CG-based Swedish-Danish MT

Eckhard Bick

University of Southern Denmark

What's this about?

- A Swedish-Danish MT system (Swe2Dan) that's not just guessing like all the others
  - Imagine, using linguistic rules in the 21 century!
  - Imagine, actually understanding what your system does!
  - Imagine being able to fix errors as they arise!
- A Constraint Grammer-based framework for MT (GramTrans):
  - lexica, lexica, rules, rules, rules ...
  - context at your fingertips, as wide as you like, as linked as you like
  - imagine using not trigrams but 25grams ...
- How to compete with people (Google & Gates) that are a million times richer than you, and have a million times more data

STMT

- Statistical MT (STMT) has all but replaced Rule-based MT (RBMT) in research, so it must be better. Or must it?
- STMT profits from (needs!) training data
  - especially if you have copy-right-free acces to the whole worlds online language data (Google or Bing)
  - especially if you are a main provider in a narrow domain (e.g. manuals, subtitles)
- STMT is cost-efficient
  - .... for companies and programmers, because the tax-payer has already paid linguists to build treebanks or the EU to translate documents and discussion transcripts
- STMT harvests fluency from target language corpora
- STMT allows you to work on languages you don't speak
  - and therefore can't do an error-analysis of ...

STMT problems

- For less-ressourced languages STMT may lack sufficient trainting data, or experience quality loss using English as an interlingua
- Without linguistic analysis, STMT has difficulties with morphologically rich languages: Individual forms are rare, n-grams can't handle features such as clause part agreement, anaphora, reflexivity
- Difficult to systematically adapt to a different domain than the one there's training data for
- Difficult to fix individual errors, even when identified, or to integrate lexical knowledge from external sources
- Without dictionaries, there's a risk of semantic confusion of words sharing the same context (currency words, antonyms)

Advantages of RBMT

- can build symbolic language models without training data
- RBMT (and hybrid systems) support deep module integration and have system-wide access to a full linguistic analysis
- Rules can be amended "locally" to handle errors, sense distinction or domain migration
- STMT can be improved by hybrid add-ons, e.g. syntactic reordering after dependency analysis, but it is an advantage of RBMT to handle this in one formalism (here: CG), with a shared category set and shared lexical information

Our approach

- High-quality Constraint Grammar dependency analysis as an integrative framework for MT
- Rule-transparency for agreement, compounding, disambiguation etc.
    - all information available at all levels
- Context-driven transfer rules
- Context-driven movement rules
- Long-distance relations
- Possibility for language-contrastive rules
- Possibility for domain-based rules
- Possibility for lexical heuristics

Underlying System Architecture

- Focus on good & deep SL analysis: SweGram parser
- Constraint Grammar system
  - uses our own, open source CG3 rule compiler
  - uses a 70,000 lemma lexicon
  - uses 8,500 tagging and disambiguatin rules
- Provides global MT access to the following information:
  - tokenization with recognition of abbreviations, numerical and scientific expressions, complex functions words, NER
  - morphological analysis, incl. compounding analysis, derivation and endings-based out-of-vocabulary heuristics
  - syntactic function tags (subject, object, predicative)
  - dependency trees
  - semantic classification of common and proper nouns
  - valency classification of verbs

Token based, distinct tag fields

```
De              [den] <*> ART nG P DEF @>N  #1->3
senaste         [sen] <jtemp> ADJ SUP nG nN DEF NOM @>N  #2->3
dagarnas        [dag] <dur> <temp> N UTR P DEF GEN @>N  #3->4
snöstorm        [snöstorm] <event> <wea> <F:snö+storm> N UTR S IDF
                    NOM @SUBJ>  #4->7
i               [i] <np-close> PRP @N<  #5->4
New=York        [New=York] <civ> <*> PROP NEU S NOM @P<  #6->5
fick            [få] <vt+INF> <mv> V IMPF AKT @FS-STA  #7->0
myndigheterna   [myndighet] <HH> <aci-subj> N UTR P DEF NOM
                    @<ACC  #8->10
att             [att] INFM @INFM  #9->10
inviga          [inviga] <mv> V INF AKT @ICL-<OA  #10->7
en              [en] ART UTR S IDF @>N  #11->13
splitterny      [splitterny] <heur> <F:splitter+ny> ADJ UTR S IDF NOM
                    @>N  #12->13
snösmältningsmaskin     [snösmältningsmaskin] <good-compound>
                    <N:snösmältning~s+maskin> <mach> N UTR S IDF
                    NOM @<ACC  #13->10
$.              [.] PU @PU  #14->0
```

