Introducing XAIRA…

An XML-aware tool for corpus indexing and searching

Lou Burnard

Tony Dodd

Research Technology Services, OUCS

What is XAIRA?

- XML Aware Indexing and Retrieval Architecture
- Developed from the British National Corpus SARA program
- Components include an Index Toolkit and a Windows Client
- Open Source: available from **http://xaira.sf.org**

Historical Background (c.1994)

- Original design goals
  - robust searching of very large (c. 1 Gb) amount of SGML data
  - re-use available indexing tools
  - usable by researchers in CL, NLP, lexicography
- Original assumptions
  - client/server architecture
  - index build once only
  - one specific corpus (the BNC) only

Historical Background (c.2004)

- Design goals
  - robust searching of any amount of XML data
  - offload processing to other components wherever possible
  - assume nothing about input DTD/Schema
- Architecture
  - client/server still valid
  - expect to re-index often
  - expect multiple interfaces

Why another search engine?

- Can't you do all this with Google?
  - Digital texts are not just for discovery and display
  - The methods of corpus linguistics have a wider relevance
- Can't you do all this with eXist?
  - Probably, but only if you have a team of programmers at your disposal!

Xaira: the key features

- Indexes the PCDATA
- Supports word search, concordance generation and manipulation, collocation, lexical analysis
- Uses XML annotation to the max
- Supports XML-aware complex queries
- Leverages existing standards
  - TEI/XCES; Unicode; CSS and XML; SOAP
- Uses efficient and compact indexing appropriate to small or huge corpora

First catch your corpus…

- any collection of well-formed XML documents
  - if a DTD is supplied, the corpus must be valid
  - if no TEI header is present, one will be created
- the more you put in, the more you get out
- "texts" are defined independently of file structure, as are the relevant units within them
- all indexing information is stored in the corpus header

Loading a corpus into Xaira

i. Put the corpus into XML
   - convert it to Unicode
   - index tools can help
     - Adding minimal tagging
     - Converting existing tagging

ii. Use the Index Tools wizard

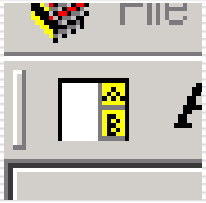
The more markup you put in, the more Xaira can do with it.

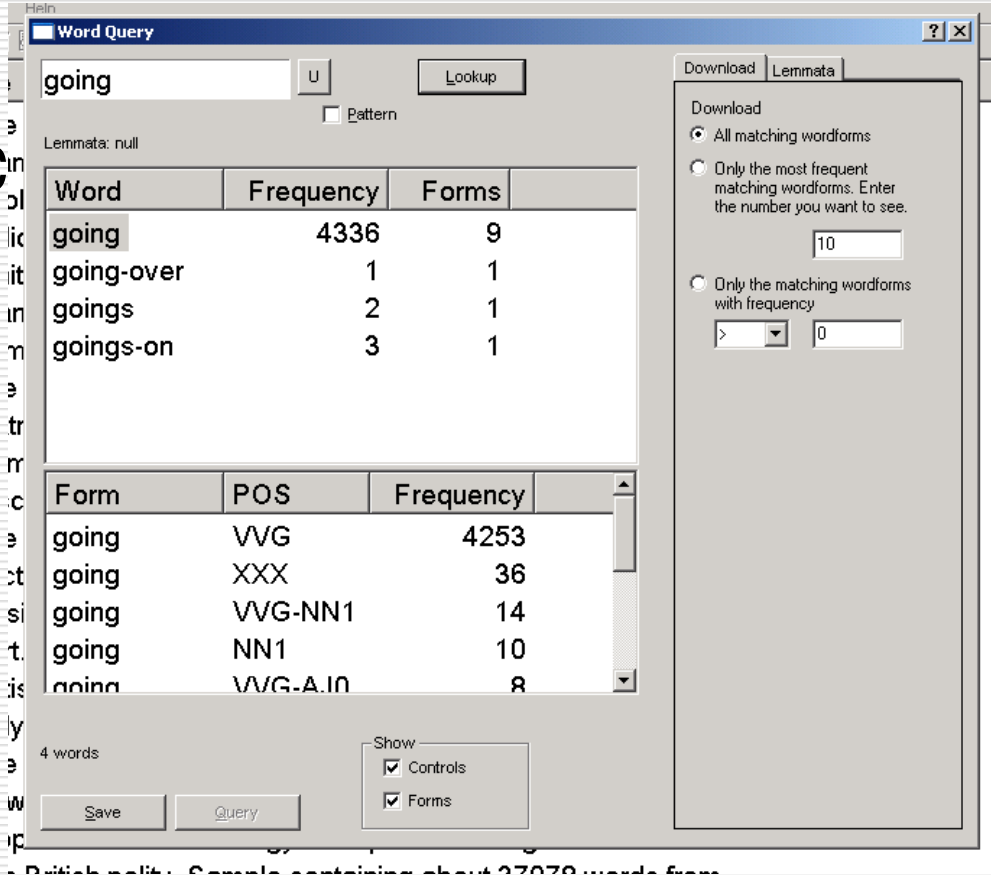
Hoorah for Unicode

- All data is held internally as Unicode
  - this allows us to defer most problems (e.g. tokenization, case-folding, line-breaking, character normalization, glyph composition) to someone else!
- User interface issues
  - For output, use one or more appropriate fonts
  - For input, we provide a keyboard definition utility

A word query

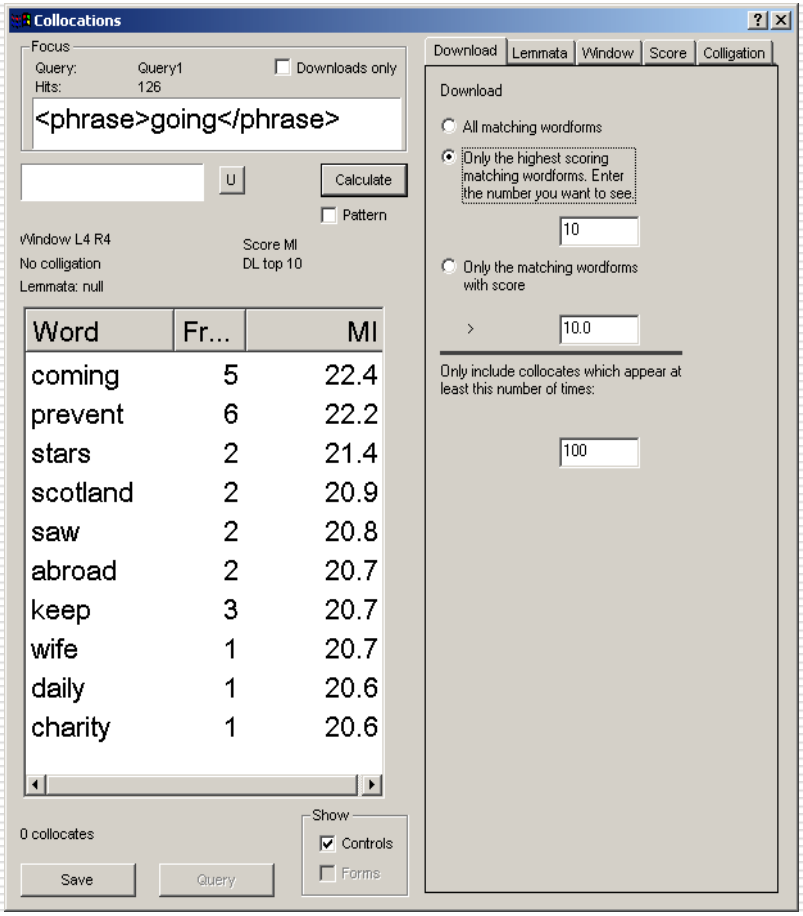
- Enter a stem or pattern
- See forms and frequenc
- See selected variants
- Look up hits
- Save wordlists



