Functional linguistics and human language technology: new opportunities — or has SFL missed the boat?

John Bateman

ISFC30 Lucknow, Friday 12th December 2003

Overview of talk

- Interactions between SFL and computational approaches to language
- The emergence of "human language technology" as a development area
- After corpora: parsed corpora and multi-treebanks
- Tools for the working linguist: open architectures and interoperability

SFL and computation: interactions

Halliday (1956) "The linguistic basis of a mechanical thesaurus..."
Halliday (1962) "Linguistics and machine translation"

Henrici (1966) "Notes on the systemic generation of a paradigm of the English clause"

Fawcett (1973) "Generating a sentence in systemic functional grammar"

Davey (1974) "Discourse production: a computer model of some aspects of a speaker"

McCord (1977) "Procedural systemic grammars"

SFL and computation: interactions

Mann/Matthiessen/Halliday (1983) The Penman text generation system and the Nigel grammar

Cummings (1985) "A PROLOG parser-generator for systemic analysis of Old English nominal groups"

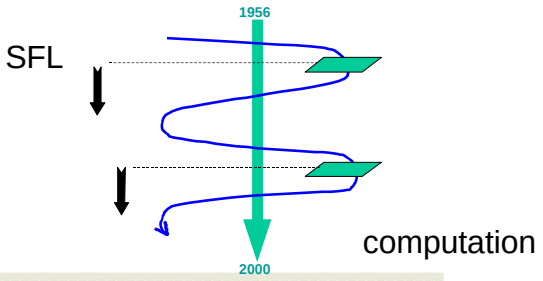
Fawcett/Tucker (1988) Communal text generation system and the Cardiff Grammar

SFL and computation: interactions

Marilyn Cross (1992) 'Horace'
Mick O'Donnell (1994) 'Wag'
Chris Nesbitt (1994) 'HyperGrammar'
Petie Sefton (1995) interaction
Licheng Zeng (1993-96) 'Multex'
Elke Teich (1999) 'Komet'
Liesbeth Degand (1996) Dutch
Brigitte Grote (1996-) German
Gordon Tucker (1995) adjectives

multilinguality
multimodality
analysis
detailed descriptions
development and documentation environments

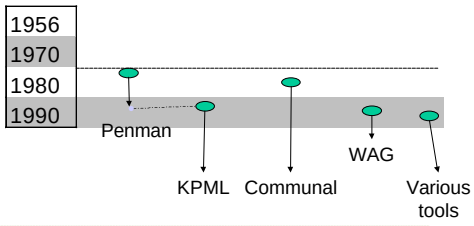
Interactions



The state of computation at each point of interaction

1956	theoretical notion of an algorithm, virtually no memory, extremely limited interaction with user
1970	limited memory, slow, basic user interaction
1980	memory available (but expensive), slow, beginnings of varied user interaction: graphical interfaces and debugging environments
1990	more memory available (price dropping), speed increasing, varied possibilities for interaction

Computational SFL systems



The state of computation at each point of interaction: 2000

2000	**memory**: for all practical purposes, limitless **speed**: for natural language processing tasks, fast **interaction styles**: graphical, textual, touch, ...

- Human Language Technology
- Language Engineering
- Linguistic Engineering

Human Language Technology

- industrial interest in language applications
- substantially larger budgets
- many research and development groups in both universities and companies
  - large lexicons
  - large thesauri (e.g., EDR, Wordnet)
  - ever larger corpora of different kinds of language

Human Language Technology

- industrial interest in language applications
- substantially larger budgets
- many research and development groups in both universities and companies
  - large lexicons
  - large thesauri (e.g., EDR, Wordnet)
  - ever larger corpora of different kinds of language

Human Language Technology: Impacts

- There are a considerable number of applications and tasks that can be addressed by a combination of:
  - relatively simple techniques
  - very large scale source data
- Examples
  - speech synthesis
  - information retrieval (also multilingual)

Human Language Technology: emphases

- much more interest (need) for large-scale handling of material **automatically**
- very substantial efforts (EU, India, ...) on collecting **multilingual** language data
- great concern with **evaluation** and **evaluation criteria**
- notions of 'best practice' and **standardization** (both actual and de facto)

Information vs. Meaning

- Scale alone does not create meaning
- Result:
  - particular ways of structuring information in order to make aspects of its 'meaning' more accessible
  - particular techniques for processing such 'structured data'

An example of 'adding meaning': corpora and annotations

- Just placing a few hundred million words in a computer file does not mean that one has a useful research resource
- Necessary to support the search for significant patterns
- **Development**: combination of corpora and 'mark-up' or annotation technology

An example of 'adding meaning': corpora and annotations

- Just placing a few hundred million words in a computer file does not mean that one has a useful research resource
- Necessary to support the search for significant patterns
- **Development**: combination of corpora and 'mark-up' or annotation technology

Three steps...

