClass and an 18-strong team of skilled lexicographers, using the Sketch Engine (www.sketchengine.co.uk/) for corpus-querying, and IDM's Dictionary Production System (DPS: www.idm.fr) for entry-building. The resulting database records the semantic, grammatical, combinatorial, and text-type characteristics of over 42,000 single-word lemmas and 23,000 compounds and phrasal verbs, and includes over 27,000 idioms and phrases, underpinned by over 600,000 sentence examples from the corpus.

Though DANTE's primary function was to provide an 'English framework' for the development of a new English-Irish dictionary (www.focloir.ie/english.asp), it was designed from the start to be a linguistic resource of more general utility. It offers publishers a launchpad for the development or updating of monolingual or bilingual dictionaries, and provides rich data for researchers, language engineers, software developers, and materials writers.

2. The DANTE User Guide

A large section of the DANTE database (the complete text from M to R) is freely available online at www.webdante.com. As well as searching for individual headwords, users can employ a query-builder in 'Advanced Search' mode to create a wide range of specialised searches: to find, for example, every verb that takes a particular construction, or every American English word with the style label 'humorous'. The complete database is available through research or commercial licences (http://dante.sketchengine.co.uk/).

The DANTE User Guide has been prepared to facilitate searches for users of the public website or of the complete database. It has two functions:

- to describe the search parameters available on the DANTE website
- to explain the components of entries in the database, and the editorial policies underlying them.

The website's Help function provides all the information needed to specify a search. But entries in DANTE are often complex; the lexicographers' Style Guide runs to well over 100 pages. So it is important to stress that those entry components available as search options on the website represent only a subset of all possible entry components. One of the objectives of the Guide, therefore, is to explain the content and rationale of every information-type you see when viewing entries returned by a search.

3. The 'lexical unit' in DANTE

One of the key features of DANTE is that each of the main entry components (such as a part of speech label, a grammar code, or a style label) is associated with a particular 'lexical unit'. In DANTE, a 'lexical unit' (or LU) approximates to a 'dictionary sense', and it is the principal 'currency' of the database. More precisely, it is an umbrella term for describing any use of a word that carries its own discrete meaning or function: single-sense headwords, individual senses of polysemous words, idioms, compounds, and phrasal verbs are all lexical units. And (just as with polysemous headwords), if an idiom, compound, or phrasal verb has more than one sense, each counts as a lexical unit.

4. Definitions and examples in DANTE

4.1 Definitions

Since DANTE is a lexical database rather than a dictionary, it does not have conventional definitions. Rather, every LU has a 'meaning statement', whose function is not to 'define' the item in detail but to provide enough semantic information to enable the user to recognise which meaning area or 'dictionary sense' the LU relates to. DANTE's meaning statements are thus closer to the 'sense indicators' used in bilingual dictionaries (Atkins and Rundell, 2008: 503-504) – though they are generally fuller.

4.2 Examples

All the lexical data in DANTE is driven by the corpus, and it is a fundamental design feature that every linguistic fact recorded in the database is illustrated by one or (usually) several corpus examples. Consequently, DANTE includes well over 600,000 example sentences.

Take, for example, the relatively infrequent word *recollection*. In DANTE, *recollection* has two LUs: an uncountable use ('the act of remembering'), and a countable one ('a memory'). The second of these has no fewer than 25 examples: the first four exemplify its basic use; the next five illustrate its most frequent adjective collocates (see 8.2.2), and the rest are examples of the various syntactic contexts (see 5.3) in which the noun regularly participates.

The vast majority of examples are taken directly from corpus data, without modification. Occasionally, corpus examples are supplemented by short, formulaic examples, inserted to illustrate the full range of possible contexts for members of a particular semantic set: for example, colour terms include formulaic examples for the noun use like these ones found at *pink*:

- *dressed in pink*
- *wearing pink*
- *available in pink*
- *she likes pink*
- *a shade of pink*

5. Grammar and Syntax in DANTE

This section explains the following search conditions which are available in the 'Build a search' drop-down lists on the DANTE website:

- part of speech
- inherent grammar
- syntactic context

5.1 Part of speech (POS)

Table 1 lists the items that appear in the drop-down menu for a 'part of speech' search. Most are self-evident, but an explanation is provided in cases where there might be doubt as to how the part of speech is used in DANTE.

Searches on any of the four main parts of speech – adjective, adverb, verb (all types), and noun – can be refined using the 'inherent grammar' condition. Inherent grammar is explained in section 5.2.

Part of speech in drop-down list	Part of speech in DANTE entries	Explanation	Examples
adjective	adj		
adverb	adv		
conjunction	conj		
determiner	det	Used for definite and indefinite articles, quantifiers, demonstratives, possessives.	'a', 'both', 'either', 'my', 'some'
interjection	interj		
noun	n		
numeral	num	Used for cardinal and ordinal numbers, and also for numerical uses of lemmas such as *nothing* and *nought*	*She's five foot nothing; nought point three*
prefix	pref	Used for lemmas that combine freely and can generate closed forms	'macro-' as in *macroclimate, macrostructure etc.*
preposition	prep		
pronoun	pron		
suffix	suff	Used for lemmas that combine freely and can generate closed forms	'-made' as in *homemade, homemade*; '-phone' as in *francophone etc.*
verb: auxiliary	v_aux	auxiliary verb	There are only three: 'be', 'do', 'have'
verb: lexical	v	straightforward lexical verb: the default verb type	'maintain', 'navigate', 'operate', 'persist', 'run'….
verb: modal	v_mod	modal verb	'may',' might', 'must' etc.
verb: phrasal	phr_v	phrasal verb	more details at 3.1.3

Table 1: Parts of speech

5.1.1. Parts of speech not available as search options

DANTE also uses the parts of speech **prp_adj** and **ptp_adj**, and these are not available in the POS search. They refer, respectively, to present participle adjectives and past participle adjectives, and are used only in SUBFORMs (5.1.2), not at headword level. They are applied to adjectival participles which are not sufficiently frequent to qualify for full headword status. Examples include: (prp_adj) 'a *mesmerising* story', 'the *roaring* jet engines'; (ptp_adj) '*maddened* with pain', 'a non-slip *rubberised* surface'.