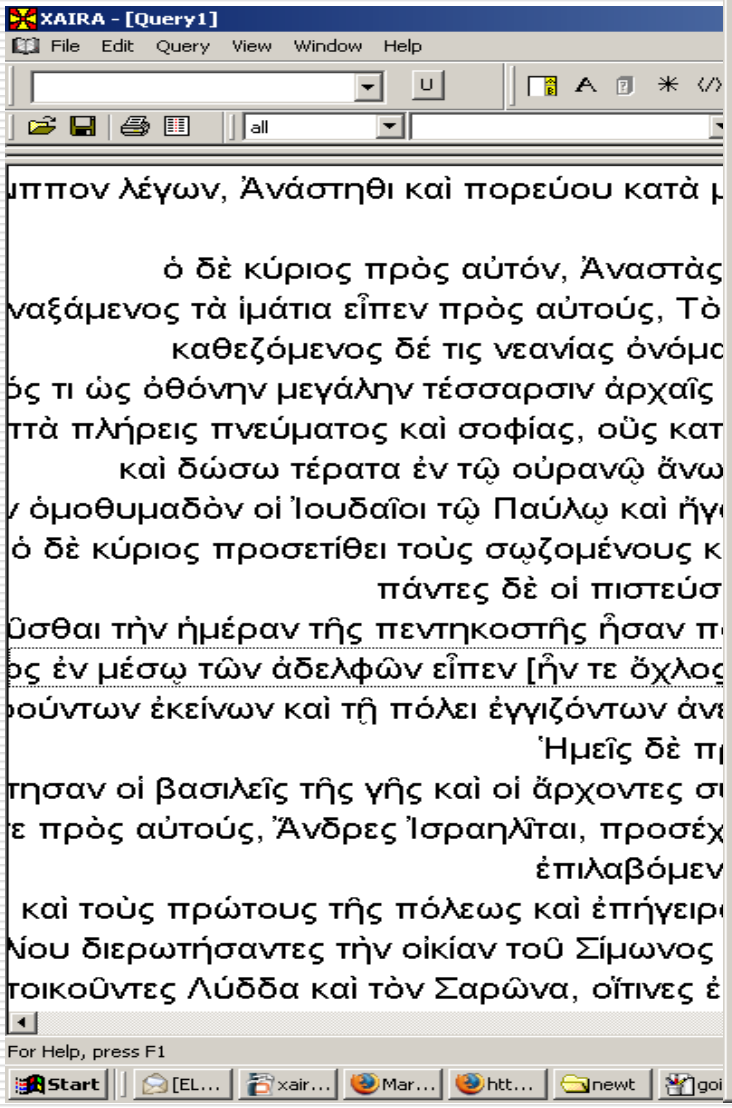


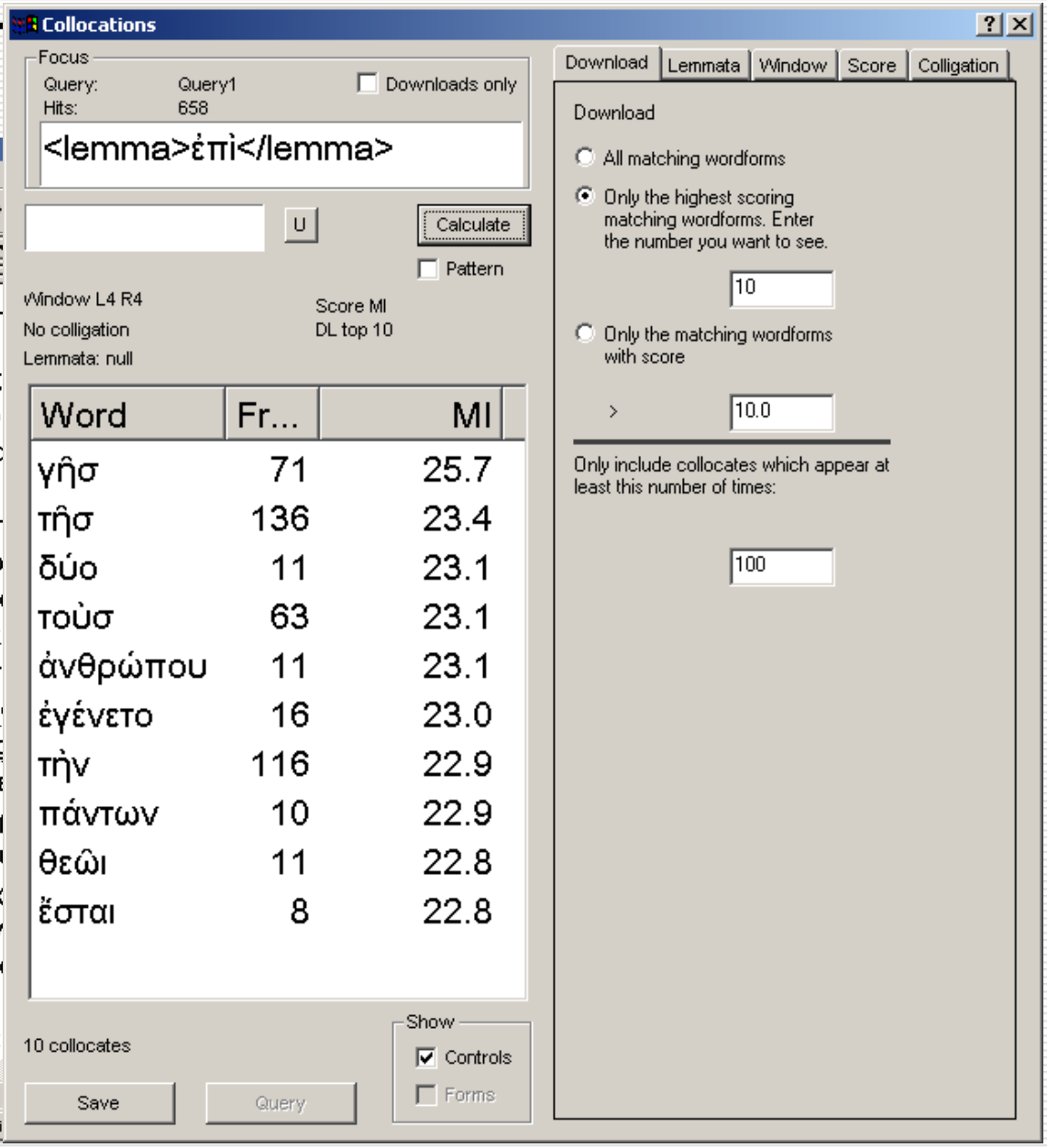
A collocation list

- After any query...
- See word forms that collocate with it
- Adjust window, filter by frequency, select score
- Look up collocates
- Save wordlists

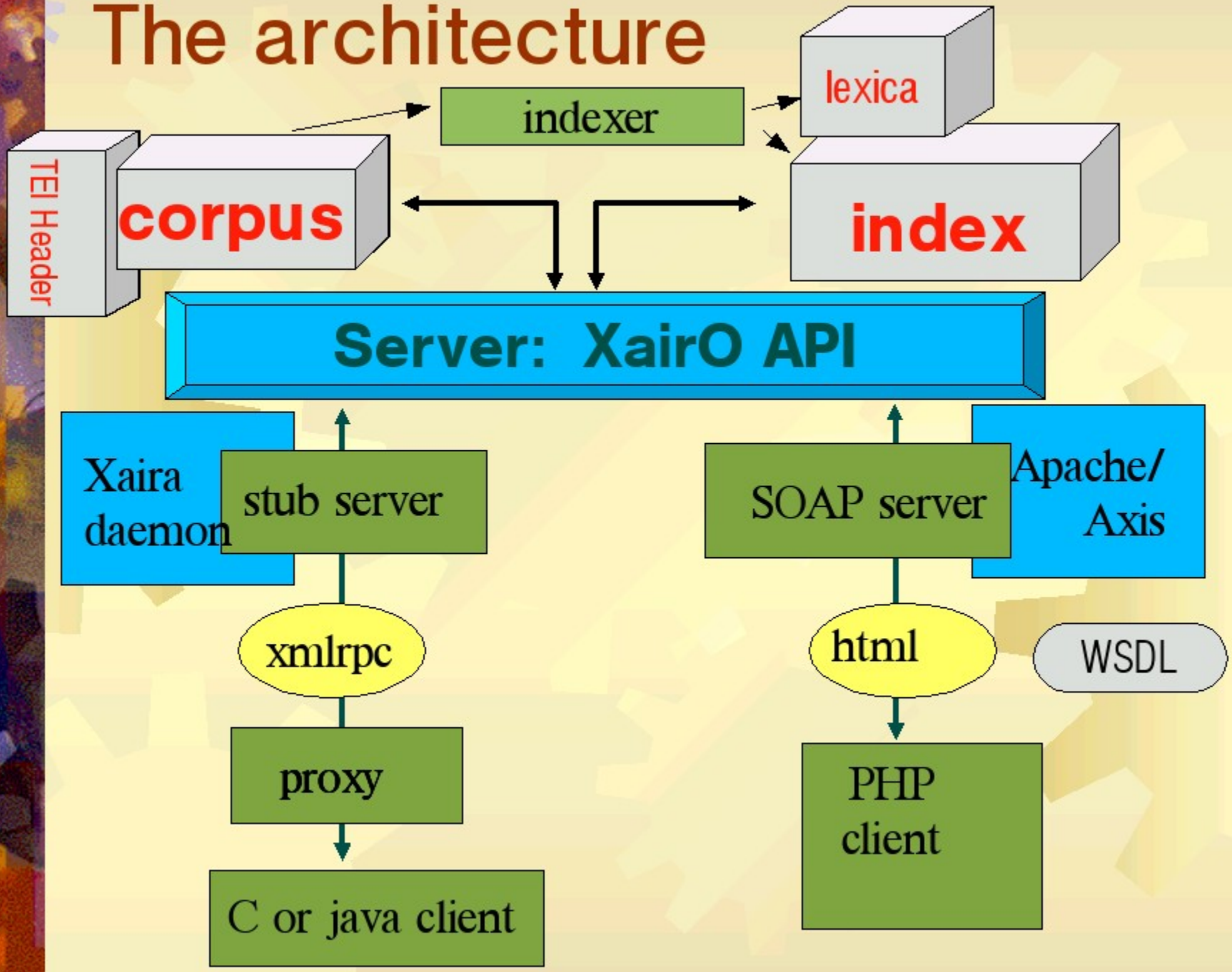


A Keyword in Con





The architecture



XXQ: the Xaira XML Query Language

- Simple XML vocabulary for searching
  - word, punctuation mark, substring
  - word + secondary keys (e.g. POS)
  - XML start- or end-tag, plus attributes
  - Unicode-compliant regular expressions
- Facilities include
  - usual Boolean operations
  - sequence, disjunction, join
  - scoped searching, gaps
- Special lexical features

