Corpora - from language description and lexicography to language teaching and learning

The ACORN (Aston Corpus Network) project

Ramesh Krishnamurthy

Aston University

Abstract

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

Why bother with corpora?

- “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)

Brief history of corpora

- **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 – language description (1)

- Distribution: The 1900 most **frequent** words in English make up 75% of all texts; the top 15,000 words make up 95%
- 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**’ (*<u>take</u> a bath, <u>make</u> a decision*)
- Reversal of **grammar/lexis**: common/grammar words are unique, not classes; but lexical classes

Corpora – language description (2)

- **‘*would*’** is used generally to talk about **hypothetical events**: e.g. *I think The Tempest would make a wonderful film*. This makes up almost half of its corpus occurrences. As a sub-category, **would** is used in **conditional sentences**: e.g. *It* would *surprise me if sterling strengthened*. **But EFL courses present *would* as a part of the second conditional. (Willis)**
- **the 'rules' for reported speech** (‘present simple becomes past simple, present perfect and past simple become past perfect’ etc) **are totally unnecessary**. Differences in tense, person, phrases of time and place occur because we are taking a different standpoint from the original writer or speaker. **This is a feature of language as a whole, NOT a feature of reported speech**. (Willis)

Corpora – lexicography (1)

- **Omit rare / outdated items**: *desuetude; yuppie*
- **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	(adjective headword: stative, gradable)

Corpora – lexicography (2)