- Corpora: from raw text to marked-up text
- Text encoding in general
- Corpora: from marked-up text to structured data

Step 1

- Corpora: from raw text to marked-up text
- Text encoding in general
- Corpora: from marked-up text to structured data

The problems of searching...

Question: search for bad weather in the novel X...

(1) select some useful words: *storm, rain, gale, wind*
(2) search and count
(3) Results: *storm (32), rain (108), gale (75), wind (345)*

strained, restraint, drain

windlass, windward, tradewinds

But: if we were looking for 'to **wind'** in a different sense, we would not then find "**wound**".

Example of tagged text from the BNC

Garside, R.G., Leech, G.N., and Sampson, G.R. (eds) (1987). *The Computational Analysis of English: A Corpus-based Approach*. Longman, London.

<s c="0000002 002" n=00001>When&AVQ-CJS; Captain&NP0; Pugwash&NP0; retires&VVZ; from&PRP; active&AJ0; piracy&NN1; he&PNP; is&VBZ; amazed&AJ0-VVN; and&CJC; delighted&AJ0-VVN; to&TO0; be&VBI; offered&VVN; a&AT0; Huge&AJ0; Reward&NN1; for&PRP; what&DTQ; seems&VVZ; to&TO0; be&VBI; a&AT0;simple&AJ0; task&NN1;.&PUN;

Corpora+POS

- most corpora nowadays are tagged at least with 'part of speech' information
- this can then be used in queries asked of the corpus
- POS-tagging for English is quite reliable

Morphological analysis:

e.g. "These were only some simplest sample sentences."

These	these+Det+Pl
These	these+Pron+NomObl+3P+Pl
were	be+Verb+PastTense+Pl
only	only+Adj
only	only+Adv
only	only+Conj+Sub
some	some+Pron+NomObl+3P+Pl
some	some+Det+SP
simplest	simple+Adj+Sup
sample	sample+Noun+Sg
sample	sample+Verb+Pres+Non3sg
sentences	sentence+Noun+Pl
sentences	sentence+Verb+Pres+3sg
.	.+?

Results from the Xerox morphological analyser and tagger

... a typical HLT result

Step 2

- Corpora: from raw text to marked-up text
- Text encoding in general
- Corpora: from marked-up text to structured data

Text Encoding Initiative (1995)

- a large effort by the Association for Computing and the Humanities, the Association for Literary and Linguistic Computing, and others.
- published guidelines for encoding electronic forms of documents for exchange and research
- based on SGML (an existing standard)
- attempts to make the structural details of text clear for archival of editions, contrasting editions, etc.

Text Encoding Initiative: example

"Have you, miss? Well, for sure!"

A short time after she pursued, "I seed you go out with the master, but I didn't know you were gone to church to be wed"; and she basted away. John, when I turned to him, was grinning from ear to ear.

"Have you, miss? Well, for sure!"

A short time after she pursued, "I seed you go out with the master, but I didn't know you were gone to church to be wed"; and she basted away. John, when I turned to him, was grinning from ear to ear.

Original

<p>
<q>Have you, miss? Well, for sure!</q></p>
<p>A short time after she pursued,
<q>I seed you go out with the master, but I didn't know you were gone to church to be wed</q>;
and she basted away. John, when I turned to him, was grinning from ear to ear.
</p>

XCES-conform markup

"Have you, miss? Well, for sure!"

A short time after she pursued, "I seed you go out with the master, but I didn't know you were gone to church to be wed"; and she basted away. John, when I turned to him, was grinning from ear to ear.

Original

<p>
<q>Have you, miss? Well, for sure!</q></p>
<p>A short time after she pursued,
<q>I seed you go out with the master, but I didn't know you were gone to church to be wed</q>;
and she basted away. John, when I turned to him, was grinning from ear to ear.
</p>

XCES-conform markup

"Have you, miss? Well, for sure!"

A short time after she pursued, "I seed you go out with the master, but I didn't know you were gone to church to be wed"; and she basted away. John, when I turned to him, was grinning from ear to ear.

Original

<p>
<q>Have you, miss? Well, for sure!</q></p>
<p>A short time after she pursued,
<q>I seed you go out with the master, but I didn't know you were gone to church to be wed</q>;
and she basted away. John, when I turned to him, was grinning from ear to ear.
</p>

XCES-conform markup

"Have you, miss? Well, for sure!"

A short time after she pursued, "I seed you go out with the master, but I didn't know you were gone to church to be wed"; and she basted away. John, when I turned to him, was grinning from ear to ear.

Original

<p>
<q>Have you, miss? Well, for sure!</q></p>
<p>A short time after she pursued,
<q>I seed you go out with the master, but I didn't know you were gone to church to be wed</q>;
and she basted away. John, when I turned to him, was grinning from ear to ear.
</p>

XCES-conform markup

TEI base tag sets

- sets of standardized tags for encoding:
  - prose
  - verse
  - drama
  - transcriptions of speech
  - print dictionaries
  - terminological databases

... and many more extensions and details...