Lexical Transfer

- CG tags are used as either local or contextual disambiguators, without the need for statistics
  - polysemy resolution (distinguish, not define meaning)
  - usage differences (synonym picking)
- local distinctors
  - POS, number, syntactic function, domain
- contextual distinctors:
  - features of arguments, attributes, heads etc.
  - dependency links to the 2nd degree
  - relative positions left/right + their dependents

Translation picking: *när*

- när_ADV
  - :nær *[near]*;
  - S=(<interr>) :hvornår *[when?]*;
  - S=(<rel>) :når *[(at the time) when]*;
  - S=(@>A) :næsten *[almost]*;
  - P-1=('hart') :næsten *[almost]*
- när_KS
  - :når *[when(ever)]*;
  - H=(IMPF) :da *[when ..ed]*

How close are Swedish and Danish?

- both Eastern Scandinavian languages
- but one-on-one translation is only possible in 50% of tokens
- traditionally regarded as mutually intelligible,
  - but only with considerable training
  - not true for many young people
- lexicon entries with transfer rules:
  - 4% of ca. 100.000 lexemes in the bilingual dictionary
  - but 53% of running tokens
  - verbs account for 40% of all contextual transfer rules

Transfer rules, verb example

- fräsa_V
  - :hvæse *(to hiss like a cat);*
  - D1=("åt") GD1=(<H>) D2=(<H> @SUBJ) :vrisse *(to snap at sb)*;
  - D=(<[HV].*> @SUBJ) D=("(iväg|förbi)") D!=(@ACC) :rase *(tear/speed along)*;
  - D=(<food.*> @ACC) :stege, :brune, :brase, :lynstege *(to fry)*;
  - D=(@ACC) D=(<H> @SUBJ) :fræse *(to mill, to cut a material or tool);*

Relations: D=dependent, GD=granddaughter, H=head

Tags: POS, @function, <semantics>, "lemma", 'word'

Adding transfer-specific features

- Not all useful contextual features are automatically marked by the SweGram parser, for instance
  - because they have not "local" (morphological) marker or structural role
  - because they are relevant only for the Danish TL
- Such features can be added *before* transfer by a separate CG module
  - reflexivity
  - article insetion
  - feature propagation from dependent to head or vice versa
    - definiteness
    - number
    - ±human
- Second round of feature propagation *after* transfer
  - e.g. gender/number agreement where different for Danish

Compounds

- both Swedish and Danish are rich in productive compounds and derivations, so a large portion of the lexicon will always escape ready-made treatment
- part-by-part translation is therefore necessary
- first, second and middle parts can be addressed separately by the transfer rules
    - lock_N (25)
        - :lok, :hårlok *[curl];*
        - S=(<second>) :låg *[cover]*;
        - S=(NEU) :låg;
        - S=(<first>) :lokke *[luring]*

Out-of-vocabulary Words

- Treatment at parser level (SweGram)
  - simple misspellings (low Levenshtein distance)
  - live analysis of productive compounding
- Treatment at the transfer level
  - rule-based part-for-part translation of compounds
  - fallback: heuristic transformation, treating Swedish as "misspelled Danish"
    - translation of only endings and affixes
    - systematic letter replacement
      -ism --> -isme, -skap --> -skab, ö/ä -> ø/æ

Vocabulary coverage: SweGram

- evaluation corpus of 144,456 non-punctuation tokens
- SweGram parser out-of-vocabulary
  - 7,120 "live recognized compounds
    - not already in the lexicon (often with known compound analyses)
    - but with recognized and allowed parts
    - names were not counted in this group
  - 1,245 outright heuristic analyses
    - 69% misspellings, non-letter characters or missing spaces

