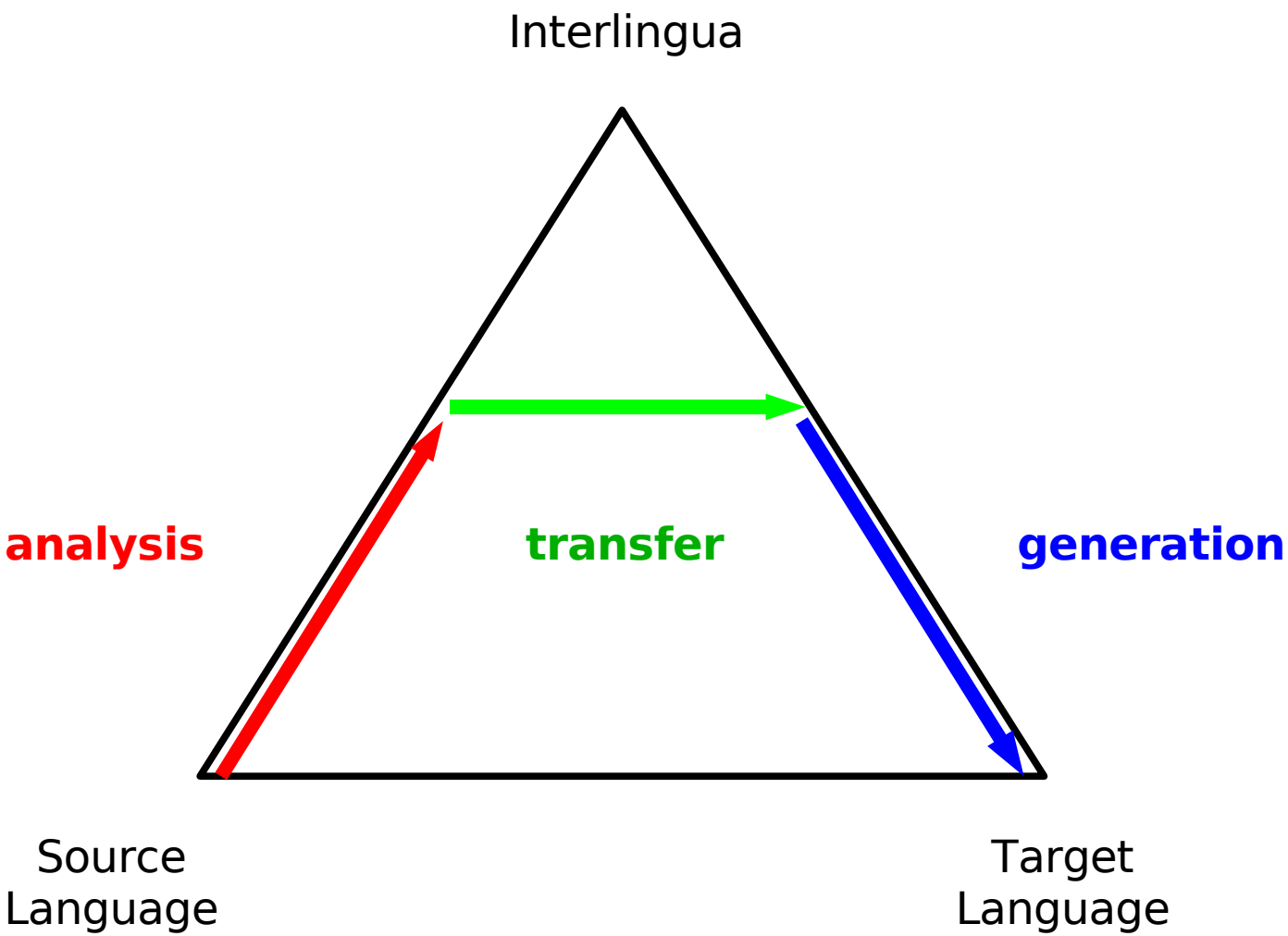
Eckhard Bick

Constraint Grammar based Machine Translation

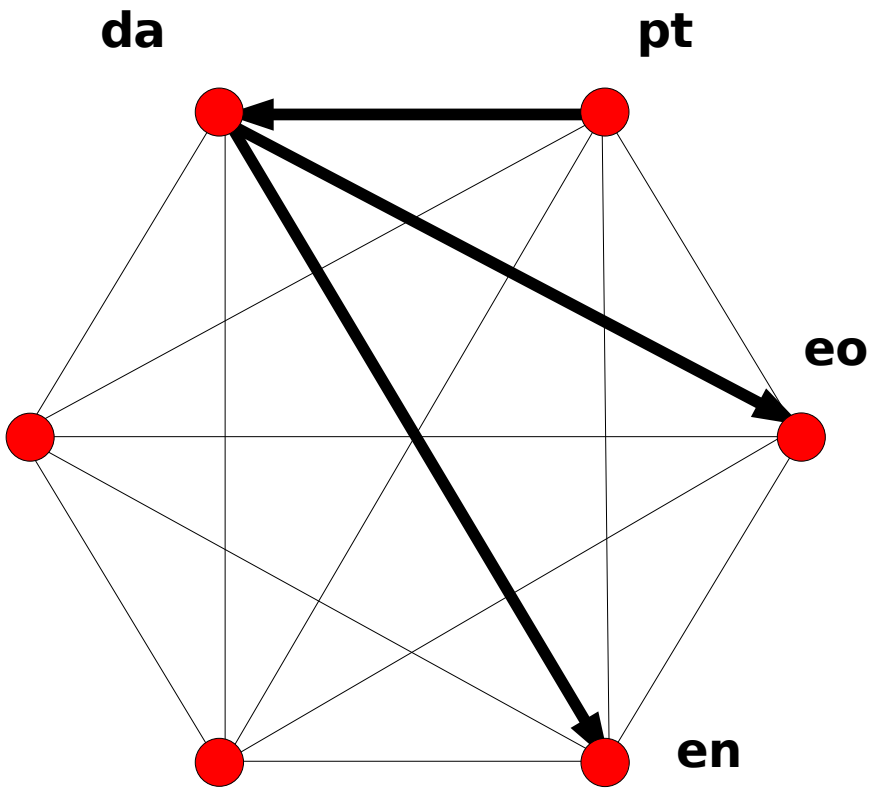
Core principles

- The crucial idea is to reduce MT to good SL-analysis - i.e. to express as much of polysemy-disambiguation, syntactic movement etc. in terms of dependency links, syntactic function tags and semantic prototype selection restrictions
- Don’t guess if you know:
  - choose rule based methods as long as possible
