Using Danish as a CG Interlingua: A Wide-Coverage Norwegian-English Machine Translation System

Eckhard Bick
Institute of Language and Communication
University of Southern Denmark
Odense, Denmark
eckhard.bick@mail.dk

Lars Nygaard
The Text Laboratory
University of Oslo
Oslo, Norway
lars.nygaard@iln.uio.no

Abstract

This paper presents a rule-based Norwegian-English MT system. Exploiting the closeness of Norwegian and Danish, and the existence of a well-performing Danish-English system, Danish is used as an «interlingua». Structural analysis and polysemy resolution are based on Constraint Grammar (CG) function tags and dependency structures. We describe the semiautomatic construction of the necessary Norwegian-Danish dictionary and evaluate the method used as well as the coverage of the lexicon.

1 Introduction

Machine translation (MT) is no longer an unpractical science. Especially the advent of corpora with hundreds of millions of words and advanced machine learning techniques, bilingual electronic data and advanced machine learning techniques have fueled a torrent of MT-project for a large number of language pairs. However, the potentially most powerful, deep rule-based approaches still struggle, for most languages, with a serious coverage problem when used on running, mixed domain text. Also, some languages, like English, German and Japanese, are more equal than others, not least in a funding-heavy environment like MT.

The focus of this paper will be threefold: Firstly, the system presented here is targeting one of the small, «unequal» languages, Norwegian. Secondly, the method used to create a Norwegian-English translator, is ressource-economical in that it uses another, very similar language, Danish, as an «interlingua» in the sense of translation knowledge recycling (Paul 2001), but with the recycling step at the SL side rather than the TL side. Thirdly, we will discuss an unusual analysis and transfer methodology based on Constraint Grammar dependency parsing. In short, we set out to construct a Norwegian-English MT system by building a smaller, Norwegian-Danish one and piping its output into an existing Danish deep parser (DanGram, Bick 2003) and an existing, robust Danish-English MT system (Dan2Eng, Bick 2006 and 2007).

2 The MT system

The Bokmål standard variety of Norwegian is a language historically so close to Danish, that speakers of one language can understand texts in the

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 21–28

other without prior training - though the same does not necessarily hold for the spoken varieties. It is therefore a less challenging task to create a Norwegian-Danish MT system than a Norwegian-English or even Norwegian-Japanese one. Furthermore, syntactic differences are so few, that lexical transfer can to a large degree be handled at the word level with only part of speech (PoS) disambiguation and no syntactic disambiguation, allowing us to depend on the Danish parser to provide a deep structural analysis. Furthermore, the polysemy spectrum of many Bokmål words closely matches the semantics of the corresponding Danish word, so different English translation equivalents can be chosen using Danish context-based discriminators.

2.1 Norwegian analysis

As a first step of analysis, we use the Oslo-Bergen Tagger (Hagen et al. 2000) to provide lemma disambiguation and PoS tagging, the idea being to translate results into Danish, using a large bilingual lexicon, and feed them into the syntactic and dependency stages of the DanGram parser. However, though both the OBT tagger and DanGram adhere to the Constraint Grammar (CG) formalism (Karlsson 1990), a number of descriptive compatibility issues had to be addressed. Since categories could not always be mapped one-to-one, we had to also use the otherwise to-be-skipped syntactic stage of the OBT tagger in order to further disambiguate a word's part of speech. Thus, the Danish preposition-adverb distinction is underspecified in the Norwegian system where the 2 lexemes have the same form, using the preposition tag even without the presence of a pp. The same holds for about 50 words that in Danish are regarded as unambiguous adverbs, but in Norwegian as unambiguous prepositions.

2.2 The Norwegian-Danish lexicon

The complexity of a Norwegian-Danish dictionary can be compared to Spanish-Catalan language pair addressed in the open source Apertium MT project (Corbí-Bellot et al. 2005), where a 1-to-1 lexicon was deemed sufficient (with a few polysemous cases handled as multi-word expressions), avoiding the disambiguation complexity of many-to-many lexica necessary for less-related languages. Even without extensive polysemy mismatches, the productive compounding nature of Scandinavian languages, however, increases lexical complexity as compared to Romance languages - an issue reflected in the transfer evaluation in chapter 2.3.

