



# ICALL and what it's good for



*Elena Volodina, Research Engineer, Språkbanken, Göteborgs universitet*



## Self-presentation

- Linguist/English (PhD, 1998) och computational linguist (MA, 2008)
- Research engineer at Språkbanken, GU (2010-...)
  - ✓ Development of Lärka (ICALL application)
  - ✓ Semantic linking av lexicon resources in Karp
  - ✓ Parallel corpora
  - ✓ Frequency wordlists
  - ✓ ICALL research and development
  - ✓ Learner essays/corpora

# Lecture plan

- Computer-Assisted Language Learning (CALL) versus Intelligent CALL
  - Definitions and short historical overview
  - Supportive NLP components
- Language Learning (L2) – short introduction
  - L2 main steering document (CEFR)
  - L2 skills, proficiency levels and L2 activities
- NLP for Language Learning (L2)
  - Demos
- Designing an ICALL application – methodology and problems
  - Defining aims: end-user perspective first
  - Reuse of NLP components versus creating new ones
  - Standardization, architecture principles
  - Evaluation
- ICALL platform development at Språkbanken: Lärka



# Computer-Assisted Language Learning

- Drill-and Kill era
- Multimedia and graphics
- Authoring tools as a way out of CALL determinism
- Web-based materials, item banks --> need for standards
- Criticism

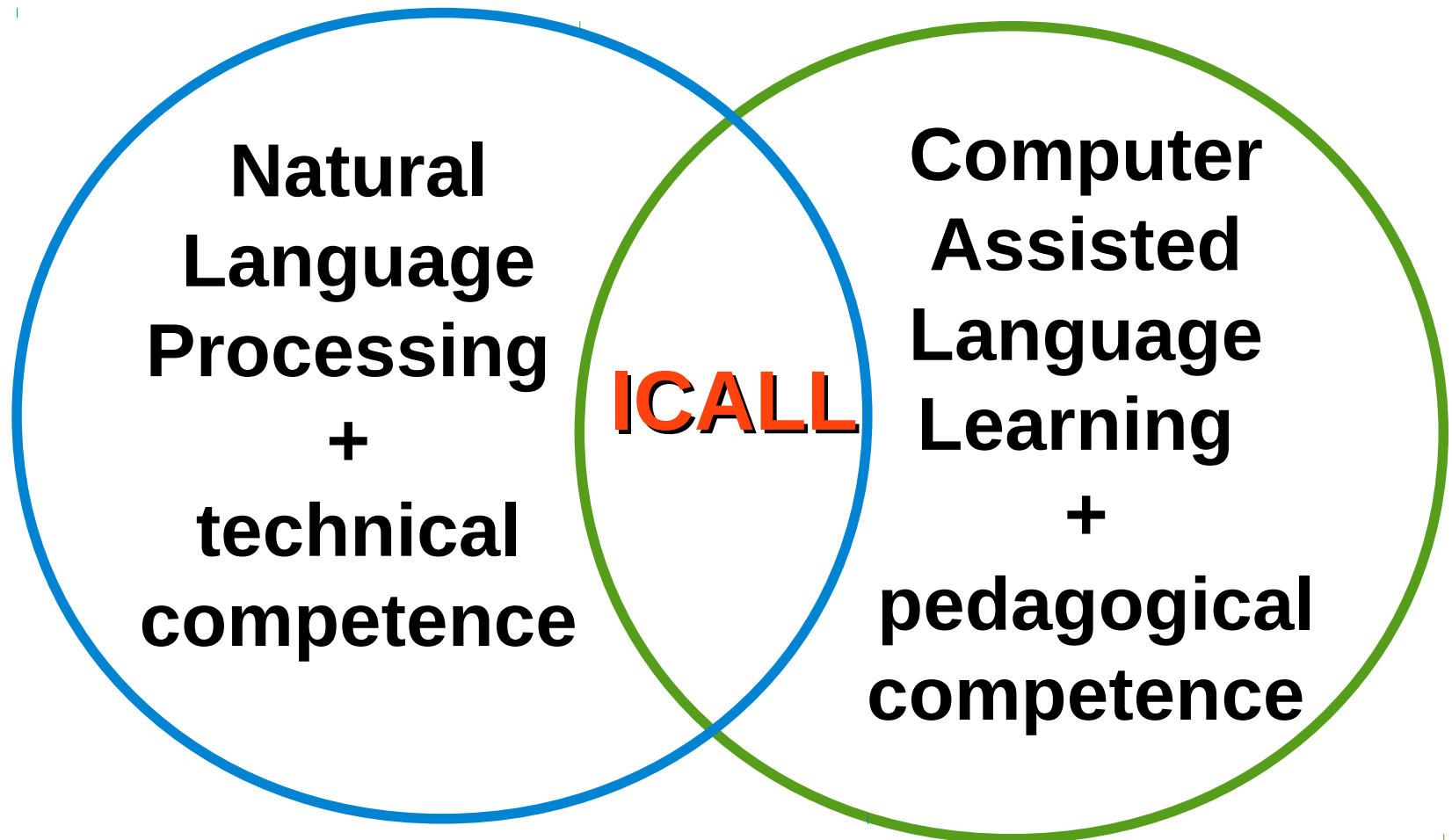


# Intelligent Computer-Assisted Language Learning

- AI-based CALL (Artificial Intelligence)
- Intelligent tutoring systems as surrogate teachers
- NLP-based CALL
  - Annotation-based CALL
  - Parser-based CALL
  - Any NLP component as intelligence in the application



**NLP + CALL = ICALL**





## NLP for CALL: an active area

- **Conferences:**

- EuroCALL, CALICO, ICCE – pedagogical conferences with ICALL/AI strands
- ACL, NAACL, EACL, COLING, Nodalida – NLP conferences with workshops in ICALL

- **Special Interest Groups:**

- EuroCALL ICALL SIG, CALICO ICALL SIG
- NEALT SIG-ICALL, SLaTE SIG

- **Special journals:**

- CALL, CALL-EJ, LLT, ReCALL, System, CALICO, etc.





