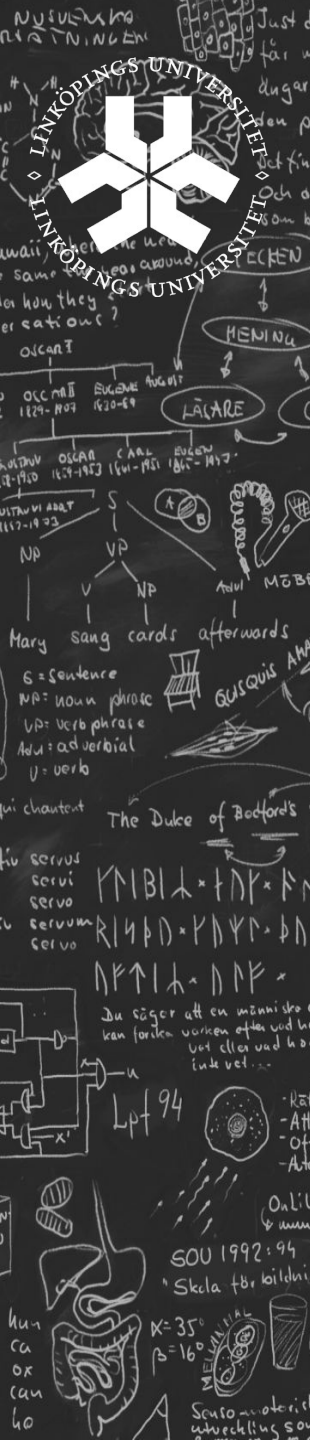


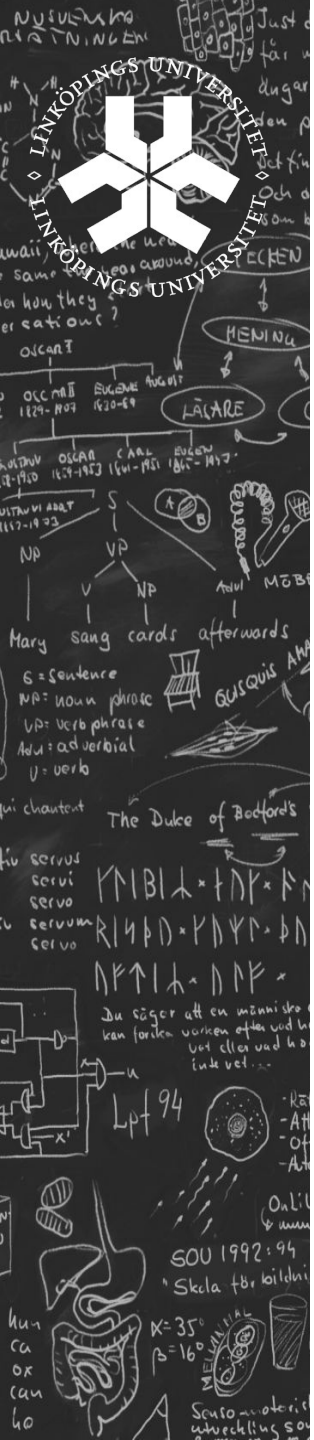
NLP for the translation class

Lars Ahrenberg and Ljuba Tarvi

2nd workshop on NLP for Computer-Assisted Language Learning

Oslo, May 22, 2013

LiU expanding reality

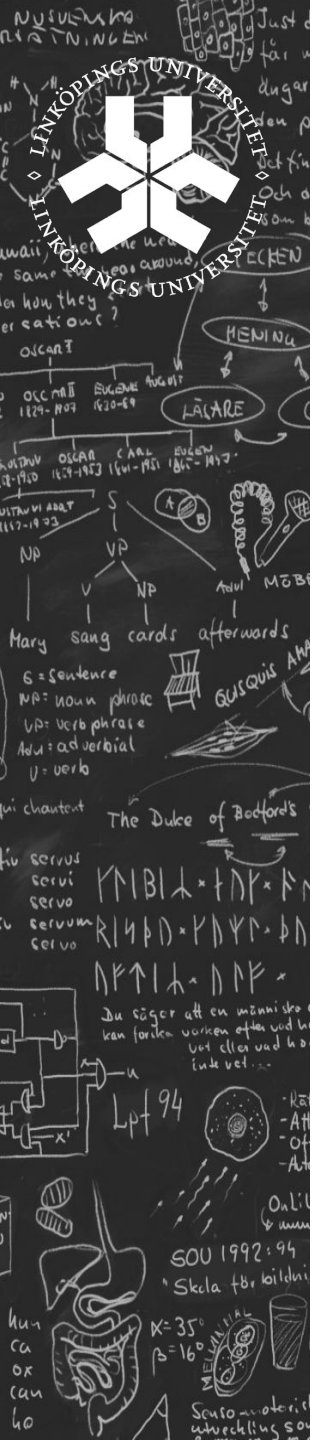


Overview

- Background
- Our proposal
- The Token-Equivalence Method (TEM)
- Alignment experiments
- Conclusions

LiU expanding reality

Process of a translation exercise



	Before class		In class	After class
Teacher Tasks:	Define exercise (text + brief) Preprocessing Distribution	Reading for assessment and grading, Preparation	Classroom presentation and discussion	Final grading