In a project with virtually zero funding, like ours, it can be difficult to build or buy a lexicon, not to mention the general lack of wide-coverage Norwegian-Danish electronic lexica to begin with. So with only a few thousand words from terminology lists or the like available, creative methods had to be employed, and we opted for a bootstrapping system with the following steps:

(a) Create a large corpus of monolingual - Norwegian text and lemmatize it automatically. Quality was less important in this step, since frequency measures could be employed to weed out errors and create a candidate list of Norwegian lemmas.

(b) Regard Norwegian as misspelled Danish, and run a Danish spell checker on the lemma-list obtained from (a). Assume translation as identical, if the Norwegian word is accepted by a Danish spell checker. Use correction suggestions by spell checkers as translations suggestions. Because differences could be greater than Levenshtein distance 1 or 2, a special, CG-based spell checker (OrdRet, Bick 2006) was used, with a particular focus on heavy, dyslexic spelling deviations

and a mixed graphical-phonetic approach.

(c) Produce phonetic transmutation rules for Norwegian and Danish spelling to generate hypothetical Danish words from Norwegian candidates, and than check if a word of the relevant word class was listed in either DanGram's parsing lexicon or its spell checker fullform list.

Methods (a-c) resulted in a list of 226,000 lemmas with translations candidates in Danish. Only 20,000 low-frequency words were completely unmatchable. In a first round of manual revision, all closed-class words, all polylexical matches were checked, and a confidence value from DanGram's spell checker module was used to grade suggestions into safe, unsafe and none. Next, a compound analyzer was written and run on all Norwegian words, accepting compound splits as likely if the resulting parts both individually existed in the word list, finally creating a Danish translation from the translations of the parts, and checking it and its epenthetic letters against the Danish lexicon. This step not only helped to fill in remaining blanks, but was also used to corroborate spell checker suggestions as correct, if they matched the translation produced by compound analysis - or replace them, if not. After this, 13.800 lemmas had no translation, 23.500 lemmas were left with an «unsafe» marking from the spell checker stage, and in 20.700 cases, compound analysis contradicted spell checker or list suggestions otherwise deemed safe. Allowing overrides in the latter case, and removing the two former cases, we were left with a bilingual lemma list of 188.500 entries.

Finally, a dual pass of manual checking was directed at all items with a frequency count over 10, corresponding to about 12.5%. In obvious cases, related low-frequency words in neighbouring positions on the alphabetical list were corrected at the same time.

In order to evaluate our method of lexicon-creation, we extracted all words with frequency 9 - the most frequent group without prior manual revision - and inspected all suggested translations (1544 cases).

type	***n***	***%***
non-word	33	2.1 %
wrong PoS	8	0.5 %
etymology =	161	10.4 %
transparent[1]	6	0.4
intransparent[2]	20	1.3 %
all corrected	**187**	**12.1 %**
all	228	14.8 %

Table 1

As can be seen from table 1, ignoring the 2.6 % of non-words from the corpus-based lemma-list, about 12% of the unrevised translations were wrong. However, in most of these cases (10.4%, over 4/5), the Danish translations were still etymologically - and thus spelling-wise - related to their Norwegian

[1] brise (blæse), spenntak (spændloft), stabbe, strupetak (strubelåg), villastrøk (villakvarter), vårluft (forårsluft)

[2] guttete (drengete), havert (slags sæl), hengemyr (hængedynd), kraftsektor (energisektor), koring, kvinneyrke, langdryg, låtskriver, lønnsnemnd, malingflekk, omvisning (rundvisning), purke (so), sauebonde, smokk (sut), strikkegenser, søppelbøtte (affaldsbøtte), tukle (fumle), tøyelig (fleksibel), vassdrag (vandløb), yrkesutdanning

counterparts, and should thus be accessible to improved automatic matching-techniques.

