From Corpus-Driven-Dictionaries to Corpus-Driven-Language-Learning

Ramesh Krishnamurthy

Aston University

Dedicated to

John Sinclair

1933-2007

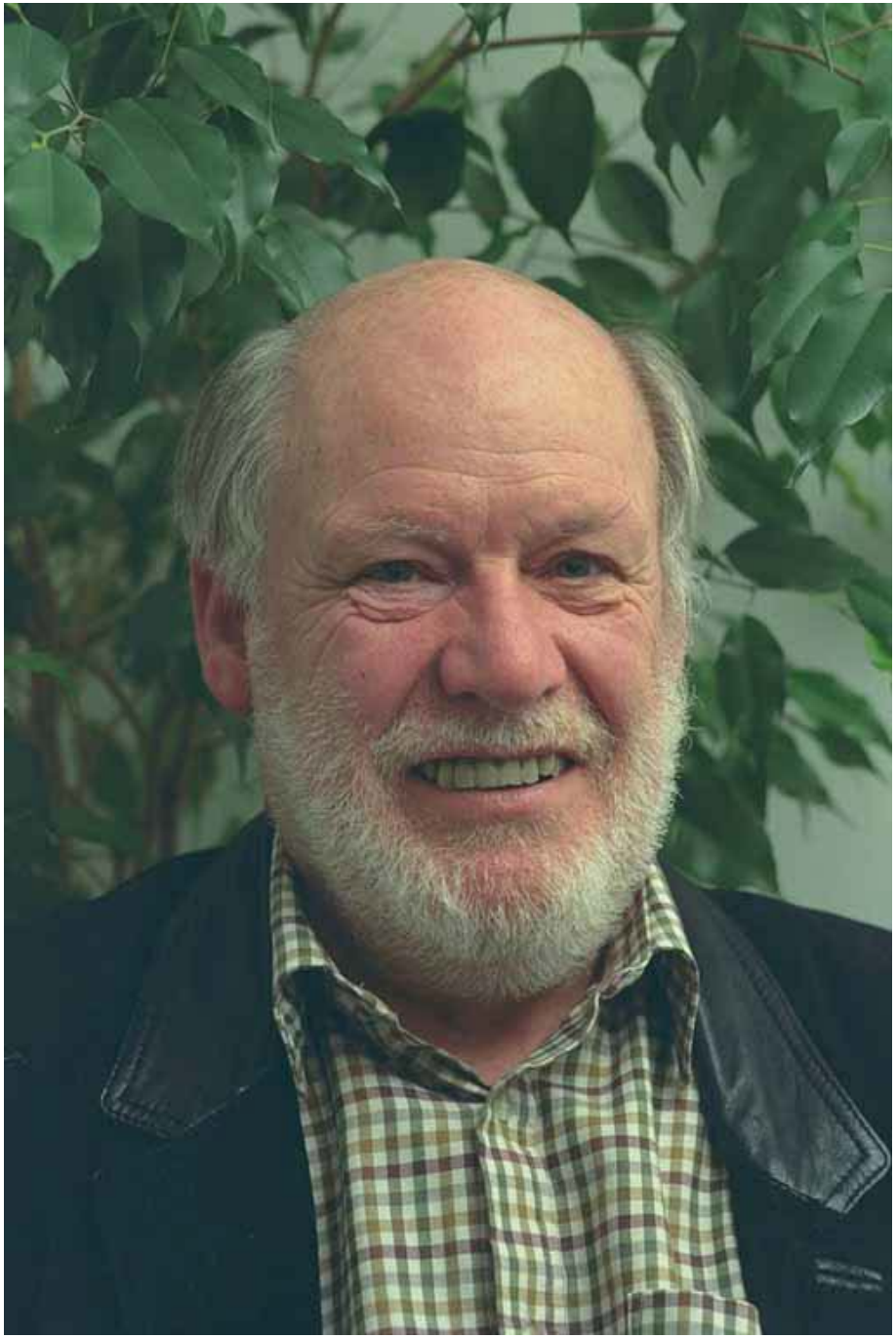

Photo by Primoz Jacopin

From Corpus-Driven-Dictionaries to Corpus-Driven-Language-Learning

From the 1960s to the 1990s, corpora were used mainly for research in **language description and lexicography**.

