The DANTE Database

What it is, how it was created, and what it can contribute to dictionaries and lexicons of the future

Michael Rundell and Sue Atkins
Lexicography MasterClass
Outline

☐ What’s in DANTE

☐ Lexicographic innovations
  ■ Project planning, software customization, quality control

☐ Availability of DANTE
  ■ website access, research, and commercial licences

☐ The value of DANTE
  ■ for publishers, NLP applications, linguists …
  ■ for e-lexicography
Origins of DANTE: reminder

Foras na Gaeilge

a modern bilingual English-Irish dictionary

Phase 1: infrastructure

Phase 2a: source language analysis

Phase 2b: translations into database

Phase 2c: dictionary editing

New English-Irish Dictionary
Progress report on NEID

☐ English framework: complete (=DANTE)
☐ Phase 2b Translations: 80% done
☐ Phase 2c Editing, entry finalisation: 20% done (on track)
☐ First publication, end 2012  
  ■ electronic dictionary  
  ■ highest-frequency lemmas plus 10% of rest
What’s in DANTE: some numbers

- 42,000 headwords
- 27,000 phrases
- 23,000 compounds and phrasal verbs
- Rich syntactic and collocational analysis:
  - 42 patterns for vbs, 16 for nouns, 15 for adjs
  - 133,000 collocates
- 156 domain (subject-field) labels
- **622,000** corpus examples
Richness of data

- *recollect*
  - medium-to-low frequency word
    - not in BNC’s top 10,000 lemmas (BNC’s 9999th lemma has 362 hits, *recollect* has 220)
  - sense 1 (of 2) has
    - 7 ‘structures’ (aka constructions, syntax patterns)
    - 26 corpus examples
Richness of data

- **peace**
  - high frequency word
    - ranked 1145\textsuperscript{th} in BNC
    - 6 senses, 6 phrases, 17 compounds
  - collocates (sense 1 only)
    - 8 adj collocates
    - 7 verb collocates
    - 4 noun collocates (where *peace* is modifier)
    - ...and 20 corpus examples
Lexicographic innovations

- Project planning: classification of headwords
- Streamlining the compilation process:
  - Customizing the software (CQS and DWS)
  - Using ‘Proforma’ entries
- Editing text: a new approach to quality control
New classification of headwords: method

- Headwords assigned to one of 12 levels, based on complexity
  - level 1: single-sense, referential (e.g. gastropod)
  - levels 10-12: big, complex words (light verbs, grammar words, etc)

- Sources
  - existing dictionaries, past experience, theoretical work (FrameNet, Regular Polysemy etc)

- Discussed in Atkins & Grundy 2006.1097ff
Headword classification: goals and outcomes

☐ Goals
  ■ improve scheduling (not an exact science)
  ■ tailor work-packs to editors’ skills/preferences

☐ Outcomes: did it work?
  ■ for scheduling: *moderately* successful – but no existing dictionary had comparable granularity
  ■ for future projects: information in database a more reliable guide to scheduling
Software customization

Goals and outcomes

- completeness: ensure nothing is missed (in terms of syntax, collocation, labelling ...)
- ease of use: let lexicographers focus on the hard parts
- maximize efficiency → time savings
Software customization

- Methods
  - Harmonize Sketch Grammar and Dictionary Grammar
    - exact match between grammatical relations shown in Word Sketch and those in dictionary
  - ‘Constructions’ list as top-level summary
  - GDEX: automated example-finding
  - One-click copying
**Word Sketch showing ‘constructions’**

**decide** *(verb)*  
LEXMCI freq = 322076 (187.2 per million)

<table>
<thead>
<tr>
<th>Constructions</th>
<th>PP_X</th>
<th>34423</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vinf_to</td>
<td>132188</td>
<td>19.2</td>
</tr>
<tr>
<td>that_0</td>
<td>47886</td>
<td>8.3</td>
</tr>
<tr>
<td>wh</td>
<td>30798</td>
<td>10.3</td>
</tr>
<tr>
<td>if</td>
<td>22193</td>
<td>55.6</td>
</tr>
<tr>
<td>NP_Vinf_to</td>
<td>7175</td>
<td>4.3</td>
</tr>
<tr>
<td>it_constrn</td>
<td>5441</td>
<td>30.7</td>
</tr>
<tr>
<td>PP_for_Vinf_to</td>
<td>5374</td>
<td>51.4</td>
</tr>
<tr>
<td>PP_Vinf_to</td>
<td>5374</td>
<td>273.0</td>
</tr>
<tr>
<td>wh_Vinf_to</td>
<td>5299</td>
<td>154.6</td>
</tr>
</tbody>
</table>