frequency	***non-word***	***wrong PoS***	***corrected***
9 (all)	2.1 %	0.5 %	12.1 %
5	0.5 %	1.5 %	14 %
4	4 %	0.5 %	14 %
3	1 %	0.5 %	10.5 %
2	4 %	3 %	9.5 %
1	3.5 %	0.5 %	8.5 %
average	2.5 %	1.1 %	11.4 %

Table 2

Small checks were also conducted for other frequencies (200 words each), randomly extracting 1 out of 10 words. Results indicate that automatic translatability remains similar in general, though
there was a slight correlation between falling frequency and *less* need for correction. The proportion of non-words was high for low frequencies, possibly reflecting spelling errors and analysis problems with rare words in the corpus data. However, since having non-existing words in the SL-list, is only «noise» and not a problem for the MT system, we conclude from their translatability that low-frequency words are at least as safe a contribution to the lexicon as high-frequency words.

2.3 Norwegian-Danish transfer

Analysed input from the Oslo-Bergen-tagger is danified by substituting Danish base forms for Norwegian ones. Even with an extensive bilingual word list, the transfer program is not, however, a mere lookup procedure. Due to the compounding structure of the languages involved, compound analysis has to be performed both on the Norwegian and the Danish side - the former to achieve a part-by-part translation for words not listed in the bilingual lexicon, the latter to permit assignment of secondary *Danish* information (valency, semantics) to Danish translations not covered by the DanGram monolingual lexicon.

The Norwegian-Danish transfer module was evaluated on 1,000 mixed-genre sentences from the Norwegian web part of the Leipzig Corpora Collection[3] and a 6.500 word chunk from the ECIcorpus[4].

	Web	***Litterature***
words	15,641	6,521
N, ADJ, V, ADV	8,976 (57.4%)	3,098 (47.5%)
not in noda-lex	991 (6.3%)	182 (2.8%)
compounds	458 (2.9%)	78 (1.2%)
not in dan-lex	127 (0.8%)	32 (0.5%)

Table 3

The failure rate for Norwegian words was 6.3% in the web corpus, in part compensated by the fact that almost half of these (2.9%) could still be compound-analyzed. The coverage rate of the Danish lexicon was very high - only 0.8% of suggested translations were not found. Figures for the literature corpus were almost twice as good - even when taking into account that the percentage of open-class inflecting words was 10 percentage points lower in this corpus.

2.4 Danish generation

Finally, Danish full-forms are generated from the translated base-forms, based

[3] http://corpora.uni-leipzig.de
[4] European Corpus Initiative, http://www.elsnet.org/resources/eciCorpus.html

both on the filtered OBT morphological tag string, and inflexional information from the Danish lexicon.

"*[hus] N NEU S DEF GEN*", for instance, will be inflected as *hus -> NEU DEF huset -> GEN husets.* Irregular forms are stored in full in a separate file, and compound stems are constructed, prior to inflexion, using rules for the insertion of epenthetic s or epenthetic e.

(1) *agurk+tid -> agurketid*
(2) forbud+stat -> forbudsstat

Alas, Danish and Norwegian morphology are not completely isomorphic, and in order to handle differences in a context-dependent way, a special CG grammar is run before generation. This grammar handles, for instance, the Norwegian phenomenon of double definiteness:

(3) NOR: ***den*** *store bil****en*** *-> DAN:* ***den*** *store bil*

Here, so-called substitution rules are used, replacing the tag DEF with IDF in the presence of definite articles (example below) or pre- or post-positioned determiners and attributes (syntactic tags @<ADJ, @<DET, @ADJ>, @DET>):

```
SUBSTITUTE (DEF) (IDF) TARGET (N)
IF (*-1 ART BARRIER NON-PRE-N/ADV) ;
```

2.5 Structural analysis

Syntactic-functional analysis was based not on the Norwegian OBT-analysis, but on a from-scratch analysis of the translated Danish text, in part because of the high syntactic accuracy of the Danish parser (Bick 2000), in part to ensure compatibility with the descriptive conventions used in the next syntactic stage, dependency analysis, and the Danish-English MT system itself. The Dependency grammar in question (described in Bick 2005) consists of a few hundred rules targeting CG function tags, supported by attachment direction markers and close/long-attachment markers from a special CG layer run as a last step before dependency.