# **NLP-based CALL : Advantages**

- ensures linguistic analysis of the input data
- adds generative power of applying the same analysis model to different (authentic) language samples, e.g. for generating exercises or detect errors in text production
- enables reuse of NLP tools and resources for practical use in language learning and that way
  - relieves teachers of monotonous tasks that can be modeled by computers
  - supports self-learning for students where it is feasible and motivated
  - popularizes NLP among CALL end-users





## **What is necessary?**

- available reliable NLP tools/algorithms, e.g.:
  - sentence segmenters, tokenizers, pos-taggers, lemmatizers, syntactic parsers, error parsers, spell-checkers, etc.
- available reliable annotated resources, e.g.:
  - corpora, lexicons, learner-oriented word lists, etc.

## **Where do we get them?**

- re-use existing reliable NLP tools and resources
- develop/create lacking ones



## Refreshing... segmentation

Shall I look up that Sofie K., whoever she is, and try to explain to her that I am not dangerous? Warnings in style with "Look out for the stairs!", "Watch your head!" are seen everywhere.

End of sentence... Where?



## Refreshing... segmentation

Shall I look up that Sofie K., whoever she is, and try to explain to her that I am not dangerous? Warnings in style with "Look out for the stairs!", "Watch your head!" are seen everywhere.

End of sentence... Here?



## Refreshing... segmentation

Shall I look up that Sofie K., whoever she is, and try to explain to her that I am not dangerous? Warnings in style with "Look out for the stairs!", "Watch your head!" are seen everywhere.

End of sentence... What about here?





## Refreshing... segmentation

- <s> Skall jag titta upp till den där Sofie K.</s>  
<s>, vem det nu är, och försöka förklara för henne att jag inte är farlig?</s>
- <s> Warnings in style with "Look out for the stairs!"</s>  
<s>, "Watch your head!"</s>  
<s> are seen everywhere.<s>



# Refreshing... tokenization & lemmatization

Num	Mening	Ditt svar	Rätt svar	Länkar
1	I andra ändan fanns ett tvättställ och en liten fyrkantig hall .	Välj ordklass		  <a href="#">JSON</a>

Word, token or lemma?



## Refreshing... tokenization & lemmatization

Num	Mening	Ditt svar	Rätt svar	Länkar
1	I andra ändan <b>fanns</b> ett tvättställ <b>och</b> en liten fyrkantig hall .	Välj ordklass		  <a href="#">JSON</a>

token

part of speech

base form (lemma)

lemgram

morpho-syntactic info  
(inflected form): verb,  
preteritum/past, s-form

fanns -> finnas -> verb -> finnas (verb) -> VB.PRT.SFO





# Language Learning Framework

- CEFR – Common European Framework of References for assessment of language proficiency (Council of Europe 2001)
- The CEFR is a document which describes in a comprehensive manner i) the competences necessary for communication, ii) the related knowledge and skills and iii) the situations and domains of communication as well as provides guidelines.

The six proficiency levels are named as follows:

C2	Mastery	}	Proficient user
C1	Effective Operational Proficiency		
B2	Vantage	}	Independent user
B1	Threshold		
A2	Waystage	}	Basic user
A1	Breakthrough		

# Language learning framework 2

## “can-do” statements

- **CEFR** “can-do” statements for each competence or skill and each level of proficiency.

*Can collate short pieces of information from several sources and summarise them for somebody else. Can paraphrase short written passages in a simple fashion, using the original text wording and ordering.*

CEFR descriptor for **B1**, for **ability to process text**

*Can understand familiar names, words and very simple sentences for example in notices, posters or in catalogues.*

CEFR descriptor for **A1**, **overall reading skills**



# NLP components in support of individual language skills

**L-ge skill**

**NLP components... Ideas?**

- Vocabulary
- Grammar
- Pronunciation
- Spelling
  
- Reading
- Writing
- Listening
- Speaking



# NLP components in support of individual language skills

## L-ge skill

## NLP components... Ideas?

- |                 |   |
|-----------------|---|
| • Vocabulary    | • Tokenizers, lemmatizers, spell-checkers, etc    |
| • Grammar       | • Parsers, taggers, error detectors, etc.         |
| • Pronunciation | • Text-to-speech modules, speech recognizers, etc |
| • Spelling      | • ...   |
| • Reading       | • ...   |
| • Writing       | • ...   |
| • Listening     | • ...   |
| • Speaking      | • ...   |

# Vocabulary

- Spelling, morphology, pronunciation, exercises (translation-based, semantic-based, sentence-based, text-based), testing
- Potential NLP components:
  - Learner lists (freq-based, level-based, text-based, etc.)
  - Spell-checkers
  - Lexicons
  - Morphological analyzers, PoS-taggers, Lemmatizers
  - Corpora – general, domain-specific, learner, written, spoken, etc. (collecting, annotating, using)
- Demos:
  - Wordrobe: [www.wordrobe.org](http://www.wordrobe.org); (<http://www.lrec-conf.org/proceedings/lrec2012/index.html> )
  - Multidict: <http://www2.smo.uhi.ac.uk/multidict/>
  - Lärka: [www.spraakbanken.ge.se/larka](http://www.spraakbanken.ge.se/larka)



## Vocabulary – research questions

- Vocabulary scope per CEFR level. How to identify? How many words per level? Which ones?
- Go by frequency... On what texts? Where to get texts? Copyright restrictions?
- Go by domain... Again – which words for which level? Manual work? Intuitions?
- Multiple choice: selection of distractors. Feasible for automatic approaches?
- Automatic selection of sentences for training... Procedures for testing sentence for complexity per level. Problems.
- Semantic disambiguation of polysemous words: lexeme rather than lemma... What do we need for that? (to fire)
- .....