Vocabulary coverage: Swe2Dan

- Swe2Dan coverage
  - 99.71% Swe2Dan translation dictionary coverage for ordinary, parser-sanctioned words
  - 99.1% of compounds had a non-heuristic translation
  - 368 words known to the parser, but not to Swe2Dan
    - 51.6% left as-is (many were foreign words)
    - 11.7% correct Swedish-->Danish transformations
    - 36.7% wrong translations (as is or transformed)
  - Translation of parser (non-compound) heuristicals
    - 62.1% Swe2Dan lexicon coverage
    - 27% left as-is, 10% correct, 17% wrong
    - 11% transformations: 10% correct, 1% wrong
  - Regular + compound coverage: 99.33%
  - Regular + compounds + transformations/as-is: 99.62%

Target Language Generation

- Swedish and Danish are more different than you'd think
  - E.g. Both have 2 genders, but those may be different even for cognates
  - And because gender propagates across np's, and to predicatives even in relative clauses etc., even the Danish generator profits from dependency analysis
- Coverage (10,000 sentences news text test chunk):
  - 94.49% of generator lookups successful as full lexemes
  - 3.03% compound second part was used
  - 0.78% heuristic "compound" analysis by the generator
  - 1.84% last resort: most common paragdim for the POS (including 0.14% for compounds), with NO errors (probably because irregular forms are frequent and therefor lexicon-covered
- i.e. almost 100% generator coverage, so errors occur only because of lower level errors (POS/inflexion)

Structural Transfer: Problems

- Swedish and Danish share basic syntax
- However, some issues have a scope beyond n-grams:
  - definiteness differences ask for interdependent (i.e. simultaneous) treatment of left and right np-edges
  - Adverb placement varies, e.g.
    - phrasal verb particles before/after the object (Danish after, Swedish before)
    - V A S --> V S A
    - adverbs before/after infinitive marker
  - Special Swedish "supine" verb forms, often indistinguishable from active voice participles --> translation ambiguity: past tense finite verb / participle clause

Structural transfer: Solutions

- For individual words, transfer rules
  - i.e. Swedish definite article as "nil", if H=(N DEF)
- For constituent ordering, transfer rules are cumbersome
  - e.g. nilling a phrasal adverb, while (re)adding it to the object translation
    packa_V D=("ut")_nil D=(@ACC)_[+ud] :pakke *[unpack]*
- Movement rules
  - w(@MV<|@OA), g(<right> @ACC) -> 2,1
  - works independently of constituent size, because dependents are automatically "carried along"
  - all in all, 61 movement rules
- Parser support + structural changes:
  - (1) Supine forms need to be recognized as finite verbs rather than participles (parser task)
  - (2) At transfer, there's ambiguity as to tense, auxiliary insertion and subject/averb movement

Evaluation

- 100 random sentences (1,500 words)
  (Leipzig Corpus Collection)
- Comparison with
  - Google Translate (STMT)
  - Bing Translator (STMT)
  - Apertium (open source RBMT but with HMM taggers)
- Gold standards
  - Manual translation
  - System post-edits
- Methods
  - BLEU/NIST (Papinieni et al. 2002)
  - TER (Snover 2006)

Individual system evaluation

BLEU / NIST	Manual reference (1)	edited system reference	multi-reference (all minus self)
GramTrans	**0.645** / 8.515	**0.838** / 9.817	**0.757** / 10.050
Google	0.387 / 6.300	0.645 / 8.361	0.539 / 8.150
Apertium	0.390 / 6.391	0.516 / 7.361	0.468 / 7.418
Bing	0.342 / 6.006	0.600 / 8.064	0.492 / 7.793

- GramTrans best, statistical systems profit from self-edit comparison (which tends to retain n-gram order)
- GramTrans advantage over Google bigger for BLEU than NIST
  - NIST downplays short/common words and small length differences, both of which ignore RBMT strengths (function words and inflexion/agreement)

Cross-system BLEU evaluation

Reference: Test:	GramTrans edited	Google edited	Apertium edited	Bing edited
GramTrans	(0.838)	0.497	**0.666**	0.501
Google	0.387	(0.645)	0.384	**0.478**
Apertium	**0.426**	0.330	(0.516)	0.325
Bing	0.358	**0.446**	0.353	(0.600)