Simple 'TEI-conform' examples: prose

```
<body>
<p>I fully appreciate Gen. Pope's splendid achievements with their invaluable results; but you must know that Major Generalships in the Regular Army, are not as plenty as blackberries.
</p>
</body>
```


TEI-conform examples: verse

```
<lg n=I>
<l>I Sing the progresse of a
deathlesse soule ,</l>
<l>Whom Fate, with God made,
but doth not controule ,</l>
<l>Plac'd in most shapes; all times
before the law</l>
<l>Yoak'd us, and when, and since,
in this I sing.</l>
<l>And the great world to his aged evening;</l>
<l>From infant morne, through manly noone I draw.</l>
<l>What the gold Chaldee, of silver Persian saw,</l>
<l>Greeke brass, or Roman iron, is in this one;</l>
<l>A worke t'out weare Seths pillars, bricke and stone,</l>
<l>And (holy writs excepted) made to yeeld to none,</l>
</lg>
```


TEI-conform examples: prose and edition-specific information

```
<p>I wrote to Moor House and to Cambridge immediately, to say what I had done: fully explaining also why I had thus acted. Diana and <pb ed=ED1 n='475'> Mary approved the step unreservedly. Diana announced that she would <pb ed=ED2 n='485'>just give me time to get over the honeymoon, and then she would come and see me.
```

This markup records the differing pagination of two editions

Motivation for adoption of SGML

- a standard already agreed upon in the print industry for re-use of content
- formal specification allows **validation** of documents marked up as TEI-conformant documents
- aspects of an interpretation of a document are explicitly represented and so can be used for indexing and retrieval

SGML documents must have a

Document Type Definition

```
<act><title>Act I</title>
    <scene><title>Scene I. Elsinore. A platform before the castle.
            </title>
    <stagedir>FRANCISCO at his post. Enter to him BERNARDO.
    </stagedir>
    <speech>
            <speaker>BERNARDO</speaker>
            <line>Who's there?<line>
    </speech>
```

DTD:

```
<!ELEMENT play (title, personae, scndesc, playsubt,
                prologue?, act+, epilogue?)>
<!ELEMENT act (title, subtitle*, prologue?, scene+, epilogue?)>
<!ELEMENT scene (title, subtitle*, (speech | stagedir | subhead)+)>
<!ELEMENT speech (speaker+, (line | stagedir | subhead)+)>
```


Step 3

- Corpora: from raw text to marked-up text
- Text encoding in general
- Corpora:
  from marked-up text to structured data

Going beyond POS-tagging

- the more linguistic information that a corpus provides, the greater its utility:
  - searching for particular grammatical configurations is possible
  - using the information for training parsers is possible
  - evaluating linguistic accounts by larger-scale comparison of predicted and observed is encouraged

The Penn Treebank (1994)

- 1 million words of newspaper text
- syntactically annotated

```
(TOP (S (NP-SBJ my best friend)
    (VP gave
        (NP me)
        (NP chocolate)
        (NP-TMP yesterday))
    .))
```


The Prague Dependency Treebank (1997)

- full morphological tagging
- syntactic analysis using dependency syntax (Panevová, Bémová)
- 'tectogrammatical' level ("linguistic meaning": e.g., participant roles)
- initial goal: 200,000 sentences to be annotated

The "International Corpus of English"

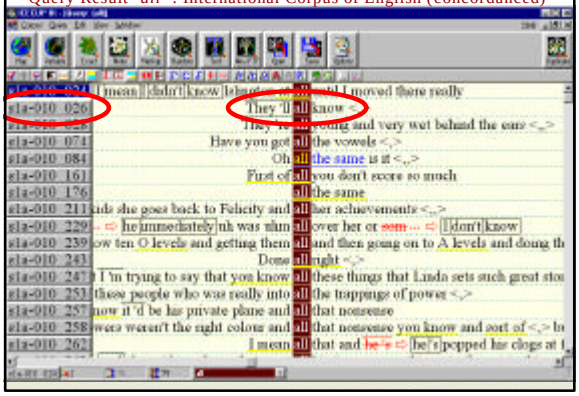
- Each ICE Corpus is divided into 2,000 word text samples representing various kinds of spoken and written English
- 500 texts: 200 written, 300 spoken
- the texts in ICE-GB were collected between 1990 and 1996
- A fully tagged and parsed corpus is only as useful as the tools that are provided to access it!

Greenbaum, Sidney (1988) 'A Proposal for an International Corpus of English', *World Englishes* 7: 315.