5.1.2. The SUBFORM field

The SUBFORM field is not available as a search option on the website, but you will sometimes see it in an entry. A SUBFORM is a specific form of the headword which is itself a lexical unit. SUBFORMs include:

- present participle adjectives and past participle adjectives (5.1.1)
- plural forms of nouns with their own distinct meanings and uses (*marbles, dealings*)
- nouns with obligatory 'the' (*the Madonna*)
- capitalized forms of lower case headwords (*King*)
- hyphenated forms (the verb *slam-dunk* at the noun entry *slam dunk*)
- combining forms (*-haired* at *hair, -metre* (as in 'a 1000-metre race') at *metre*).

5.1.3. Phrasal verbs

There is no watertight definition of the category 'phrasal verb'. In DANTE, we recognise three types of phrasal verb:

- verbs with an adverb particle: *get up, point out*
- verbs with a preposition particle: *see through* (someone's plans), *part with* (your money)
- verbs with both types of particle: *make off with, refer back to.*

Phrasal verbs in DANTE have the part-of-speech label **phr_v**. To search on the website for a phrasal verb, select 'verb:phrasal' in the drop-down list of parts of speech. As with other verbs, phrasal verb searches can be further refined using the 'inherent grammar' condition (5.2).

5.2 Inherent grammar

When searching for an adjective, adverb, verb, phrasal verb, or noun, you can refine your search by specifying the item's 'inherent grammar'. For example, the POS label 'adverb' will find *any* type of adverb, but if you add an inherent grammar condition you can narrow your search to find (for example) only adverbs of degree. If using the Advanced search mode on the website, you will see that the drop-down list for 'inherent grammar' is tailored to each of the relevant parts of speech. The codes are explained in Tables 2-6. The following parts of speech have no inherent grammar options in DANTE: conjunction, determiner, interjection, prefix, preposition, pronoun, suffix.

In the database, inherent grammar codes appear in a field called GRAM, but in the entries shown on the website, you will see only the inherent grammar code itself (following the part of speech label), not the field name GRAM.

5.2.1. Inherent grammar: adjectives, adverbs, verbs, phrasal verbs

The available codes are explained, respectively, in Tables 2, 3, 4 and 5. In DANTE policy, phrasal verbs *always* have an inherent grammar code, but for nouns, adjectives, adverbs and 'standard' verbs, inherent grammar codes are not always required.

Code	Explanation	Examples
(no code)	default: an adjective that can occur in both attributive and predicative uses	*small, happy, green*
attr_only	an adjective that is attributive only	*mere* (a *mere* mortal)
comb	combining form: a form of a headword which can combine with other words to produce an adjectival compound. Combining forms appear in the SUBFORM field (above, 2.1.1)	-*conscious* (health-*conscious* consumers), -*maintained* (a poorly-*maintained* building)
pertnm	a pertainym adjective: an adjective that means 'pertaining to X'; pertainyms are attributive only, have no comparative or superlative form, and are never modified	*marital* bliss, *political* acuity, *racial* sensitivity
post_mod	post-modifier adjective	mayor *elect*, heir *apparent*
predic_only	an adjective that is predicative only	*alone, mindful*

Table 2: Inherent grammar: adjectives

Code	Explanation	Examples
(no code)	default: a manner adverb	*accidentally, jauntily, patiently*
deg	a degree adverb	*seriously* ill, *unbelievably* stupid
sent	a sentence adverb (typically sentence-initial, but can appear in any position)	*hopefully,* it won't rain; *personally* I think he's lost it; we could take the train or, *alternatively*, we could drive
view_pt	a viewpoint adverb ('from the X point of view')	*politically* serious, *socially* inept

Table 3: Inherent grammar: adverbs

Code	Explanation	Examples
(no code)	default: a verb whose use is not restricted	*say, walk, accuse*
imper_inf	a verb used only in the imperative or infinitive	something we need to *beware* of*, let* the ceremony begin
impers	an impersonal verb	it *rains* a lot in April*,* it's *snowing*
passive	a verb occurring in only in the passive	it is *rumoured* that...
reciproc	a reciprocal verb	John and Mary *marry,* John *marries* Mary

Table 4: Inherent grammar: verbs

Code	Explanation	Examples
v_adv	a phrasal verb consisting of a verb with an adverbial particle	*pass* the message *on, pass on* the message, the custom *died out*
v_adv_prep	a phrasal verb consisting of a verb with adverbial and prepositional particles	*come up with* an idea
v_prep	a phrasal verb consisting of verb with a prepositional particle	*look at* the screen, *ran through* all his money

Table 5: Inherent grammar: phrasal verbs

5.2.2. Inherent grammar: nouns

Nouns exhibit a wide range of grammatical behaviours. Categorising them is notoriously difficult, and we are not aware of any system that accounts for all possible cases. The approach used in DANTE is a pragmatic one, and DANTE lexicographers used a flowchart to determine which (if any) of ten possible codes should be applied to a noun in a given lexical unit. The inherent grammar codes used for nouns are explained in Table 6, and the lexicographers' 'noun flowchart' is in Table 7.

