The ACORN (Aston Corpus Network) project

http://acorn.aston.ac.uk

- helps you to learn language, academic skills... And Computer Science?

Ramesh Krishnamurthy

Steven Chen

What is a Corpus?

- A large collection of electronic texts, analysed by software
- Developed since 1960s, mainly for linguistic research
- 1980s – used to write dictionaries, grammar books etc (Cobuild – Birmingham)
- Recently – use in translation, forensic linguistics, information retrieval, data mining, ontologies, semantic web

Why not use the Web and Google?		
	WEB and GOOGLE	**CORPUS**
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, OK
- but why ACORN?

- Many corpora exist, but not easily accessible
- Software and interface are too research-oriented
- Corpora not designed for non-linguists
- Corpus systems not designed for students and teachers
- Pedagogical research not adequately catered for

ASTON : Pre-ACORN

- **No Data**
- **ATA (Aston Text Analysis) Software**; (Mark 1: MS-DOS, 16-bit); (Mark 2: Windows 95, NT4, etc)
- **Software Problems**
  - PCs only
  - filesize and format constraints
  - slow
  - research tool, poor user interface
  - mainly for English
  - no software support (Peter Roe - retired)

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 (Aston **Corpus** Network) **FUTURE**

- **Increase Teaching of Corpus Linguistics**: UG and PG programmes
- **Learning and Teaching of Languages**
- **Learning and Teaching of Academic English**
- **Extend to Other Disciplines**: Social Sciences, other Aston Schools (ABS, EAS, LHS)
- **Enable Wider Access**: FE colleges, schools, libraries, local community organizations
- **Initiate Institutional Cooperation**: Midlands Corpus Network (UCE, Birmingham, Warwick, Wolverhampton), AHRC/AHDS, JISC, national/EC/international networks

ACORN: LIVE DEMO

- Login
- Language selection
- Corpus selection
- Frequency
- N-grams
- Concordance

PROGRAMMING ACORN (1) when I started

A dynamic web system:

- 40 PHP pages
  - Configure, validation, search and display
- 70 MySQL databases
  - Total size of 27,547.9 MB
- 73 Java programs
  - Cleanup & indexing
- A few Perl programs
  - File format conversion

(2) My contribution so far (6 months)

- Reading & research
  - Various text books, websites and ACORN presentations.
- Data collection > cleanup > indexing > testing
- Recoding
  - The parallel text system was rewritten (Java & PHP)
  - Multiword search

(3) What has been difficult so far?

- Lack of PHP programming skills
- Programs/code that lacked comments
- Using the appropriate programs in the appropriate sequence.

(4) Plans for the rest of my placement (6 months)

Programming for:

- Multicorpus search
- Collocation

Uploading more data

L2	L1	Centre	R1
THE	TESTING	DATA	CLASS
OF	TRAINING		THE
ON	THE		AND
TO	NETWORK		FOR
1	OF		IS
AMOUNT	LOGIN		STRUCTURES
I	FOLLOWING		NORMALIZATION
LOSS	AUDIO		PROTECTION
MUSICAL	SET		TO
THEIR	RAW		IN
PROCESSING	SOME		IT
3	ENTERING		FROM
6	4		SHOULD
AND	WINDOWED		HANDLES
SOME	3		INTO
THAT	TEST		MODIFICATION
SYSTEM	5		THEREFORE
USE	FILING		FILE
WHEN	DETAILED		STORAGE
5	INPUT		SET
TIME	DISCARDING		WAS
SET	REDUNDANT		SAMPLES
IN	ARCHIVED		AS

Vocabulary in Academic Texts

- **80%** - 2000 word families (**High frequency words**)
- **8.5%-10%** - 570 word families (**Academic vocabulary**)
- **Up to 5%** - 1000 word families per discipline (**Technical vocabulary**)
- **Proper nouns/Latin forms** (**Low frequency words**)

(Coxhead & Nation 2001)

Academic: *abandon, access, method, denote, revenue*

Technical: algorithm, class, classifier, platform,concordance

General English: *say* = “speak”