Student tasks:	Translation, Delivery			Revision

LiU expanding reality

Examples of computational aids for the translation exercise

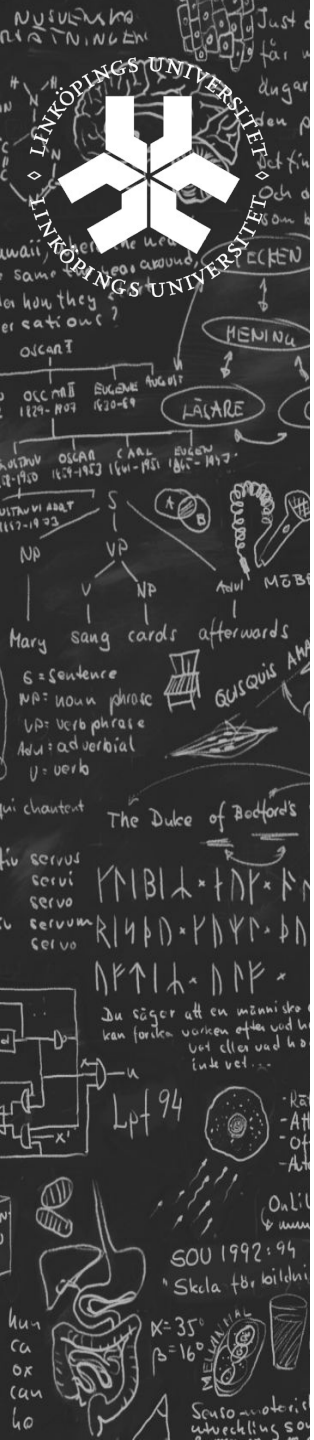
- E-learning environments
  - Fictumová, 2004, 2007
- Corpora
  - Lopez-Rodriguez and Tercedor-Sanchez, 2008;
  - Pastor and Alcina, 2009
- CAT tools
- Assessment of translations as literal or liberal
  - Shei and Pain, 2002

Our idea

- Computer-aided support for the Token-Equivalence Method (TEM; Tarvi, 2004)
- A new application area for alignment technology