Code	Explanation	Examples
(no code)	default: a countable noun. Includes concrete objects, countable senses of words with uncount senses, and 'type' and/or 'unit' senses of mass nouns	*cup, dog, idea, teacher, organisation, risk, coffee* ('three *coffees*, please'), *wine* ('the *wines* of New Zealand')
c_u	a noun that is usually countable but has (1) generic-type uncountable uses, or (2) senses which combine the ideas of 'the act or an instance of X'	(1) *dinner* (three *dinners*, *dinner* is at 7), *bus* (three *buses*, go by *bus*); (2) *realignment, rationalisation* (a series of *realignments/rationalisations*; in need of *realignment/rationalisation*)
mass	a mass noun: applied to conceptually mass items, which are typically substances of some kind, including fabrics, foods, liquids, chemical elements and compounds, etc. Names of colours are also coded as 'mass' in DANTE, but (like many mass nouns) they can also have separate senses to cover 'type' or 'unit' uses	*blood, sand, sewage, bedding, pasta, oxygen, titanium, heroin, purple, wine*
pl_0	a noun with unchanged plural form	*sheep, gasworks*
pl_also_0	a noun which has a regular plural form, but which can also have a 'collective' or 'hunting' plural where the form is unchanged	five *herrings,* go fishing for *herring*; equipped with four 20mm *cannons*, we could hear the sound of *cannon*
pl_only	a noun occurring in the plural only	*ablutions, algae, scissors*
proper	a proper noun: typically a name; always capitalised; most operate normally without a definite or indefinite article; some require a definite article; some can be pluralised	*Edinburgh, Christmas, Monopoly, Napoleon, the White House, the Maldives*
u_c	a noun that is usually uncountable but has countable uses in certain predictable contexts:. This code is used mainly for infrequent, non-core lemmas or LUs, when conflating uncountable and countable uses, (typically covering 'the act/process of X' and 'an instance of X' or 'the result of X').	*nominalisation, popularisation*
uncount	an uncountable noun, rarely if ever found in the plural form. Applied to: abstract nouns (typically with definitions that start with any of: 'an act, state, quality, feeling etc'); academic subjects; schools of thought; medical conditions; sports; musical genres, etc. The code 'uncount' is also applied in DANTE to items which are coded 'singular' or 'singular only' in some dictionaries, as in: a *riot* of colour, grab a *bite*, the test was a *breeze*	*anger, maintenance, surrealism, jazz, hockey, weather, geography, bronchitis, asthma*
v_sg	a noun denoting a group of people but taking a singular verb	*government*, *team* (especially in American English)
v_sg_pl	a noun denoting a group of people and taking a singular or plural verb	*government*, *team* (especially in British English)

