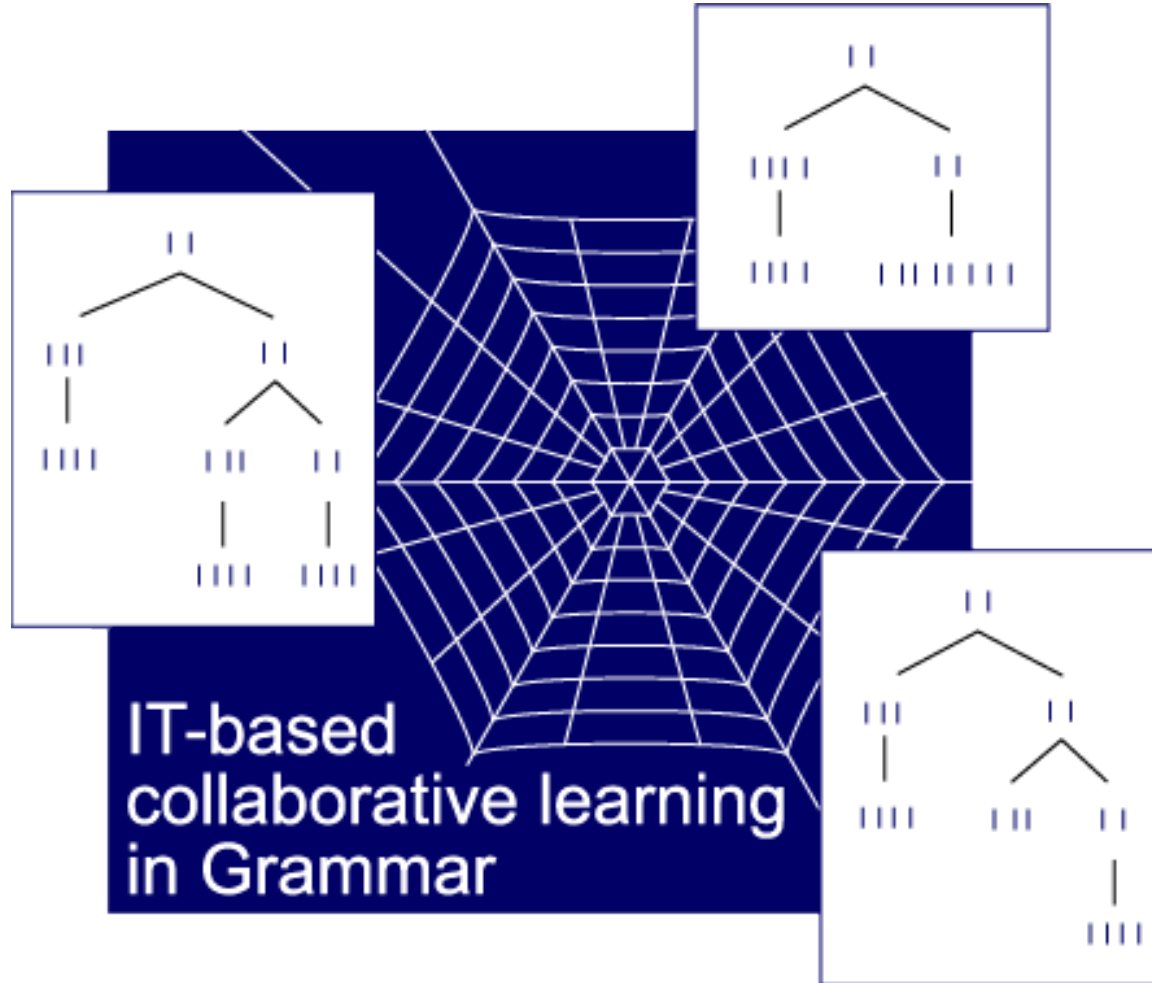




# Does the ITG platform exist



Leif-Jöran Olsson - Lars Borin  
Språkbanken, Svenska språket, GU

XML Prague 2006  
Praha, June 18. 2006



# ITG - brief project && platform facts

**Principal investigator:** Anju Saxena, UU  
**Cooperation between** UU - SU - GU  
**funding:** Distum (Swedish Agency for Distance Education)/Council of Higher Education (2002-2005); Magnus Bergvalls stiftelse, Rausings stiftelse (2005); Svenska språket/Språkbanken, GU 2005-  
**Corpus-based grammar exercises in a specially designed interface**



# functionality in ITG

Word class and part of sentence exercises

Corpus management (create a special corpus)

(corpus) viewer (text)

Corpus search with concordance, "hit map", clickable hit contexts

Linguistic "encyclopedia"

Net-based (Java Web Start)

Backend XML database (eXist) for corpus data in standardised format (TEI)

Possibility to add annotations



# challenges

grammar writing environment

Format and annotation differences

collaboration module

licensing and source code

availability/maturity



# compatibility issues

Differences in storage and text markup formats

Differences in POS tagging and syntactic annotation

Differences in grammatical analyses among corpora and grammar development tools/parsers



# other areas of use

diachronous (S)LL studies, ASU  
parallel corpora incorporation  
lexical visualisation



# ITG, ASU and the future [ 1 ]

need for a new interface for ASU  
(and other learner corpora)