  Supporting
  - teacher's assessment and grading
  - discussion in class

Process of a translation exercise

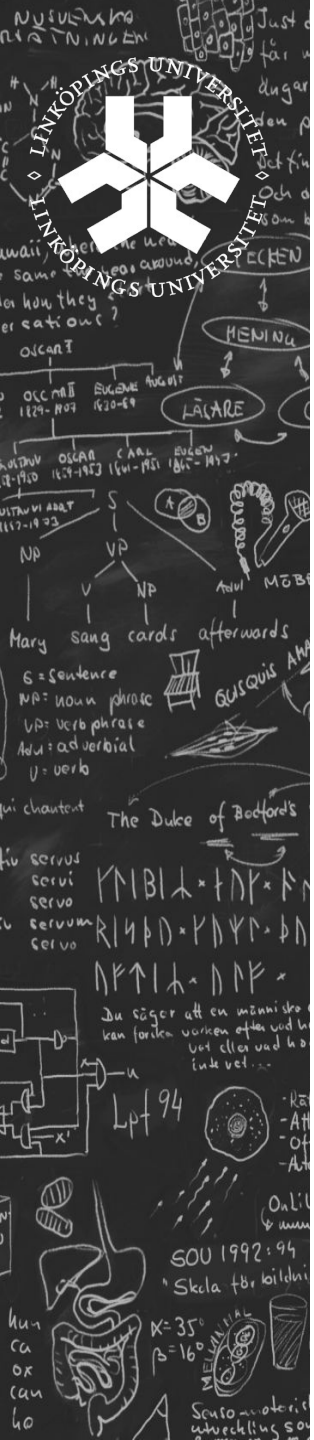


	Before class		In class	After class
Teacher Tasks:	Define exercise (text + brief) Preprocessing Distribution	Reading for assessment and grading, Preparation	Classroom presentation and discussion	Final grading
Student tasks:	Translation, Delivery			Revision

LiU expanding reality

Token alignment as a basis for instruction in class

- Segment views
  - display of different translations of the same source segment
- Token views
  - display of different translations of the same token(s)
- Type views
  - e.g. frequency tables of translations of words and phrases
- Global views
  - metrics and grades computed for the full text or parts thereof



The Token-Equivalence Method (TEM)

- Token correspondences, based on
  - content words
  - denotational meaning
- Frames
  - metrics that quantify relations between source and translation
  - combined to rank translations

LiU expanding reality

An example (RU - EN)

Pushkin, Eugene Onegin, stanza LIX: 1-2

- *Proshla lyubov, yavilas' muza, i projasnilsya tyomnyi um.*

Translation (by Nabokov)

- *Love passed, the Muse appeared, and the dark mind cleared up.*

Indexing tokens

- *1:Proshla 2:lyubov, 3:yavilas' 4:muza, 5:i 6:projasnilsya 7:tyomnyi 8:um.*

  [passed] [love] [appeared] [muse] [and] [cleared up] [dark] [mind]

- *1:Love 2:passed, 3:the 4:Muse 5:appeared, 6:and 7:the 8:dark 9:mind 10:cleared 11:up.*

”Standard” alignment representation

- 1-2 2-1 3-5 4-4 5-6 6-10 6-11 7-8 8-9 0-3 0-7

An example (RU - EN)

Pushkin, Eugene Onegin, stanza LIX: 1-2

- *Proshla lyubov, yavilas' muza, i projasnilsya tyomnyi um.*

Translation (by Nabokov)

- *Love passed, the Muse appeared, and the dark mind cleared up.*

Indexing tokens

- *1:Proshla 2:lyubov, 3:yavilas' 4:muza, 5:i 6:projasnilsya 7:tyomnyi 8:um.*

  [passed] [love] [appeared] [muse] [and] [cleared up] [dark] [mind]

- *1:Love 2:passed, 3:the 4:Muse 5:appeared, 6:and 7:the 8:dark 9:mind 10:cleared 11:up.*

”Standard” alignment representation

- 1-2 2-1 3-5 4-4 5-6 6-10 6-11 7-8 8-9 0-3 0-7

multiword correspondence

LiU expanding reality

An example (RU - EN)

Pushkin, Eugene Onegin, stanza LIX: 1-2

- *Proshla lyubov, yavilas' muza, i projasnilsya tyomnyi um.*

Translation (by Nabokov)

- *Love passed, the Muse appeared, and the dark mind cleared up.*

Indexing tokens

- *1:Proshla 2:lyubov, 3:yavilas' 4:muza, 5:i 6:projasnilsya 7:tyomnyi 8:um.*

  [passed] [love] [appeared] [muse] [and] [cleared up] [dark] [mind]

- *1:Love 2:passed, 3:the 4:Muse 5:appeared, 6:and 7:the 8:dark 9:mind 10:cleared 11:up.*