Table 6: Inherent grammar: nouns

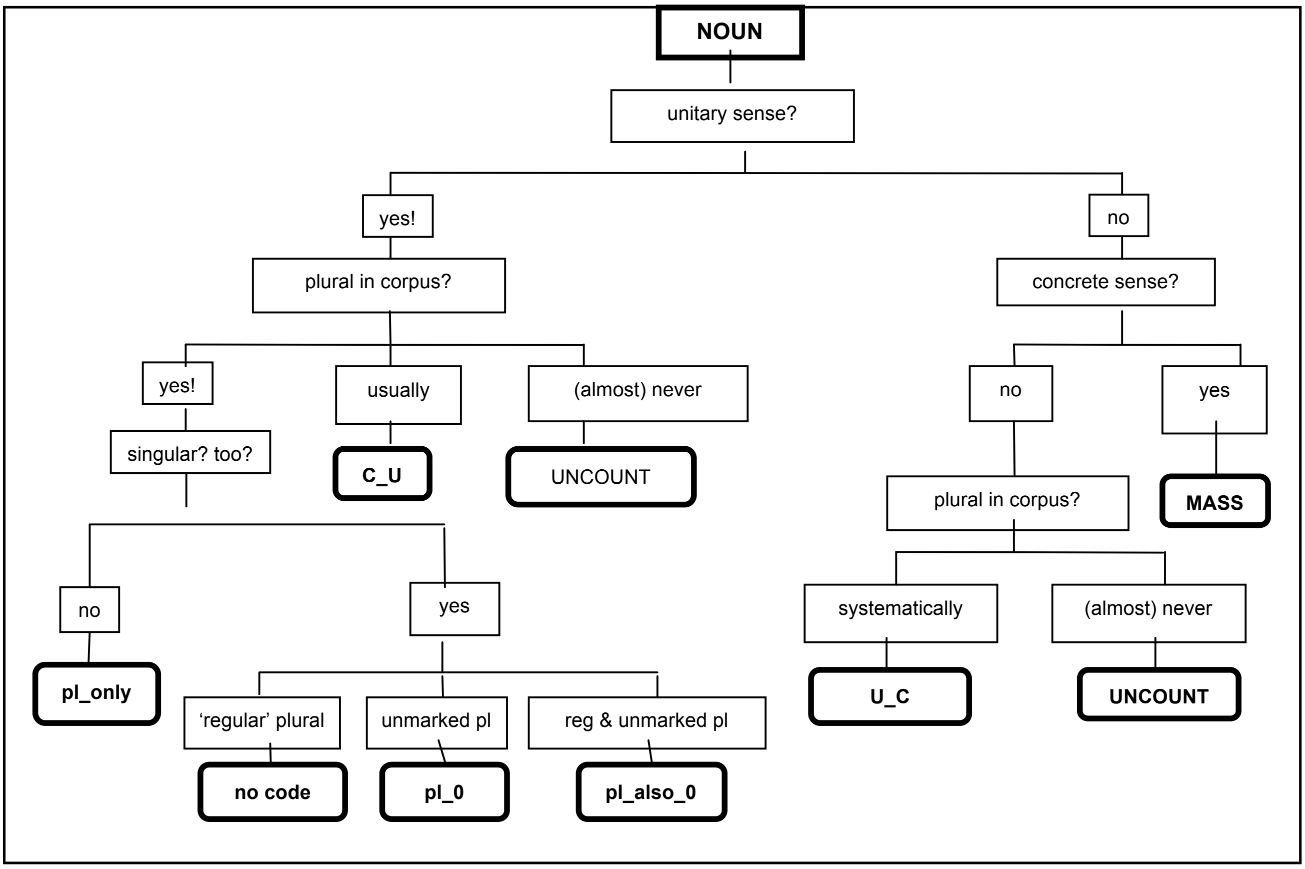

Table 7: Noun flowchart (for determining inherent grammar)

5.3 Syntactic context

When searching for an adjective, noun, or verb, you can refine your search by specifying a particular 'syntactic context'. Syntactic context codes are used for describing those syntax patterns (or constructions) which – according to the evidence in our corpus – are associated with a particular lexical unit. Adding the search condition 'syntactic context' brings up a drop-down list which includes all the codes available for the part of speech you are searching on. Syntactic context codes are optional for adjectives and nouns, but all verbs have at least one. The syntactic context codes are explained in Tables 8 (adjectives), 9 (nouns), and 10 (verbs).

In entries viewed on the website, a syntactic context code is preceded by the word **STRUCTURE** (in red). In entries viewed in the database itself, the code is preceded by: STRADJ (for adjective codes), STRN (for nouns), or STRV (for verbs).

Code	Explanation	Examples
AVP_premod	pre-modified by adverbial	*minimally* cooperative, *significantly* different
PP_X	preposition phrase, with named preposition, e.g. PP_X **at**	amazed *at all this*, happy *with what he had*, delighted *for all of you*
PP_X-NP-Ving	preposition phrase + noun phrase + gerund	aware *of him laughing*
PP_X-Ving	preposition phrase + gerund	tired *of living*, interested *in knowing about it*
PP_X-cl_wh	preposition phrase + wh-clause	curious *about where I can find that information*, absorbed *in what she was reading*
PP_for-Vinf_to	copula + for + 'to' infinitive	is it possible *for you to do it*

Code	Explanation	Examples
Vinf_to	'to' infinitive	happy *to know*, eager *to do it*
Ving	gerund	busy *repairing her bicycle*
it-PP_X-Vinf_to	'it' preposition phrase + 'to' infinitive	*it* was necessary *for her to go*
it-Vinf_to	'it' + copula + 'to' infinitive	*it is/seems etc.* imperative *to do…*
it-that_0	'it' + copula + that-clause with or without explicit 'that'	*it is/seems etc.* clear *(that)…*
it-wh	'it' + copula + wh-clause	*it is/seems etc.* clear *why/how etc. …*
that_0	indicative that-clause with or without explicit 'that'	I'm sure *(that) you will understand*
that_0_subj	subjunctive that-clause with or without explicit 'that'	they were insistent *(that) he join in*
wh	wh-clause	curious *where he was*, curious *what to expect*

Table 8: Syntactic context: adjectives

Code	Explanation	Examples
AJ_pert	pre-modified by pertainym adjective	*educational* institution, *chemical* reaction
AVP_post_mod	post-modified by adverb	the journey *home*
N_mod	pre-modified by a noun	*sea* view
N_premod	pre-modifying a noun	rose *petal*
PP_X	preposition phrase, with named preposition, e.g. PP_X **at**	a look *at the screen,* a letter *from home,* an alliance *between the two parties*
PP_X-NP-Ving	preposition phrase + noun phrase + gerund	the thought *of him going*
PP_X-PP_X	preposition phrase x 2	an argument *with John about money*
PP_X-Ving	preposition phrase + gerund	the thought *of going*
PP_X-cl_wh	preposition phrase + wh-clause	questions *about what courses are offered*, concerns *about which to support*
PP_for-Vinf_to	for + 'to' infinitive	his wish *for them to be there*, her anxiety *for him to get better*
Vinf_to	'to' infinitive	his desire *to be present*, her need *to behave well*
if	whether/if clause	the question *whether he would go*
it_constrn	'it' + copula + 'to' infinitive	*it's* a mistake *to think about it*, *it's* fun *to swim in the sea*
that_0	indicative that-clause	the news *(that) he had arrived*
that_0_subj	subjunctive that-clause	their request *(that) he go with them*
wh	wh-clause	the reason *why he left*, the question *when to go*

