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

LEXMC

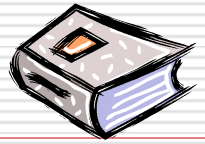
Phase 1 : infrastructure

Phase 2a: source language analysis

Phase 2b: translations into database

Foras na Gaeilge

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 1145th 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)

Constructions		
Vinf_to	132188	19.2
that_0	47886	8.3
wh	30798	10.3
if	22193	55.6
NP_Vinf_to	7175	4.3
it_constrn	5441	30.7
PP_for_Vinf_to	5374	51.4
PP_Vinf_to	5374	273.0
wh_Vinf_to	5299	154.6

PP_X	34423	
PP_on-i	11547	4.7
PP_whether-i	4721	157.6
PP_by-i	4595	2.8
PP_to-i	3173	0.9
PP_in-i	2521	0.4
PP_if-i	1616	15.8
PP_at-i	1350	0.9
PP_upon-i	1088	21.3
PP_for-i	506	0.2
PP_against-i	498	3.3

PP_Vinf_to	5374	273.0
whether	5341	9.6

PP_for_Vinf_to	5374	51.4
whether	5341	9.6

PP_NP_Vinf_to	1413	7.1
upon	53	2.78
whether	36	2.41
on	603	0.77

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)* **LEXMCI freq = 134412** (78.1 per million)

displaying only: **PP_X** whole word sketch

PP_for-i	6246	5.3
abolition	27	6.04
centrality	9	5.25
interpretation	85	4.94
retention	29	4.76
legalisation	6	4.63
superiority	10	4.52
primacy	6	4.52
reassessment	6	4.51
rethink	5	4.3
necessity	20	4.2
importance	89	4.16
inclusion	39	4.15
unity	22	4.13
impossibility	5	4.11
existence	37	4.08
boycott	7	3.98

PP_in-i	2207	1.0
favour	429	8.27
favor	75	7.91
Prospect	5	4.37
pamphlet	8	3.91
Guardian	22	3.38
court	91	3.23
chapter	33	3.18
past	31	3.17
vein	7	3.15
essay	18	3.14
circle	17	2.65
Independent	13	2.55
Brussels	5	2.48
defence	19	2.36
paper	85	2.33
Chapter	10	2.31

PP_with-i	2025	2.7
umpire	16	5.97
referee	44	5.57
ref	22	5.44
idiot	15	5.33
eloquence	5	5.25
justification	14	4.25
interviewer	5	3.88
passion	10	2.81
logic	10	2.78
Him	7	2.51
anybody	5	1.92
neighbour	5	1.76
someone	28	1.49
God	31	0.99
anyone	17	0.88
official	22	0.87

PP_against-i	1879	33.6
notion	28	4.05
proposition	14	3.92
ban	15	3.01
hypothesis	8	2.95
existence	14	2.77
adoption	6	2.57
motion	14	2.42
thesis	5	2.05
penalty	10	1.83
belief	13	1.71
identification	6	1.69
proposal	27	1.64
interpretation	8	1.6
inclusion	6	1.54
view	52	1.51
possibility	11	1.41