# Grammar

- Recognition, use, exercises, testing
- Potential NLP components:
  - Morpho-syntactic description (corpus annotation & tagging)
  - Syntactic parsing (dependency relations)
  - Tree visualization
  - Error parsing
  - Formal grammars (phrase grammar, context-free grammar, etc.)
  - Corpora & corpus search applications
- Demos:
  - VISL: <http://beta.visl.sdu.dk/>
  - GF: <http://cloud.grammaticalframework.org/>
  - GRASP: <http://www.grammaticalframework.org/~peter/grasp>
  - Lärka: [spraakbanken.gu.se/larka](http://spraakbanken.gu.se/larka)





## Grammar – research questions

- Grammar scope per CEFR level. Which grammar phenomena?
- Go by frequency... On what texts? Where to get texts? Copyright restrictions?
- Manual work? Intuitions?
- Multiple choice: selection of distractors. Feasible for automatic approaches?
- Acceptable alternative forms – how to solve the problem?
- Automatic selection of sentences for training... Procedures for testing sentence for complexity per level.
- Error typology: special corpus of learner texts? For each CEFR level? Corrected by teachers? How to annotate?
- Error parsing of incorrect texts



## Writing

- Essay writing, letter writing, free responses
- Potential NLP components:
  - Error parsing
  - Error detection
  - Error annotation
  - Spell checking
  - Lexicons
  - Specific corpora
  - Assessment
  - Feedback generation



## Writing: grammar and spell-checking

- I remember having difficulties with physics and especially with the rule of thumb (if I name it correctly). I have been fighting and grudging over a course book for a while, torturing my parents and friends. It all happened before the exam in physics we were supposed to take. Finally, I broke down at one of the class meeting weeping over the "ridiculous rule without any logics". After the class one of the classmates offered me help which I suspiciously accepted: so many people have been trying to explain that rule to me without any success, that I lost any hope. Andrew and I spent approximately two hours over the rule, taking a break now and then for clearing our minds and letting out exasperation. We both were exhausted by the end. Ironically, it ended up with the classmate questioning whether there was any logics in the rule, while I - finally - understood how it worked. I still remember how happy I felt!

Non-existent words

Real word errors

Grammar errors



# Grammar and spell-checking

## Let's test it!

<http://www.pearsonschool.com/index.cfm?locator=PS1f8e>

<http://spellcheckplus.com/>

<http://www.onlinecorrection.com/>

<http://www.grammarly.com/>

Recommend to have a look at:

- <https://www.ets.org/Media/Products/Criterion/tour2/critloader.html>



# Grammar and spell-checking

## Essay text

I remember having difficulties with physics and especially with the rule of thumb (if I name it correctly). I have been fighting and grudging over a course book for a while, torturing my parents and friends. It all happened before the exam in physics we were supposed to take. Finally I broke down at one of the class meetings weeping over the "ridiculous rule without any logics". After the class one of the classmates offered me help which I suspiciously accepted: so much people have been trying to explain that rule to me without any success, that I lost any hope. Andrew and I spent approximately two hours over the rule, taking a break now and then for clearing our minds and letting out exasperation. We both were exhausted by the end. Ironically, it ended up with the classmate questioning whether there was any logic in the rule, while I - finally - understood how it worked. I still remember how happy I felt!

# Reading

- Understanding, question asking and answering, grammar+vocabulary
- Potential NLP components:
  - Corpora (+ annotation)
  - Automatic text selection
  - Information retrieval
  - Readability assessment (text & sentences)
  - Text-to-Speech synthesis
  - Question generation
  - Semantic disambiguation
- Demos:
  - <http://www2.smo.uhi.ac.uk/clilstore/>
  - <http://www.let.rug.nl/glosser/Glosser/>
  - <http://sifnos.sfs.uni-tuebingen.de/WERTi/index.jsp>



# Pronunciation, Listening, Speaking

- Recognition in speech, use in speech, training exercises
- Potential NLP components:
  - Lexicons with recordings of words
  - Text-to Speech synthesis
  - Speech recognition
  - Dialogue-based systems
- Demos:
  - <http://imtranslator.net/translate-and-speak/>
  - Lärka: [spraakbanken.gu.se/larka](http://spraakbanken.gu.se/larka)





## **Developing ICALL applications**

- End-user needs versus technological solutions.
  - Technology-driven or pedagogically driven?
  - NLP community versus L2 teachers – how aware are they of each other?
- Reliability of NLP components. Some linguistic teasers.
- Designing an ICALL application – methodology and problems
- Architecture principles
- Evaluation
- ICALL platform development at Språkbanken: Lärka, its scope and future



# End-user needs versus technological solutions

- Technology-driven or pedagogically driven?
  - Expectations of technology vs what it can perform
- NLP community versus L2 teachers
  - Are they aware of each other?
  - Communication problems? Different cultures?  
Misunderstandings?
  - Linguists & teachers = technophobes and keep to their practices?
  - IT developers = arrogant and ignorant of pedagogy and language learning?



# Use of NLP components in CALL: Challenges

- Not reliable; cannot promise 100% correctness.

## Linguistic teaser 1

*Where should a sentence boundary go?*

”What have you done to your breast?” she asked.