| PP_Vinf_to        | 5374   | 273.0 |
|                  | whether | 5341  | 9.6   |
|                  | whether | 5341  | 9.6   |

| PP_for_Vinf_to    | 5374   | 51.4  |
|                  | whether | 5341  | 9.6   |

| subj_NP           | 88907  | 4.3   |
| judge             | 611    | 6.61  |
| court             | 1082   | 6.45  |
| committee         | 802    | 6.17  |
| government        | 2588   | 6.16  |
| jury              | 222    | 5.9   |
| Committee         | 1159   | 5.86  |
| Court             | 609    | 5.75  |
| tribunal          | 211    | 5.74  |
| council           | 586    | 5.68  |
| adjudicator       | 138    | 5.53  |
# PPs in Word Sketch

**argue** *(verb)*  \( \text{LEXMC freq} = 134412 \) (78.1 per million)

<table>
<thead>
<tr>
<th>PP_for-i</th>
<th>6246</th>
<th>5.3</th>
<th>PP_in-i</th>
<th>2207</th>
<th>1.0</th>
<th>PP_with-i</th>
<th>2025</th>
<th>2.7</th>
<th>PP_against-i</th>
<th>1879</th>
<th>33.6</th>
<th>PP_about-i</th>
<th>979</th>
<th>7.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>abolition</td>
<td>27</td>
<td>6.04</td>
<td>favour</td>
<td>429</td>
<td>8.27</td>
<td>umpire</td>
<td>16</td>
<td>5.97</td>
<td>notion</td>
<td>28</td>
<td>4.05</td>
<td>merit</td>
<td>15</td>
<td>3.99</td>
</tr>
<tr>
<td>centrality</td>
<td>9</td>
<td>5.25</td>
<td>favor</td>
<td>75</td>
<td>7.91</td>
<td>referee</td>
<td>44</td>
<td>5.57</td>
<td>proposition</td>
<td>14</td>
<td>3.92</td>
<td>validity</td>
<td>5</td>
<td>3.23</td>
</tr>
<tr>
<td>interpretation</td>
<td>85</td>
<td>4.94</td>
<td>Prospect</td>
<td>5</td>
<td>4.37</td>
<td>ref</td>
<td>22</td>
<td>5.44</td>
<td>ban</td>
<td>15</td>
<td>3.01</td>
<td>meaning</td>
<td>15</td>
<td>2.22</td>
</tr>
<tr>
<td>retention</td>
<td>29</td>
<td>4.76</td>
<td>pamphlet</td>
<td>8</td>
<td>3.91</td>
<td>idiot</td>
<td>15</td>
<td>5.33</td>
<td>hypothesis</td>
<td>8</td>
<td>2.95</td>
<td>existence</td>
<td>7</td>
<td>1.79</td>
</tr>
<tr>
<td>legalisation</td>
<td>6</td>
<td>4.63</td>
<td>Guardian</td>
<td>22</td>
<td>3.38</td>
<td>eloquence</td>
<td>5</td>
<td>5.25</td>
<td>existence</td>
<td>14</td>
<td>2.77</td>
<td>definition</td>
<td>10</td>
<td>1.59</td>
</tr>
<tr>
<td>superiority</td>
<td>10</td>
<td>4.52</td>
<td>court</td>
<td>91</td>
<td>3.23</td>
<td>justification</td>
<td>14</td>
<td>4.25</td>
<td>adoption</td>
<td>6</td>
<td>2.57</td>
<td>composition</td>
<td>5</td>
<td>1.44</td>
</tr>
<tr>
<td>primacy</td>
<td>6</td>
<td>4.52</td>
<td>chapter</td>
<td>33</td>
<td>3.18</td>
<td>interviewer</td>
<td>5</td>
<td>3.88</td>
<td>motion</td>
<td>14</td>
<td>2.42</td>
<td>politics</td>
<td>9</td>
<td>1.22</td>
</tr>
<tr>
<td>reassessment</td>
<td>6</td>
<td>4.51</td>
<td>past</td>
<td>31</td>
<td>3.17</td>
<td>passion</td>
<td>10</td>
<td>2.81</td>
<td>thesis</td>
<td>5</td>
<td>2.05</td>
<td>religion</td>
<td>8</td>
<td>1.07</td>
</tr>
<tr>
<td>rethink</td>
<td>5</td>
<td>4.3</td>
<td>vein</td>
<td>7</td>
<td>3.15</td>
<td>logic</td>
<td>10</td>
<td>2.78</td>
<td>penalty</td>
<td>10</td>
<td>1.83</td>
<td>extent</td>
<td>6</td>
<td>0.91</td>
</tr>
<tr>
<td>necessity</td>
<td>20</td>
<td>4.2</td>
<td>essay</td>
<td>18</td>
<td>3.14</td>
<td>Him</td>
<td>7</td>
<td>2.51</td>
<td>belief</td>
<td>13</td>
<td>1.71</td>
<td>bill</td>
<td>7</td>
<td>0.39</td>
</tr>
<tr>
<td>importance</td>
<td>89</td>
<td>4.16</td>
<td>circle</td>
<td>17</td>
<td>2.65</td>
<td>anybody</td>
<td>5</td>
<td>1.92</td>
<td>identification</td>
<td>6</td>
<td>1.69</td>
<td>direction</td>
<td>8</td>
<td>0.38</td>
</tr>
<tr>
<td>inclusion</td>
<td>39</td>
<td>4.15</td>
<td>Independent</td>
<td>13</td>
<td>2.55</td>
<td>neighbour</td>
<td>5</td>
<td>1.76</td>
<td>proposal</td>
<td>27</td>
<td>1.64</td>
<td>importance</td>
<td>6</td>
<td>0.32</td>
</tr>
<tr>
<td>unity</td>
<td>22</td>
<td>4.13</td>
<td>Brussels</td>
<td>5</td>
<td>2.48</td>
<td>someone</td>
<td>28</td>
<td>1.49</td>
<td>interpretation</td>
<td>8</td>
<td>1.6</td>
<td>everything</td>
<td>9</td>
<td>0.16</td>
</tr>
<tr>
<td>impossibility</td>
<td>5</td>
<td>4.11</td>
<td>defence</td>
<td>19</td>
<td>2.36</td>
<td>God</td>
<td>31</td>
<td>0.99</td>
<td>inclusion</td>
<td>6</td>
<td>1.54</td>
<td>merit</td>
<td>15</td>
<td>4.01</td>
</tr>
<tr>
<td>existence</td>
<td>37</td>
<td>4.08</td>
<td>paper</td>
<td>85</td>
<td>2.33</td>
<td>anyone</td>
<td>17</td>
<td>0.88</td>
<td>view</td>
<td>52</td>
<td>1.51</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>boycott</td>
<td>7</td>
<td>3.98</td>
<td>Chapter</td>
<td>10</td>
<td>2.31</td>
<td>official</td>
<td>22</td>
<td>0.87</td>
<td>possibility</td>
<td>11</td>
<td>1.41</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
GDEX: ‘best’ examples promoted