- **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
```

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

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)**
- **<u>to increase flexibility</u>: wider range of texts, variety of new approaches, access the resources at any time (via web interface)**

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**?

PROJECT TEAM

- **Ramesh Krishnamurthy (Lecturer, English, LSS)**
- **Iztok Kosem (Research Student, LSS)**
- **Husman Ahmed (Placement Student, EAS, 2006-07)**
- Chris Martin (Placement Student, EAS, 2005-06)
- Sylwia Jaworska (ASO, German, LSS)
- Stefan Baumgarten (PhD Student, Translation, LSS)
- Constantin Orasan (Consultant; Wolverhampton Uni)
- Irina Benzel (Erasmus Student, LSS)
- Carolina Gonzalez-Gonzalez (Erasmus Student, LSS)
- Kieran Connell (Undergraduate Student, Bristol Uni)

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**

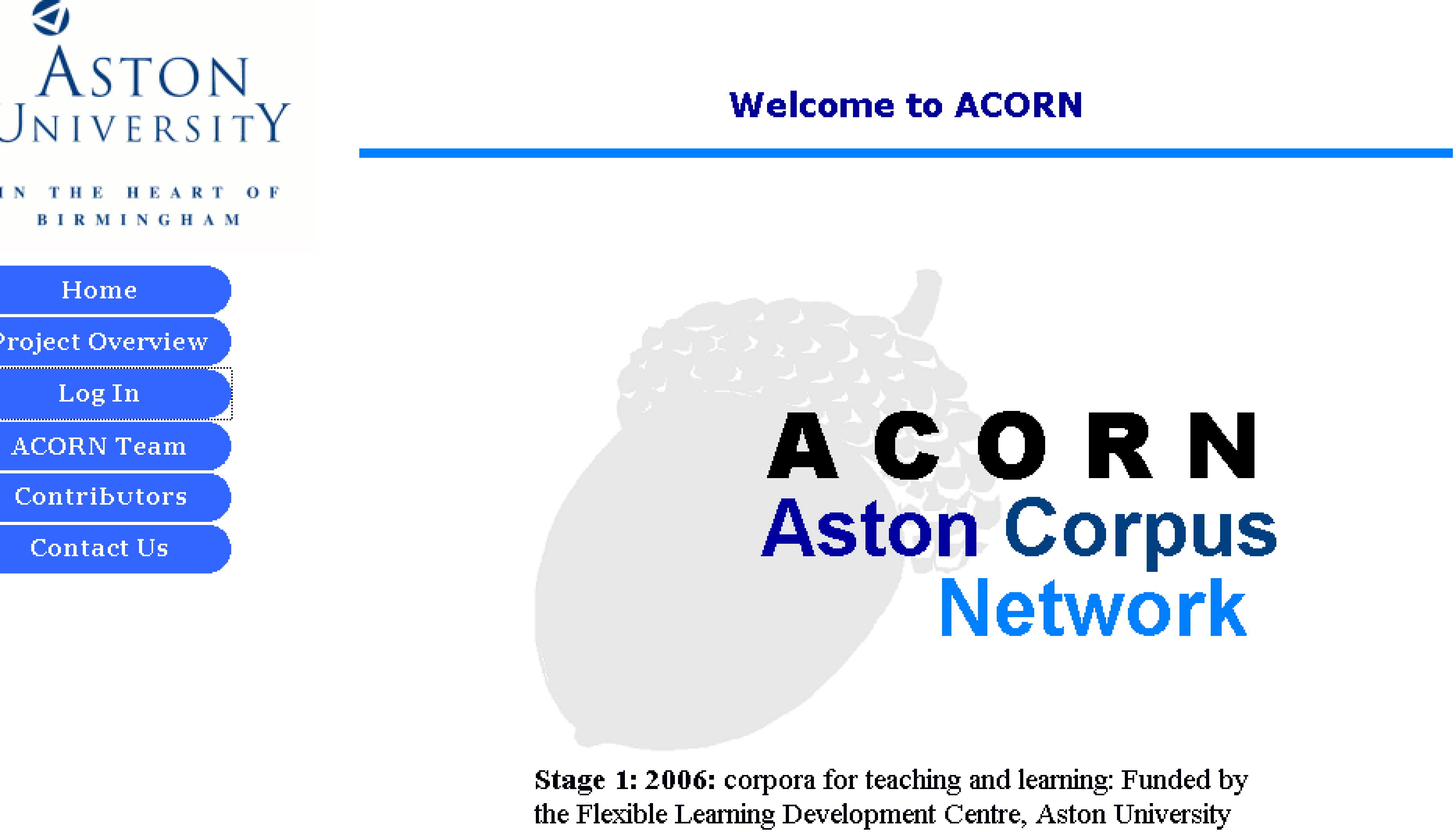
ACORN Software: current analytical functions and displays

Frequencies: words and phrases (N-grams)	Is the word or phrase common or rare? (to decide to pursue your query or not)
Distribution: i.e. which texts/authors use the word/phrase	Is it relevant to the text/context you are working in (reading/writing)
Concordances: examples of use	grammatical, phraseological and contextual behaviour of words/phrases