Table 9: Syntactic context: nouns

Code	Explanation	Examples
AJP	adjective phrase	you seem *sad*, he looks *taller than you*
AVP	adverb phrase	he had aged *badly*, act *responsibly*
NP	noun phrase	I like *him*, I heard *a story*, I dropped *the metal lid*
NP-AJP	noun phrase + adjective phrase	paint *the wall green*, we found *it very dull*
NP-AVP	noun phrase + adverb phrase	allow *them through*, they floated *it downstream*
NP-NP	noun phrase x 2	crown *him king*, show *me your essay*, give *her a book*, I cooked *her a curry.*
NP-PP_X	noun phrase + preposition phrase	change *the colour to white*
NP-PP_X-Ving	noun phrase + preposition phrase + gerund	admire *somebody for doing*
NP-Part	noun phrase + directional or locative particle	push *it away/into the room*, put *it down/under the table*
NP-Part_X	noun phrase + specific particle	push *it in*, lever *it up* gently
NP-V_ptp	noun phrase + past participle	get *your hair cut*, have *the house valued*
NP-Vinf	noun phrase + bare infinitive	make *him leave*, she let *him go*
NP-Vinf_to	noun phrase + 'to' infinitive	we want *you to leave, she dared him to do* it
NP-Ving	noun phrase + gerund	she watched *the children playing,* I heard *him leaving*
NP-cl_that_0	noun phrase + that-clause	tell her *(that) he's here*
NP_refl	reflexive	cross *oneself*
NP_refl-PP_X	reflexive + preposition	abandon *oneself to a life of pleasure*, dedicate *oneself to a cause*
NP_refl-PP_X-Ving	reflexive + preposition + gerund	dedicated *himself to caring* for her
NP_refl-Vinf_to	reflexive + 'to' infinitive	he wouldn't demean *himself to apologise*
NP_refl-Ving	reflexive + gerund	they enjoyed *themselves swimming*
PP_X	prepositional phrase	they looked *at the screen*, that depends *on the situation*, we thought *of you*
PP_X-NP-Ving	prepositional + noun phrase + gerund	we objected *to him going*
PP_X-PP_X	prepositional phrase x 2	argued *with John about money*
PP_X-Vinf_to	prepositional + 'to' infinitive	I would prefer *for him to go*, they looked *to him to do* it
PP_X-Ving	prepositional phrase + gerund	don't insist *on doing* it, I thought *of going*
PP_X-cl_wh	prepositional phrase + wh-clause	he enquired *about which train I was taking*, it depends *on what you mean*

Code	Explanation	Examples
Part	directional or locative particle	Run *away / into the room*, sit *there / in that chair*
Part_X	specific particle	work *away*, chatter *on*
Quo	quote (direct speech)	*'Get out of here!'* she shouted, he mumbled, *'Why should I?'*
Quo-NP	quote + noun phrase	'*You said you would*,' she reminded him
Vinf	bare infinitive	he dared not *do* it
Vinf_to	'to' infinitive	I love *to visit* them, I tried *to go*
Ving	gerund	she likes *ironing*, I hate *washing* dishes
_0	no complement	*it disappeared, she shouted*
if	whether/if clause	she asked *if I knew,* he wondered *whether she'd found it*
it_constrn	'it' construction	*it seemed there was a mistake, it rests with him to do this*
subj_NP	specifying the noun subject of the headword	(only when needed for sense distinction) *mountain* looms, *disaster* looms
that_0	indicative that-clause	I hear *(that) he's arrived*
that_0_cond	conditional that-clause	he wishes *(that) she would go away*
that_0_subj	subjunctive that-clause	they demanded *(that) he obey them*
wh	wh-clause	I forgot *what I'd said*, she guessed *when you had arrived*, I know *how you feel*
wh-Vinf_to	wh-word + 'to' infinitive	I didn't know *what to say*, watch *how to do it*

Table 10: Syntactic context: verbs

5.3.1. The use of 'X' in syntactic context codes

You will notice that many of the codes include an 'X'. This is used in codes which include a particle or preposition, and the X indicates that a *particular* particle or preposition is required. All codes which include an X are followed by a specific preposition or particle.

For example, the pattern illustrated in the sentence:
It takes a while to ***acclimatise to*** *the humid conditions*
is coded as PP_X (preposition phrase with named preposition) and is followed by '**to**'.

5.3.2. The use of 'Part' in syntactic context codes

The verb codes include four which contain 'Part'. These are:

- Part
- NP Part
- Part_X
- NP Part_X

Part' is used for recording the use of a particle (an adverb or preposition) in a verb which is not a phrasal verb. For example, the verb *amble* can occur in sentences like 'we ambled *across* the wide lawn' or 'they were ambling *along*', and in cases like this, the verb is coded as 'Part'. The addition of 'NP' indicates that a noun phrase comes after the verb and before the particle (as in 'roll it *across* the floor/*up* the hill'). In these codes, various particles are possible, and the examples will show a range of typical instantiations.

The addition of 'X' (cf. 5.3.1. above) indicates that *one specific* particle is required (and the particle is named). For example 'we all filed *in*' (coded 'Part_X in') or 'nail the planks *together*' (coded 'Part_X together'). Where there is no 'X' (in other words, where a range of particles is possible), the code is followed by a 'subcode' which indicates whether the particles show direction (Part_dir) or location (Part_loc). For example, the verb *rush* includes the code 'Part' and subcode 'Part_dir', to cover uses such as 'she rushed *past* me/rushed *up* to him/rushed *into* the room'.

6. Inflections in DANTE

Inflections are not shown explicitly in the Dante database. Lexicographers were briefed to include different forms of nouns or verbs in the example sentences, including:

- singular and plural forms of nouns
- singular and plural uses of nouns with 'zero

plural' forms

- a range of verb forms

Irregular inflected forms (such as *children, mice, addenda, throve*) have their own (short) entry in the database, consisting of a cross-reference to the base form of the lemma.

7. Labels in DANTE

Labels are applied to any item in the database which is not part of the core, 'unmarked' vocabulary of English. Items which – on the evidence of the corpus – are characteristic of a particular text-type or speech community will attract a label. Such items may include entire headwords or lower-level components like grammar codes or examples. but labels are most frequently associated with individual lexical units.