XXQ more formally

- XXQ is a pattern matching language
- Returns manipulable streams of matching *locations* in a corpus
- XXQ query components
  - simple: <form>, <addKey>, <lemma>, <element>, <pattern>
  - complex: <seq>, <and>, <or>
- matching may be literal or regexp

XXQ examples

```
<form>fish</form>
```

returns locations of the form *fish*

```
<lemma scheme="x">fish</lemma>
```

returns locations of forms which have the lemma *fish* in scheme x

```
<addKey name="POS">VB</addKey>
```

returns locations of forms which have an additional key of type *POS* with the value *VB*

Element vs pattern queries -1

- element query
  - Finds XML elements
    - Independent of their location
    - Restriction by attribute value
- pattern query
  - Restricts locations in terms of their XML context (by using Xpath)

Element vs pattern queries -2

```
<element name=”book”>
 <attribute name=”n”>acts</attribute>
</element>
```

```
<pattern match=”book[@n=Acts]”>
<form>Paul</form>
</pattern>
```

combinations -1

- <seq>: ordered sequence (possibly containing <gap>s)

```
<seq>
 <form>fish</form>
 <gap max=”2”/>
 <form>chips</form>
</seq>
```

combinations -2

- <and>: union operator

```
<and>
 <form>fish</form>
 <addkey name=”pos”>VERB</addkey>
</and>
```

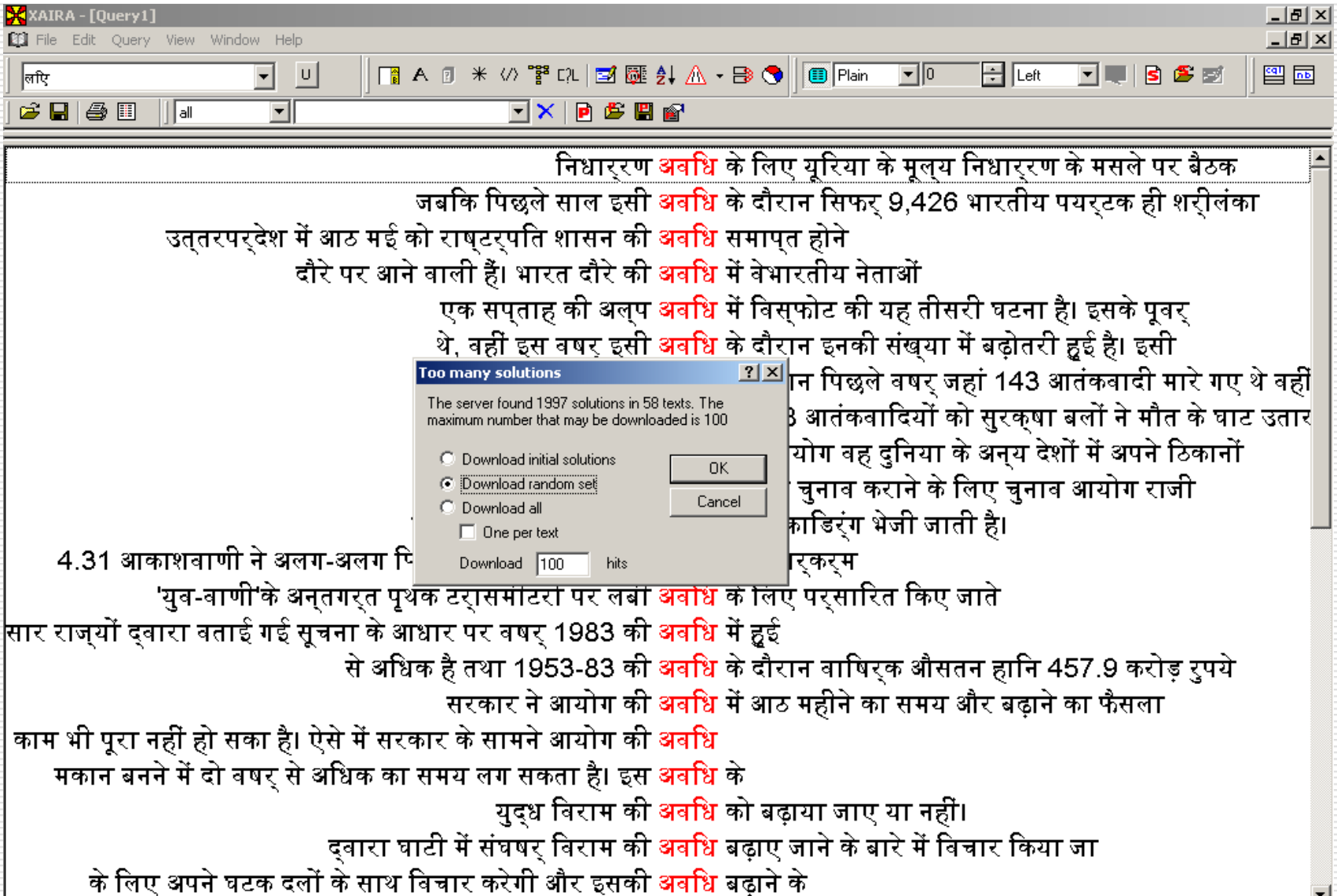
- <or>: disjunction operator

```
<or>
<seq><gap/><form>fish</form></seq>
<seq><gap/><form>chips</form></seq>
</or>
```

This is all hidden behind the GUI, of course

- Word and lemma query
- User-configurable display
  - plain, XML, user-defined stylesheets
- Texts, Results, Browse windows
- Results can be exported in XML
- “visual interface” for complex queries

Exploring a text



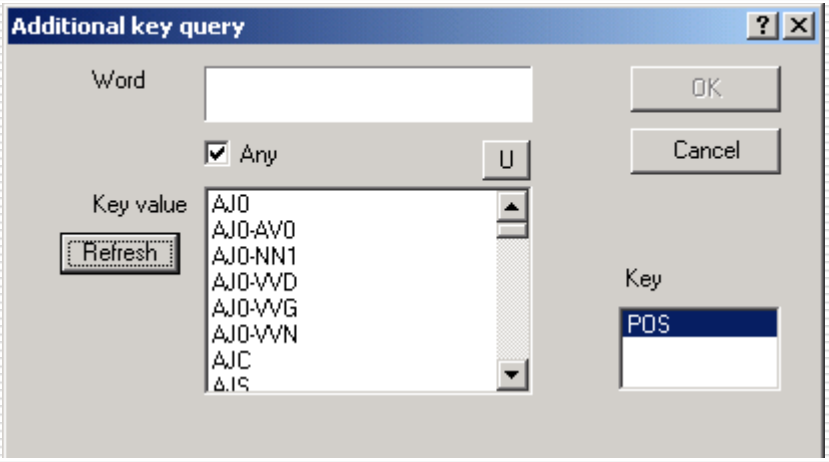
निधार्रण अवधि के लिए यूरिया के मूल्य निधार्रण के मसले पर बैठक
जबकि पिछले साल इसी अवधि के दौरान सिफर् 9,426 भारतीय पयर्टक ही शर्ीलंका
उत्तरपर्देश में आठ मई को राष्टर्पति शासन की अवधि समाप्त होने
दौरे पर आने वाली हैं। भारत दौरे की अवधि में वेभारतीय नेताओं
एक सप्ताह की अल्प अवधि में विस्फोट की यह तीसरी घटना है। इसके पूवर्
थे, वहीं इस वषर् इसी अवधि के दौरान इनकी संख्या में बढ़ोतरी हुई है। इसी
ान पिछले वषर् जहां 143 आतंकवादी मारे गए थे वहीं
3 आतंकवादियों को सुरक्षा बलों ने मौत के घाट उतार
योग वह दुनिया के अन्य देशों में अपने ठिकानों
चुनाव कराने के लिए चुनाव आयोग राजी
काडिर्ंग भेजी जाती है।
4.31 आकाशवाणी ने अलग-अलग पि र्कर्म
'युव-वाणी'के अन्तगर्त पृथक ट्रासमीटरों पर लबी अवधि के लिए पर्सारित किए जाते
सार राज्यों द्वारा बताई गई सूचना के आधार पर वषर् 1983 की अवधि में हुई
से अधिक है तथा 1953-83 की अवधि के दौरान वाषिर्क औसतन हानि 457.9 करोड़ रुपये
सरकार ने आयोग की अवधि में आठ महीने का समय और बढ़ाने का फैसला
काम भी पूरा नहीं हो सका है। ऐसे में सरकार के सामने आयोग की अवधि
मकान बनने में दो वषर् से अधिक का समय लग सकता है। इस अवधि के
युद्ध विराम की अवधि को बढ़ाया जाए या नहीं।
द्वारा घाटी में संघषर् विराम की अवधि बढ़ाए जाने के बारे में विचार किया जा
के लिए अपने घटक दलों के साथ विचार करेगी और इसकी अवधि बढ़ाने के