2.6 Danish-English transfer

Though the Danish-English MT system (Dan2eng, fBick 2007) is not the focus of this paper, and used *as is* in a black box fashion, a short description is in order - not least because of the perspective of ultimately creating a similar system for direct Norwegian-English transfer.

The core principle of Dan2eng is to rely as much as possible on deep and accurate SL analysis. In this spirit, the selection of translation equivalents is based on lexical transfer rules exploiting *syntactic* relations in a semanticised way. The way in which Dan2eng semanticizes syntax, differs significantly from many older rule-based MT systems designed in the 80's and 90's. First, it uses dependency rather than constituent analyses, and second, it is the first MT system ever to be based on Constraint Grammar, a combination that provides it with a robust way of progressing from shallow to deep analyses (Bick 2005) without the high percentage of parse failures inherent to many generative systems when run on free text[5].

As an example, let us have a look at the translation spectrum Danish verb *at regne (to rain),* which has many other, non-meteorological, meanings *(calculate, consider, expect, convert ...)* as well. Here, Dan2eng simply uses *grammatical distinctors* to *distinguish* between translations, rather than *define* sub-senses.

[5] Even today, MT systems using deep syntax, may find it cautious to restrict their domain or structural scope, like the LFG- and HPSG-based LOGON system (Lønning et al. 2004).

Thus, the translation *rain (a)* is chosen if a daughter/dependent (D) exists with the function of situative/formal subject (@S-SUBJ), while most other meanings ask for a human subject. As a default[6] translation for the latter *calculate (f)* is chosen, but the presence of other dependents (objects or particles) may trigger other translations. *regne med (c-e),* for instance, will mean *include,* if *med* has been identified as an adverb, while the preposition *med* triggers the translations *count on* for human «granddaughter» dependents (GD = <H>), and *expect* otherwise. Note that the *include* translation also could have been conditioned by the presence of an object (D = @ACC), but would then have to be differentiated from (b), *regne for ('consider').*

regne_V[7]
(a) D=(@S-SUBJ) :rain;
(b) D=(<H> @ACC) D=("for" PRP)_nil :consider;
(c) D=("med" PRP)_on GD=(<H>) :count;
(d) D=("med" PRP)_on :expect;
(e) D=(@ACC) D=("med" ADV)_nil :include;
(f) D=(<H> @SUBJ) D?=("på" PRP)_nil :calculate;

The example shows how information from different descriptive layers is integrated in the transfer rules. Structural conditions may either be expressed in n-gram fashion (with P+n or P-n) positions, or dependency fashion (reference to daughters, mothers, granddaughters and grandmothers independent of distance). Semantic conditions can either be inferred with regular expressions from word or base forms, or exploit DanGram's semantic prototype tags in a systematic way, e.g. <tool>, <container>, <food>, <Hprof> etc. for nouns (160 types in all). Adjectives and verbs have fewer classes (e.g. psychological adjective, move, speech or cognitive verbs), but make up for this with a rich annotation of argument/valency tags.

The rule-based transfer system is supplemented by a dictionary of fixed expressions and a (so far sentence-based) translation memory. The Danish-English bilingual lexicon was built to match the coverage of the DanGram lexicon (100.000 words plus 40.000 names), but does not yet have the same coverage for compounds. In any case, compounds are productive, and therefore covered by a special back-up module that combines part-translations, affix-translations. Rules may be used to force a different translation for a lexeme if used as first or second part in compounds, e.g. *FN-styrke,* where *styrke* should be '*force*', not '*strength*'. The compound module is doubly important for our Nor2eng interlingua approach, since secondary Danish lookup-failures may be caused by Norwegian lookup-failures.