In the past decade, they are increasingly being used for **language teaching and learning**.

This talk will discuss some of the changes involved, from the practical perspective of the **ACORN project at Aston University**:

http://corpus.aston.ac.uk/

Pre-Corpus Linguistics (1)

- **Linguistics** = grammar (= phonology, morphology, syntax)
- **Grammar** = sentence-level
- **Data** = invented sentences, used as basis for intuitive grammaticality judgments
- **Lexis** (**form**) = morphology
- **Lexis** (**meaning**) > semantics > philosophy

The study of language can be conducted without special assumptions so long as we pay no attention to the meaning of what is spoken.
(Bloomfield 1933:75)

Pre-Corpus Linguistics (2)

- ***The lexicon*** *"requires particular and different statements for each item" and is therefore "an appendix of grammar and the list of basic irregularities"* (Bloomfield 1933:274)
- ***Linguistic theory*** *is concerned with an ideal speaker/listener in a completely homogeneous speech community who knows language perfectly and is not affected by factors such as memory limitations or distractions.* (Chomsky 1965:4)
- *Everyone agrees that if we take* ***language*** *to be I-language, that is the internal state of the language faculty, then of course the use of language is language external.* (Chomsky in Andor 2004:101)

Pre-Corpus Linguistics (3) Firth: the ‘London School’

- *Indeed, the main aim of descriptive linguistics is to make statements of meaning.* (**Firth** 1957:190)
- *At a time when few linguists, other than lexicographers themselves, devoted much attention to the study of lexis, and outlines of linguistics often contained little reference to dictionaries or other methods in lexicology, J.R. Firth repeatedly stressed the importance of lexical studies in descriptive linguistics. He did not accept the equation of ‘lexical’ with ‘semantic’, and he showed that it was both possible and useful to make formal statements about lexical items and their relations*. (**Halliday** 1966:14)
- Halliday MAK (1966) ‘Lexis as a Linguistic Level’ and **Sinclair** JM (1966) ‘Beginning the Study of Lexis’, both in Bazell CE, Catford JC, Halliday MAK & Robins RH (eds). *In Memory of J.R. Firth*. London: Longman

Corpora (1): Why bother?

- “Language users cannot accurately report language usage, even their own” (**Sinclair**, 1987)
- “There are many facts about language that cannot be discovered by just thinking about it, or even reading and listening very intently” (**Sinclair**, 1995)
- As language teachers and professionals, we often have strong intuitions about language use… *Corpus-based research, however, shows us that our intuitions are often completely wrong.* (Biber 2005)

Corpora (2): Brief history

- A **corpus** is a large collection of texts stored in a computer and accessed by retrieval software
- Developed alongside **computer** technology
- **1960s-80s**: ‘small’ corpora of English (Brown Uni; LOB; 1m words): word frequencies, concordances, mainly **grammar** studies
- **1980s-1990s**: ‘larger’ corpora (COBUILD – Birmingham Uni, 18m; Longman, Cambridge): major impact on **language description**, and (EFL) **dictionary** compilation; corpora in other languages
- **1990s-2000s**: ‘huge’ corpora (BNC – 100m; Bank of English – 450m; Oxford English Corpus – 1bn): greater depth of analysis, variety of texts, and statistical accuracy; corpora in most other languages; specialist corpora (business, translation, language learners, etc)

Corpora (3): Corpus methodology

- **Tokenization** – to identify the objects of study (what is a word? ‘sequence of characters bounded by spaces’ – OK for many languages): words, phrases
- **Frequency** lists – to establish hierarchy of importance (NB salience? How to measure?); single words,

  n-grams, lemmatization (NB pre-corpus classes)
- **Collocation –** co-occurrence of words within contextual ‘window’, establishing senses and uses
- **Concordances** – observe behaviour of words/phrases in detail, patterns, register, genre, pragmatics