Collocation: ‘word attraction’	Less fixed aspects of phraseology
Extended Contexts:	Examine discourse and textual features
Bibliographic information:	For quotation, referencing

ACORN: DEMONSTRATION

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

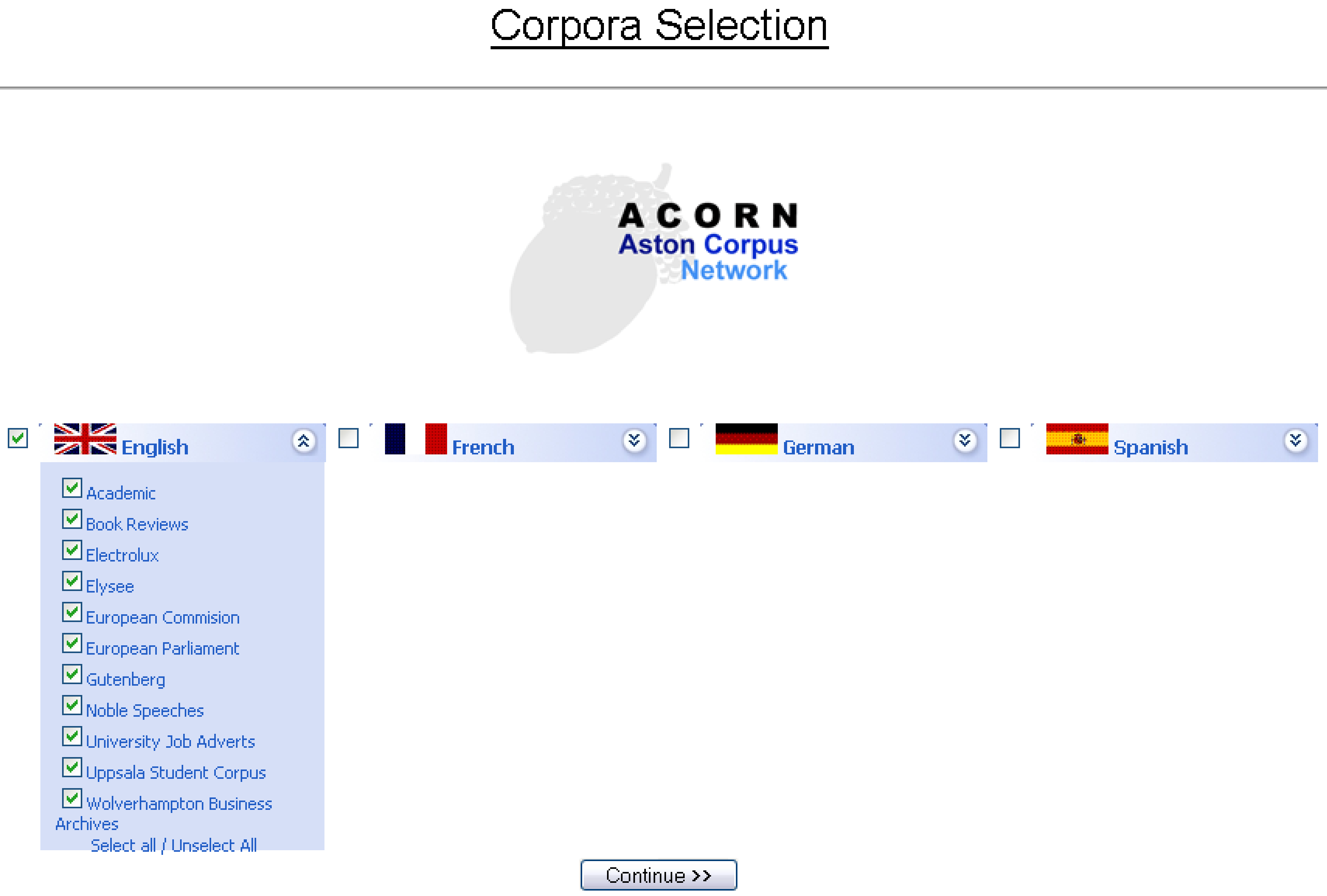
- **The system will be made available to all Aston staff and students during the next few months**
- **accompanied by training sessions**
- **Help files, guided tours, exercise templates will be added**



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

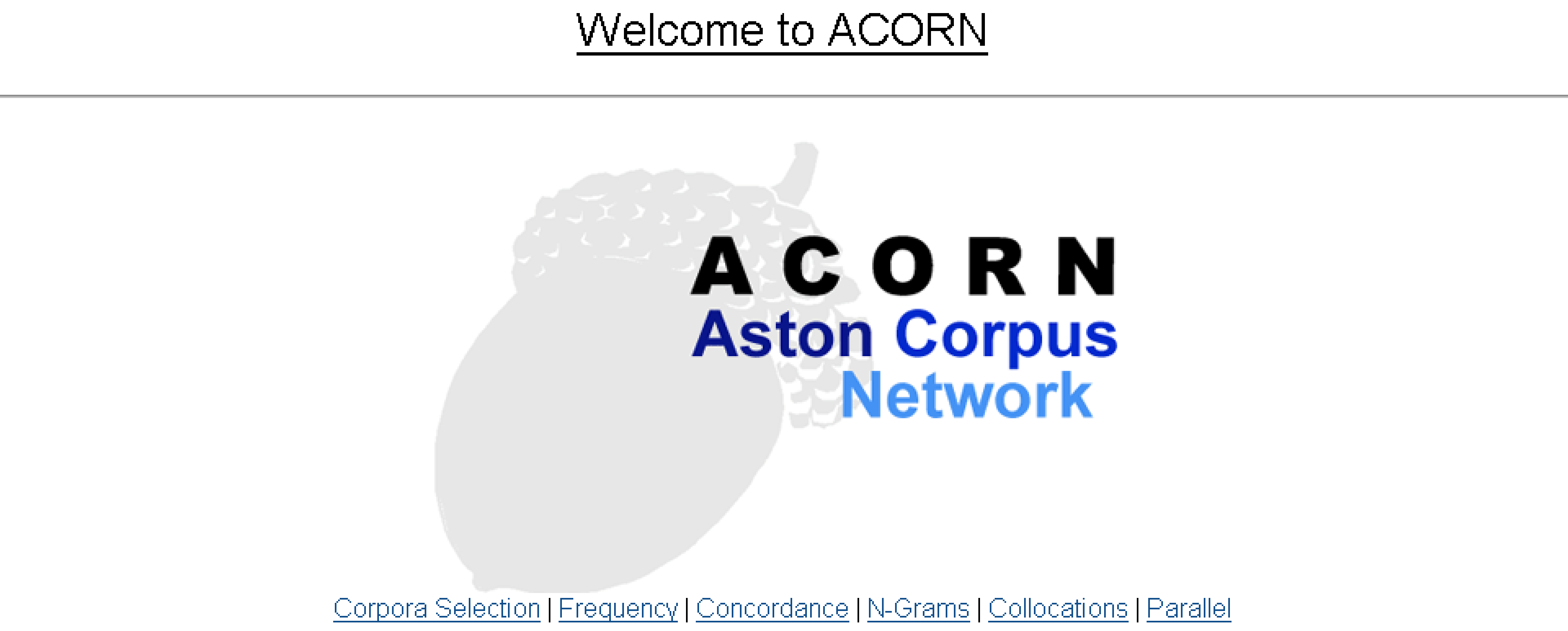
ACORN
Aston Corpus
Network

☑ English □ French □ German □ Spanish

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