2.7 English generation and syntax

English generation is handled much like Danish generation, drawing on CG morphological tags, a lexicon of irregular forms and some phonetic/stress heuristics to inflect translated base forms - again supported by a special CG layer performing systematic substitutions (for instance plural translations of singular words) and insertions (certain modals, or articles). Differences in syntax are handled by successive transformation rules, which may move either words or whole dependency tree sections if

[6] The ordering of differentiator-translation pairs is important - readings with fewer restrictions have to come last. The example lacks the general, differentiator-free default provided with all real lexicon entries.
[7] The full list of differentiators for this verb contains 13 cases, including several prepositional complements not included here *(regne efter, blandt, fra, om, sammen, ud, fejl ...)*

certain tags, tokens or sequences are found.

In the following example, two movement rules were applied. The first changes the Scandinavian VS order into SV after a filled front field, placing the fronted adverbial between S and V. The other rule, classifying the adverbial, decides on a better place for it - between auxiliary and main verb.

NOR: *På 1980-tallet ble sammenhengen mellom sosiale faktorer og helse i stor grad avskrevet.*

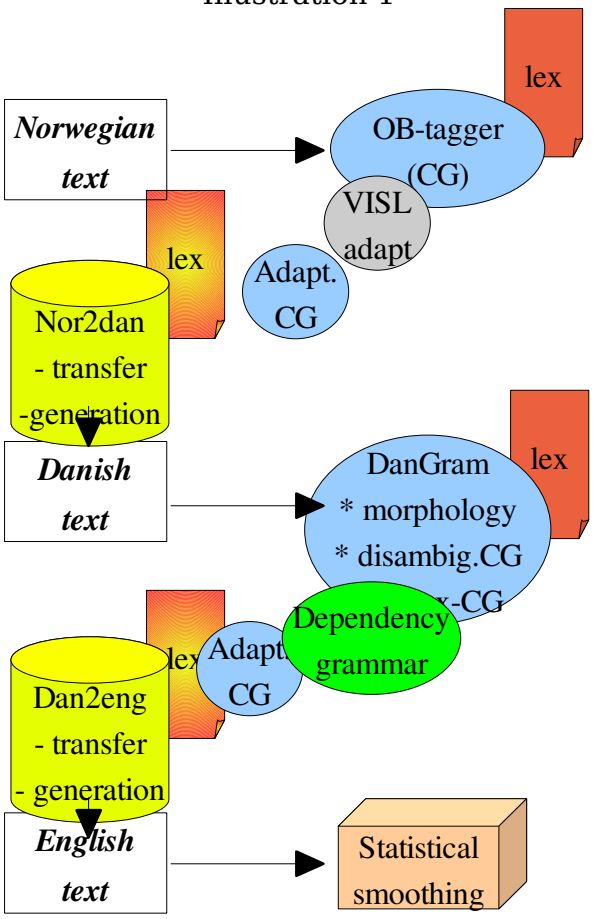
DAN:

I PRP @ADVL #1->13
1980'erne N @P< #2->1
blev V @STA #3->0
sammenhængen N @SUBJ #4->3
mellem PRP @N< #5->4
sociale ADJ @>N #6->7
faktorer <cjt1> N @P< #7->5
og KC @CO #8->7
helse <cjt2> N @P< #9->7
i PRP @ADVL #10->13
stor ADJ @>N #11->12
grad N @P< #12->10
afskrevet V @AUX< #13->3.

ENG: *In the 1980s the connexion between social factors and health was largely written off.*

Note also the fact, that the preposition change is a difference between Norwegian and Danish, not between Danish and English, and that the subject movement acted on the *whole* NP, including its dependent PP, which again contained a coordination. The necessary dependency links are marked in the Danish interlingua sentence.