ASU specs influenced the development of  
ITG

the interface is now tested both as  
grammar exercise environment and as  
corpus tool (at UU and SU),  
with a lot of valuable feedback



# ITG, ASU and the future [2]

## To do/under development:

- ★user-based corpus access
- ★more information (about the project, about the corpora, tutorials, etc.)
- ★speed up some DB searches
- ★incorporate other tools and resources (lexica like SAL, Svenska ord, other corpora, etc.)
- ★implement more types of visualisation
- ★new types of exercises (frequency-based word exercises, differentiated word class and part of sentence exercises)



# eXist usage in ITG

XmlDB api

XQuery modules in the DB

xmlrpc interface:

- ★ User data: authentication, selections, concordances, settings
- ★ Corpus searches

REST interface:

- ★ Encyclopedia (DocBook 5) transforms on the fly in serialisation from DB



# Stockholm Umeå corpus

```
<text id=k101>
<body>
<p>
<s id=k101-001>
<d n=1>-<ana><ps>MID<b>-</d>
<w n=2>Vilka<ana><ps>HD<m>UTR/NEU PLU IND<b>vilken</w>
<w n=3>djävla<ana><ps>JJ<m>POS UTR/NEU SIN/PLU IND/DEF..
<w n=4>optimister<ana><ps>NN<m>UTR PLU IND NOM<b>opti...
<d n=5>,<ana><ps>MID<b>,</d>
<w n=6>frustade<ana><ps>VB<m>PRT AKT<b>frusta</w>
<name type=person>
<w n=7>Lasse<ana><ps>PM<m>NOM<b>Lasse</w>
</name>
<d n=8>.<ana><ps>MAD<b>.</d>
</s>
```



# Talbanken [ 1 ]

P21803012001	0000	<<	GM
P21803012002	*DET	POOP	SS
P21803012003	RÖR	VVPS	FV
P21803012004	SIG	POXP	AAOO
P21803012005	ALLTSÅ	ABKS	+A
P21803012006	OM	PR	OAPR
P21803012007	FALL	NN	OA
P21803012008	1000	RC	OAET
P2180301200910002	DÄR	ABRA	RA
P2180301201010002	ORSAKEN	NNDD	SS
P2180301201110002	TILL	PR	SSETPR
P2180301201210002	PATIENTENS	NNDDHHGGSSETDT	
P2180301201310002	SYM TOM	NN	SSET
P2180301201410002	INTE	ABNA	NA
P2180301201510002	PRIMÄRT	AJ	AA
P2180301201610002	ÄR	AVPS	FV
P2180301201710002	ÅDERFÖRKALKNING	VN	SS SP
P2180301201810002	1100	+F	+F
P2180301201911002	UTAN	++MN	++
P2180301202011002	I	ABMN	+A
P2180301202111002	STÄLLET	ID	+A
P2180301202211002	BEROR	VVPS	FV

...



# Talbanken [2]

0000	Det	rör	sig	alltså	om	fall	1000		
<<	PO OP	VV PS	PO XP	AB KS		PR	NN	RC	
GM	SS	FV	AA OO	+A		OA PR	OA	OA ET	
där	orsaken		till		patientens		symtom		
AB RA	NN DD		PR		NN DD HH GG		NN		
RA	SS		SS ET PR		SS ET DT		SS ET		
inte	primärt		är	åderförkalkning			1100		
AB NA	AJ		AV PS	VN -- SS			+F		
NA	AA		FV	SP			+F		
utan	i	stället		beror	på	en			
++MN	AB MN	ID		VV PS		PR	EN		
++	+A	+A		FV		OA PR	OA DT		



# POS meta data

```
<?xml version="1.0" encoding="UTF-8"?>
<posEntries>
  <posHeader>
    Ordklasser enligt SUC (Ejerhed och Källgren).
  </posHeader>
  <posEntry type="open" id="AB">
    <posExplanation lang="sv" class="adverb">
      Adverb
    </posExplanation>
    <posExplanation lang="en" class="adverb">
      Adverb
    </posExplanation>
  </posEntry>
  <posEntry type="open" id="DT">
    <posExplanation lang="sv" class="determinerare">
      Determinerare
    </posExplanation>
    <posExplanation lang="en" class="determiner">
      Determiner
    </posExplanation>
  </posEntry>
  ...
</posEntries>
```



# POS meta data (distractors)

```
<?xml version="1.0" encoding="UTF-8"?>
<posDistractors>
  <posHeader>
    Distraktorer för ordklassövningar i ITG för SUC.
  </posHeader>
  <distractorEntry lang="sv" class="adverb">
    <posDistractor lang="sv">preposition</posDistractor>
    <posDistractor lang="sv">subjunktion</posDistractor>
    <posDistractor lang="sv">pronomen</posDistractor>
    <posDistractor lang="sv">adjektiv</posDistractor>
  </distractorEntry>
  <distractorEntry lang="sv" class="determinerare">
    <posDistractor lang="sv">pronomen</posDistractor>
    <posDistractor lang="sv">adverb</posDistractor>
    <posDistractor lang="sv">konjunktion</posDistractor>
    <posDistractor lang="sv">subjunktion</posDistractor>
  </distractorEntry>
  ...
</posDistractors>
```



# eXist bottlenecks in ITG

frequency calculation handling

POS tag extraction from selection

collection walking in general

★2,300+ collections

★depending on dynamic user selections