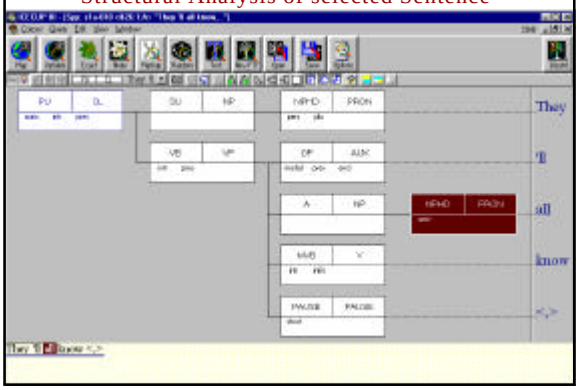
Nelson, Gerald (1996a) 'The Design of the Corpus', in S. Greenbaum (ed.), *Comparing English Worldwide: The International Corpus of English*, Oxford: Clarendon Press, 27-35

Nelson, Gerald (1996b) 'Markup Systems', in S. Greenbaum (ed.), (*op.cit.*), pp36-53

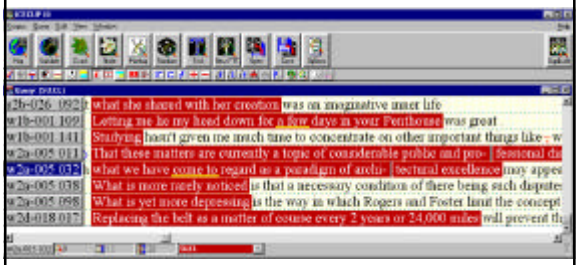
Query Result *'all'* : International Corpus of English (concordanced)



Structural Analysis of selected Sentence

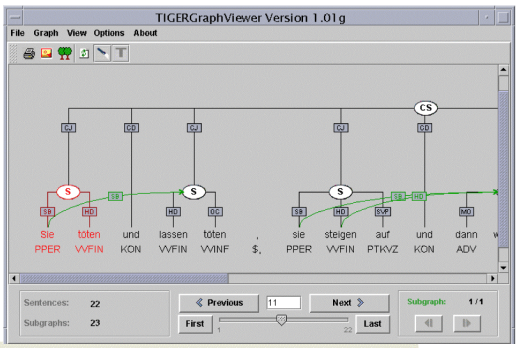


Search Results for Tree Fragment:
Subject filled by Clause



The TIGER treebank (2003)

- German
- 35,000 newspaper sentences
- target 80,000 sentences by end of project
- automatic parsing of corpus (broad coverage LFG)
- conversion of parse results into more 'neutral' form



Other treebanks / parsed corpora

- Susanne corpus (Sampson, 1995)
- Lancaster parsed corpus (Leech, 1992)
- Under development:
  - French
  - Spanish
  - Italian
  - Bulgarian
  - Russian

...three steps.

- Corpora: from raw text to marked-up text
- Text encoding in general
- Corpora: from marked-up text to structured data

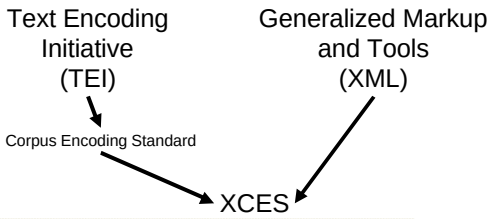
Several problems and a solution

- how to overcome the one corpus - one tool syndrome?
- how to move between different representations of similar information?
- how to increase the complexity and breadth of linguistic annotation?

Solution: XML and related technology

- XML - the extensible markup language - replaces SGML as the markup language of choice
- strongly supported by software developers (W3C, Java, ...)
- advanced tools becoming freely available: including **multilayer annotation**
- already the representation language of choice for corpora

Current Corpus Annotation Standards



Problem of intersecting hierarchies

- XML allows only 'balanced' bracketting
- Brackets may not cross each other
- But: many kinds of information cannot be combined into single hierarchies...
- This is a very common problem: also well-known to us in linguistics

Basic 'stand-off' annotation

XML 'base' document

```
<w id="u-01">Have</w>
<w id="u-02">you</w>
<punc type="comma" id="u-03">,</punc>
<w id="u-04">miss</w>
<punc type="question" id="u-05">?</punc>
<w id="u-06">Well</w>
<punc type="comma" id="u-07">,</punc>
<w id="u-08">for</w>
<w id="u-09">sure</w>
<punc type="exclamation"
      id="u-10">!</punc>
```

XML document for page breaks

```
<page id="page-01" from="..." to="u-07"/>
<page id="page-02" from="u-08" to="..."/>
```