- `<s>`”What have you done to your breast?`</s>`
- `<s>`” she asked. `</s>`
- `<s>`”What have you done to your breast?” she asked.`</s>`



## Linguistic teaser 2

She moved a bit

she = pronoun

moved = verb

a = determiner

bit = noun

she = pronoun

moved = verb

a\_bit = adverb, multiword unit



## Reusing NLP components

- NLP components are
  - Monolithic and inflexible; need to be individually adapted to every new application
  - Not readily available as the rights are held by individuals or institutions all over the world
  - Physically located in different places
  - Not interoperable via standardized interfaces
- What are the strategies for making use of them?
  - Rewrite in another programming language
  - Find chunks of similar code and build upon it using open-source initiatives (e.g. Free Software Foundation: <http://www.fsf.org>)
  - Any other ideas?

# Reusing NLP components, 2

- Service Oriented Architecture (SOA) principles:
  - Modular services that can be reused by others
  - Communication layer with a well-defined interface for sending a request and getting a response
  - Standardized data output format
  - Well-documented interface and its service
  - Services loosely coupled and can be recombined
- Web services as an implementation technology
  - Wrapper around a program defining a port of access to it
  - Can reuse other web services, databases, resources, etc.
  - Access over internet; the original software can still be residing on its original server
  - Standardization initiative... trying to attract software and resource owners to provide web services

# Web-service request / response

- Url request - example:
  - <http://spraakbanken.gu.se/ws/larka?exetype=pos1&indent=1>
- Url request is a call to a web-service i.e. a script that does something depending on the variables in the request:
  - **exetype=pos1**
  - **exetype=synt1**
- The web-service returns a response



# Output from the backend (training syntactic relations)

```
{
  "corpus": "TALBANKEN",
  "distractors": ["AG", "FV", "IO", "IV", "OO", "SP", "SS"],
  "distractors_en_sv": {
    "AG": {"en": "adverbial", "sv": "adverbial"},
    "FV": {"en": "finite verb", "sv": "finit verb"},
    "IO": {"en": "indirect object", "sv": "indirekt objekt"},
    "IV": {"en": "nonfinite verb", "sv": "infinitt verb"},
    "OO": {"en": "object", "sv": "objekt"},
    "SP": {"en": "predicative", "sv": "predikativ"},
    "SS": {"en": "subject", "sv": "subjekt"}
  },
  "exetype": "synt1",
  "sent_index": 3440,
  "sentence_left": "Den ena är att man har en förebild som visar hur ",
  "sentence_right": "ska vara : enheten och kärleken mellan Kristus och  
de kristna .",
  "target": "äktenskapet ",
  "target_deprel": "SS",
  "target_index": 11
}
```



## Evaluation

- Group vs individual
- Qualitative vs quantitative
- Student vs teacher opinions
- Pedagogical effect (teacher vs student perspectives)
- Reliability
- Time effectiveness
- Retaining rate
- User-friendliness
- etc.

# Lärka

- ✓ Lärka: **Lär** språket via **KorpusAnalys**
- ✓ ICALL platform for **L2 learners** of Swedish and **students of Linguistics**
- ✓ **ICALL**: Intelligent Computer-Assisted Language Learning
- ✓ Pedagogal framework: **CEFR** (European proficiency scale nivåskalan)
- ✓ Web-based, **open**, no installation





# Resources

- Korp: corpus search infrastructure
  - Swedish corpora, partly manually annotated
  - Different genres: newspapers, novels, etc.
- Karp: lexicon search infrastructure
  - Saldo morphology
  - Lexin





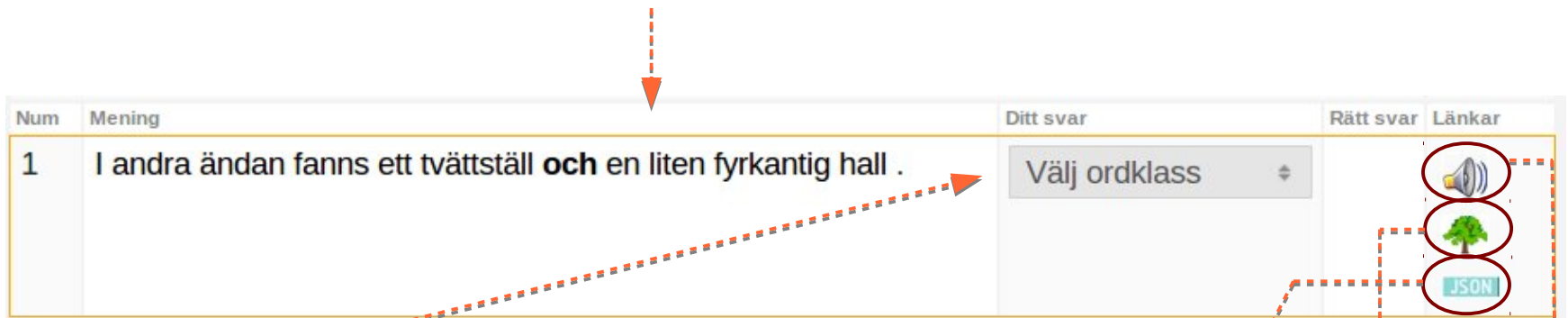
# Exercise types




- Students of **Linguistics** can train:
  - Parts of speech
  - Syntactic relations
  - Semantic roles
- **Learners of Swedish** can train:
  - Word knowledge
  - Word inflections
  - Spelling and listening comprehension



# Common features

- Main exercise context is sentence (one sentence at a time)



Num	Mening	Ditt svar	Rätt svar	Länkar
1	I andra ändan fanns ett tvättställ <b>och</b> en liten fyrkantig hall .	Välj ordklass		  