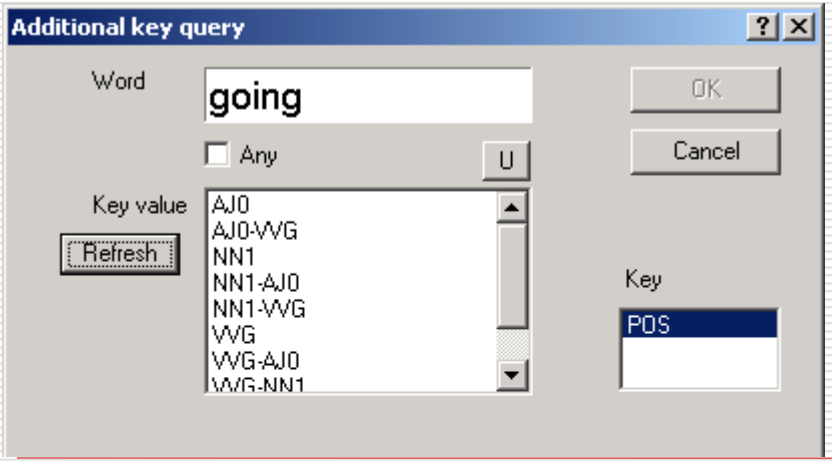
Referencing system

- Corpus
- Subcorpus
- Hit number : total hits (texts)
- Text identifier
- Context identifier

Additional keys (e.g. POS codes)

- “find a POS code, whatever word it’s attached to”

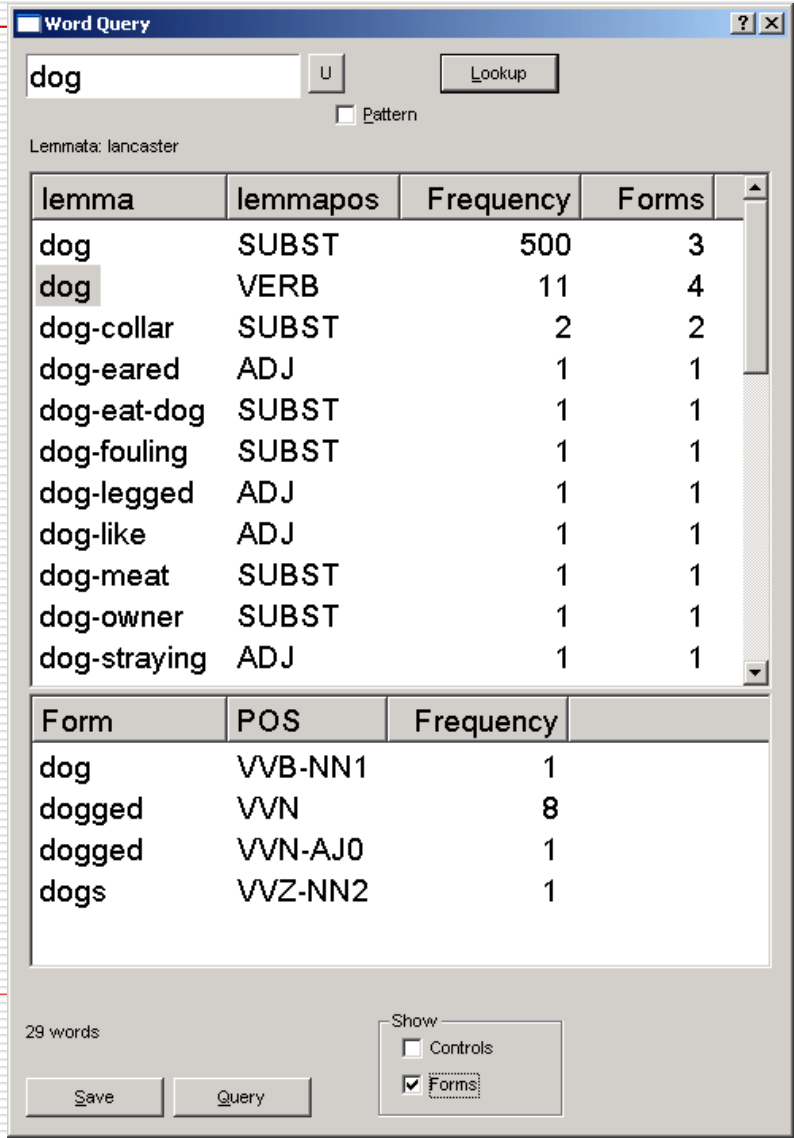




- “find a POS code only when it’s attached to ‘going’”

Multi-part lemmatization

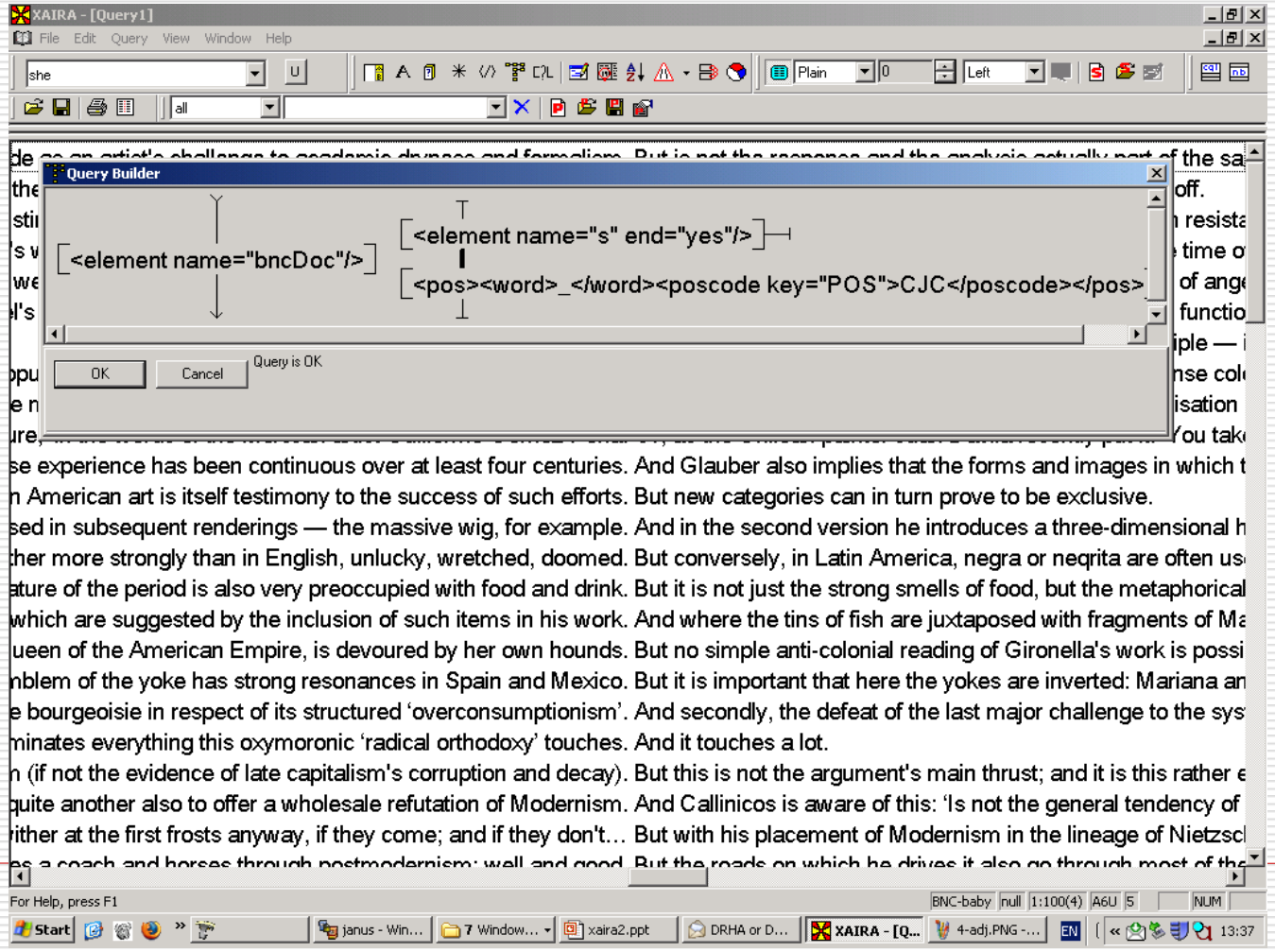
- DOG (verb)
  - Dog VVB
  - Dogged VVN
  - Dogs VVZ
  - Dogging VVI
- DOG (substantive)
  - Dog NN1
  - Dogs NN2
  - Dog NP1