Corpus-driven dictionaries (1)

- **Distribution**: The 2500 most frequent words in English make up 80% of all texts; the top 15,000 words make up 95%
- the **most frequent words** are function words
- **most frequent lexical/content words** (nouns): *time, people* Rosamund Moon 1988:110;
  http://www.askoxford.com/worldofwords/wordfrom/?view=uk
  01/06/2006 English Uncovered: the hundred commonest English words
  Catherine Soanes: "The commonest nouns are time, person"
- **Rare words:** half the words in a corpus occur only once
- **Chunking**: pervasive influence of collocation and phraseology – we don't choose individual words, but co-select word pairings (*hard work, work hard; innocent bystanders*) and 'chunks of language' (*this, that and the other*)
- **Semantics**: Common verbs have become 'delexicalised' (*take a bath, get warm, give a talk, make a decision*)
- **Stable classes**: reversal of traditional ideas: grammar words are unique, not classes; but regular lexical classes

Corpus-driven dictionaries (2)

LDOCE 1978 (pre-corpus) – 1987 edition:

out: *desuetude, the deuce you will, petticoat government* [18th-19thC], *phalarope* [bird], *phlebotomy, teaching machine*

in: *destabilize, detectable, phantom pregnancy, phase he's going through, phone-tapping, taxman*

- Fewer rare words, more phrases and collocations for common words

Corpus-driven dictionaries (3)

- **Omit rare / outdated items**: *yuppie (1987-2001)*
- **Omit rare forms**:

(traditional dictionaries):

en·crust (ĕn-krŭst') also in·crust (ĭn-) tr.v., - crust·ed, -crust·ing, -crusts. 1. To cover or coat with or as if with a crust 2. To decorate by inlaying or overlaying with a contrasting material

(corpus dictionaries):

encrust	9	(not included)
encrusts	1	(not included)
encrusting	13	(not included)
encrusted	**645**	**(headword: adj: stative, gradable)**

Corpus-driven dictionaries (4)

- **omit rare/outdated uses**:
- e.g. **fissiparous** adj (of cells) reproducing by fission

propaganda put out by the fissiparous and endearing tribes of Anc
the insipiently fevered and fissiparous cultural world of 1899 coul
During the transformation of a fissiparous dictatorship, a special dan
the often antagonistic and fissiparous exiled groups who sip mint
managed to hold together the fissiparous federal state of Yugoslavi
lingness of unpredictable and fissiparous foreign terrorist groups to
It is worth comparing the fissiparous Labour Party, kept together

e.g. **pylon** n gateway to an Egyptian temple

- **give accurate information**:
- e.g. **lame** (lām) adj., lam·er, lam·est. 1. Disabled so that movement, especially walking, is difficult or impossible 2. Marked by pain or rigidness 3. Weak and ineffectual; unsatisfactory; **lame'ly adv**.; lame'ness n.

moving backwards and forwards in a lamely seductive manner. In her mi
without a licence fee, Dyke answered lamely: `Who knows?" He seemed to
its part in persuading the deputies lamely to consent to Mussolini's
Quiet Flows the Don # cf1 and I said lamely that it was very educationa

Corpus-driven dictionaries (5)

- ‘**Core/historical** meaning’ is not **commonest** usage:

see = NOT ‘vision’, but ‘understand’ (esp spoken: *I see, You see*);
surge = NOT *sea/waves/tides* – rare; BUT figures (*imports, shares*), abstract (*demand, investment*), emotion (*joy, pity*), movement (*people, animals, cars*)

- **LDOCE 1978 > 1987 editions**

fewer rare senses

determination 5 > 3 [`fixing limits', `formal decision' **omitted**]
detract from 2 > 1 [`say evil things about' **omitted**]
pettifogging 3 > 2 [`using dishonest tricks; esp. lawyers' **omitted**]
phew 3 > 1 [a (relief), b (tired), & c (shocked) **merged**]

more common senses

teach 1 > 5 [1 **split** into 3; 2 phrases + usage note **added**]
destined 1 > 2 [`having as a destination‘ **added**]
tear (2) V 10 > 12 [`limb from limb' & `That's torn it' **added**]
technical 4 > 5 [`needing special knowledge‘ **added**]
technician 1 > 2 [`SB who has a good technique' **added**]

Corpus-driven dictionaries (6)

- All **forms do not behave in the same way**; not homogeneous/uniform

  e.g. *set* (80%), *sets* (9%), *setting* (11%)

- **Many 'deficient' verbs – or not verbs at all?**

  e.g. *encrust (given earlier)*

  e.g. overstaff:

  vb. (tr.) to provide an excessive number of staff for (a factory, hotel, etc.)