- Multiple-choice format (1 correct + 2-4 distractors)
- Result tracker:

Resultatsamlare	
Övningstyp	Korrekt/totalt
Lingvister/POS1, självstudier	0/1

pronunciation  
syntactic structure  
further information



## Common features

- “Modes” :
  - self-study  
(possible to change the answer)
  - diagnosis
  - test
  - timer 

tid (sek)  kvar
- “Tips” (reference information):
  - Saldo morphology (inflections)
  - Wikipedia, Wiktionary →
  - Pronunciation with an avatar
- “Statistics” (information on user performance)

▼ Saldo morfologi: värld

**VÄRLD**  
*lemgram:* värld..nn.1;  
*ordklass:* nn;  
*saldo-paradigm:* nn\_2u\_mening;  
*inherent:* u;

<i>sg indef nom</i>	värld
<i>sg indef gen</i>	världs
<i>sg def nom</i>	världen
<i>sg def gen</i>	världens
<i>pl indef nom</i>	världar
<i>pl indef gen</i>	världars
<i>pl def nom</i>	världarna
<i>pl def gen</i>	världarnas
<i>ci</i>	världs-
<i>ci</i>	världs
<i>cm</i>	världs-
<i>cm</i>	världs

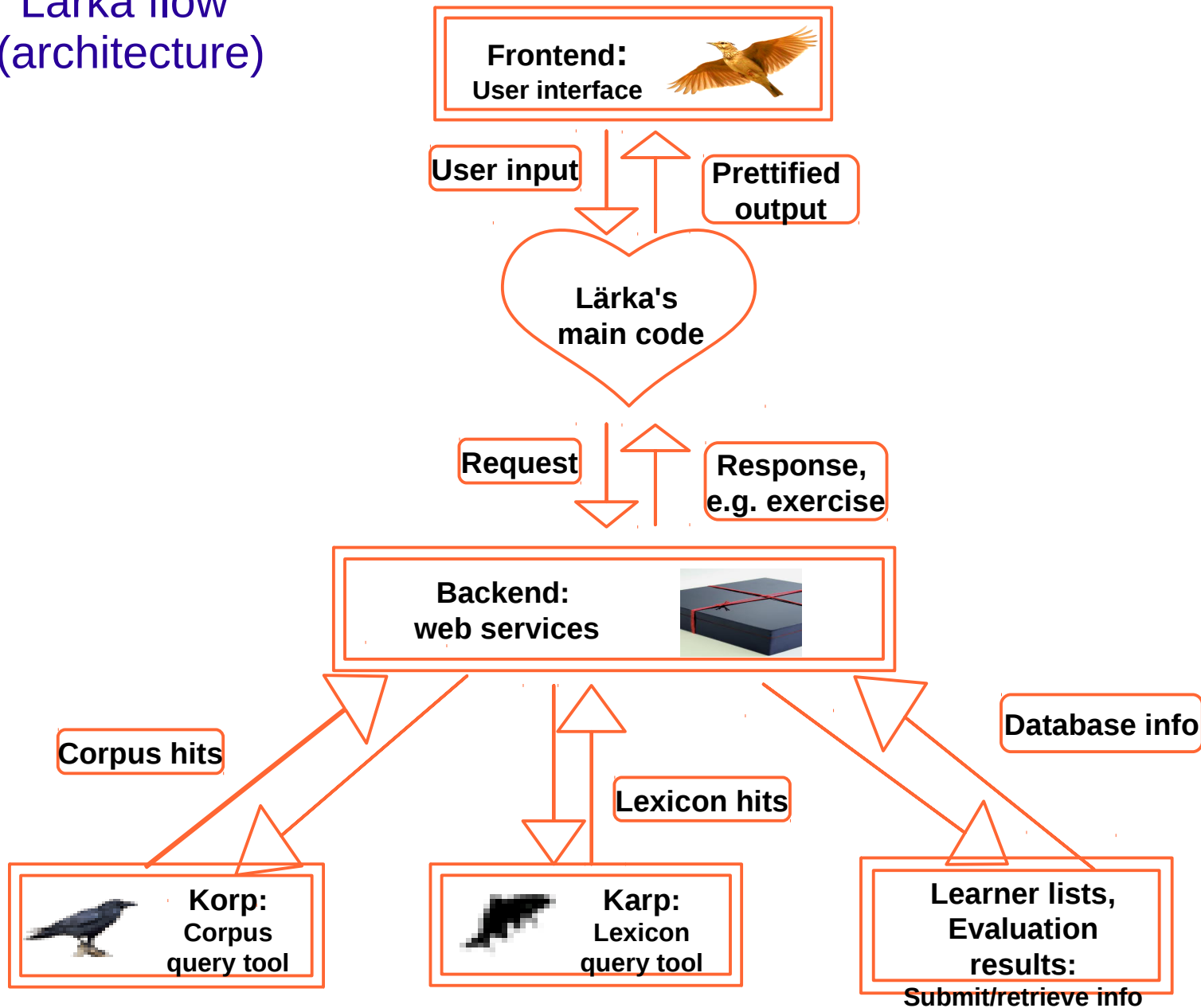
► Wikipedia: värld

► Wiktionary: värld

► Monica: lyssna på uttalet



# Lärka flow (architecture)



# Output from the backend (training syntactic relations)

```
{
  "corpus": "TALBANKEN",
  "distractors": ["AG", "FV", "IO", "IV", "OO", "SP", "SS"],
  "distractors_en_sv": {
    "AG": {"en": "adverbial", "sv": "adverbial"},
    "FV": {"en": "finite verb", "sv": "finit verb"},
    "IO": {"en": "indirect object", "sv": "indirekt objekt"},
    "IV": {"en": "nonfinite verb", "sv": "infininit verb"},
    "OO": {"en": "object", "sv": "objekt"},
    "SP": {"en": "predicative", "sv": "predikativ"},
    "SS": {"en": "subject", "sv": "subjekt"}
  },
  "exetype": "synt1",
  "sent_index": 3440,
  "sentence_left": "Den ena är att man har en förebild som visar hur ",
  "sentence_right": "ska vara : enheten och kärleken mellan Kristus och  
de kristna .",
  "target": "äktenskapet ",
  "target_deprel": "SS",
  "target_index": 11
}
```