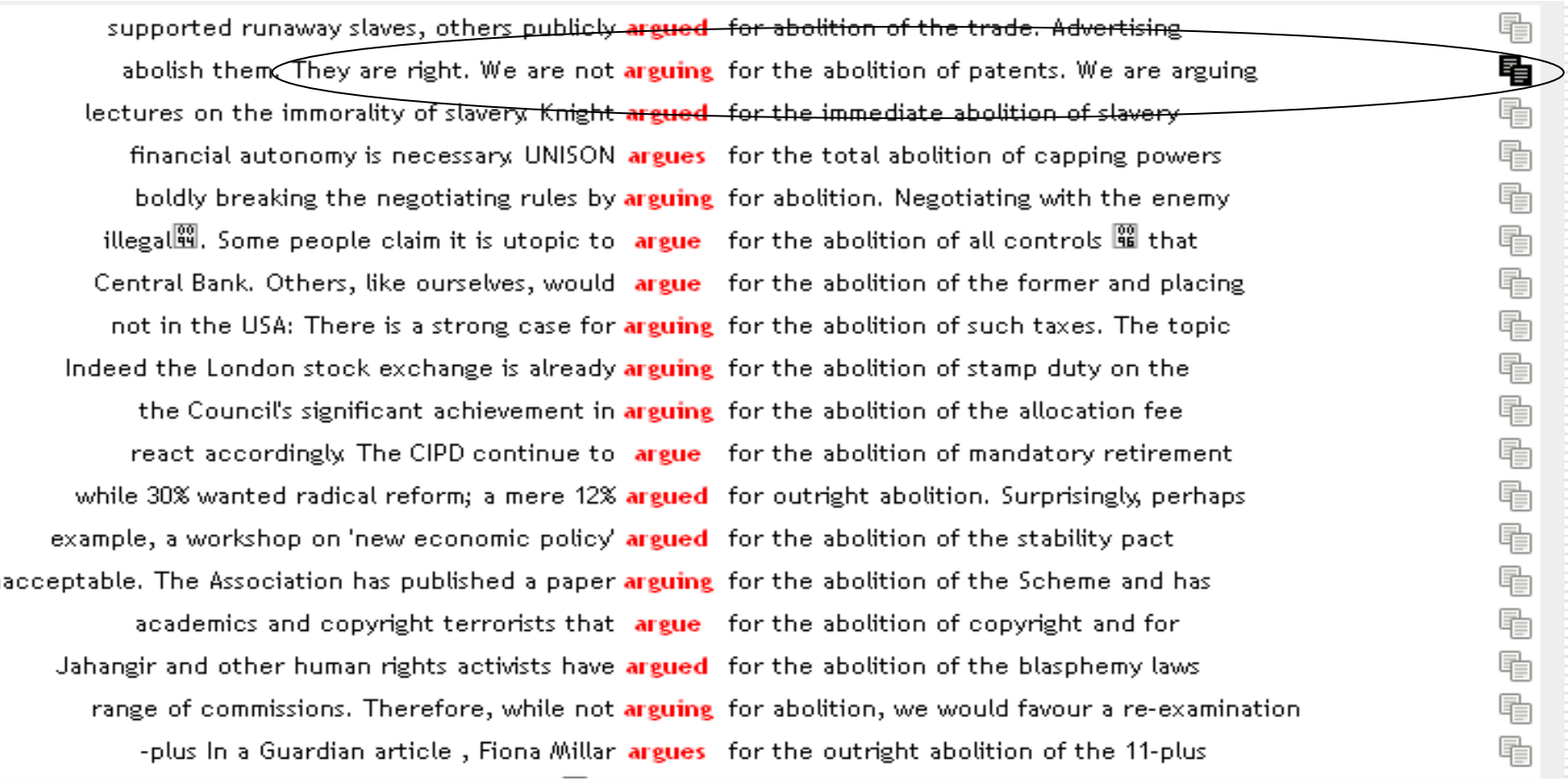
PP_about-i	979	7.1
merit	15	3.99
validity	5	3.23
meaning	15	2.22
existence	7	1.79
definition	10	1.59
composition	5	1.44
politics	9	1.22
religion	8	1.07
extent	6	0.91
bill	7	0.39
direction	8	0.38
importance	6	0.32
everything	9	0.16

PP_over-i	519	5.2
merit	15	4.01

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
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
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



3 v 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*

GDEX and one-click: from corpus to database in one easy move

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)

<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

DocID	label	xml_0
Select All		
Edit View	macadam	/POS><GRAM code="mass"></GRAM><MEANING> broken-up stones bound together by a substance such as tar for us
Edit View	macaroni	abel="food"></DOMAIN></LabelGp><MEANING>tube-shaped pasta</MEANING><ExCnt><EX>Cook the <b>macaron
Edit View	macaroni	abel="food"></DOMAIN></LabelGp><MEANING>a dish of macaroni in melted cheese</MEANING><ExCnt><EX>It makes
Edit View	mace	/POS><GRAM code="mass"></GRAM><MEANING> spice </MEANING><LabelGp><DOMAIN label="cook">
Edit View	mace	rCnt><GRAM code="mass"></GRAM><MEANING> substance used as a weapon or for crowd control </MEANING><LabelG
Edit View	machair	/POS><GRAM code="mass"></GRAM><MEANING> particular type of low-lying grassland on the coasts of Ireland a
Edit View	machinery	/POS><GRAM code="mass"></GRAM><MEANING> machines, equipment </MEANING><ExCnt><EX>It was he who devis
Edit View	machinery	/POS><GRAM code="mass"></GRAM><MEANING> working parts of a machine </MEANING><ExCnt><EX>He looked a look
Edit View	mackerel	belGp><GRAM code="mass"></GRAM><MEANING>the fish as food</MEANING><ExCnt><EX>She decided to make