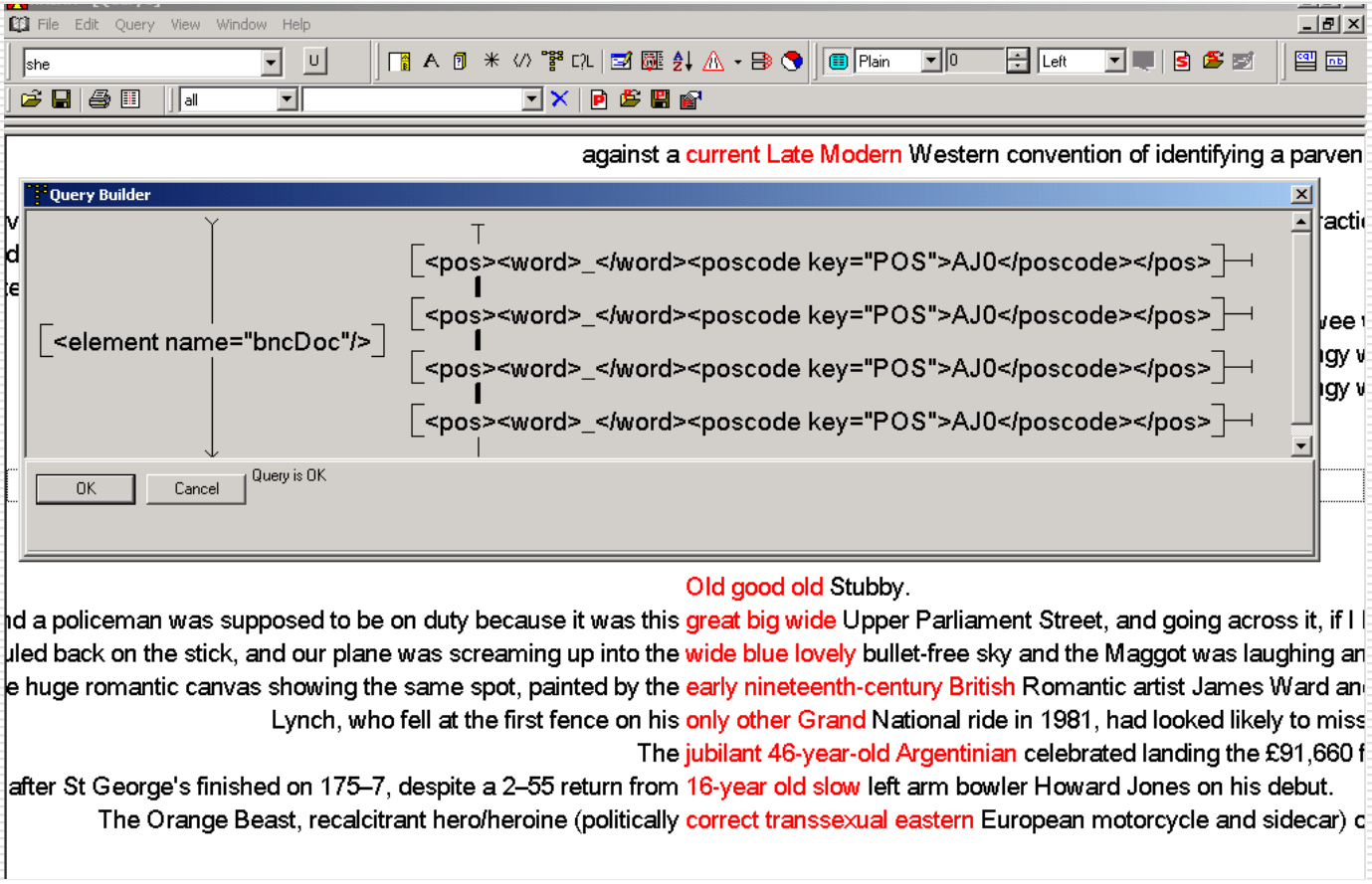
Building complex queries

- visual interface
- *scope node* defines where to look
  - an XML element
  - by span
- *query node*s define what to look for
  - word, phrase, addkey, pattern, XML
- *link type*s define sequence in which query node targets should occur
  - next, one-way, two-way