  - use relational context (functional dependencies) rather than n-grams
  - use translation memory where it makes sense, i.e. for fixed expressions
  - use statistics (only) where no grammatical distinctors can be found
- Wherever possible, express semantics in terms of syntax
  (Halliday: Ever more fine-grained syntax becomes semantics)

The MT Triangle

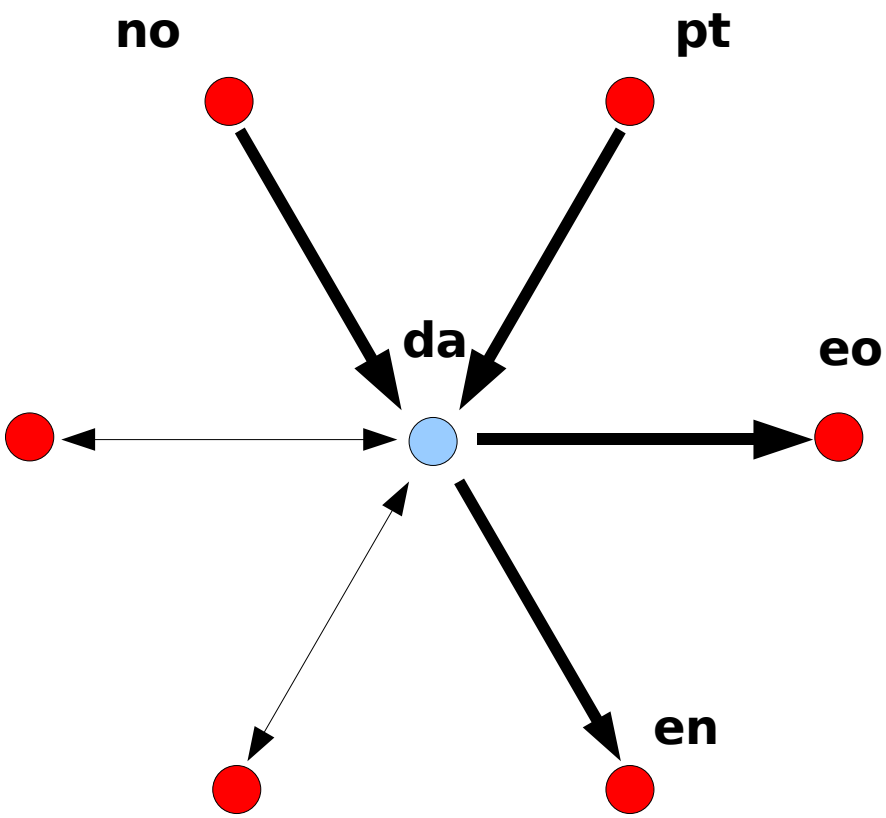


Direct transfer
n (n-1) systems
tailored, but slow to build
faster at run-time



6 * 5 = 30
(now covering 3)