DANTE has six categories of label, allowing us to mark any item according to:

- attitude
- regional variety
- register
- style
- time
- subject field (or domain)

In the entries you see as the result of a search on the website, labels appear in italics within (italicised) square brackets: for example *[offens], [AmE], [journ]*. The exception is domain (or subject-field) labels: see 7.6.

7.1 Attitude

Three labels are available for indicating the attitude of the speaker or writer:

- apprec ('appreciative'), as in: *tireless, slender*
- offens ('offensive'), as in *half-caste, poof*
- pej ('pejorative'), as in *moralize, paltry, bimbo*

7.2 Regional variety

Some of the regional-variety labels in DANTE are idiosyncratic and require explanation. The DANTE database was developed for Foras na Gaeilge as a launchpad for its *New English-Irish Dictionary*, and this has implications for vocabulary coverage and labelling. On the one hand, many major World Englishes (such as South African English or Indian English) are not described systematically (though high-profile usages from any variety will be covered). On the other hand, DANTE provides extensive coverage of Hiberno-English (the variety of English spoken in Ireland), and the corpus resources used by DANTE lexicographers included a specially-created 25-million-word corpus of Hiberno-English.

Consequently, the label 'BrE' (British English) has a different application in DANTE than in standard dictionaries. Conventionally, a 'BrE' label is applied to usages typical of the British Isles as a whole (including Ireland), in contrast to 'AmE' usages. But in DANTE, such usages are labelled 'non_AmE', while 'BrE' is reserved for items typical of usage in mainland Britain but not found in Hiberno-English. Conversely, items specific to Hiberno-English and not familiar to British speakers are labelled 'HibE'.

Examples of items labelled 'BrE' in DANTE include *barnet* (someone's hair or hairstyle), *comp* (a comprehensive school) and *stopping train* (a slow train that stops at every station on the route); none of these is used in Hiberno-English. Examples of items labelled 'HibE' include *brutal* (in the sense of 'awful')*, eejit* (idiot), and *deadly* (very fine or attractive).

The complete set of regional labels in DANTE is:

- AmE: American English
- non_AmE: explained above
- BrE: explained above
- HibE: Hiberno-English
- AusE: Australian English
- Scot: Scottish English

7.3 Register

The four register labels in DANTE are used for indicating the formality level of the labelled item:

- fml ('formal'): words, senses, or expressions that are typical of formal usage: examples include *admonitory, remediable, munificence, ameliorate*
- inf ('informal'): examples include *bolshie, hellish, megabucks, on the razzle*
- vinf ('very informal'): while 'informal' uses are found in a wide range of text types, this label denotes items found only in very informal discourse, typically between people who know each other well or belong to the same social grouping. This equates to what many dictionaries label as 'slang' (a contentious term). Examples include: *charlie* (cocaine), *pants* (as an adjective: very bad), *fanny about* (mess around).
- vulg ('vulgar'): vulgar uses may cause offence and equate to what some dictionaries label as 'taboo' (which we see as an outmoded term and concept). Examples include the familiar four-letter words, and items like *motherfucker, prick, shitless* (as in 'scared/bored shitless').

7.4 Style

7.4.1. Labels for foreign borrowings

The labels in this category are used to mark borrowings that remain noticeably foreign and are often not pronounced in an English manner. Well-integrated borrowings such as *macaroni* are not labelled – though, as is well known, the boundary between the two types is impossible to draw precisely.

The following labels are available:

- Fr (French): *mangetout, ménage à trois, objet d'art*
- Ger (German): *schadenfreude, Zeitgeist*
- Lat (Latin): *modus operandi, non sequitur, obiter dictum*
- Span (Spanish): *mojito, mestizo, paso doble*
- For (any borrowing not covered by the labels above): *feng shui, edamame, perestroika*

7.4.2. Other style labels

This is a somewhat heterogeneous category, and these labels often occur in combination with register, attitude, or domain labels. It is important to stress the difference between a particular *style* of speech or writing and the *domain* which a text belongs to (see 7.6 for domain labels). For example, words like *abatement, predecease* and *heretofore* belong to a legal 'style' of writing ('legalese') and get the style label 'leg'; words like *alibi, foreman* (of a jury), and *bail* are words belonging to the subject field of law, so get the domain label 'Law'.

The following labels are available:

- TM (trademark): *Blu-ray, frisbee, Portakabin, Prozac*
- child (child language): *grown-up, poo*
- drugs (drug-users' slang): *mainline, charlie, re-up*
- euph (euphemism): *nether regions* (genitals), *economical with the truth*, *pass away*
- hum (humorous): *mugshot, nookie, bridezilla*
- iro (ironic): *princely* (sum), *dulcet* (tones) *pearl of wisdom*
- journ (journalese): *beleaguered, probe, wed*
- leg (legalese): *abatement, predecease, heretofore*
- lit (literary): *bounteous, morrow, asunder*
- pc (politically correct language): *person of colour, challenged* (mentally, visually etc)
- prov (proverb): *pride comes before a fall, too many cooks*
- spok (spoken: rarely found in written English): *anyways, and your point is? bro*
- tech (technical usage: often used in combination with a domain label): *macromolecular, meiosis, anisotropic*
- youth (young people's slang): *rad, boyf, respect!*

7.5 Time

DANTE is essentially synchronic, but it includes some items which are in the process of losing their currency and are now rarely heard (labelled 'dat', or dated), and others that are virtually never found in contemporary discourse (labelled 'obs', or obsolete). The latter are included only if users are likely to come across them in classic works of literature.