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

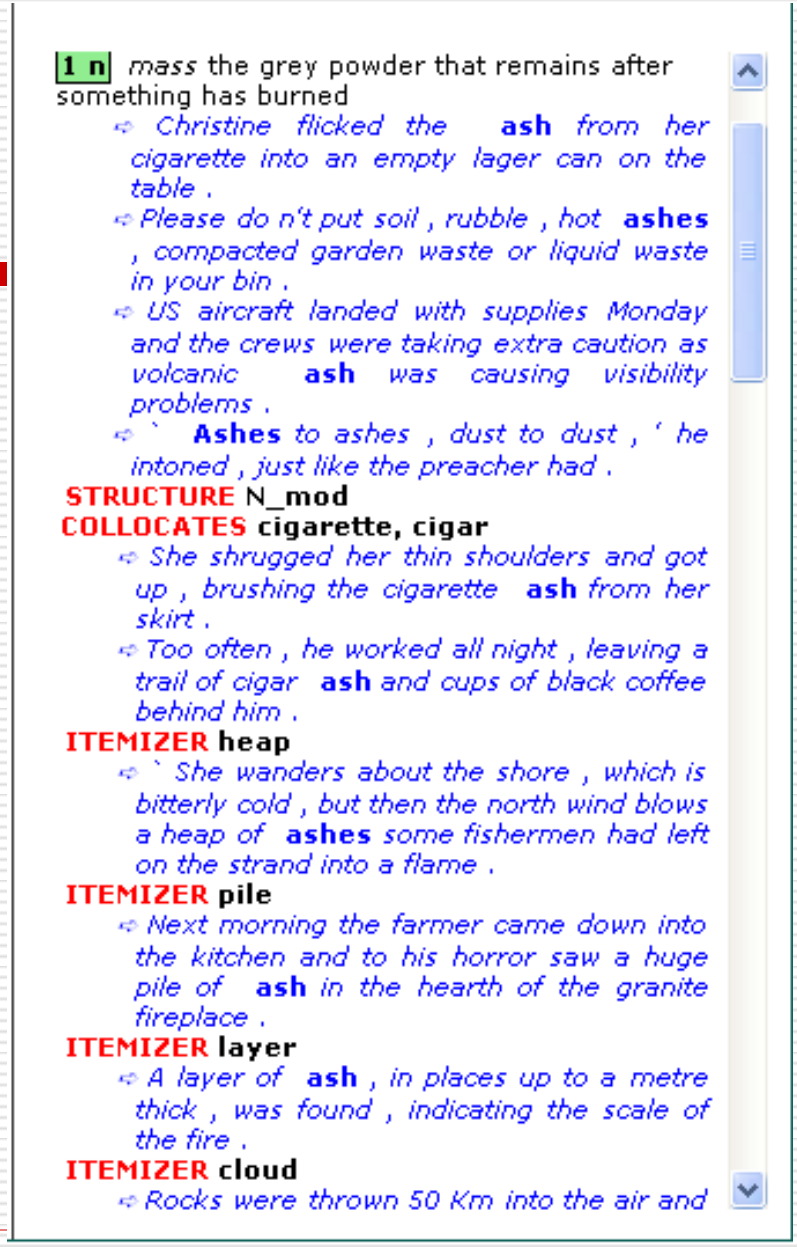
- Public website: www.webdante.com
  - free access to complete DANTE entries, with simple and advanced search options, in range M-R
- Register for free A-Z web access
  - personal, research, and teaching purposes only
- Get licence for free download of full database
  - for research groups in univs or 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*



1 n *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 do n'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**

- *` 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 .*

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 ofr 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

- Atkins, B. T. S. and V. Grundy. (2006) Lexicographic profiling: An aid to consistency in dictionary entry design. In Corino E., Marello C., Onesti C. (eds.), *Proceedings of 12th EURALEX International Congress, Euralex 2006* (Torino, Italy September 6-9, 2006), Alessandria: Edizioni Dell'Orso.
- Convery, C., P. Ó Mianáin, M. Ó Raghallaigh, B.T.S. Atkins, A. Kilgarriff, and M. Rundell (2010). Database of ANalysed Texts of English (DANTE): the NEID database project. In A.Dykstra and T. Schoonheim (Eds). *Proceedings of the XIV Euralex International Conference*. Leeuwarden, Netherlands: Fryske Akademy.
- Rundell, M. and Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end?, in *A Taste for Corpora. A tribute to Professor Sylviane Granger*. Meunier F., De Cock S., Gilquin G. and Paquot M. (eds), Benjamins.