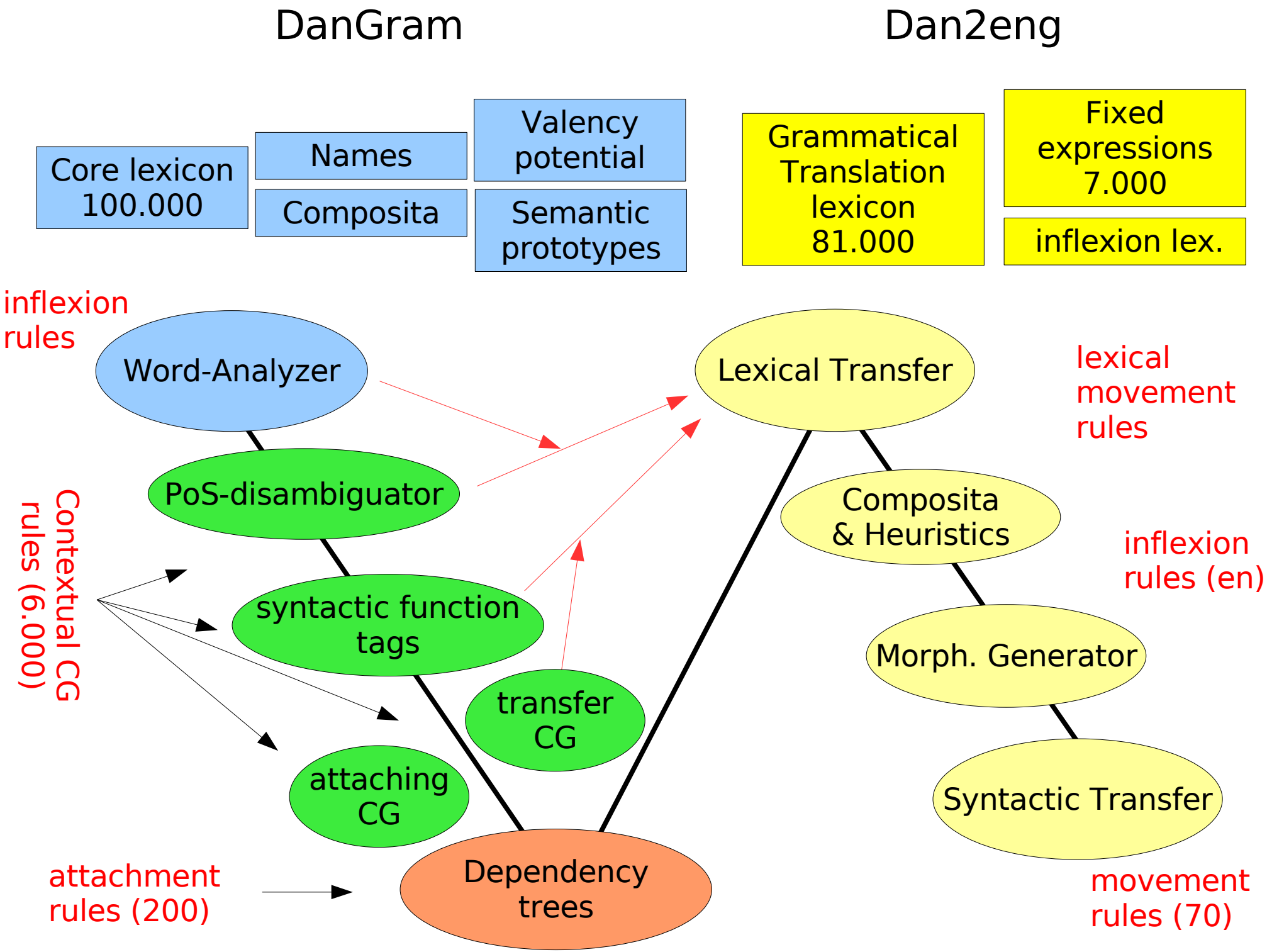
Interlingua
2n systems
less specific, but faster to build
slower at run-time