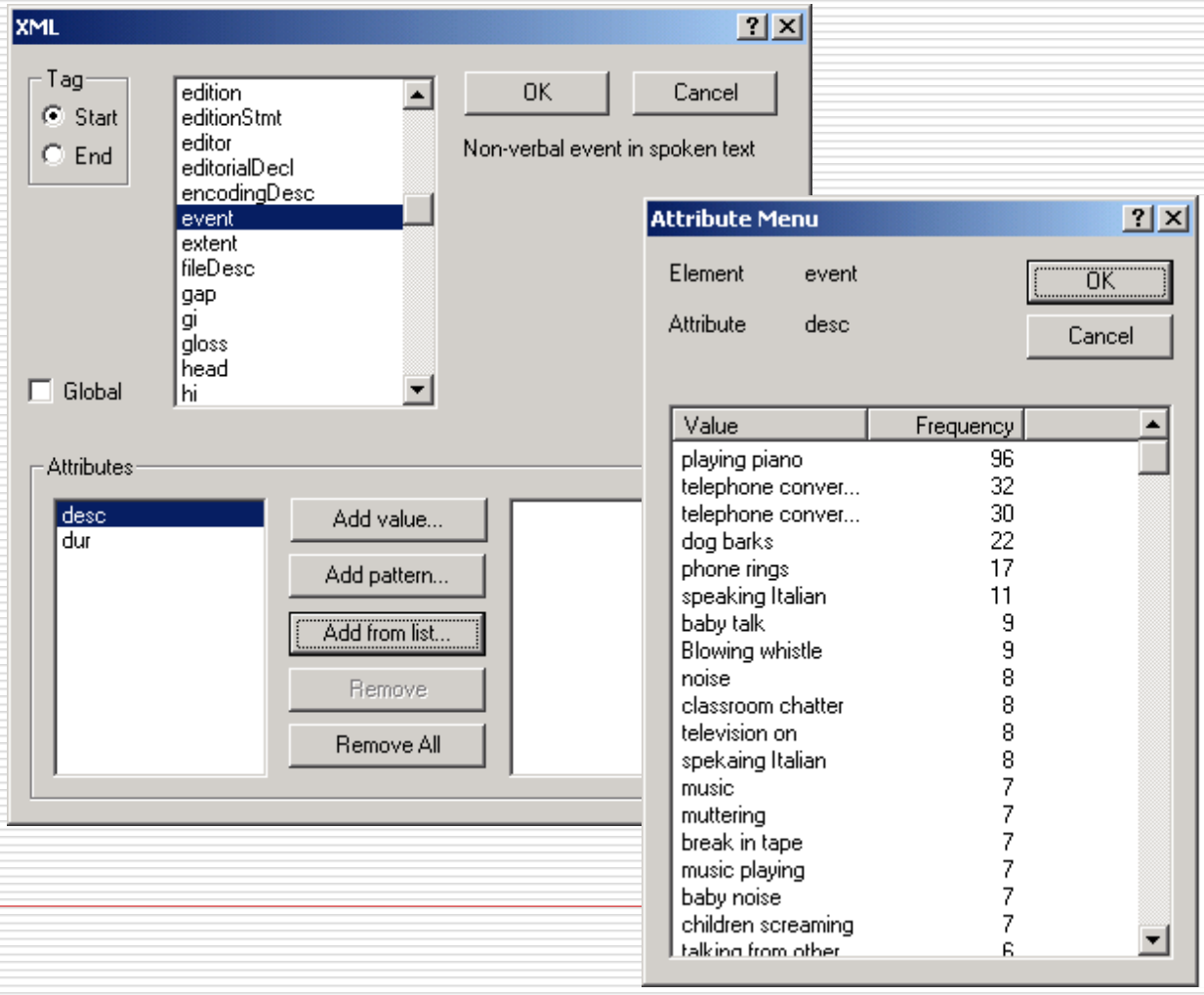
Sentences beginning with conjunctions



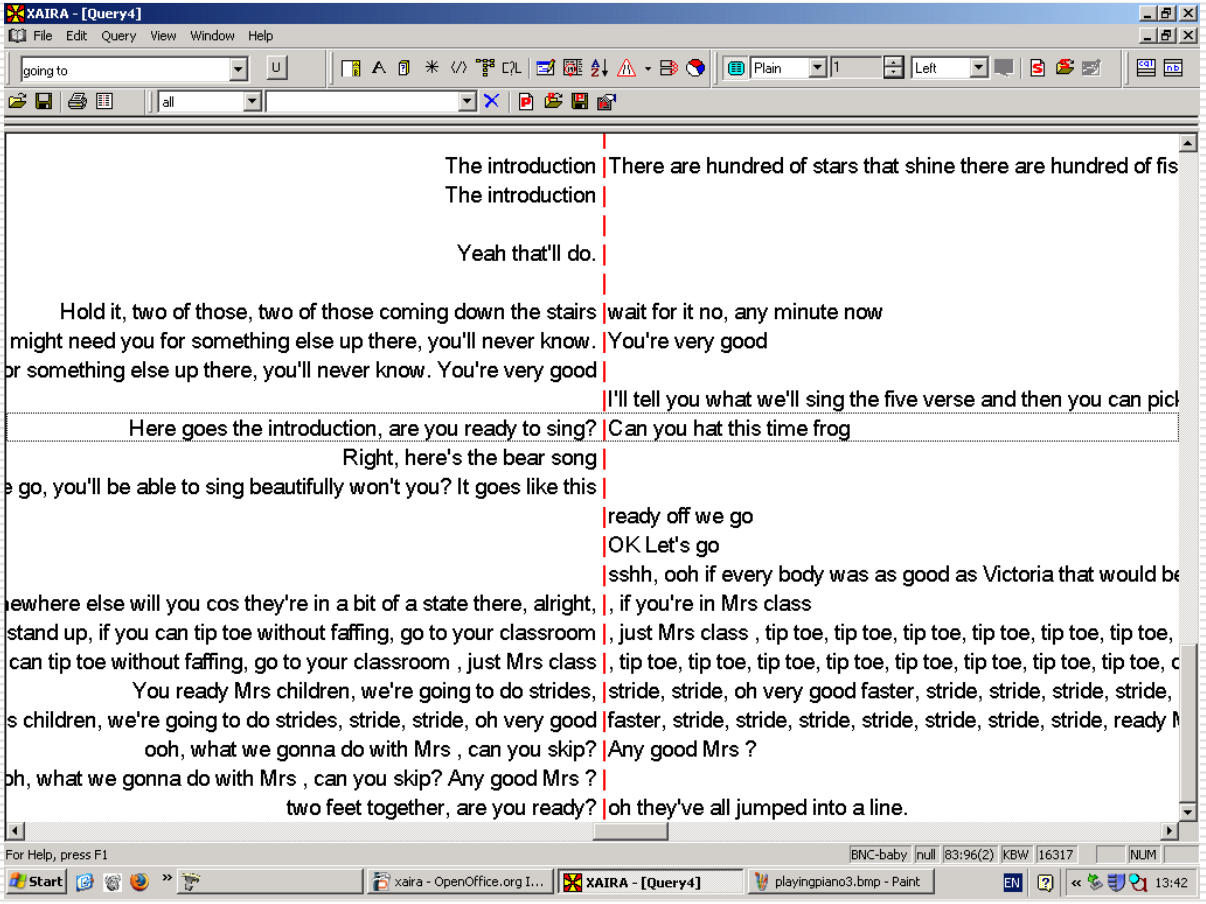
Sequences of four adjectives



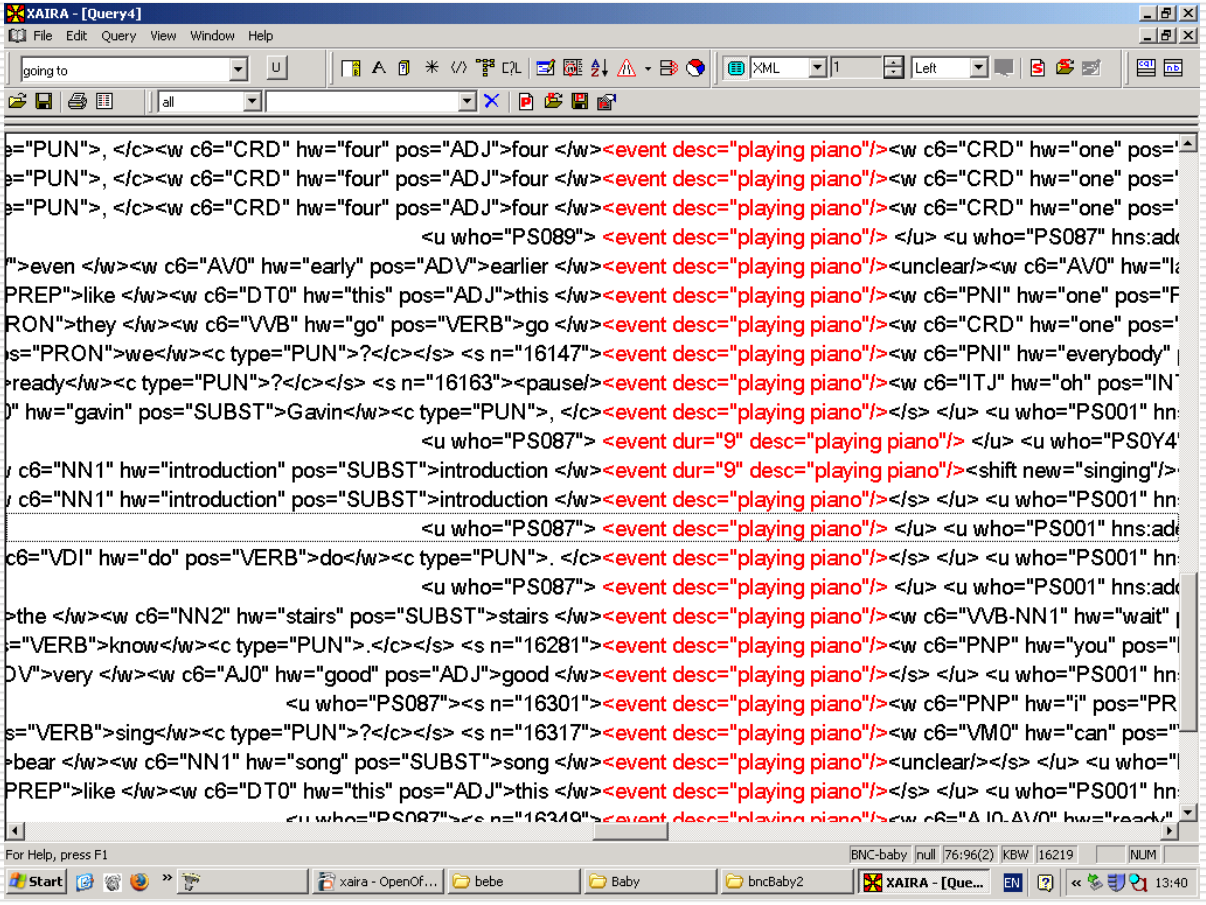
XML Queries


XML
Tag
Start
End
Global
edition
editionStmt
editor
editorialDecl
encodingDesc
event
extent
fileDesc
gap
gi
gloss
head
hi
OK
Cancel
Non-verbal event in spoken text
Attributes
desc
dur
Add value...
Add pattern...
Add from list...
Remove
Remove All
Attribute Menu
Element event
Attribute desc
OK
Cancel
Value
Frequency
playing piano 96
telephone conver... 32
telephone conver... 30
dog barks 22
phone rings 17
speaking Italian 11
baby talk 9
Blowing whistle 9
noise 8
classroom chatter 8
television on 8
spekaing Italian 8
music 7
muttering 7
break in tape 7
music playing 7
baby noise 7
children screaming 7
talking from other 6