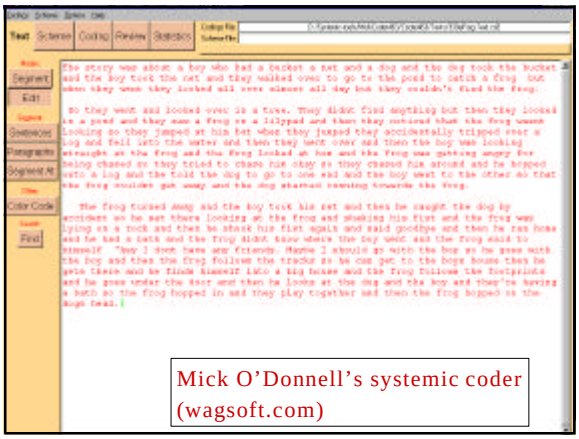
XML document for sentences

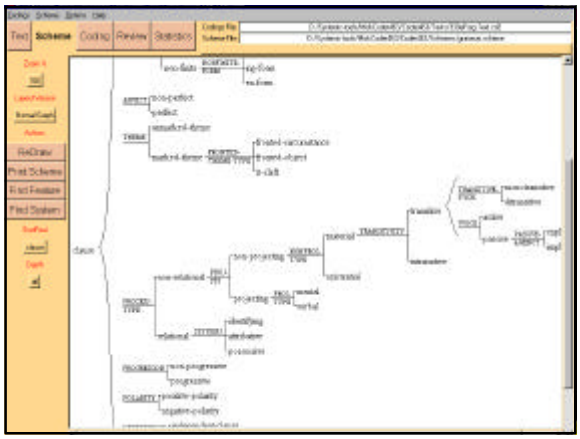
```
<s id="s-01" from="u-01" to="u-05"/>
<s id="s-02" from="u-06" to="u-10"/>
```

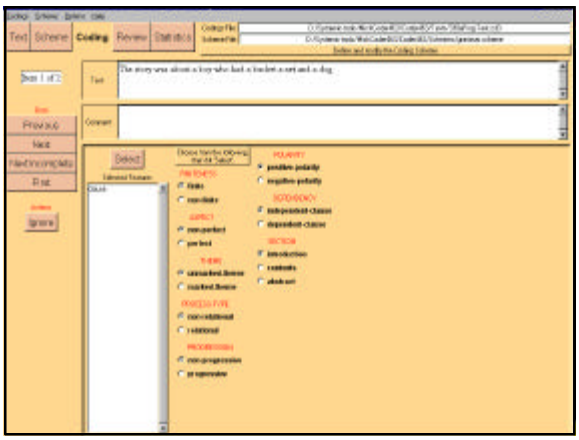

Technology developments

- XML is set to replace HTML as the basic language of the World-Wide Web
- XML extensions provide increasing functionality
  - translations between XML schemes
  - transparent interfaces between XML and DB
  - flexible rendering: graphical, typesetting, ...
- the technology is already moving inside Web-browsers...

Implications





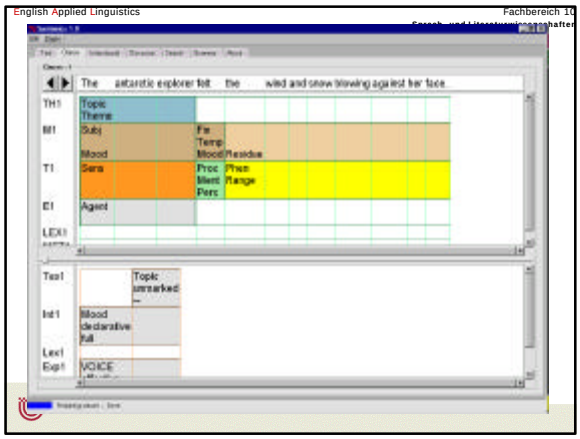


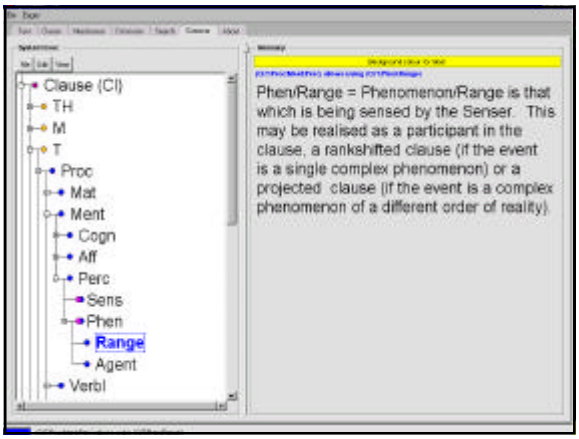
Interoperability problems

- coding scheme may only be changed with the tool itself
- use of coding schemes from elsewhere then not possible
- use of coding scheme elsewhere not possible

Kay O'Halloran: Systemics
(University of Singapore)

- Very nice tool
- Easy to prepare complex multilayered functional analyses
- Covers conjunctive relation, exchange structure, and cohesion analyses
- But: what do you do with the results?
- Technology limited (printing methods)

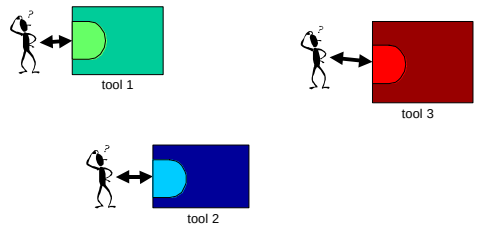




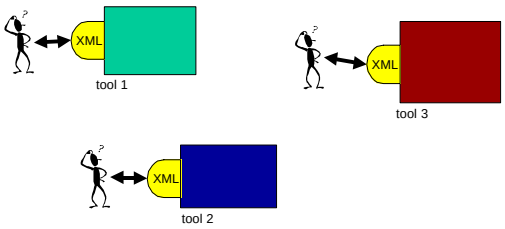
Interoperability problems

- no standardized output forms for analyses
- mixture of display and content
- no standardized output forms for grammar
- mixture of display and content
- coding scheme may only be changed with the tool itself