# Lärka's research agenda

- Automatic generation of exercise items:
  - ✓ *for L2 vocabulary training, dictation & spelling*
  - ✓ *in sentence-long context (at the moment)*
- Practical needs:
  - ✓ *receptive vocabulary scope per level*
  - ✓ *sentence readability measure per level*
  - ✓ *text readability per level*
- How?
  - ✓ *e.g. study texts used for teaching CEFR-based courses, per level*

# Pros and cons of a corpus compilation

## Disadvantages

- ✗ *Time-consuming...  
...and therefore expensive*
- ✗ *Based on subjective judgements*

## Advantages

- ! *Provides answers to the questions... (...that we have at the moment)*
- ✓ *Contains coursebooks by many authors, accepted for teaching by many teachers, i.e. represent collective “objective” picture*

## Besides:

- ✓ *Receptive vocabulary scope*
- ✓ *Readability tests (sentence & text levels)*
- ✓ *Facilitates automatic interpretation of CEFR descriptors*
- ✓ *Presents empiric evidence*
- ✓ *...and more*

# CEFR-corpus

project financed by the Department of Swedish

- Gold standard for CEFR-based text research
- Text types: normative (input) and learner-produced (output)



- Focus in this project: normative texts

# CEFR-corpus 2

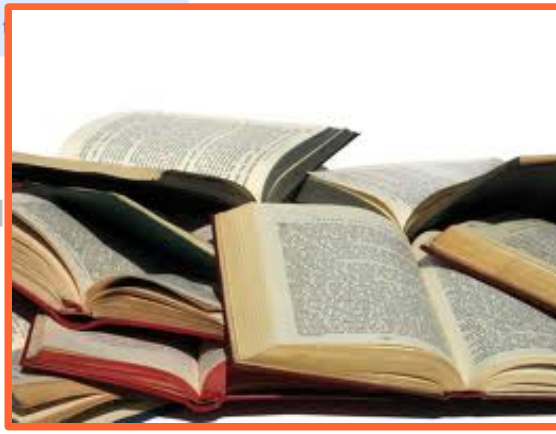
## identifying relevant sources

- Interviews with teachers on relevant course books & novels used in CEFR-based teaching
  - ✓ *resulted in a list of 15+ titles*
  - ✓ *that contain 3187+ pages;*
  - ✓ *with an estimated corpus size of approx. 3 mln tokens*
- Contacts with publishers
  - ✓ *Folkuniversitets förlag, Studentlitteratur, Natur och Kultur, Svenska institutet – negative to sharing electronic materials*
  - ✓ *Liber – positive to collaboration; provided e-texts for research*

Sök efter bok (substantiv) Sök även som ☐ förled ☐ efterled och ☐ skiftlägesoberoende

Relaterade ord

bok (substantiv)



ragit med goda råd och synpunkter under arbetet med boken för kaffefläcken, står det om en man som har skrivit en

CEFR

**boken**

**bok**

**Boken**

**bok**

**bok**

**böcker**

**böckerna**

**böckerna**

**böcker**

**böcker**

**böcker**

**Boken**

**Boken**

**böcker**

om hur man bäst undviker att bli irriterad över

som Åsa läste om heter Hetsa inte upp dig över

om sitt liv.

som fanns i hemmet.

om de svenska emigranter som reste från Småland

på 1970-talet och de blev förlagan till en musikal

var en protest mot nazismen?

blev han mest känd för?

?

var mycket, mycket intressant.

var \_\_\_\_\_intressant.

har lockat läsare i alla åldrar, fastän hon främst

titel:

På svenska! 2. Svenska som främmande språk. Lärobok  
författare: Ulla Göransson, Annika Helander, Mai Parada  
datum: 2002

## ordattribut

ordklass: substantiv

grundform:

bok

lemgram:

bok<sup>2</sup> (substantiv)

bok (substantiv)

betydelse:

bok<sup>2</sup>

bok

skapade av...



Sök efter bok (substantiv) Sök även som ☐ förled ☐ efterled och ☐ skiftlägesoberoende

Relaterade ord

bok (substantiv) ▼

uppsats artikel läsa läsa\_upp dokument magasin läsare läsning skrift tidskrift rabbla  
lättläst text tidning

KWIC: träffar per sida: 25 ▼ sortera inom korpus på: förekomst ▼ Statistik: sammanställ på: ord ▼

KWIC

Statistik

Ordbild

Antal träffar: 58

Föregående

1

2

3

Nästa

Visa KWIC

# CEFR

Vi vill tacka alla dem som på olika sätt bidragit med goda råd och synpunkter under arbetet med **boken**.