”Standard” alignment representation

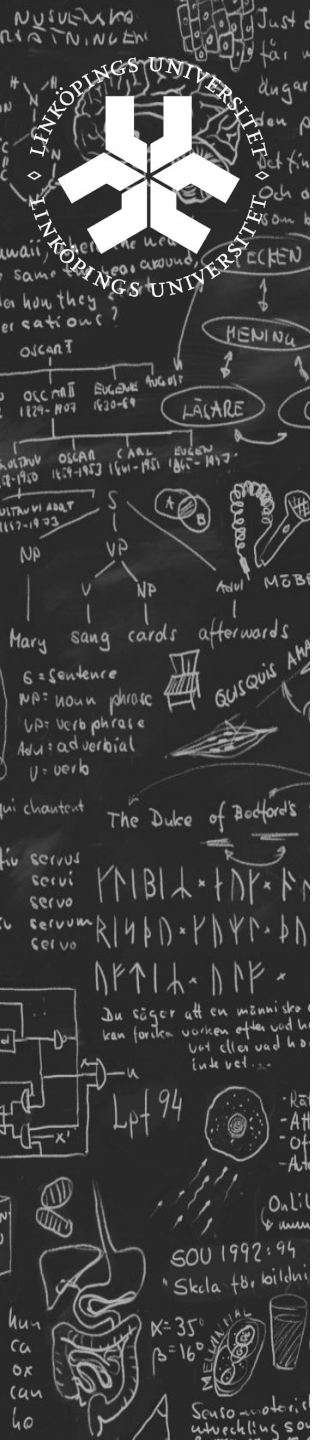
- 1-2 2-1 3-5 4-4 5-6 6-10 6-11 7-8 8-9 0-3 0-7

null links

LiU expanding reality

Differences TEM and standard SMT alignment

Aspect	TEM	SMT
Approach	manual	automatic
Punctuation	ignored	tokenized
Token types	two types	one type
Multiword units	single tokens	several tokens



Some definitions

- t_S: a source token
- t_T: a target token
- null(t): a token without correspondent
- nonnull(t): a token with at least one correspondent
- cont(t): a content token
- gram(t): a grammar token

TEM frames

- Basic content frame
  - ”the percentage of source content tokens that have received a translation”
  - BCF = 100 * | { t_s | cont(t_s)∧nonnull(t_s) } | / |{t_s}|
- Optional content frame
  - OCF = | { t_T | cont(t_T)∧null(t_T) } |

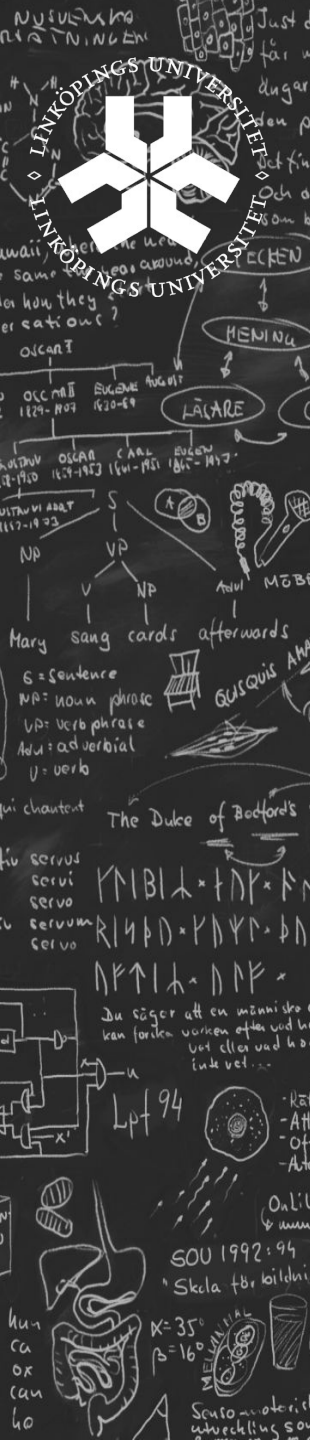
TEM frames (cont.)