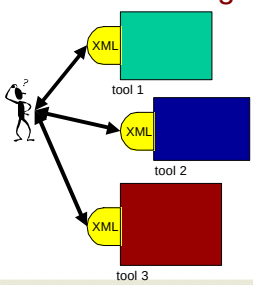
1990s view of linguistic tools



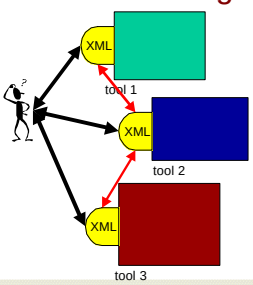
2000s view of linguistic tools



2000s view of linguistic tools



2000s view of linguistic tools



Current project

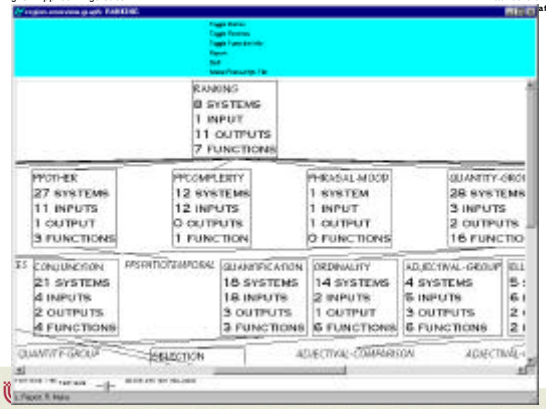
- to provide XML schema definitions for the main theoretical constructs in SFL
  - system networks
  - realization statements
  - instantiated syntagmatic structures
- to provide XML wrappers around existing tools to improve their interoperability

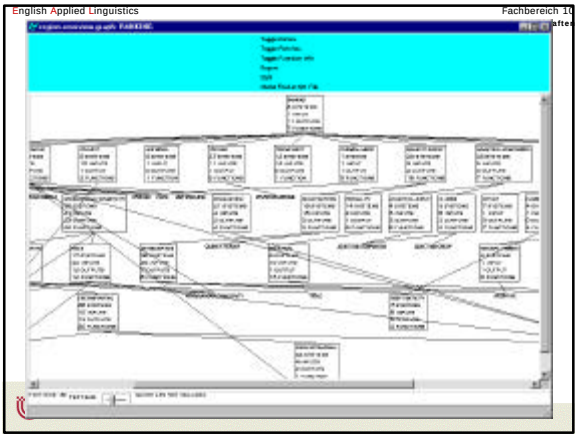
Example 1

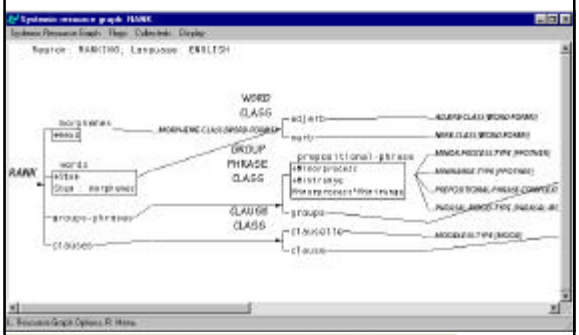
Grammar debugging with KPML and coding with Mick's coder

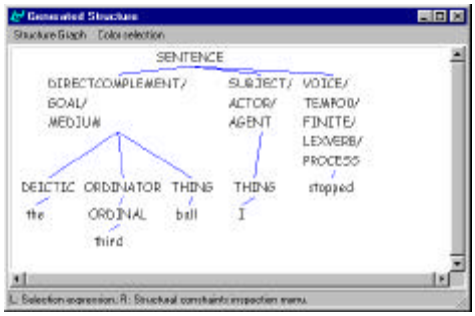
KPML

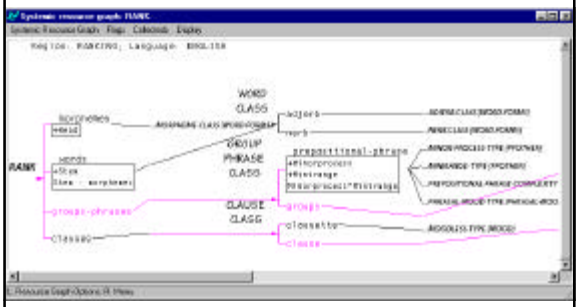
- a graphical development environment for large-scale systemic grammars built on top of Penman
- allows views of resources and their instantiations
- allows views according to axis, rank, metafunction, functional region and stratum
- strongly multilingual