- Se här, vännen min, sa ordentliga Åsa och pekade på tidningen. Här, strax nedanför kaffefläcken, står det om en man som har skrivit en **bok** om hur man bäst undviker att bli irriterad över småsaker. Han menar att man ska tänka på allt man har, i stället för allt man inte har! Det står att det gäller att skaffa sig perspektiv på livet, och att det är bättre att ta det lugnt och fokusera på det som är viktigt än att bli arg och frustrerad. Han säger också att vi blir stressade för att vi är rädda att inte lyckas. Och att man varje dag ska tala om för någon att man beundrar honom eller henne. Det

## Korpus

CEFR

## textattribut

titel: På svenska! 2. Svenska som främmande språk. Lärobok  
författare: Ulla Göransson, Annika Helander, Mai Parada  
datum: 2002

## ordattribut

ordklass: interpunktion  
grundform: [tom]  
lemgram: [tom]  
betydelse: [tom]  
förled: [tom]  
efterled: [tom]  
dependensrelation: Punkt  
msd: MAD ⓘ

[Visa dependensträd](#)

# CEFR-corpus

## teaser

- What is the genre?
  - ✓ facts / instruction?
  - ✓ evaluation / personal reflection?

Ulf Frövi är psykolog och arbetar som konsult med personer som ska flytta utomlands. Här ger han några råd som kan underlätta processen:

- När det känns jobbigt, tänk på att livet till stor del består av små och stora problem som ska lösas, även i ditt hemland.
- Försök lära dig så mycket som möjligt om den nya kulturens historia, politik, geografi, konsthistoria osv. Ju mer du vet desto mer förstår du.
- Glöm inte bort dina intressen och hobbyer. Om du förut var medlem i någon klubb, försök då att hitta motsvarande klubb i ditt nya land. Det kan vara svårt att få nya vänner, men ett bra sätt är garanterat att försöka hitta personer som har samma intressen som du.
- Värdera inte och jämför inte olika länder. Konstatera bara att man gör på olika sätt i olika länder.
- Om du är frustrerad över hur dina nya landsmän beter sig, tänk då på att du inte kan ändra på ett helt folk. De är ganska nöjda med sakernas tillstånd. Du kan bara ändra på din egen attityd och ditt eget beteende.
- Om du har en partner från ett annat land, försök då att vara flexibel utan att glömma bort värderingar och principer som är viktiga för dig. Välj dina krig. Ta bara upp diskussioner om sådant som du tycker är extra viktigt. Det är kanske viktigare t ex att vara överens om hur barnen ska uppfostras, än exakt vilka jultraditioner ni ska ha.



# CEFR-corpus

## present-day status

- Two course books for B1 levels (+ 3 books for A1/A2)
  - ✓ *scanned*
  - ✓ *annotated*
  - ✓ *uploaded into Korp*

<http://spraakbanken.gu.se/eng/korp>

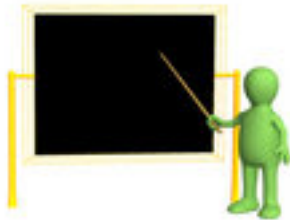
- Tests on sentence readability for B1 level
  - ✓ *master thesis project by Ildikó Pilán*
  - ✓ *to be presented at EuroCALL 2013*



[http://spraakbanken.gu.se/larka/larka\\_hitex\\_index.html](http://spraakbanken.gu.se/larka/larka_hitex_index.html)

## MT on sentence readability: Purpose

- Automatically **select** and **rank** sentences from Swedish native language texts.
- Sentences should be:
  - **understandable** by students of Swedish as a second language (L2), especially at B1 level
  - suitable **exercise item**
  - appropriate **examples** to illustrate a new lexical item
- Target users:



Teachers of  
L2 Swedish



Students of  
L2 Swedish



Lexicographers

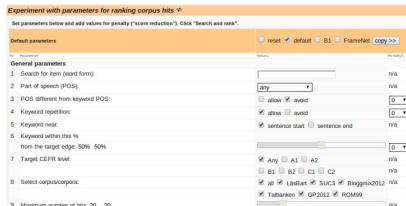
Most informative features	Logistic Reg. (RFE)	Decision Tree
Sentence length	✓	✓
Average token length	✓	✓
Percentage of words longer than 6 characters		✓
Modifiers	✓	
Average dependency depth		✓
Average number of senses per word	✓	✓
Nominal Ratio	✓	✓
Average frequency in the Wikipedia list	✓	
Average frequency in Kelly list	✓	✓
Percentage of difficult words	✓	✓
Number of difficult words	✓	
Adverb variation	✓	
Noun / Verb ratio	✓	
Model Verb / Verb ratio	✓	



# The readability module



preferences



**FRONTEND**  
(user interface)

keyword



**CORP**

sentences

parameters




**BACKEND**  
(web service)

**filtered and  
ranked  
sentences**

# The user interface in Lärka

[http://spraakbanken.gu.se/larka/larka\\_hitex\\_index.html](http://spraakbanken.gu.se/larka/larka_hitex_index.html)

**Experiment with parameters for ranking corpus hits** 

Set parameters below and add values for penalty ("score reduction"). Click "Search and rank".

**Default parameters** ☐ reset ☒ default ☐ B1 ☐ FrameNet

No	Parameter	Value1	Penalty1
<b>General parameters</b>			
1	Search for item (word form):	<input type="text"/>	n/a
2	Part of speech (POS):	<input type="text" value="any"/>	n/a
3	POS different from keyword POS:	<input type="checkbox"/> allow <input checked="" type="checkbox"/> avoid	<input type="text" value="0"/>
4	Keyword repetition:	<input checked="" type="checkbox"/> allow <input type="checkbox"/> avoid	<input type="text" value="0"/>
5	Keyword near:	<input checked="" type="checkbox"/> sentence start <input type="checkbox"/> sentence end	n/a
6	Keyword within this % from the target edge: 50% 50%	<input type="text" value="50"/>	<input type="text" value="0"/>
7	Target CEFR level:	<input checked="" type="checkbox"/> Any <input type="checkbox"/> A1 <input type="checkbox"/> A2 <input type="checkbox"/> B1 <input type="checkbox"/> B2 <input type="checkbox"/> C1 <input type="checkbox"/> C2	n/a
8	Select corpus/corpora:	<input checked="" type="checkbox"/> all <input checked="" type="checkbox"/> LäsBart <input checked="" type="checkbox"/> SUC3 <input checked="" type="checkbox"/> Bloggmix2012 <input checked="" type="checkbox"/> Talbanken <input checked="" type="checkbox"/> GP2012 <input checked="" type="checkbox"/> ROM99	n/a
9	Maximum number of hits: 20 20	<input type="text" value="20"/>	n/a