- GramTrans/Apertium (RBMT) and Google/Bing (STMT) both profit from being evaluated against a system of the same research paradigm (RBMT/STMT)
- But even when tested against STMT post-edits, GramTrans outperforms the STMT systems

TER edit distances (low is good)

Reference: Test:	Manual	GramTrans edited	Google edited	Apertium edited	Bing edited
GramTrans	**20.84**	**(8.57)**	32.12	***19.77***	30.98
Google	45.05	44.40	(23.60)	45.20	***37.56***
Apertium	34.54	***31.13***	41.96	(24.51)	41.75
Bing	48.62	46.98	***40.70***	48.03	(28.05)

- In relative terms, the GramTrans advantage is even bigger in TER than in BLEU/NIST
- GramTrans has a 3 times "lower editing cost" than the other systems
- TER confirms the in-paradigm similarity of systems

Qualitative Inspection

- GramTrans (and Apertium) completely avoid certain "STMT traps":
  - confusion of "ontological sister terms"
    - Bing: "dollar/kroner", "Svensson/Smith"
    - Google: "Solbergaskolen/Solbergaleden" (org/place)
  - literal name translation caused by case folding
    - Bing: "Huge chockstartade"
      --> "Kæmpe [=big/huge] chockstartade"
  - big-data training language used as intermedia (English) for the small languages Swedish and Danish
    - Google: "styrelse" --> "board" --> "bord" (table)
    - Google: "Se upp för elgen" --> "Watch out for elg"
  - compounding:
    - "säkerhetsexpert" --> "security expert" --> "sikkerhed ekspert" (should be "sikkerhedsekspert" as in Swedish)

Long distance relations: Reflexivity

- Long distance relations are a strength of RBMT
  - e.g. reflexive possessiv "sin" versus non-reflexive "hans"

		refl	num-ber	lex	com-pound
Swedish original	Nicole talar ut om sina viktproblem *Nicole speaks out about her weight problems*				
GramTans	Nicole taler ud om sine vægtproblemer	ok	ok	ok	ok
Google	Nicole taler ud om hendes vægt problemer	err	ok	ok	err
Apertium	Nicole taler ud om sine viktproblem	ok	(ok)	err	(ok)
Bing	Nicole taler ud om hans vægt problem	err	err	ok	err

- Only GramTrans gets the whole reflexive np right
- STMT messes up the long distance relation (refl), BING also close agreement (num), Google the compound
- Apertium gets the relations right, but the lemma wrong

A real-life application: WikiTrans

Titan (måne)

> Nedenstående er en automatisk oversættelse af artiklen *Titan (måne)* fra den svenske Wikipedia, udført af GramTrans den 2014-08-24 10:20:31. Eventuelle ændringer i den svenske original vil blive fanget igennem regelmæssige genoversættelser. Du har mulighed for at redigere oversættelsen til brug i den originale danske Wikipedia.

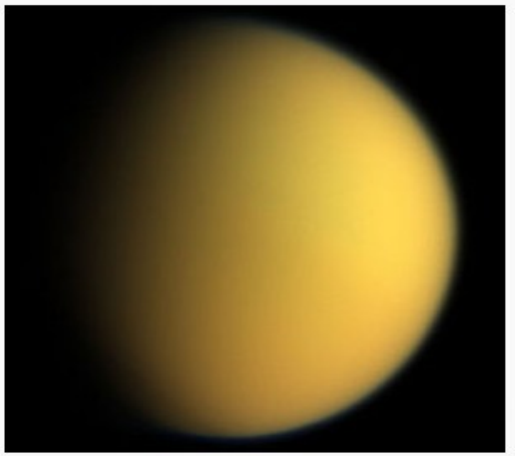
Titan er Saturns største måne og den den næststørste måne i solsystemet , efter Jupiters måne Ganymedes . Titan blev opdaget den 25 marts 1655 af den nederlandske astronom Christiaan Huygens og var den første satellit i solsystemet som blev opdaget efter Jupiters galileiske måner . Titan er den eneste måne i solsystemet som har en tæt atmosfære [4] . Den tætte atmosfære har forhindret nærmere studier af månens overflade, men i øjeblikket undersøges Titan af rumsonden Cassini-Huygens og ny viden tilføjes hele tiden.