supported runaway slaves, others publicly argued for abolition of the trade. Advertising abolish them. They are right. We are not arguing for the abolition of patents. We are arguing lectures on the immorality of slavery. Knight argued for the immediate abolition of slavery financial autonomy is necessary. UNISON argues for the total abolition of capping powers boldly breaking the negotiating rules by arguing for abolition. Negotiating with the enemy illegally. Some people claim it is utopic to argue for the abolition of all controls that Central Bank. Others, like ourselves, would argue for the abolition of the former and placing not in the USA: There is a strong case for arguing for the abolition of such taxes. The topic Indeed the London stock exchange is already arguing for the abolition of stamp duty on the the Council’s significant achievement in arguing for the abolition of the allocation fee react accordingly. The CIPD continue to argue for the abolition of mandatory retirement while 30% wanted radical reform; a mere 12% argued for outright abolition. Surprisingly, perhaps example, a workshop on ‘new economic policy’ argued for the abolition of the stability pact unacceptable. The Association has published a paper arguing for the abolition of the Scheme and has academics and copyright terrorists that argue for the abolition of copyright and for Jahangir and other human rights activists have argued for the abolition of the blasphemy laws range of commissions. Therefore, while not arguing for abolition, we would favour a re-examination.
One-click copying: example and XML mark-up links straight to DWS

supported runaway slaves, others publicly argued for abolition of the trade. Advertising abolish them. They are right. We are not arguing for the abolition of patents. We are arguing lectures on the immorality of slavery. Knight argued for the immediate abolition of slavery.