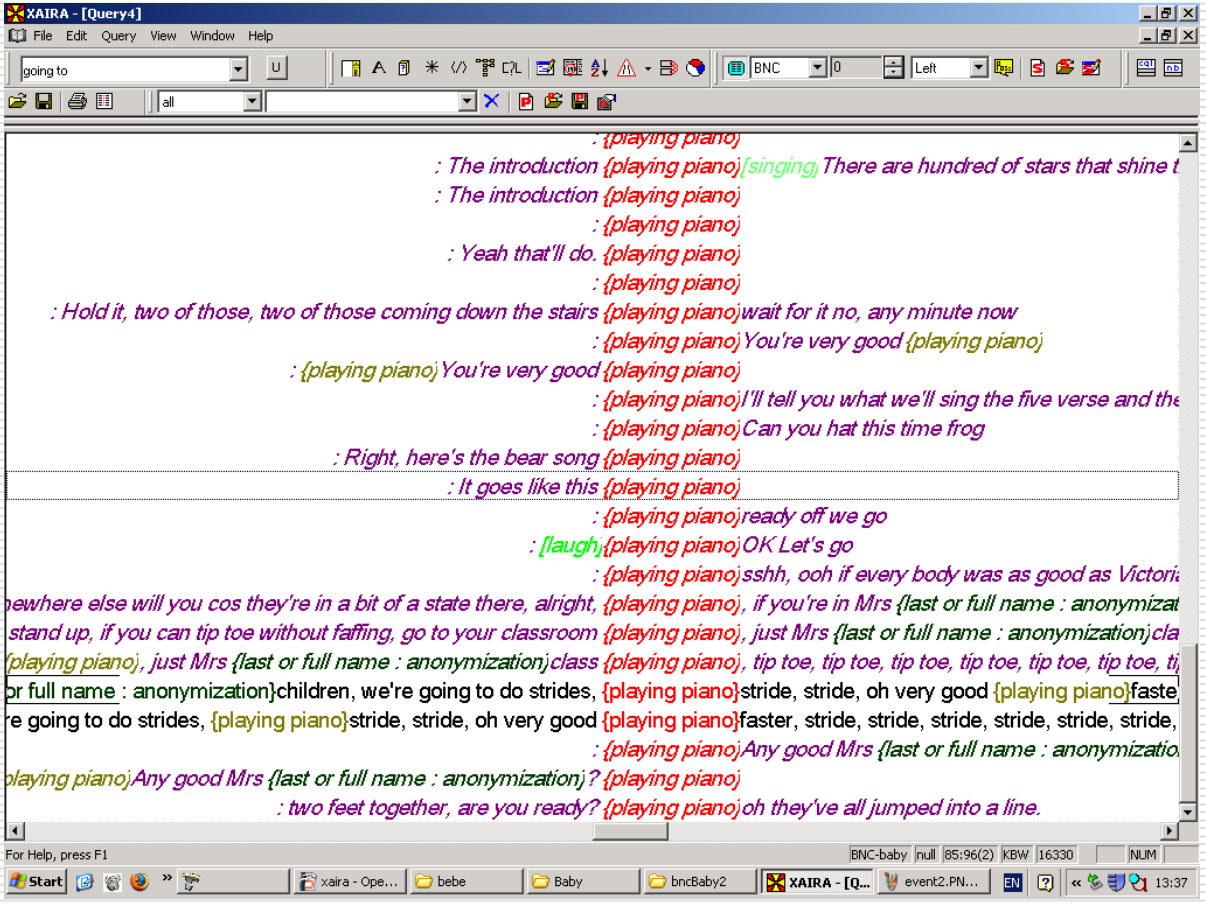
“playing piano” – plain view



“playing piano” – xml view



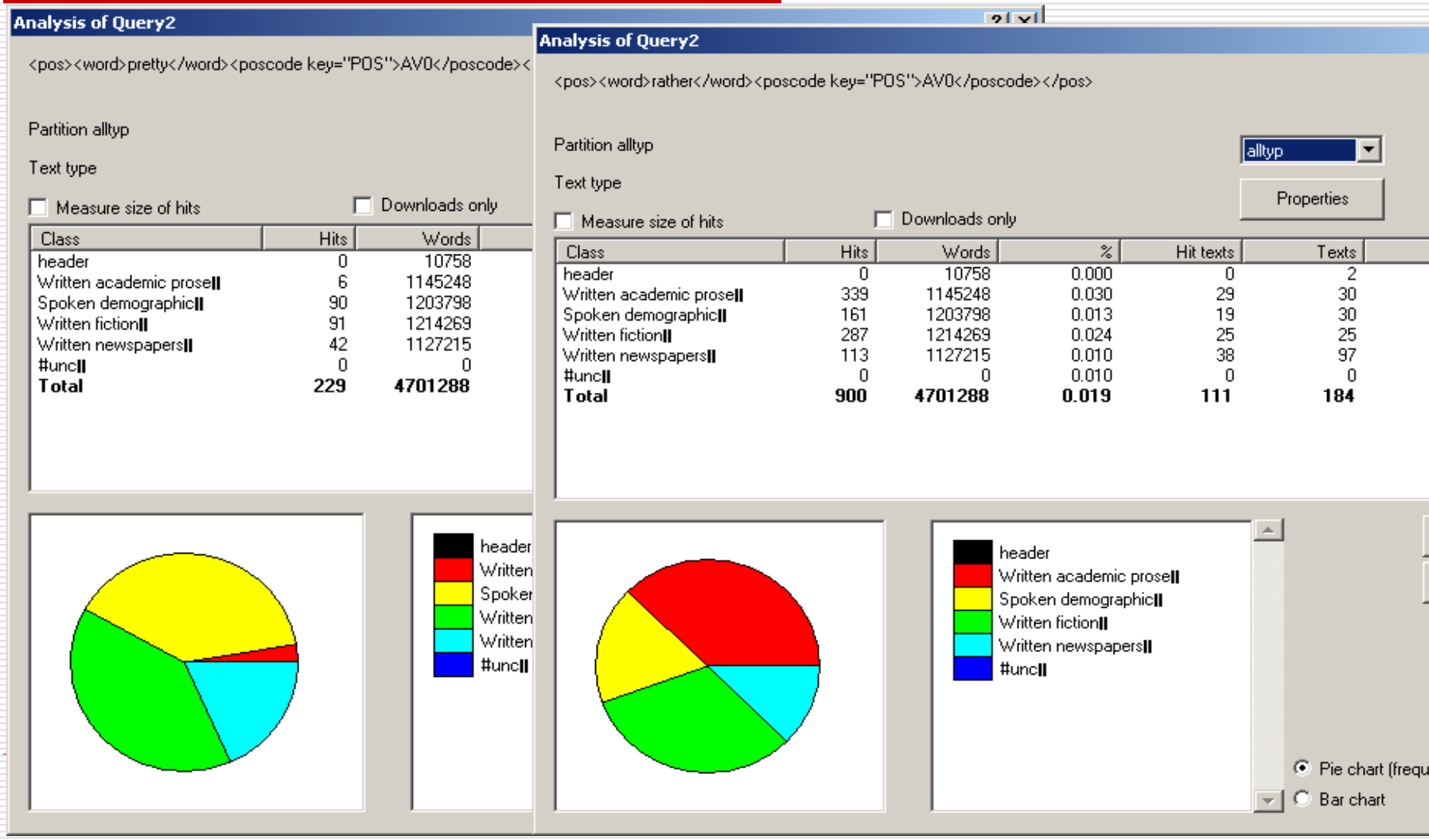
“playing piano” – custom view



Categorizing texts in a corpus

- A corpus can be *partitioned* in many ways
  - By pre-assigned category, domain, text-type, topic etc. specified in the markup
  - Manually or by running a query
- Searches can be restricted to texts of a given classification (subcorpora)
- Or results compared across the partition

“pretty” vs “rather” in BNC Baby



What is XAIRA's niche?

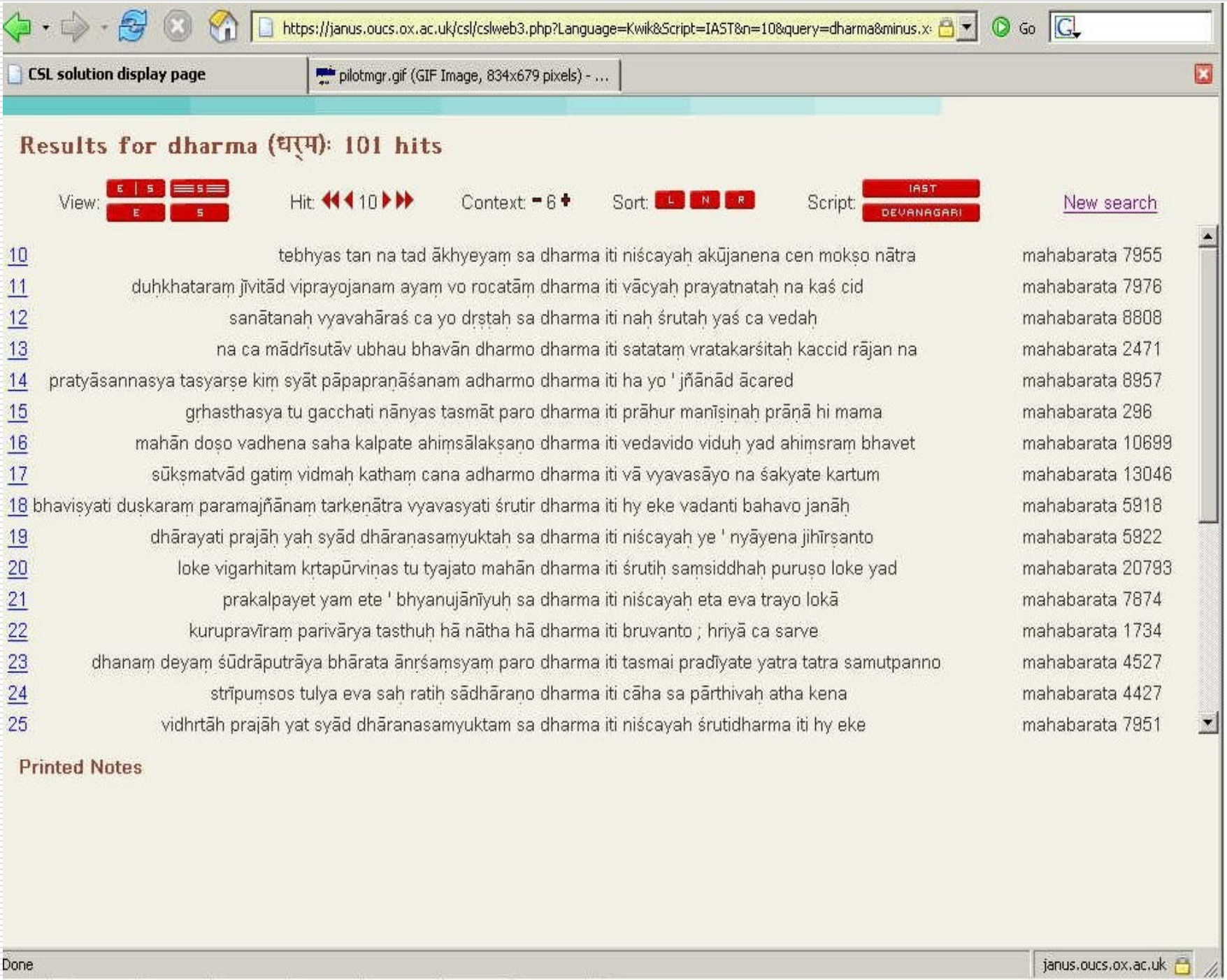
- Web search engines
  - patchy and unknowable coverage
  - designed to recover content, not word forms
  - hard to cite, harder to process
- XML display engines
  - expensive, geared to reader not searcher
  - focus on presentation rather than content
- As a back end for your next generation web application