  [who would deliberately do this? **Mistaken derivation?** Paradigm = poorly staffed, under-staffed, well-staffed, overstaffed (Corpus: overstaff 1, overstaffs 0, overstaffing (n) 22, overstaffed (adj) 68)]

Corpus-driven dictionaries (7)

- **Non-corpus dictionaries give misleading or unnatural examples**:

  He argued well… [Corpus: *successfully*, *passionately*, *forcefully*, *convincingly*]

  Don't hold the gun by the business end.

```
e example at the business end of the season
 it comes to the business end of the meal,
oard home is the business end of the Brisbane River
ailenders at the business end of proceedings
e moment, at the business end of the game,
is calf with the business end of his lathi.
nvolved with the business end of things.
as well. <p> The business end of the upper deck
```

Collocation - Firth

Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of 'night' is its collocability with 'dark', and of 'dark', of course, collocation with 'night'. (Firth 1957)

Collocation: ‘dark night’?

- There are 40,309 examples of *dark* and 203,524 examples of *night*
- **Nights are usually dark, so why do we say ‘dark night’?**
- There are 59 examples of *starry night*, 45 of *moonlit night*, and 7 of *moonlight night* and *bright night*.
- **There are many near-synonyms for ‘dark’, so why don’t we use them?**
- There are 73 examples of *black night*, 31 of *moonless night*, 7 of *dull night*, 6 of *bleak night* and *gloomy night*, 4 of *murky night* and *overcast night*, 3 of *starless night*, and 2 of *dreary night*.
- **BUT there are still 292 examples of *dark night***, so the attraction between *dark* and *night* (292 examples) is stronger than all the other collocates put together (248 examples).
- **AND neither grammar nor semantics can fully explain this phenomenon… hence collocation**

Collocation: disambiguates near-synonyms

- ***electric/electrical*** *screwdriver?*

electric:15630; electrical:7731
electric + shock, guitar, light, chair, cars, motor, windows, current
electrical + engineering, equipment, goods, appliances, power, system

- ***ethnic*****,** ***racial*** **and** ***tribal****:*

a) 4 news articles: Yugoslavia (*ethnic* violence); Kenya (*tribal* killings); UK (*ethnic*=black and/or Asian); South Africa (*racial*)
b) dictionaries: similar definitions (race, culture, religion, language, customs, etc), except that only ***tribe*** has pejorative (primitive) and 'humorous' uses
c) corpus: **ethnic**: 5128, **racial**: 2924, **tribal**: 1362
ethnic *+ groups, minorities, minority,* ***violence****,* ***cleansing****, [Europeans]*
racial *+* ***discrimination****, equality, commission, South Africa*
tribal *+ assembly, leaders, chiefs, groups,* ***killings****, Navajo [Africans]*

Collocation: Semantic prosody

(Sinclair 1987) *set in, break out, happen* [NEG: bad things]

(Louw 1993) *bent on, symptomatic of, utterly, build up of;*
NB *build up a* [POS]; Breaking prosody = irony or insincerity

(Stubbs 1995) *cause ; provide* [POS]

Full exposition: collocation, colligation, semantic preference, semantic prosody
naked eye (Sinclair 1996)