Den 27 juli 2006 meddelte NASA at man har fundet søer af kulbrinter ved Titans nordlige polarregioner.

Indhold

Indholdsfortegnelse

Titan



Titan

Opdaget	
Opdaget af:	Christiaan Huygens
Upptäcksår:	25 marts 1655
Omløbsbanens egenskaber	
Afstande fra planeten:	1.221.870 km
Excentricitet :	0,0288
Omløbstid :	15,945 d
Middelbanhastighed:	km/s
Inklination :	0,34854°
Satellit til :	Saturn
Fysiske egenskaber	
Radius :	2576± 2,00 km (0,404 jorde) [1]
Areal:	8,3× ^{107} km²
Masse :	1,3452± 0,0002× ^{1023} kg

Navn

Huygens kaldte helt enkelt sin opdagelse *Saturni Luna* (som er latin for "Saturnus måne") (*De Saturni Luna observatio nova* , 1656; XV). Senere navngav Giovanni Domenico Cassini de fire måner som han opdagede (Tethys , Dione , Rhea og Japetus) til *Sidera Lodoicea* ("Ludvigs stjärnor") for at hædre kong Ludvig XIV av Frankrig . Astronomer begyndte at kalde dem for *Saturnus I* til *Saturnus V* . Da Mimas og Enceladus blev opdaget i 1789 kaldtes Titan sommetider for *Saturns sjette satellit* eller *Saturnus VI* (benævnelsen anvendes fortsat) som kommer af i hvilken orden månerne befinder sig fra Saturn. Navnet "Titan" og navnene på de andre da kendte syv månerne kom fra John Herschel (søn til William Herschel som havde opdaget Mimas og Enceladus) som foreslog at månerne skulle navngives efter titanerne som var søskende af Kronos (den græske pendent til guden Saturn).

Titan kan ikke ses fra jorden med det blotte øje, men kan ses med små teleskoper (med diameter større end 5 centimeter) og stærke kikkerter.

Language discriminsion in Wikipedia

- The amount of information available in Wikipedia is very language-dependent

Language	Articles	Words (2013)	Binaries, e.g. pict (2013)	Editors / M speakers	Stanisław Lem (scrolling pages)
English	~ 4,600,000	?	1.8 M (2009)	21	9
Swedish	> 500,000 (1,800,00)	316 M	157 k	69	4
Esperanto	~ 190,000	48 M	160 k	98	2
Danish	~ 200,000	62 M	12 k	36	1 (stub)

Stanisław Lem: Danish stub

Stanisław Lem

Fra Wikipedia, den frie encyklopædi

Stanisław Lem (12. september 1921, Lviv – 27. marts 2006, Kraków) var en polsk forfatter, digter, filosof og futurolog.

Han er kendt mest for sine science fiction-romaner, der er oversat til 41 sprog og solgt i 27 millioner eksemplarer.

Stanisław Lem i 1966

Lem på dansk [redigér | redigér wikikode]

- *Solaris* (*Solaris*, 1961)
- *Fremtidskongressen* (*Kongres futurologiczny*, 1971)
- *Hvordan verden blev reddet* (*Jak ocalał świat*, ?)

Se også [redigér | redigér wikikode]

- detaljeret bibliografi af Lem på den engelske Wikipedia

Eksterne henvisninger [redigér | redigér wikikode]

- Wikimedia Commons har flere filer relateret til **Stanisław Lem**
- Stanisław Lems hjemmeside
- Nekrolog i The Times

Stanisław Lem, Kraków (Poland), 30.10.2005

Stub

Denne forfatterbiografi er kun påbegyndt. Du kan hjælpe Wikipedia ved at tilføje mere.