2 * 6 = 12
(no covering 8)

VISL research languages, parsers & treebank tools

	revised syntactic trees (tokens)	**morphological analysis**	**syntactic analysis**	**semantics**
	200.000* 4 subcorpora	lexicon and rule based analyzer + CG	CG + DEP	semantic prototypes Po-Da MT, NER
	40.400 13 subcorpora	integrated TWOL/CG (lingsoft) + add-on	CG + PSG or DEP	WordNet based tagging
	425.000* 9 subcorpora	lexicon and rule based analyzer + CG	CG + PSG or DEP or topol.	semantic prototypes Da-En/Eo MT, NER
	8.400 3 subcorpora	lexicon and rule based analyzer + CG	CG + tree-generator	-
	16.000 3 subcorpora	integrated TWOL/CG (lingsoft) + add-on	CG + PSG	semantic prototypes (experimental)
	30.000 4 subcorpora	Decision Tree Tagger (H.Schmid & A.Stein)	CG + PSG or DEP	-
	1.000 2 subcorpora	Decision Tree Tagger (H.Schmid & A.Stein)	CG	-
	-	morpheme based analyzer + CG	CG (experimental)	Da-Esp MT


DanGram
Dan2eng
Core lexicon 100.000
Names
Composita
Valency potential
Semantic prototypes
Grammatical Translation lexicon 81.000
Fixed expressions 7.000
inflexion lex.
inflexion rules
Word-Analyzer
PoS-disambiguator
syntactic function tags
transfer CG
attaching CG
Contextual CG rules (6.000)
attachment rules (200)
Dependency trees
Lexical Transfer
lexical movement rules
Composita & Heuristics
inflexion rules (en)
Morph. Generator
Syntactic Transfer
movement rules (70)

Lexical transfer

- naïve assumption: 1 to 1, word to word
  - ‘many to one’ doesn’t hurt, ‘one to many’ does: *spalte - crack/column*
  - running lists: most prototypical translation, most robust, most frequent
- non-naïve: list of translation equivalents with discriminators for all but the first, most robust one: meget_ADV :a lot; S=(@>A) :very; D=(@>A) :much
- transfer ambiguity: *så - V-PAST:saw, V-INF:sow, ADV:so, KS:so_that, KC:thus*
- compound words (composita): recognise parts and their PoS, translate part-by-part: *oversættelsesudvalg <N:oversættelse~s+udvalg> translation committee, kapitel 4-fly -> chapter 4 plane*
  - specific translation can be chosen for prefix, suffix, first and second parts: *FN-styrke -> UN-force (not: UN-strength)*
- lexical gaps: *nænne (magte, orke), gid, mon, efterløn, grundskyld*

Polylexicals

- "words" with spaces, treated by the system as single tokens:
  for god ordens skyld, en gang imellem
- names -
  - Semantic classification: person, sted, hændelse, organisation
  - translate or not: *Peter Madsen < Den Danske Bank < Det Danske Sprog- og Litteraturselskab*
- "memory based list" (enhanced by true live translation memory)
  - Fixed terms: *Aflåst sideleje – recovery position*
  - Prepositional, nominal or adjectival phrases: *af gammel vane – from habit, bleg om næsen – green about the gills, i mands minde – in living memory*
  - Fixed expressions: *alle kneb gælder – no holds are barred*
- Variable, non-compositional expressions:
  e.g. *skøn sild* (inflects, allows lexical variation: *smuk, dejlig*)

skønne sild: sild_N :herring; D=("(dejlig | skøn | smuk)") :girl

male byen rød

male_V :paint; .... D=("by" DEF @ACC)_nil D=("rød" @OC)_nil :have=some=serious=fun

Local polysemy resolution

- word class: *så*
- number: *slægt S = family, slægt P = generation*
- gender: *rod UTR = root, rod NEU = mess*
- tense: *måtte PR = must, måtte IMPF = have_to*
- “Local” features may be instantiated through non-local relations, e.g. dependent agreement (gender or number of articles or modifiers for nouns)

Structural transfer

- rearrange, delete or insert words
  - Danish s-passive -> English analytical passive
- rearrange the order of constituents (defined as mother-daughter-chunks in dependency grammar)
  - Front field does not trigger S-V inversion in English
    I dag drikker vi vin. = Today we drink wine.