Wynne (corpora-list 2003) (*pay)…personal price*

Collocation: metaphor

graveyard *1278 (political 7; aircraft 3; television 2; shift 36; slot* (TV)*19, slots 4; spiral* (aircraft); *of 150 (champions, christians, ambition, welsh, daytime television)* [**NOT plural** *graveyards*]

cemetery *3401:* almost all LITERAL, a few exceptions:
In the *mid-1980s Latin America became a cemetery of failed stabilisation programmes*

treadmill *(Rundell 1999): metaphorical for centuries, now literal again (gyms)*

Idioms, longer Sayings, Proverbs – rarely quoted in full; elements
new broom (situation > person; *new … broom:* policy, location, person)
every cloud/silver lining
silver spoon (OR *silver X …in mouth*, or *born… silver… in*)
mother of all [*battles* - Saddam Hussein 1990]

Corpus: lexicogrammar

like + to-INF OR –ing
enjoy + -ing

(Sinclair 2005 TWC course):
lexis forms stable classes, grammar words do not (they are one-offs: e.g. *of* is not a preposition; Sinclair 1991)

NB changes: *comprise > comprise of (consist of, be composed of)*

NB *would of, should of* [in written texts!]

Corpus: Language change

deep freeze > deep-freeze > deepfreeze

technology > new vocabulary
social/cultural changes

karaoke – speed of nativisation (prons)
low-carb

apostrophe
would of etc

Corpus: new technology – new words

	1985 18m words	1995 121m words
camcorder	0	1214
email	0	39
mobile phone	0	455
satellite dish	0	236

Corpus: other new words of 1990s

alcopops, bull bars, chaos thory, clone, gridlock, karaoke, listeriosis, pro-choice, shell suit, snail mail, snowboarding, TESSA, white-knuckle rides

Meanwhile…changes in language teaching and learning

- Shift in goals: from intellectual/cultural to practical (business, media, tourism)
- Shift in language focus: from classroom to real world; made-up to authentic language
- Shift in teaching methods: from passive to active; teacher to student; use of computer technology
- Corpora are well-suited to these shifts

Pedagogic interest in corpora

- **CALL** (Computer-Assisted Language Learning) and now **CorpusCALL**
- **Data-Driven Learning** (Tim Johns)
- **Lexical Syllabus** (Willis 1990)
- **TALC** (Teaching and Language Corpora) conferences (1994 – 2006)

  http://talc7.eila.jussieu.fr/previous_sites.en.shtml
- **CLLT** (Corpus Linguistics and Language Teaching) newsgroup

Corpora vs Coursebooks

- Coursebooks often use made-up or heavily edited, unnatural text
- Glossaries and grammar explanations are severely restricted to specific context
- Often out-of-date
- Follow a grammatical syllabus
- Limited varieties of text

Corpora vs Dictionaries

- **Printed dictionaries**: limited by space; information is always partial, mediated, summarized, interpreted (sometimes wrongly!); internally inconsistent, or contradict each other; out of date; suffer from inertia, 'legacy' effect; under pressure from publishing deadlines, marketing, competition, etc;
- **Dictionary users**: receive no training and are impatient, so often miss the information, misinterpret it, or misuse it
- **Electronic dictionaries**: EFL ones are still in their infancy – copies of printed ones, with a few extra features; bilingual ones are often poor quality

Corpora vs Web and Search Engines		
	Web and Search Engines	**Corpora**
SIZE	**Vast**	**Manageable**
PROCESSING SPEED	**Slow**	**Fast**
ANALYSES	**Coarse-grained, General**	**Fine-grained, Detailed, Specific**
CONTENT RANGE	**No Overview; Diffuse, Uncategorized**	**Selected, Documented, Categorized**
CONTENT STABILITY	**Volatile/Dynamic: cannot replicate analyses**	**Stable: can replicate analyses**
CONTENT QUALITY	**uncontrolled**	**Controlled by selection**
SOFTWARE	**complex, 'black box'**	**simple, fully documented**

Corpora and language learning/teaching: the challenge

- We learn our **mother tongue** by experiencing thousands of similar examples of natural language use, in a wide range of texts and situations, over a long period of time
- But we learn **foreign languages** on the basis of more abstract information about the language: grammar rules, dictionary definitions, etc; with less exposure to the language itself, and fewer opportunities to experience the variety of texts and situations; over a much shorter period of time
- **Corpora** can help by providing **more exposure** to natural language use…
- …but can corpora also help to create genuine **new additional paradigms for language learning**?
- …and how can corpora be used in **other disciplines**?