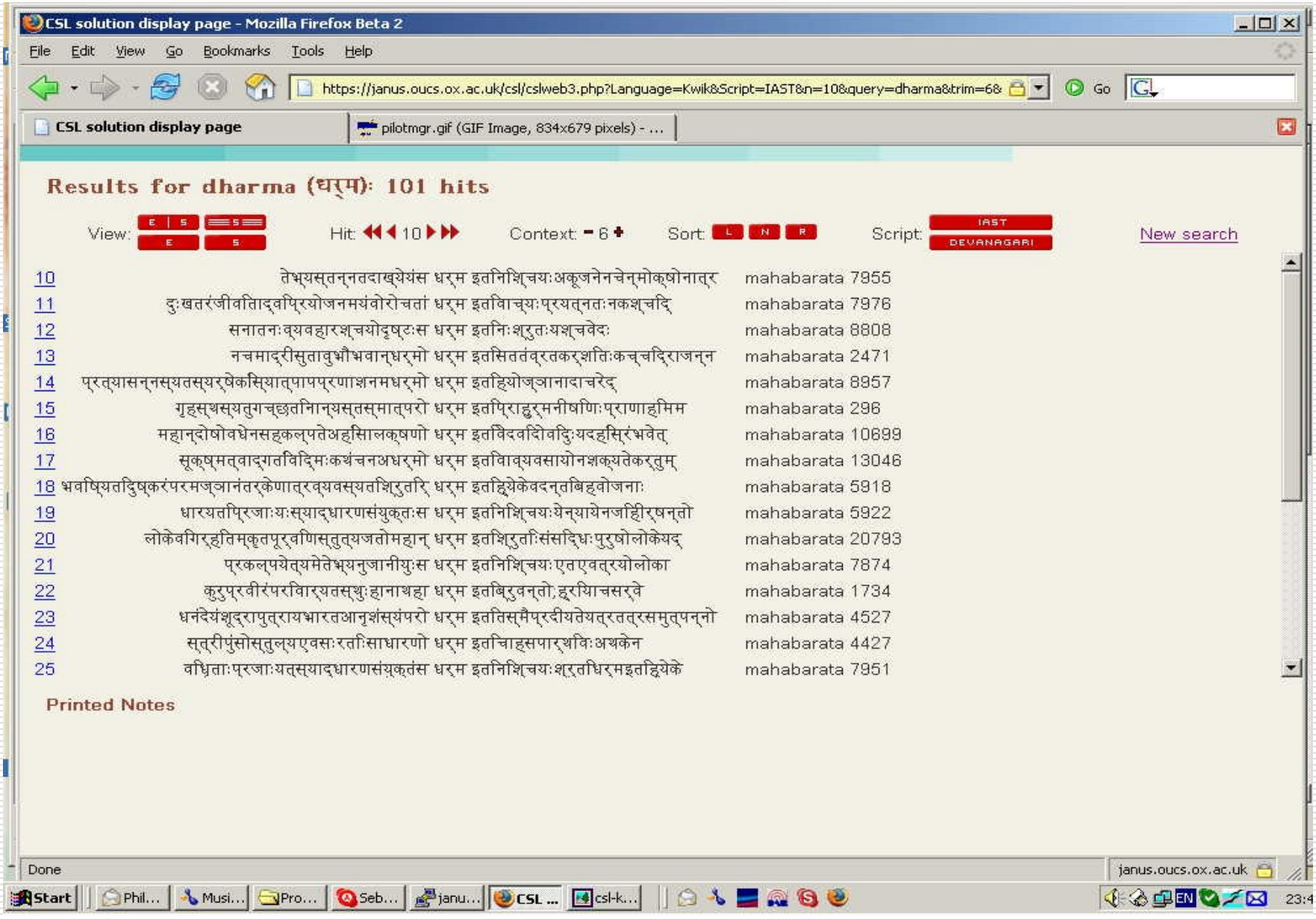
Interfaces available now

- Full-featured Windows client
  - can be run standalone or as interface to Windows, Unix, or Mac OSX server
  - includes simple corpus building utility
- Cross-platform
  - SOAP interface
  - Simple PHP and Java clients
  - Server and indexer
- Xpath engine

Using Xaira as a back end

- PHP interface
  - supports all calls to the Xaira object model
  - suitable for any web programmer
  - platform independent
- For example...


https://janus.oucs.ox.ac.uk/csl/cslweb3.php?Language=Kwik&Script=IAST&n=10&query=dharma&minus.x:
Go
CSL solution display page
pilotmgr.gif (GIF Image, 834x679 pixels) - ...
Results for dharma (धर्म): 101 hits
View:
Hit: 10
Context: 6
Sort:
Script: IAST DEVANAGARI
New search
10 tebhyas tan na tad ākhyeyaṃ sa dharma iti niścayaḥ akūjanena cen mokṣo nātra mahabarata 7955
11 duḥkhataraṃ jīvitād viprayojanam ayaṃ vo rocatāṃ dharma iti vācyaḥ prayatnataḥ na kaś cid mahabarata 7976
12 sanātanaḥ vyavahāraś ca yo dṛṣṭaḥ sa dharma iti naḥ śrutaḥ yaś ca vedaḥ mahabarata 8808
13 na ca mādrīsutāv ubhau bhavān dharmo dharma iti satataṃ vratakarśitaḥ kaccid rājan na mahabarata 2471
14 pratyāsannasya tasyarṣe kiṃ syāt pāpapraṇāśanam adharmo dharma iti ha yo ' jñānād ācared mahabarata 8957
15 gṛhasthasya tu gacchati nānyas tasmāt paro dharma iti prāhur manīṣiṇaḥ prāṇā hi mama mahabarata 296
16 mahān doṣo vadhena saha kalpate ahiṃsālakṣaṇo dharma iti vedavido viduḥ yad ahiṃsraṃ bhavet mahabarata 10699
17 sūkṣmatvād gatiṃ vidmaḥ kathaṃ cana adharmo dharma iti vā vyavasāyo na śakyate kartum mahabarata 13046
18 bhaviṣyati duṣkaraṃ paramajñānaṃ tarkeṇātra vyavasyati śrutir dharma iti hy eke vadanti bahavo janāḥ mahabarata 5918
19 dhārayati prajāḥ yaḥ syād dhāraṇasaṃyuktaḥ sa dharma iti niścayaḥ ye ' nyāyena jihīrṣanto mahabarata 5922
20 loke vigarhitam kṛtapūrviṇas tu tyajato mahān dharma iti śrutiḥ saṃsiddhaḥ puruṣo loke yad mahabarata 20793
21 prakalpayet yam ete ' bhyanujānīyuḥ sa dharma iti niścayaḥ eta eva trayo lokā mahabarata 7874
22 kurupravīraṃ parivārya tasthuḥ hā nātha hā dharma iti bruvanto ; hriyā ca sarve mahabarata 1734
23 dhanaṃ deyaṃ śūdrāputrāya bhārata ānṛśaṃsyaṃ paro dharma iti tasmai pradīyate yatra tatra samutpanno mahabarata 4527
24 strīpuṃsos tulya eva saḥ ratiḥ sādhāraṇo dharma iti cāha sa pārthivaḥ atha kena mahabarata 4427
25 vidhṛtāḥ prajāḥ yat syād dhāraṇasaṃyuktam sa dharma iti niścayaḥ śrutidharma iti hy eke mahabarata 7951
Printed Notes
Done
janus.oucs.ox.ac.uk


CSL solution display page - Mozilla Firefox Beta 2
File Edit View Go Bookmarks Tools Help
https://janus.oucs.ox.ac.uk/csl/cslweb3.php?Language=Kwik&Script=IAST&n=10&query=dharma&trim=6&
Go
CSL solution display page
pilotmgr.gif (GIF Image, 834x679 pixels) - ...
Results for dharma (धर्म): 101 hits
View:
Hit: 10
Context: 6
Sort:
Script: IAST DEVANAGARI
New search
10 तेभ्यस्तन्नतदाख्येयंस धर्म इतिनिश्चयःअकूजनेनचेन्मोक्षोनात्र mahabarata 7955
11 दुःखतरंजीवितिाद्वपि्रयोजनमयंवोरोचतां धर्म इतिवाच्यःप्रयत्नतःनकश्चिद् mahabarata 7976
12 सनातनःव्यवहारश्चयोदृष्टःस धर्म इतिनिःश्रुतःयश्चवेदः mahabarata 8808
13 नचमाद्रीसुतावुभौभवान्धर्मो धर्म इतिसततंव्रतकर्शतिःकच्चिद्राजन्न mahabarata 2471
14 प्रत्यासन्नस्यतस्यर्षेकस्यिात्पापप्रणाशनमधर्मो धर्म इतिह्योज्ञानादाचरेद् mahabarata 8957
15 गृहस्थस्यतुगच्छतनिान्यस्तस्मात्परो धर्म इतपि्राहुर्मनीषणिःप्राणाहुमिम mahabarata 296
16 महान्दोषोवधेनसहकल्पतेअहसिालक्षणो धर्म इतविदवदिोवदिुःयदहसि्रंभवेत् mahabarata 10699
17 सूक्ष्मत्वाद्गतविदिमःकथंचनअधर्मो धर्म इतिवाव्यवसायोनशक्यतेकर्तुम् mahabarata 13046
18 भवषि्यतदिुष्करंपरमज्ञानंतर्केणात्रव्यवस्यतशि्रुतरि् धर्म इतिह्येकेवदन्तबिहवोजनाः mahabarata 5918
19 धारयतपि्रजाःयःस्याद्धारणसंयुक्तःस धर्म इतिनिश्चयःयेन्यायेनजहिीर्षन्तो mahabarata 5922
20 लोकेवगिर्हतिम्कृतपूर्वणिस्तुत्यजतोमहान् धर्म इतशि्रुतिःसंसदि्धःपुरुषोलोकेयद् mahabarata 20793
21 प्रकल्पयेत्यमेतेभ्यनुज्ञानीयुःस धर्म इतिनिश्चयःएतएवत्रयोलोका mahabarata 7874
22 कुरुप्रवीरंपरविार्यतस्शुःहानाथहा धर्म इतब्रिुवन्तो;हूरयिाचसर्वे mahabarata 1734
23 धनंदेयंशूद्रापुत्रायभारतआनृशंस्यंपरो धर्म इतितस्मैप्रदीयतेयत्रततत्रसमुत्पन्नो mahabarata 4527
24 सत्रीपुंसोस्तुल्यएवसःरतिःसाधारणो धर्म इतिचाहसपार्थविःअथकेन mahabarata 4427
25 वधिताःप्रजाःयत्स्याद्धारणसंयुक्तंस धर्म इतिनिश्चयःश्रुतधिर्मइतह्येके mahabarata 7951
Printed Notes
Done
janus.oucs.ox.ac.uk
Start
Phil...
Musi...
Pro...
Seb...
janu...
CSL ...
csl-k...
EN
23:4

Frequently Asked Questions

- Does it work on a Mac?
- Does it work with non-XML text?
- How much does it cost?
- How do I make it do xxxx?
- Where's the documentation?