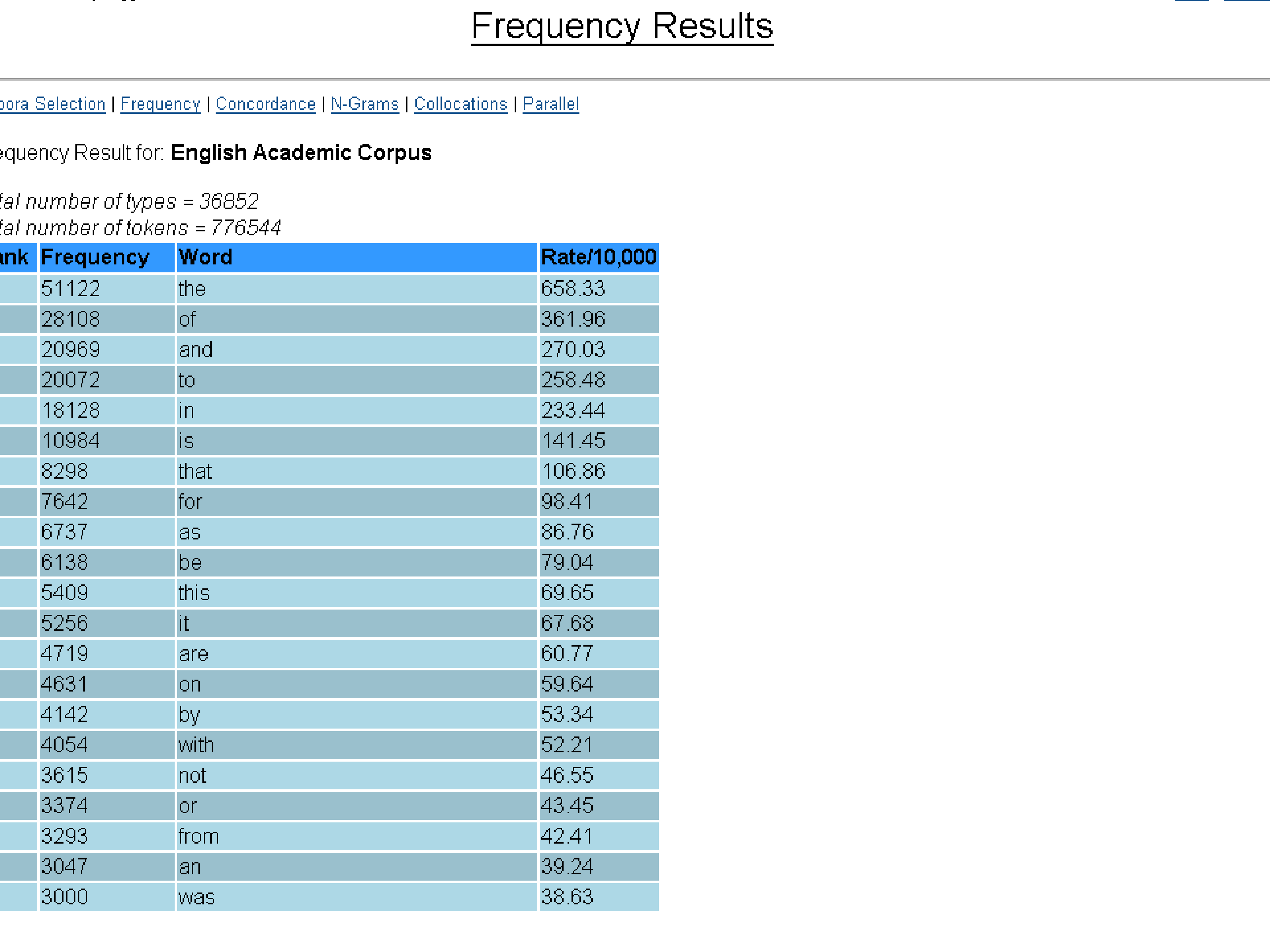
Select all / Unselect All

Continue >>



Welcome to ACORN

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



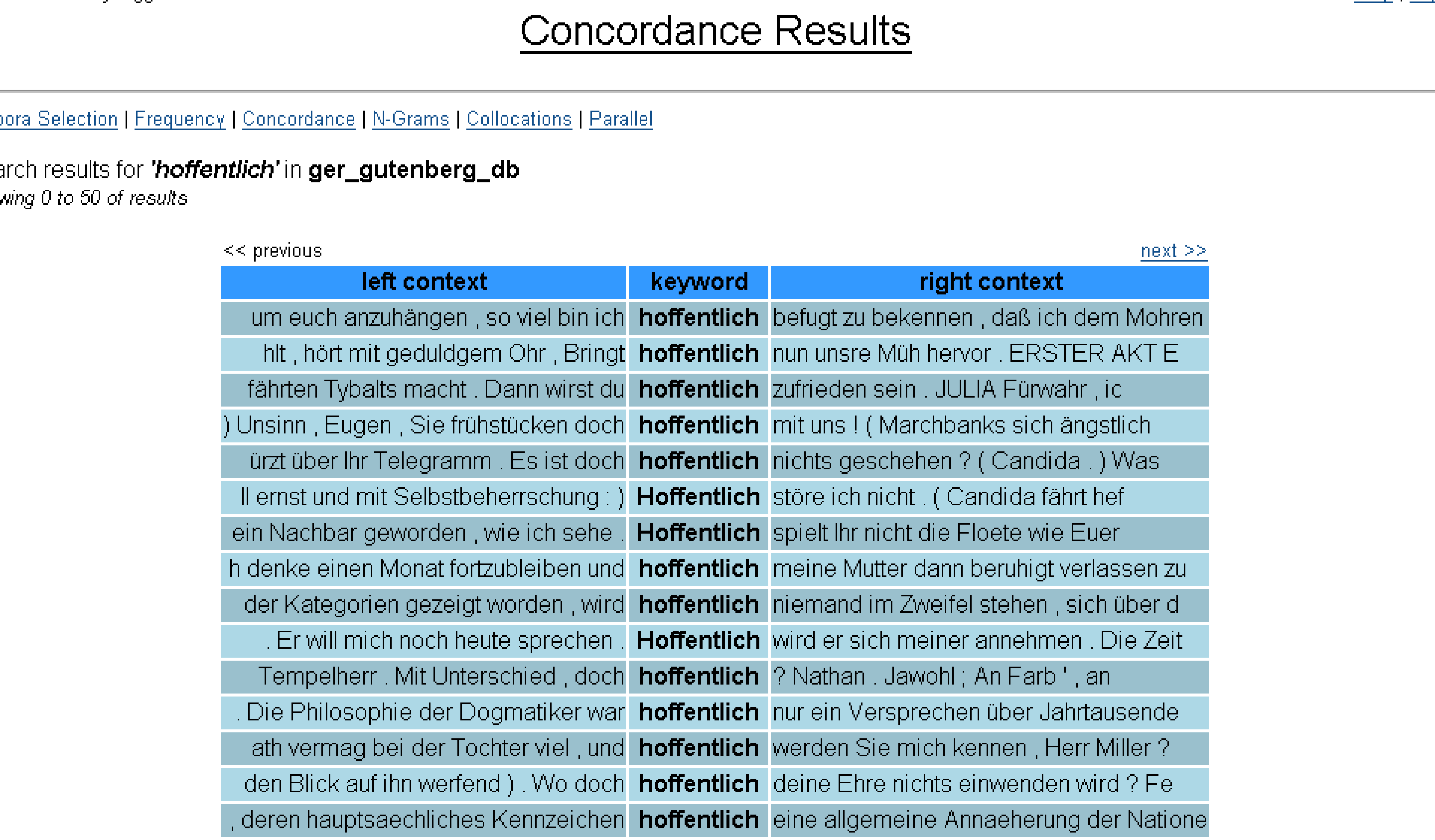
Frequency Results

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

equency Result for: **English Academic Corpus**

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

nk	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


ora Selection | Frequency | Concordance | N-Grams | Collocations | Parallel
<< previous
next >>

Concordance Results

arch results for *'hoffentlich'* in **ger_gutenberg_db**

wing 0 to 50 of results

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