ACORN
(Aston Corpus Network)
initiated 2005

- **2006-2007: Funded by the Flexible Learning Development Centre, Aston University**
- **AIMS: to provide Aston University with**
- **(a) corpora (for English, French, German, Spanish; Translation Studies)**
- **(b) customized software to analyse the texts**
- **(c) pedagogical outputs (for teaching, learning, assessment, and feedback)**
- **to increase flexibility: wider range of texts, variety of new approaches, access the resources at any time (via web interface)**

ACORN: Corpus Creation Processes

- **Design Principles** (informed by staff questionnaires)
- **Text Selection** (according to design)
- **Copyright Permission**
- **Text Acquisition** (download from Web; email attachment; copy from hard drives; keyboard; scan; transcribe speech)
- **Data Conversion** (PDF, DOC, HTML, etc to plain TXT)
- **Indexing** (making texts available for analytical software)

ACORN Design: Staff Questionnaires

- **Topics**: national identities and stereotypes, cultural differences, current affairs (social and political), institutions, international relations, history, economics, marketing, education, globalisation, cultural events, media, immigration
- **Language**: stylistics, terminology, dialects, grammar, discourse analysis, history, policy and planning
- **Text Types**: film/book reviews, commentaries, readers' letters, obituaries, abstracts, instruction manuals, résumés, academic writing, tourist brochures, recipes, fairy tales, short stories, novels, political speeches, medical texts
- **Academic Journals**: Sprachreport, Babel, Discourse and Society, Europe-Asia Studies, etc
- **Literature**: Camus, Sartre, Goethe, Schiller, Kafka, Mann, Brecht, Romanticism
- **Journalism**: Der Spiegel, Die Zeit, Le Monde, Le Figaro, Guardian, Times, Economist, Financial Times
- **Websites**: Newspapers and Journals, European Union, Governments, Red Cross, Amnesty

ACORN Data collected: details

English: Business English, Academic Writing, Instruction Manuals, Political Speeches, Emails, EU legislation, Classic Literature (Shakespeare, Bronte, Darwin, Dickens, Poe, Shaw, Wilde), Nobel Speeches, University Job Advertisements, Junk Emails, Medical Abstracts, Fairy Tales
German: also Amnesty, Der Spiegel, Die Zeit, Book Reviews, Classic Literature (Goethe, Hesse, Kant, Lessing, Nietzsche, Schiller, Storm)
French: also Spoken Corpus, Classic Literature (Balzac, Daudet, Descartes, Maupassant, Verne, Zola)
Spanish: also Classic Literature (Cervantes, Zorilla)

ACORN: data indexed

- **English: 52,673,690 words**
- **French: 43,704,693**
- **German: 47,688,703**
- **Spanish: 32,531,928**
- **TOTAL: 176,599,014 words**
- **Translation Studies: much of the data is available in translated versions**

ACORN Student Data

- **Obtaining material from students: consent forms; electronic submission (for plagiarism detection)**
- **Research will provide more information about the students:**
- general academic development; gradual mastery of topics, themes, subjects
- development in academic writing style
- language development
- successful learning and teaching strategies which can be shared by staff and students
- possible problem areas, and the need to use alternative learning and teaching strategies
- the strengths and weaknesses of the current syllabus, and the need to adjust the focus, alter the sequence, and add or omit elements

How can ACORN help language learning and teaching?

- **Flexible additional resources, with alternative methods (discovery procedures, quantitative approaches, etc), and variety of texts**
- **Provides more examples than dictionaries**
- **Allows you to see common and typical patterns of language use**
- **Enables you to discover the different ways that words/phrases are used by different speakers/writers in different contexts, text genres, and registers**

How can ACORN serve disciplines other than languages?