    (@ADVL|@ACC|@FS-ADVL|@FS-ACC|@>>P), w(@FMV|@FAUX|@FS-[^Q]+), w(@ICL-AUX<)?, w(@ADVL)?, (@SUBJ|@F-SUBJ|@S-SUBJ) -> 1,5,2,3,4
- fixed structures, e.g time expressions
  - *kl. 17 = (at) 17 hrs., (at) 5 o'clock*
    25. november 2005 = November 25th, 2005
- TL-constraints:
  - no "naked" negation of main verbs (do-insertion)
  - tense-inflected main verbs can't start a sentence (questions need 'do', conditionals need 'if')

Choosing the right translation: contextual polysemy discriminators

```
udsætte_V
    {opsætte} :postpone, :put=off;
    D=(@ACC) D=("for")_to :expose
    D=("(belønning|præmie)" @ACC) :offer;
    S=(INF) M=(<quant>) :criticize
    D=("vagt")_sentry :post;
    D=(<Vwater> @ACC) :put=out;
    D=("lejer" @ACC) :evict;

fylde_V
    :fill;
    D=("år" @ACC) GD=(NUM) :will=be=ACC=old;
    D=("benzin" @ACC)_nil D=("på" ADV)_nil :fill=up=the=tank;
    D=(@ACC) D=("på" ADV)_nil :pour;
    D=(@ACC) D=("på" PRP)_nil GD=(@P<)_[@P<->@MOVE] :fill=MOVE=with;
    D=(@ACC) D=("op" ADV)_up :fill;
    D=("op" ADV)_nil D!=(@ACC) :take=up=a=lot=of=room;
```

Chosing the right translation: PoS

Participiples interpreted as ADJ or N according to function:

```
        @>N, @SC = ADJ
boligsøgende_ADJ :house-hunting
        @SUBJ, @ACC, @P< = N
boligsøgende_N :house-hunter
```

```
fylde_V
    :fill;
    D=("år" @ACC) GD=(NUM) :will=be=ACC=old;
    D=("benzin" @ACC)_nil D=("på" ADV)_nil :fill=up=the=tank;
    D=(@ACC) D=("på" ADV)_nil :pour;
    D=(@ACC) D=("på" PRP)_nil GD=(@P<)_[@P<->@MOVE] :fill=MOVE=with;
    D=(@ACC) D=("op" ADV)_up :fill;
    D=("op" ADV)_nil D!=(@ACC) :take=up=a=lot=of=room;
```

best practise compromise

- 4 translations or one?

  det er slut med læger, der
  det skal være slut med læger, der
  det er slut med at køre over grænsen
  det skal være slut med at køre over grænsen

- Only "be curtain time for" can be used with all 4 constructions. The humanly optimal translation will often involve turning parts of the rest of the sentence into a subject, thus changing word order in complex ways, - not exactly ideal at the lexicon lookup stage!

- slut_N
  :end;
  P-1=("til")_in :end[IDF->DEF]; *til slut* -> *in the end*
  H=("være") B=(@S-SUBJ)_there P1=("med")_no=more P2=(N S DEF)
  _[DEF->IDF] :nil; *det er slut med kagerne* -> there are no more cakes
  D=("med" PRP)_for S=(S IDF) H=("være") :SIC-curtain=time
  det er slut med (at ..., PROP, P IDF) -> *it is curtain time for ...*

Movements

- through lexical patterns using “morphological” additions to neighbouring words
  - familie_N :family; P1=(PROP)_[+family] S=(S DEF) :SIC-the
    familien Petersen -> the Petersen family
- through lexical patterns instantiated by movement rules
  - magt_N ... H=("tage")_make=MOVE=lose=control P1=("fra")_nil
    P2=(@P<)_[@P<->@MOVE] :nil
    "Børnene tog magten fra ham." -> "The children made him lose control"
  - overhælde_V :drench; D=("med")_nil[@.*->@MOVE] :pour=MOVE=over
- through independent movement rules
  - question-like inversion in conditional subclauses
    w(@FS-ADVL), (@SUBJ|@[FS]-SUBJ) -> 'if',2,1 # *Står det til Rigspolitiet*
  - Fronting: VS -> SV (but not for formal subjects: *there is ... here is*)

(Morphological) Generation

- categories are basically the same as in Danish. There are Danish categories (definiteness, passive) that don't exist in English, but not vice-versa!
- English is an inflexion-poor, isolating language, and most morphology is regular
  - e.g. Plural: -s, rules: y -> ies, s/sh/x -> -es
  - problem: consonant doubling (verbal inflexion) is complex and depends on stress, not writing: *shopping vs. developing*
- a small inflexion dictionary is used for irregular forms
- transfer problems:
  - "semantic" number: *penge P -> money S*
  - 1-to-many translation: DEF -> 'the' + ..., s-PAS -> 'blive' + ...