Illustration 1



3 Perspectives: Statistical smoothing

In spite of the fact that Dan2Eng employs tens of thousands of hand-written lexical transfer rules, it is extremely difficult to cover all idiosyncrasies of, for instance, preposition usage or choice of synonym in a rule based way. Furthermore, mismatches are more likely when chaining two translations. On the other hand, statistical methods allow to check the probabilities of rule-suggested

translations in a given context, smoothing out translational rough spots. Given the lack of large bilingual Norwegian-Danish or Norwegian-English corpora, it is an added advantage, that such methods work with *monolingual,* target language corpora - of which there are almost unlimited amounts availabe in the case of English. To prepare for an integration of TL smoothing, we performed dependency annotation of 1 billion words, and started extracting n-gram information as well as what we call dep-grams - hierarchical chains of dependency-linked words, the former with the perspective of preposition-smoothing, the latter for argument-smoothing.

Future evaluations, to be conducted after a more complete revision of the Norwegian bilingual lexicon and the construction of a polysemy-sensitive Norwegian-Danish transfer grammar, will have to address not only the overall quality of the MT system as a whole - optimally in comparison with other systems, like LOGON (Lønning et al. 2004) -, but also the relative contributions of rule based and statistical modules.

References

Bick, Eckhard. 2001. «En Constraint Grammar Parser for Dansk», in Peter Widell & Mette Kunøe (eds.), *8. Møde om Udforskningen af Dansk Sprog, 12.-13. oktober 2000*, pp. 40-50, Århus University

Bick, Eckhard. 2003, «A CG & PSG Hybrid Approach to Automatic Corpus Annotation», In: Kiril Simow & Petya Osenova (eds.), *Proceedings of SProLaC2003* (at Corpus Linguistics 2003, Lancaster), pp. 1-12

Bick, Eckhard. 2005 «Turning Constraint Grammar Data into Running Dependency Treebanks», In: Civit, Montserrat & Kübler, Sandra & Martí, Ma. Antònia (red.), *Proceedings of TLT 2005 (4th Workshop on Treebanks and Linguistic Theory, Barcelona, December 9th - 10th, 2005)*, pp.19-27

Bick, Eckhard. 2006. «A Constraint Grammar Based Spellchecker for Danish with a Special Focus on Dyslexics». In: Suominen, Mickael et al. (ed.) *A Man of Measure: Festschrift in Honour of Fred Karlsson on his 60th Birthday*. Special Supplement to SKY Jounal of Linguistics, Vol. 19 (ISSN 1796-279X), pp. 387-396. Turku: The Linguistic Association of Finland

Bick, Eckhard. 2007. «Fra syntaks til semantik: Polysemiresolution igennem dependensstrukturer i dansk-engelsk maskinoversættelse.» (forthcoming)

Corbí-Bello, Antonio M. et al. 2005. An open-source shallow-transfer machine translation engine for the Romance Languages of Spain. In *Proceedings of the European Association for Machine Translation, 10th Annual Conference*, Budapest 2005, p. 79-86.

Hagen, Kristin, Johannessen, Janne Bondi, Nøklestad, Anders. 2000. "A Constraint-Based Tagger for Norwegian". In: Lindberg, C.-E. and Lund, S.N. (red.): 17th Scandinavian Conference of Linguistic, Odense. Odense Working Papers in Language and Communication, No. 19, vol I.

Karlsson, Fred. 1990. Constraint Grammar as a Framework for Parsing Running Text. In: Karlgren, Hans (ed.), *COLING-90 Helsinki: Proceedings of the 13th International*. Conference on Computational Linguistics, Vol. 3, pp. 168-173

Lønning, Jan Tore, Stephan Oepen, Dorothee Beermann, Lars Hellan, John Carroll, Helge Dyvik, Dan Flickinger, Janne Bondi Johannessen, Paul Meurer, Torbjørn Nordgård, Victoria Rosén, and Erik Velldal. 2004. LOGON. A Norwegian MT effort. In *Proceedings of the Workshop in Recent Advances in Scandinavian Machine Translation*, Uppsala, Sweden,

Paul, Michael. 2001. Knowledge Recycling for Related Languages. *Proceedings of MT Summit VIII.* Santiago de Compostela, Spain. pp. 265-269.