Examples include:

- dat: *poppycock, betrothed, blotto*
- obs: *apothecary, pox, dropsy*

7.6 Domain

One of DANTE's most valuable features is its extensive use of domain (or subject-field) labels. The editorial team had available to them **156** domain labels, and were encouraged to apply them whenever appropriate. In the entries you see as the result of a search on the website, domain labels appear in capitals within square brackets, in the form of (usually) self-evident abbreviations: thus [BOT] indicates an item from the domain of botany, and [SOCIOL] an item typical of texts about sociology.

IN DANTE's Advanced Search mode, you can search for lexical units with a domain label by clicking the item 'subject field' in the left-hand dropdown list. The right-hand drop-down list includes the labels themselves. Note, however, that this list does not include the full set of 156 domain labels. Instead, it provides a subset of **28** domains, most of which act as superordinates and subsume many related domains. (Obviously, licensed users of the full database can search for any of the 156 domain labels.) Table 11 shows the 28 labels available for searches on the website, and indicates any other labels which these subsume.

Label in drop-down list	**Additional labels this covers**
Agriculture	botany, horticulture
Art	ceramics, fashion, photography
Artifacts	clothing, cosmetics, furniture, tools.
Calendar	(none: this label covers items like days of the week and names of festivals)
Colours	(none: this label is applied to all colour terms)
Communication	telecommunication
Culinary	(none: this label covers all cooking vocabulary)
Education	(none)
Engineering	aerospace, automotive , chemical engineering, civil engineering, electrical engineering, machinery, mechanical engineering, mining

Finance	accountancy, economics, finance, insurance, tax
Government	(none)
Humanities	architecture, astrology, heraldry, history, sociology, philosophy, mythology
Industry/ Employment	business administration, commerce, construction, fishing, hair-dressing, plumbing, publishing, public relations, surveying, textiles, tourism, transport
IT	(none)
Law	police
Leisure	climbing, collecting, darts, do-it-yourself, table games
Linguistics	(none)
Literature	(none)
Mathematics	measurement units, statistics
Medicine	anatomy, health & fitness, pharmacology, physiology , psychology/psychiatry
Military	air force, army, navy, weaponry
Music	(none)
Nautical	(none)
Performing arts	cinema, theatre, dance
Politics	(none)
Religion	(none)
Science	anthropology, archaeology, astronomy, biochemistry, biology, chemistry, cosmology, dentistry, ecology, electronics, genetics, geography, geology, insects, meteorology, mineralogy, optics/ophthalmology, physics, veterinary science, viticulture, zoology
Sport	American football, archery, athletics, badminton, baseball, basketball, bowls, boxing, cricket, curling, cycling, equitation, fencing, football, Gaelic football, golf, gymnastics, hockey, hurling, horse-racing, hunting, ice hockey, ice-skating, lacrosse, martial arts, motor racing, polo, rowing, rugby, sailing, shooting, softball, surfing, swimming, table tennis, tennis, water sports, windsurfing, winter sports, wrestling

Table 11: Domain (or subject-field) labels

8. Multiwords in DANTE

This section explains how multiword units and expressions are treated in DANTE. Many different types of word+word combination were recorded by the lexicographers using the tags PHRASE, CHUNK, COLLOC, CPD (compound), ITEM (itemiser), and SUPPREP (support preposition). These elements are not searchable using the web interface, but they may appear in the entries returned by a search, so they are explained here.

Phraseology is a ‘scalar’ feature of language. Multiword combinations encompass a huge range of language events, from fixed, opaque idioms (‘for good measure’) to completely open combinations (‘a large house’). DANTE has a number of strategies for recording such items, but the boundaries between each type are not always clear. So the question of where a given combination fits best often comes down to editorial judgment. For example, the recurrent string ‘fit/match/answer a description’ is recorded in DANTE as a ‘phrase’ (a distinct LU with its own definition) but could, arguably, have been treated as a ‘chunk’ (8.2.1). In this case, and many similar cases, there is no ‘right’ answer.

8.1 Phrases, phrasal verbs, and compounds

Idiomatic phrases, phrasal verbs (5.1.3), and compounds are ‘nested’ : that is, they are handled in the entry for the lemma to which they are related: for compounds and phrasal verbs, the first word; for phrases, the first ‘lexical’ word. They appear after the LUs (or senses) of the base form, in a section called (in the database) the ‘Multiword Expression Block’ (or MWEBlk). Thus at the lemma *map*, the eight LUs of *map* itself (four noun senses and four verb senses) are followed (in this order) by:

- two phrases (*off the map, on the map*)
- one phrasal verb (*map out*)

- four compounds (*map maker, map projection, map reading, map reference*)

Each of these is an LU in its own right (so it has its own POS, labels, grammar codes, examples etc). In some cases items of this type consist of more than one lexical unit: the phrase *off the map*, for example, has three separate LUs, each with their own definitions and examples. In the entries shown on the DANTE website, each section (for phrases, phrasal verbs, and compounds) is signalled by a heading in red.

8.2 Multiwords which are not lexical units

DANTE records many other recurrent multiword strings which do not have the status of full lexical units. In addition to 'phases', several other options are available. These are:

- chunks
- collocations
- support verbs
- support prepositions
- itemisers

8.2.1. Phrase or Chunk?

A phrase is a full lexical unit, and (like any LU) requires a definition. Chunks, on the other hand, appear within an LU and do not have their own definitions.

Phrases in DANTE are non-transparent combinations, whose meaning or communicative function cannot be inferred from its components. Phrases span a range of types, from the stereotypical idiom *kick the bucket* (completely opaque), to cases where one or more of the component words has one of its 'usual' meanings, but the meaning of the phrase is still not retrievable: thus *look daggers at* does involve 'looking at' someone, but it is nevertheless not wholly transparent: it therefore needs a definition, and it therefore qualifies as a discrete LU to be treated as a phrase.