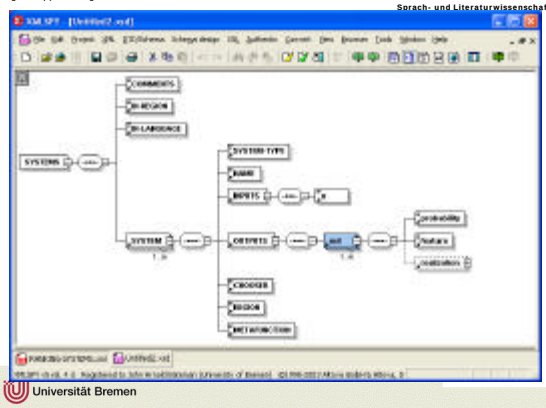




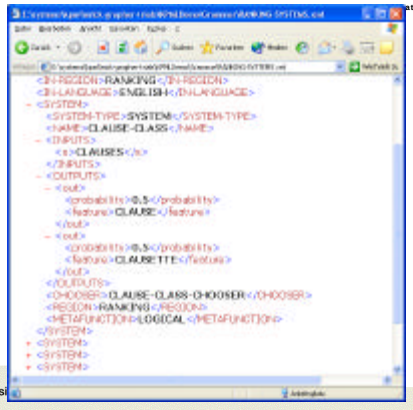


Interoperability

- XML versions of systems and structures may be both accepted as input and produced as output

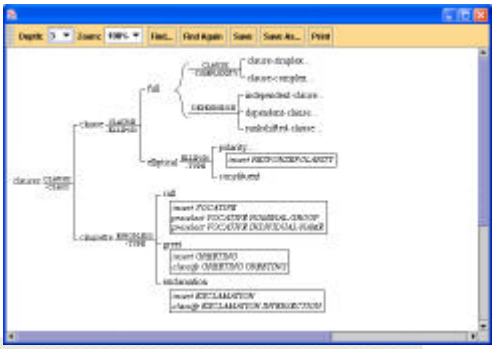



```
<SYSTEM>
  <SYSTEM-TYPE>SYSTEM</SYSTEM-TYPE>
  <NAME>CLAUSE-CLASS</NAME>
  <INPUTS><s>CLAUSES</s></INPUTS>
  <OUTPUTS>
      <out>
        <probability>0.5</probability>
        <feature>CLAUSE</feature>
      </out>
      <out>
        <probability>0.5</probability>
        <feature>CLAUSETTE</feature>
      </out>
  </OUTPUTS>
  <CHOOSER>CLAUSE-CLASS-CHOOSER</CHOOSER>
  <REGION>RANKING</REGION>
  <METAFUNCTION>LOGICAL</METAFUNCTION>
</SYSTEM>
```

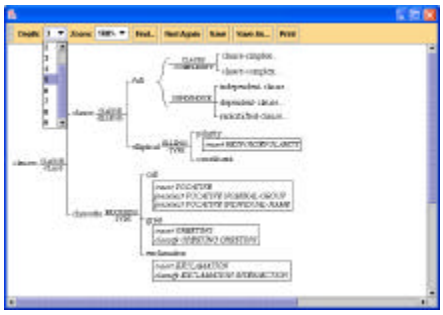



Mick's coder

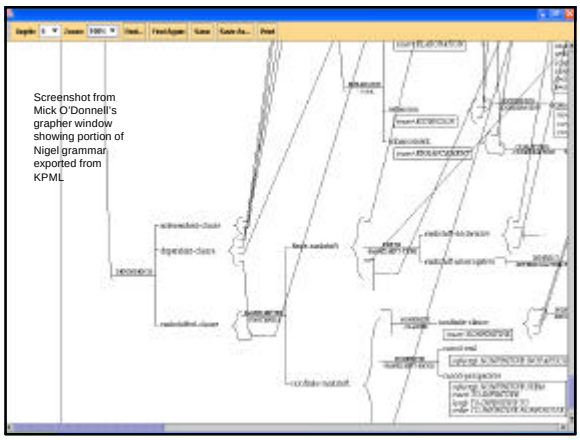
- extended to accept and produce the XML-definition of system networks



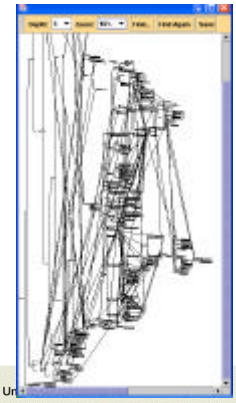
Screenshot from Mick O'Donnell's grapher window showing portion of Nigel grammar exported from KPML



Screenshot from Mick O'Donnell's grapher window showing portion of Nigel grammar exported from KPML



Screenshot from Mick O'Donnell's grapher window showing portion of Nigel grammar exported from KPML