- Basic formal frame
  - ”the number of grammar tokens in the translation”
  - BFF = $| \{ t_T \mid gram(t_T) \wedge null(t_T) \} |$
- Optional formal frame 1
  - ”the percentage of source tokens that are translated by a token of the same part-of-speech”
- Optional formal frame 2
  - ”the percentage of pairs of source tokens whose order and dependency relation is kept under translation”

The translation quotient (TQ)

- The TQ is defined as the average of all frames that are expressed as percentages:
- TQ = (BCF + OFF1 + OFF2) / 3

- All frames may be used to compute a rank for each translation



Word alignment for the translation class

- Source texts are short
- Translations, on the other hand, may be many
- Source texts are known beforehand
- Content tokens and grammar tokens should be treated differently

→ Statistical and rule-based methods may be combined

LiU expanding reality

Alignment experiments

- Russian-English data
  - 8 translations of 17 stanzas from Eugene Onegin
- English-Swedish data
  - 5 translations of two small extracts of English prose text used as exercisez in a course.
  - ***J.D. Salinger. Catcher in the rye, New York, 1951 Roddy Doyle: The Van, 1991.***
- Systems used
  - Giza++ (both corpora)
  - A ”pressure-aligner” (only EN-SE), using
    - a dictionary
    - part-of-speech patterns
    - alignment topology

Alignment results, RU-EN, Giza++ (model 4)

	Precision	Recall	F-measure
1 trl, all links	0,308	0,298	0,303
8 trls, no null links	0,434	0.467	0.450
8 trls, all links	0,482	0.480	0.481

Note: the gold standard used has some 40% added tokens, while Giza++ gives 20%.

Alignment results for EN-SE, Giza++ (model 4)

	Precision	Recall	F-value
1 trl, no null links	0.751	0.652	0.698
1 trl, all links	0.681	0.681	0.681
5 trls, no null links	0.816	0.698	0.752
5 trls, all links	0.752	0.738	0.745

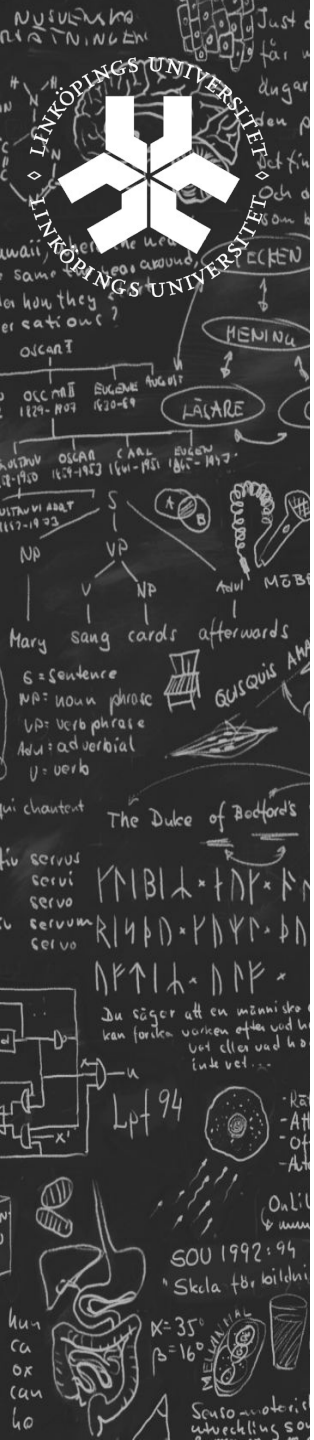
Alignment results for EN-SE, rule-based aligner

	Precision	Recall	F-value
PA1, no null links	0.815	0.492	0.614
PA1, all links	0.502	0.554	0.527
PA2, no null links	0.885	0.608	0.721
PA2, all links	0.606	0.664	0.633

PA1 has a small lexicon, while PA2 has a lexicon adapted for the corpus.