- **Improve general English academic writing skills**
- Differentiate concepts/terms of your discipline from general language use: **identification of subject terminology**
- Examine the **opinions and arguments** of **experts** in your discipline by looking at **contexts around the key concepts/terms;** how are opinions/arguments **presented/discussed**
- Compare the **discourses of different disciplines**
- Identification of suitable **quotes**

ACORN: DEMONSTRATION

- http://corpus.aston.ac.uk

- **The system will be made available to all Aston staff and students during the next few months**
- **accompanied by training sessions**
- **Help files, guided tours, exercise templates will be added**

ASTON UNIVERSITY

IN THE HEART OF BIRMINGHAM

- Home
- Project Overview
- Log In
- ACORN Team
- Contributors
- Contact Us

Welcome to ACORN

ACORN
Aston Corpus Network

Stage 1: 2006: corpora for teaching and learning: Funded by the Flexible Learning Development Centre, Aston University

Corpora Selection

- [x] English
  - [x] Academic
  - [x] Book Reviews
  - [x] Electrolux
  - [x] Elysee
  - [x] European Commision
  - [x] European Parliament
  - [x] Gutenberg
  - [x] Noble Speeches
  - [x] University Job Adverts
  - [x] Uppsala Student Corpus
  - [x] Wolverhampton Business Archives

  Select all / Unselect All
- [ ] French
- [ ] German
- [ ] Spanish

Continue >>

Welcome to ACORN

Corpora Selection | Frequency | Concordance | N-Grams | Collocations | Parallel

Frequency Results

equency Result for: **English Academic Corpus**

tal number of types = 36852
tal number of tokens = 776544

ank	Frequency	Word	Rate/10,000
	51122	the	658.33
	28108	of	361.96
	20969	and	270.03
	20072	to	258.48
	18128	in	233.44
	10984	is	141.45
	8298	that	106.86
	7642	for	98.41
	6737	as	86.76
	6138	be	79.04
	5409	this	69.65
	5256	it	67.68
	4719	are	60.77
	4631	on	59.64
	4142	by	53.34
	4054	with	52.21
	3615	not	46.55
	3374	or	43.45
	3293	from	42.41
	3047	an	39.24
	3000	was	38.63

Concordance Results

arch results for ***'hoffentlich'*** in **ger_gutenberg_db**
wing 0 to 50 of results

<< previous next >>

left context	keyword	right context
um euch anzuhängen , so viel bin ich	**hoffentlich**	befugt zu bekennen , daß ich dem Mohren
hlt , hört mit geduldgem Ohr , Bringt	**hoffentlich**	nun unsre Müh hervor . ERSTER AKT E
fährten Tybalts macht . Dann wirst du	**hoffentlich**	zufrieden sein . JULIA Fürwahr , ic
) Unsinn , Eugen , Sie frühstücken doch	**hoffentlich**	mit uns ! (Marchbanks sich ängstlich
ürzt über Ihr Telegramm . Es ist doch	**hoffentlich**	nichts geschehen ? (Candida .) Was
ll ernst und mit Selbstbeherrschung :)	**Hoffentlich**	störe ich nicht . (Candida fährt hef
ein Nachbar geworden , wie ich sehe .	**Hoffentlich**	spielt Ihr nicht die Floete wie Euer
h denke einen Monat fortzubleiben und	**hoffentlich**	meine Mutter dann beruhigt verlassen zu
der Kategorien gezeigt worden , wird	**hoffentlich**	niemand im Zweifel stehen , sich über d
. Er will mich noch heute sprechen .	**Hoffentlich**	wird er sich meiner annehmen . Die Zeit
Tempelherr . Mit Unterschied , doch	**hoffentlich**	? Nathan . Jawohl ; An Farb ' , an
. Die Philosophie der Dogmatiker war	**hoffentlich**	nur ein Versprechen über Jahrtausende
ath vermag bei der Tochter viel , und	**hoffentlich**	werden Sie mich kennen , Herr Miller ?
den Blick auf ihn werfend) . Wo doch	**hoffentlich**	deine Ehre nichts einwenden wird ? Fe
, deren hauptsaechliches Kennzeichen	**hoffentlich**	eine allgemeine Annaeherung der Natione