financial autonomy is necessary. UNISON argues for the total abolition of capping powers boldly breaking the negotiating rules by arguing for abolition. Negotiating with the enemy illegal. Some people claim it is utopic to argue for the abolition of all controls that Central Bank. Others, like ourselves, would argue for the abolition of the former and placing not in the USA: There is a strong case for arguing for the abolition of such taxes. The topic Indeed the London stock exchange is already arguing for the abolition of stamp duty on the Council’s significant achievement in arguing for the abolition of the allocation fee react accordingly. The CIPD continue to argue for the abolition of mandatory retirement while 30% wanted radical reform; a mere 12% argued for outright abolition. Surprisingly, perhaps example, a workshop on 'new economic policy' argued for the abolition of the stability pact unacceptable. The Association has published a paper arguing for the abolition of the Scheme and has academics and copyright terrorists that argue for the abolition of copyright and for Jahangir and other human rights activists have argued for the abolition of the blasphemy laws range of commissions. Therefore, while not arguing for abolition, we would favour a re-examination -plus In a Guardian article, Fiona Millar argues for the outright abolition of the 11-plus
state an opinion and give reasons for
it STRUCTURE it_constrn CORPUS
PATTERN passive
  • It can be argued convincingly that the
    Government should recompense people for
    the shoddy guidance given
  • It has been argued that it is wrong for
    such a body to impose its views
STRUCTURE that_0
  • They argue that downsized companies,
    having slimmed down, have made
    themselves more efficient
  • But the 14 other EU countries argued that
    would be unfair
COLLOCATES author, commentator
  • The author argues that these results
    evidence some shortcomings in the
    traditional method
  • Many commentators have argued that this
    dilemma has always existed
COLLOCATES critic, opponent, sceptic
  • Critics argue that private operators will
    sacrifice safety
  • Opponents argued the change would drive
    up real estate prices
  • Sceptics have argued that these results
    cannot be extrapolated
STRUCTURE PP_X for
  • We are not arguing for the abolition of
    patents
  • Mr. Gates argues persuasively for having
    a digital counterpart to the human nervous
    system
Software customization: proformas

- **Background**
  - already standard for ‘obvious’ closed sets (days of week, chemical elements, letters of alphabet)
  - greatly extended for DANTE: 68 proformas set up

- **Informed by phenomenon of ‘regular polysemy’**.
  - e.g. container/contents; tree/wood; qualification/holder

- **Goals: as before**
  - streamline compilation process
  - improve completeness and systematicity
Proformas: *drinks*

- Writing system pre-populated with data
- Sense 1 includes (inter alia)
  - POS and grammar (noun, mass)
  - domain label (DRINK)
  - slots for itemizers (*glass of*, *cup of*, *shot of* etc)
- Sense 2 is for ‘unit/container of’ meaning
  - example guidance (*would you like a beer*, *3 coffees please*, etc)
- Lexicographer’s job: fill in what’s needed, delete what isn’t
Quality control

☐ A two-pronged strategy

☐ Traditional method

☐ Editor scans text, identifies problems and

☐ fixes them

☐ gives feedback to lexicographer

☐ New addition: use of search ‘scripts’ in DPS SkXml search function)
Quality control

Search scripts

- use XML structure to identify any instance of any phenomenon, e.g.
  - `<FwkSenCnt:<(gram@code=mass,<%MEANING),<hwd:(^[m-r].*)>`
  - finds all nouns with GRAM code ‘mass’ in given range
- identify and fix problems, manually or by program

- Almost 200 scripts developed for DANTE project
### Database search: mass nouns in ‘M’

#### SkXml Box

**Evidence finder**

Logged in as: adminidm (administrator administrator, administrator)

```xml
<FwkSenCnt:<gram@code=mass,<%MEANING>,<hwd:(^[m].*)>
```

Details for query: `<FwkSenCnt:<gram@code=mass,<%MEANING>,<hwd:(^[m].*)>` (237 result(s))

Selected corpus: New English Irish Dictionary for LexMC (237 result(s))