Interoperablity is thus achieved between the two tools

Example 2

Storing and viewing results of systemic analyses in various forms

PoW Treebank

- Polytechnic of Wales parsed corpus (Fawcett, 1989)
- Child language corpus
- 65,000 words
- 11,000 trees
- Available through the International Computer Archive of Modern English (ICAME, Bergen)

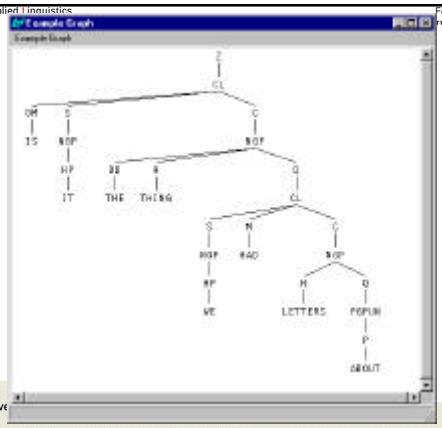
Example entry on ICAME CD

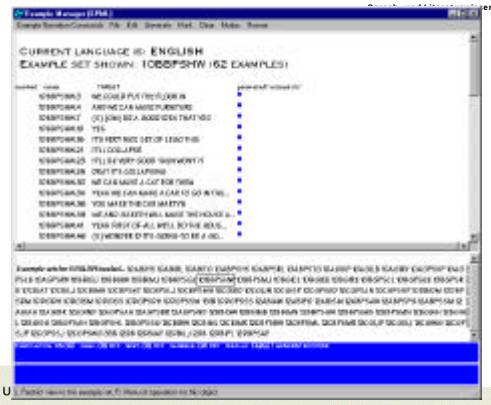
Z 1 CL 2 S NGP HP I 2 M PLAY 2 C PGP 3 P WITH [FS:MY-CHIP] 3 CV NGP 4 DD MY [RP:MY] 4 MO QQGP AX BIG 4 H TIPPER-LORRY 1 CL 5 & AND 5 S NGP HP I [RP:I] 5 M CALL 5 C PGP 6 PM FOR 6 CV NGP HN DAVID

Example entry on ICAME CD

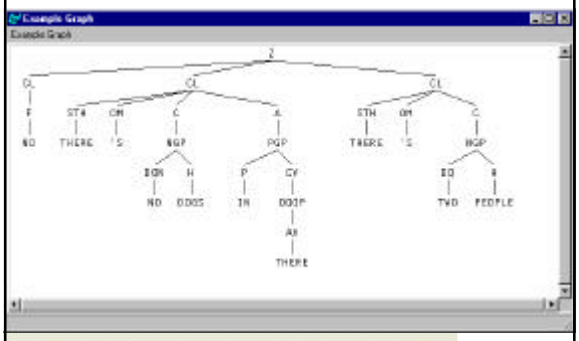
Z 1 CL 2 S NGP HP I 2 M PLAY 2 C PGP 3 P WITH [FS:MY-CHIP] 3 CV NGP 4 DD MY [RP:MY] 4 MO QQGP AX BIG 4 H TIPPER-LORRY 1 CL 5 & AND 5 S NGP HP I [RP:I] 5 M CALL 5 C PGP 6 PM FOR 6 CV NGP HN DAVID

This can be converted to the XML form defined for systemic instantiated structures and then read into any tool that supports this. For example, KPML...

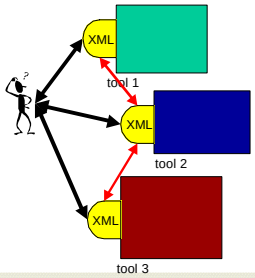




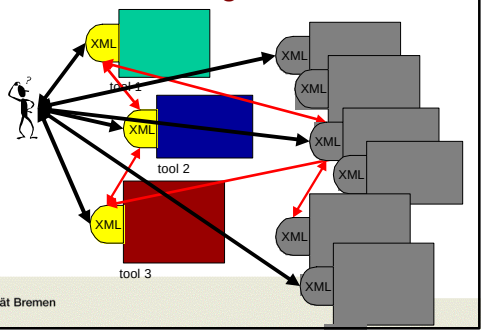
1088PSHW.191 THERE'S ANOTHER MAN
1088PSHW.192 THERE'S A MAN'S HEAD
1088PSHW.193 THERE'S ANOTHER HEAD
1088PSHW.194 NO THERE'S NO DOGS IN THERE THERE'S...
1088PSHW.198 VERY GOOD
1088PSHW.202 YEAH
1088PSHW.207 I THINK IT'S GOIN-TO TAKE A LONG TI...
1088PSHW.210 (S) (OM) NOT (M) (C) (P) THE RATE W...



2000s view of linguistic tools



2000s view of linguistic tools



Some boats to catch...

- where are the systemic-functional treebanks?
- if analyses are produced or 'written up' using a tool that can export XML structures for incorporation in a treebank, this would be of enormous value
  - supporting comparison of analyses
  - supporting collections of teaching examples
  - documenting the state of the art