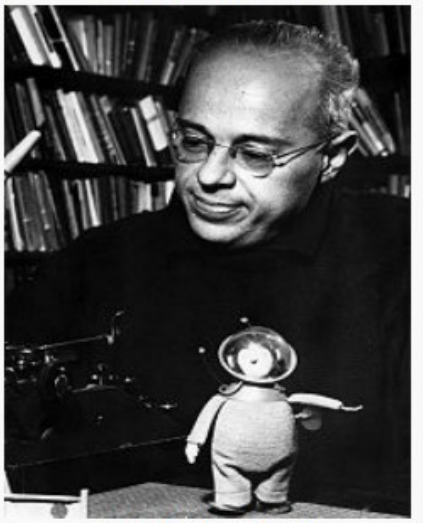
Stanisław Lem: WikiTrans

Stanisław Lem

Nedenstående er en automatisk oversættelse af artiklen *Stanisław Lem* fra den svenske Wikipedia, udført af GramTrans den 2014-09-02 23:00:59. Eventuelle ændringer i den svenske original vil blive fanget igennem regelmæssige genoversættelser. Du har mulighed for at redigere oversættelsen til brug i den originale danske Wikipedia.

Stanisław Lem , født 12 september 1921 , død 27 marts 2006 , var en polsk satirisk og filosofisk science fiction-forfatter . Hans bøger er blevet oversat til 41 forskellige sprog og solgt i 27 millioner eksemplarer.

Stanisław Lem



Stanisław Lem, 1966

Født	12 september 1921
Død	27 marts 2006 (84 år)
Nationalitet	Polen
Sprog	Polske

Biografi

Stanisław Lem blev født i Lwów , Polen i 1921 (nu Lviv , Ukraine) som søn til en læge. Han læste medicin ved Lwów-universitet, men Anden Verdenskrig afbrød hans uddannelse. I krigen og den tyske besættelse arbejdede han som mekaniker og svejser og var medlem i den polske modstandsbevægelse. 1946 flyttede Lem fra området som er blevet annekteret af Sovjetunionen til Kraków og genoptog medicinstudierne ved Jagielloniska universitetet . Efter at have gennemført sin uddannelse bestemte han sig for ikke at tage diplomer for at slippe at arbejde som militærlæger. Han arbejdede som forskningsassistent ved et forskningsinstitut og begyndte da at skrive sagaer på sin fritid. 1981 tog han et æresdiplom fra Wrocław Polyteknisk , senere fra Opole-universitet , Universitetet i Lwów , og senest fra Jagielloniska universitetet .

Andre emner indeholder ideale og utopiske civilisationer og problemet med menneskelig eksistens i en verden hvor det stort set noget findes ikke at gøre på grund af teknisk udvikling. Hans ekstraterrestriske civilisationer indeholder ofte mekaniske fluesværme (i *Segraren*) og havet (i *Solaris*). Spørgsmål om tekniske utopier findes i *Fred på jorden* , i *Observation on the Spot* , og, i mindre udstrækning, i *Cyberiaden* .

Lem blev æresmedlem i Science Fiction and Fantasy Writers of America (SFWA) i 1973 men blev fyret ud i 1976 da han kritiserede amerikansk science fictions-litteratur. Han havde beskrevet den som kitsch, meget gennemtænkt, dårligt skrevet, og at den ikke havde interesse af at udvikle nye litteraturformer og - idéer uden i stedet tjene så meget penge som muligt. Senere inviterede SFWA ham på et regulært medlemskab, hvilket han takkede nej til.

Lems bøger er fyldte af intelligent humor og ordlege og er blevet oversat til svensk af Martin von Zweigbergk og Sven Christer Swahn .

I året 1977 blev han anerkendt som æresborger af Kraków.

Bibliografi

Conclusion and Outlook

- For the Swedish->Danish language pair, a modular CG-based system with manual transfer rules can outperform STMT benchmark systems such as Google and Bing
- GramTrans/Swe2Dan is mature enough for direct application, e.g. WikiTrans
- Future research:
  - improved heuristics for out-of-vocabular words
  - more comprehensive use of domain flags
  - Danish->Swedish sister system (non-trivial in RBMT, because neither SL analysis nor transfer rules can be reused)

Swe2Dan: **http://gramtrans.com**

WikiTrans: **http://wikitrans.net**

SweGram & Constraint Grammar:
http://visl.sdu.dk

eckhard.bick@mail.dk