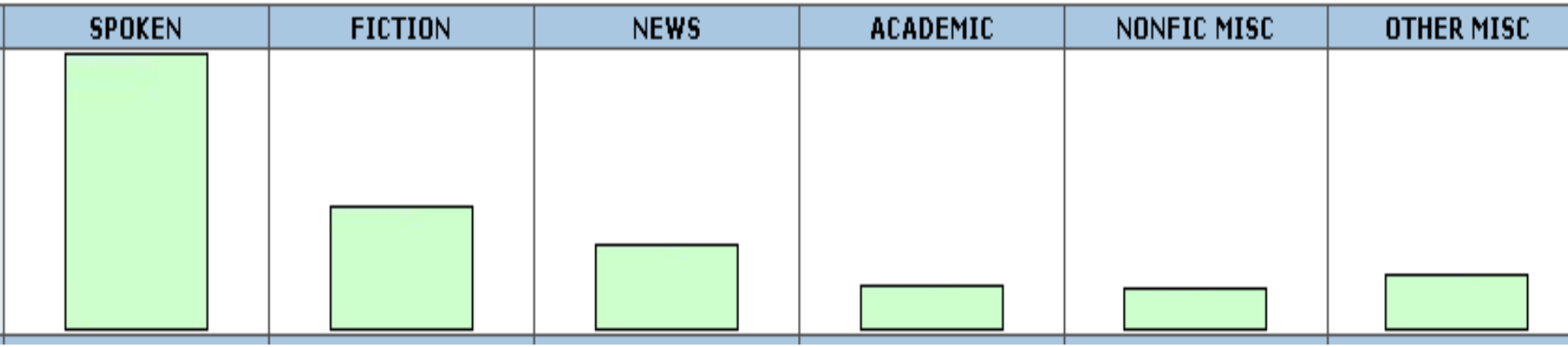
ake you for a woman. He meant to	**say**	more, but he never got the chance.
shu-tt up-pp] I've got something to	**say**	to you, and by God you're going to
ng a brainy little lady like her had to	**say**	would be plumb important, as well

Academic English: *say* = “write”

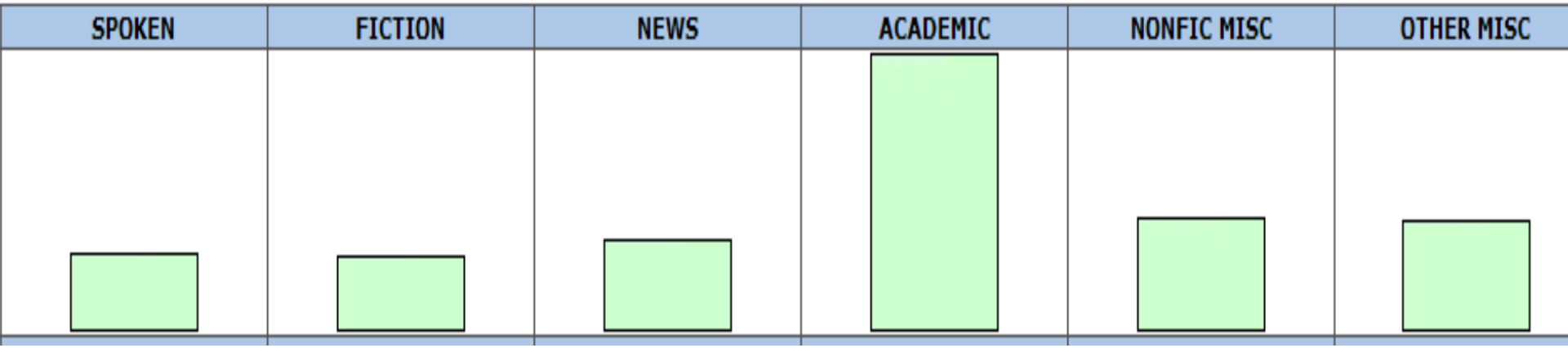
tivities . Legutke and Thomas (1991)	**say**	that the aim of courses should be
tations . McCarthy and Carter (1995)	**say**	that the grammar of spoken text
arning . Scharle and Szabo (2000,8)	**say**	that by self - evaluation learners

Source: ACORN (Aston Corpus Network)

say



argue



Source: http://corpus.byu.edu/bnc/

Lack of Computer Science texts

- British National Corpus (100 million words)
  - Computing texts (1,3 million words) – 1.4%
  - Small number of academic texts
  - All the texts produced in 1994 or earlier
- British Academic Spoken English corpus
  - 1,6 million words
  - 4 computing lectures (22,000 words) – 1.4%
- British Academic Written English corpus
  - Collecting upper-level student texts
  - Reporting on problems obtaining computer science texts

ACORN and Computer Science

- Computer Science role equal to other subjects
- Better insight into academic conventions of Computer Science academics/students
- Benefits for Computer Science students:
  - General Academic English
  - Specific Academic English (Computer Science and other disciplines)

Corpus for learning and teaching: pilot projects

- Kaori – pre-sessionals
- Steven – self-correction

- Ramesh/Iztok – corpus linguistics
- Guadalupe - Spanish grammar clinics
- Pierre – Semantics (town and city)

ACORN: student data

- Collecting student texts at LSS:
  - Initially, students contacted directly (2006)
  - Approval of consent form by SSCC
  - Electronic submission of texts via VLE
  - Once-for-all consent form
  - 2007: online version of the consent form
  - 2008: Collecting student texts at EAS, LHS, and ABS