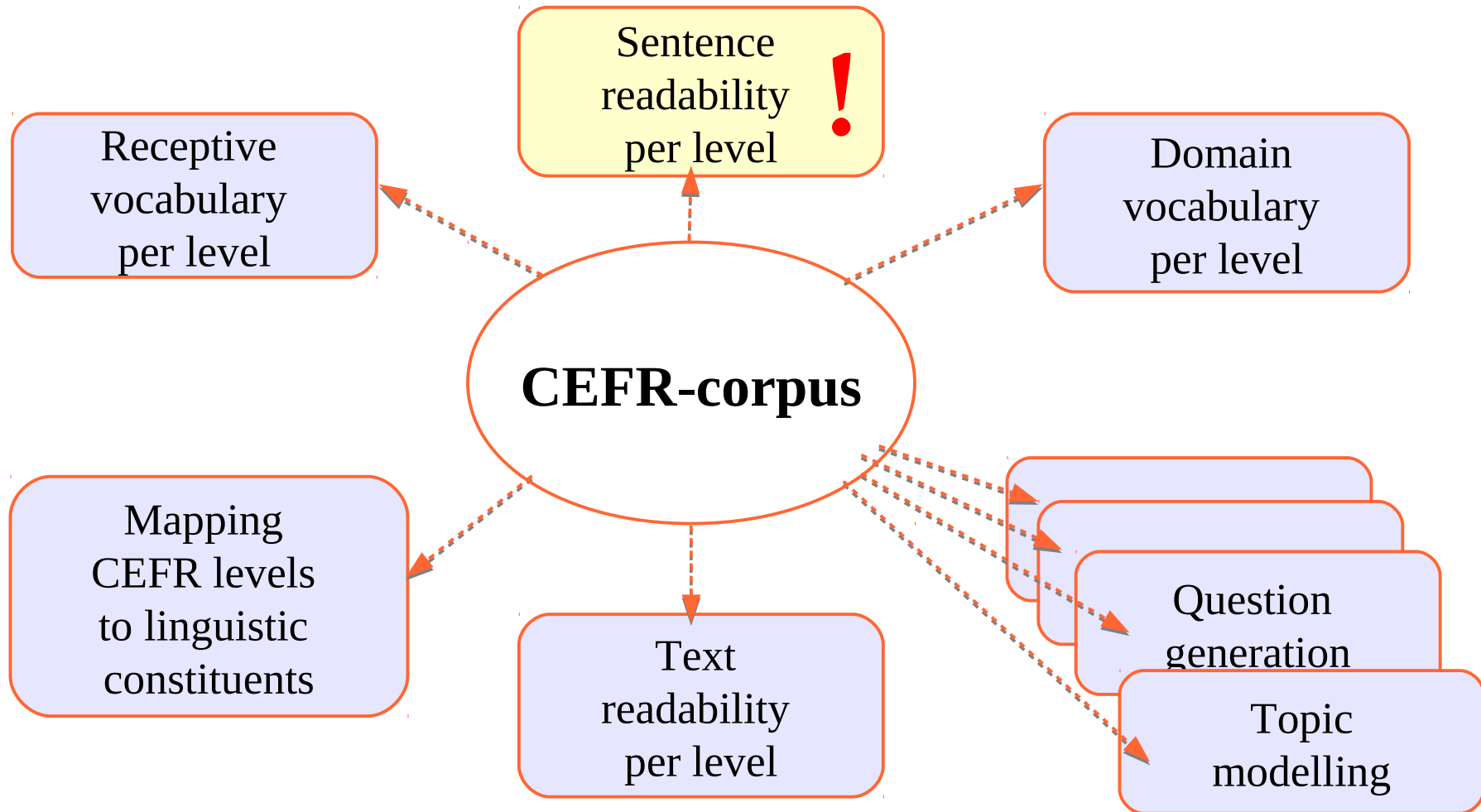
# CEFR-corpus

## intended use

- Identification of receptive vocabulary per proficiency level
- Test on *sentence* readability per proficiency level
- Tests on *text* readability per proficiency level
- Topic modeling
- Question generation
- Mapping CEFR “can-do” statements (for some of the competences) to linguistic constituents
- etc.



# CEFR corpus for Lärka's research agenda





# Lärka, near future

- Expand exercise scope:
  - (already in pipeline): gap cloze, wordbank
  - (potentially): morphological paradigm, semantic closeness, yes-no diagnostic test, etc...
  - (potentially, corpus-based): naming grammar features (past, present, etc); shuffling word-order by syntactic groups
  - word-building (compounding, affixation)
- Learner lists/lexical database: tests on receptive/productive vocabulary scope
- Enrich encyclopedia feedback
- Exercises based on syntactic trees

# Lärka, distant future, if ever

- Text readability analysis
- Half-automatic mode for exercise generation (feeding the system with the user choices/lists, etc.)
- Editable “mode” of exercise production – proofreading and modifying automatically created items; saving the items into a database
- Error typology and analysis of written texts/essays
- etc.



# FOOL STOP

*Do you mean:*

*Fools stop*

*Fool stops*

*Full stop*

*...?*