Alignment results for EN-SE, combinations of Giza++ and rule-based aligner

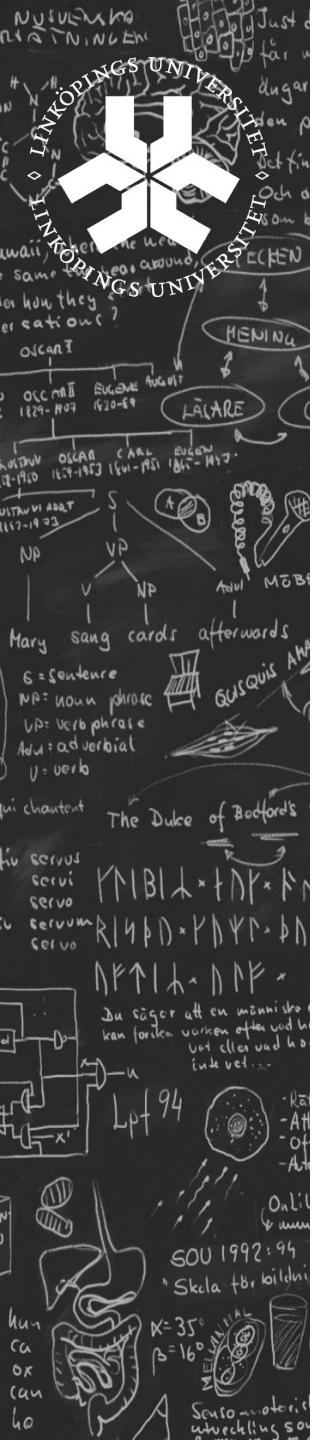
	Precision	Recall	F-value
Union, no null	0.775	0.777	**0.776**
Union, all	0.739	**0.789**	0.763
Intersection, no null	**0.980**	0.530	0.688
Intersection, all	0.875	0.543	0.670
Grown, no null	0.849	0.665	0.746
Grown, all	0.794	0.660	0.721



Observations on alignment performance

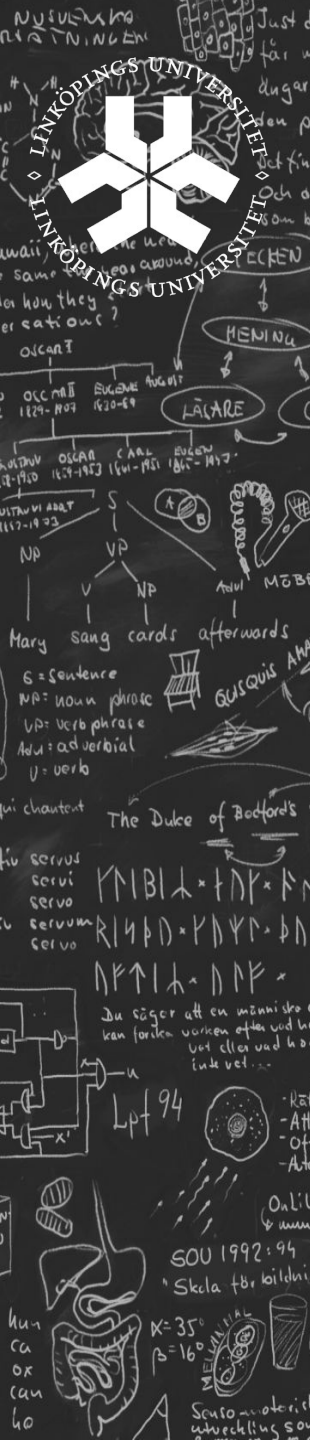
- As expected, adding more translations improves the results of the statistical aligner
- Since the source text is known, and small, creating a dictionary for the source adapted for the task is not so much work and improves the results of the pressure aligner substantially
- A combination of statistical and dictionary-based alignment can give very high precision
- All possibilities have not been explored yet...

LiU expanding reality



Conclusions

- There is much work ahead
  - implementation
  - trying it out
- Even with further improvements in the automatic tools, there will still be much to do for the teacher in reviewing and correcting token alignments
  - Need for good interactive tools!



Thank You

LiU expanding reality