Statistical perspectives

- Bilingual corpora for term extraction
  e.g. Technical: *tændrør – sparking plug*
- Dependency-annotation of large target-language corpora (1 billion words) in order to build a machine-learned database of relational probabilities:
- Choice of (1) preposition or (2) synonym based on relative n-gram and dep-gram probabilities
  appetit på Y – appetite for Y, X på Island – X in Iceland
  køre [PRP] @ACC ->
  drive [i/nil] – car,
  ride [på/nil] -bicycle,
  go [nil/by] - train
- (3) Statistical surface-smoothing for e.g. Number, definiteness and word order: *køre bil – drive **a** car*
  *Frihed**en** er dyrebar – freedom is precious*

Applications

SMS

Tale

Tale 2 Tekst

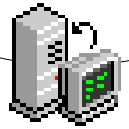
Dansk tekst

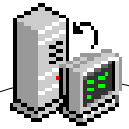
remote UI

GUI

DanGram

Dokument-integration (fx privat/firma-hjemmeside)





Informations-bearbejdning (fx DMI, QA)

Dan2eng

GUI

remote UI

English text

Text 2 Speech

Speech

SMS

Danish source text	*Dan2eng English translation*
Det er imod naturen at køre 50 km/t på tosporede veje med hegn i midten. Det mener Rådet for Større Færdselssikkerhed, som ønsker fartgrænsen hævet til 70 km/t. Bilister skal have lov til at køre 70 på indfaldsveje, hvor fartgrænsen er 50. (Kilde: Jyllandsposten/internet)	It is against nature to drive 50 km/h on two-lane roads with fences in the middle. That is what the Council for Greater Road Safety, who wishes the speed limit raised to 70 km/h, thinks. Motorists have to be allowed to drive 70 on approach roads, where the speed limit is 50.
I uge 32 åbner Tøjhusmuseet dørene for et lærerigt og morsomt show om militærets brug af heste i 1700-tallet. Her har publikum mulighed for at lære om forskellen på rytteri, dragoner og husarer, opleve heste, gamle uniformer, våben og ikke mindst noget om børneopdragelse. I forbindelse hermed, vil der hver dag i denne uge kl. 13 og 15 være opvisning af ca. 30 minutters varighed. (Kilde: Kultunaut/internet)	In week 32 Tøjhusmuseet opens its doors to an instructive and funny show about the army's use of horses in the 18th century. Here the audience has a chance to learn about the difference between cavalry, dragoons and hussars, experience horses, old uniforms, weapons and not least something about child-rearing. In connection with this, there will every day this week at 13 and 15 be a show of about 30 minutes' duration.
Skyet vejr med lidt sne af og til, og let til jævn vind fra nordøst og øst. Temperaturer mellem frysepunktet og 2 graders frost. I aften og i nat bliver vinden jævn til hård fra øst og nordøst, med sne og snefygning, men sidst på natten aftager vinden noget og der kommer kun sne af og til. Temperaturer mellem 3 og 6 graders frost. (Kilde: DMI/internet)	Cloudy weather with a little snow from time to time, and moderate to even winds from the north east and the east. Temperatures between the freezing point and 2 degrees below zero. This evening and night the wind will be medium-strong to hard from the east and the north east, with snow and snow-drifting, but late at night the wind abates a little and there occurs only snow from time to time. Temperatures between 3 and 6 degrees below zero.
Uddannelse kan sikre job En lang række danske virksomheder flytter en del af deres	Education can secure jobs A long row of Danish companies moves a part of their

Inventory: Linguistic tools

- Full morphological taggers and syntactic parsers for 6-8 languages
  da, de, en, eo, es, fr, it, pt
- Large monolingual lexica with valency and some semantic information
- NER for Danish and Portuguese
- Annotated corpora for 8 languages
- Error mapping grammar for Danish

Inventory: applications

- MT for Da-En, Po-Da, Da-Eo, No-Da, No-En
- Spell-/Grammarchecking for Danish (OrdRet)
- Interactive teaching tools and games (VISL)
- Visualisation tools for text features (TextPainter/WebPainter)
- cgi/php-interfacing, web interface
- Some experiments: QA

Demo- and development sites

beta.visl.sdu.dk
grammarsoft.com
visl.dk