The category 'chunk' was introduced for a combination which is non-idiomatic, but frequent enough in the corpus to be worth recording as a significant fact about the lemma it belongs to. In its *form*, a chunk has some of the features of a phrase, in that the selection of words may be idiosyncratic. But its meaning is readily deducible from its component parts: hence it needs no definition, and hence it does not qualify as a separate LU. Whereas phrases appear in the 'Multiword Expression Block' at the end of an entry, chunks are covered in the LU whose meaning they invoke.

Examples of chunks include:

- *go into administration* (at the LU of *administration* that refers to the disposition of insolvent companies etc)
- *by/on one's own admission*
- *out of deference to*
- *on a daily basis*
- *decide for oneself*

8.2.2. Collocation

A collocation is a two-word combination consisting of the lemma and another lexical word (a 'collocate') with which the lemma frequently occurs. For example at the lemma *pool*, the adjective collocates *outdoor, indoor, heated*, and *private* are listed at the noun LU referring to a swimming pool; and the noun collocates *resources, funds*, and *data* are recorded as typical objects of the verb LU meaning 'to put things together for collective use'.

In the entries shown on the website, collocates are listed (following the word COLLOCATES in red), and then provided with examples. Using the 'Word Sketch' function in the Sketch Engine corpus query system, DANTE lexicographers identified and recorded the most frequent collocates for each LU, and the result is a rich and systematic account of collocation in English.

8.2.3. Support verbs

A support verb is a 'light' verb in a verb+noun combination in which the verb makes little semantic contribution. DANTE recognises five support verbs: *do, give, have, make, take*. A support verb+noun combination typically paraphrases the verb cognate of the noun. For example, the combination <u>take</u> *a walk* (where *take* is the support verb) is equivalent to the verb *to walk*. Other examples include:

- *do the packing*
- *give a salute*
- *have a quarrel*
- *make a promise*
- *take a shower*

Support verbs are one of the search parameters available in the Advanced Search mode on the website.

8.2.4. Support prepositions

A support preposition is a preposition which frequently occurs *directly before* a noun. For example, when *peace* refers to 'freedom from disturbance', it is frequently found in sentences like: 'a place where you can do your work *in peace*' or 'spaces where people can walk and cycle *in peace*, away from traffic'. At the relevant LU of *peace*, the support preposition 'in' is shown after the words SUPPORT PREP (in red). Other examples include:

- *at rest*
- *in hysterics*
- *on vacation*
- *by helicopter*

8.2.5. Itemisers

An itemiser is a word that is used in conjunction with a concrete noun to instantiate the idea of 'a piece [of the noun]'. Itemisers are recorded in DANTE when corpus data indicates they are frequent. Examples of itemisers include:

- *a drop of blood*
- *a head of broccoli*
- *a piece/item/article of clothing.*

Different LUs of the same lemma may have different itemisers: thus for the 'mass' sense of *chocolate*, a common itemiser is 'bar', while for its countable use ('small sweet or piece of candy'), 'box' is often used.

9. Miscellaneous

There are four other fields which are not available as search conditions on the website, but may appear in entries retrieved by a search:

- corpus pattern
- pragmatics
- link
- xref (cross reference)

In all cases, the field is shown in red (in website entries), and the relevant information follows. These are explained here.

9.1 Corpus pattern

Corpus data sometimes reveals recurrent features of a word's behaviour which are not covered by any of the grammatical or phraseological categories described above. The corpus pattern field is used for recording such information. For example, the verb *gall* often appears in 'cleft' sentences such as:
–What galled me even more was her insistence that…
–But what galls many motorists more is the fact that…

Since this is clearly a characteristic feature of the verb, it is recorded in the corpus pattern field as: 'often in cleft sentences'.

Other examples include:

- *gag*: 'always in progressive' (*we were gagging for a drink*)
- *abide*: 'usually in negative or broad negative environment' (*I never could abide lobsters*)
- *demote*: 'often passive' (*he was demoted in the cabinet reshuffle*)

9.2 Pragmatics

Where appropriate, DANTE records information about a word's pragmatics: for example, the connotations of a word, or what it tells us about the attitude of the speaker. Sometimes, this can be conveyed through the use of attitude labels (such as 'apprec' or 'pej': 7.1) or style labels (like 'euph' or 'hum': 7.4.2). In other cases, none of the available labels is adequate, so the pragmatics field is used. Examples include:

- *charm*: 'can sometimes have connotations of manipulation'
- *constant* : 'often expresses annoyance'
- *micromanage*: 'expresses disapproval'

9.3 Links and XRefs

A 'link' is used to cross-refer to another item within the same entry. Most typically, links are used to point to a multiword expression (phrase, compound, or phrasal verb) whose meaning is closely related to the LU where the link appears. The 'link' field exists primarily to alert other lexicographers or translators using the database that there is an item relating to that LU further down in the same entry. Examples include:

- *access*: the LU referring to access to computer data or the Internet includes a link to the related compound *access provider*
- *allowance*: the LU meaning 'the fact of taking something into account' includes a link to the phrase *make allowances for*
- *auction*: the verb LU includes a link to the phrasal verb *auction off*

A 'xref' is used to cross-refer from one headword to another. The most common type of xref is from an 'empty' entry to the main entry (e.g. from *center* to *centre*). Otherwise, since most of the information in DANTE is machine-retrievable, explicit cross-references are used only sparingly.

10. References

Atkins, B.T.S. Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.