Summary | Print | Go to list module | Export

237 matches in 166 documents. Go to page: 1 2 3 4 5 6 7 8 9 10 ... last

<table>
<thead>
<tr>
<th>DocID</th>
<th>label</th>
<th>xml_0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>macadam</td>
<td></td>
</tr>
<tr>
<td></td>
<td>macaroni</td>
<td></td>
</tr>
<tr>
<td></td>
<td>macaroni</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mace</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mace</td>
<td></td>
</tr>
<tr>
<td></td>
<td>machair</td>
<td></td>
</tr>
<tr>
<td></td>
<td>machinery</td>
<td></td>
</tr>
<tr>
<td></td>
<td>machinery</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mackerel</td>
<td></td>
</tr>
</tbody>
</table>
Quality control

☐ Benefits

■ focus on known problems (e.g. mass or uncount?)

■ recurrent, team-wide problem may indicate weak policy, needs changing?

■ identify problems of individual editors

■ quality-control more systematic, less time-consuming
Lexicographic innovations: outcomes

- Towards semi-automated dictionary compilation
  - software populates (parts of) database
  - lexicographer completes, rejects as appropriate

- Processes streamlined, costs controlled:
  - DANTE completed on time, on budget: just under 3 years

- Implications for e-lexicography: see end
Who needs DANTE?

☐ Publishers, lexicographers
  - for new bilingual dictionaries (E → *): ready-made English framework of highest quality
  - for new monolingual English dictionaries: offers huge reductions in origination costs

☐ NLP community
  - word sense disambiguation, grammar checking
  - machine translation
  - information extraction
  - every word sense linked to specific contextual features

☐ Linguists, language teachers, …
Availability of DANTE

- Public website: www.webdante.com
  - free access to complete DANTE entries, with simple and advanced search options, in range M-R

- Register for free A-Z web access
  - personal, research, and teaching purposes only

- Get licence for free download of full database
  - for research groups in univs or not-for-profit institutions

- Licences available for full commercial exploitation

- Go to http://dante.sketchengine.co.uk
webdante.com

- Simple search: finds lemma (*recollect*)
- Advanced searches: find every lemma exemplifying selected information types:
  - POS=adverb, inherent grammar=degree
  - finds all ‘degree’ adverbs
  - POS=verb, syntactic content = Quo
  - finds all verbs used in direct speech
  - etc etc (domain, regional variety, register …)
- Web interface doesn’t include every information-type
  - e.g. itemizers (next)
Itemizers

Itemizers for *ash*

- **Itemizers**
  - heap
    - She wanders about the shore, which is bitterly cold, but then the north wind blows a heap of *ashes* some fishermen had left on the strand into a flame.
  - pile
    - Next morning the farmer came down into the kitchen and to his horror saw a huge pile of *ash* in the hearth of the granite fireplace.
  - layer
    - A layer of *ash*, in places up to a metre thick, was found, indicating the scale of the fire.
  - cloud
    - Rocks were thrown 50 km into the air and

**mass** the grey powder that remains after something has burned

- Christine flicked the *ash* from her cigarette into an empty lager can on the table.
- Please don’t put soil, rubble, hot *ashes*, compacted garden waste or liquid waste in your bin.
- US aircraft landed with supplies Monday and the crews were taking extra caution as volcanic *ash* was causing visibility problems.
- *Ashes* to ashes, dust to dust, ‘he intoned, just like the preacher had.

**STRUCTURE** N_mod

**COLLOCATES** cigarette, cigar

- She shrugged her thin shoulders and got up, brushing the cigarette *ash* from her skirt.
- Too often, he worked all night, leaving a trail of cigar *ash* and cups of black coffee behind him.

**ITEMIZER** heap

- *Ashes* to ashes, dust to dust, ‘he intoned, just like the preacher had.

**ITEMIZER** pile

- Next morning the farmer came down into the kitchen and to his horror saw a huge pile of *ash* in the hearth of the granite fireplace.

**ITEMIZER** layer

- A layer of *ash*, in places up to a metre thick, was found, indicating the scale of the fire.

**ITEMIZER** cloud

- Rocks were thrown 50 km into the air and
Implications for e-Lexicography

- Methodological
  - significant steps towards automation
  - many features now in wider use
    - e.g. ANW (INL Leiden), Slovene LDB

- Project management
  - better scheduling
  - improved quality at lower cost
Implications of e-lexicography

- Content
  - a platform for development or research
  - a model for other-language versions
  - one of several sources of data in grander vision for “search”
    - “information is cheap, meaning is expensive”
References

