The Fyntour Multilingual Weather and Sea Dialogue System

Eckhard Bick
University of Southern Denmark
Odense
echard.bick@mail.dk

Jens Ahlmann Hansen
University of Southern Denmark
Odense
ahlmann@voicetech.dk

1 Introduction

The Fyntour multilingual weather and sea dialogue system provides pervasive access to weather, wind and water conditions for domestic and international tourists who come to fish for seatrout along the coasts of the Danish island of Funen. Callers access information about high and low waters, wind direction etc. via spoken dialogues in Danish, English or German. We describe the solutions we have implemented to deal with number format data in a multi-language environment. We also show how the translation of free text 24-hour forecasts from Danish to English is handled through a newly developed machine translation system. In contrast with most current, statistically-based MT systems, we make use of a rule-based apporach, exploiting a full parser and context-senstitive lexical transfer rules, as well as target language generation and movement rules.

2 Number Format Data

The Fyntour system provides information in Danish, English and German. A substantial amount of data is received and handled in an interlingua format, i.e. data showing wind speed (in m/s) and precipitation (in mm) are language-neutral numbers which are simply converted into language-specific pronunciations by specifying the locale of the speech synthesis in the VoiceXML , e.g.

<prompt xml:lang="da-DK"> 1 </prompt> ”en”
<prompt xml:lang="de-DE"> 1 </prompt> ”ein”
<prompt xml:lang="en-GB"> 1 </prompt> ”one”

In Germany, wind speed is normally measured using the Beaufort scale (vs. the Danish m/s norm), while visitors from English speaking countries are accustomed to the 12-hour clock (vs. the continental European 24-hour clock). These cultural preferences can be catered for by straightforward conversions of the shared number format data – performed by the application logic generating the dynamic VXML output of the individual languages.

However, the translation of dynamic data in a free text format, from Danish to English and Danish to German, – such as the above-mentioned forecasts, written in Danish by different meteorologists – is more complex. In the Fyntour system, the Danish-English translation problem has been solved by a newly developed machine translation (MT) system. The Constraint Grammar based MT-system, which is rule-based as opposed to most existing, probabilistic systems, is introduced below.

3 CG-based MT System

The Danish-English MT module, Dan2eng, is a robust system with a broad-coverage lexicon and grammar, which in principle will translate unrestricted Danish text or transcribed speech without strict limitations to genre, topic or style. However, a small benchmark corpus of weather forecasts was used to tune the system to this domain and to avoid lexical or structural translation gaps, especially concerning time and measure expressions, as well as certain geographical references and names.

Methodologically, the system is rule-based rather than statistical and uses a lexical transfer approach with a strong emphasis on source language (SL) analysis, provided by a pre-existing Constraint Grammar (CG) parser for Danish, DanGram (Bick 2001). Contextual rules are used at 5 levels:

1. CG rules handling morphological disambiguation and the mapping of syntactic func-

tions for Danish (approximately 6.000 rules)

2. Dependency rules establishing syntactic-semantic links between words or multi-word expressions (220 rules)
3. Lexical transfer rules selecting translation equivalents depending on grammatical categories, dependencies and other structural context (16.540 rules)
4. Generation rules for inflexion, verb chains, compounding etc. (about 700 rules)
5. Syntactic movement rules turning Danish into English word order and handling subclauses, negations, questions etc. (65 rules)

At all levels, CG rules may be exploited to add or alter grammatical tags that will trigger or facilitate other types of rules.

As an example, let us have a look at the translation spectrum of the weatherwise tedious, but linguistically interesting, Danish verb *at regne (to rain),* which has many other, non-meteorological, meanings *(calculate, consider, expect, convert ...)* as well. Rather than ignoring such ambiguity and build a narrow weather forecast MT system or, on the other hand, strive to make an "AI" module *understand* these meanings in terms of world knowledge, Dan2eng chooses a pragmatic middle ground where grammatical tags and grammatical context are used as *differentiators* for possible translation equivalents, staying close to the (robust) SL analysis. 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[1] 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[2]
(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)_nil :expect;
(e) D=(@ACC) D=("med" ADV)_nil :include;
(f) D=(<H> @SUBJ) D?=("på")_nil :calculate;

It must be stressed that the use of grammatical relations as translation differentiators is very different from a simple memory based approach, where chains of words are matched from parallel corpora. First, the latter approach - at least in its naïve, lexicon-free version - cannot generalize over semantic prototypes (e.g. <H> for human) or syntactic functions, conjuring up the problem of sparse data. Second, simple collocation, or co-occurrence, is much less robust than functional dependency relations that will allow interfering material such as modifiers or sub-clauses, as well as inflexional or lexical variation.

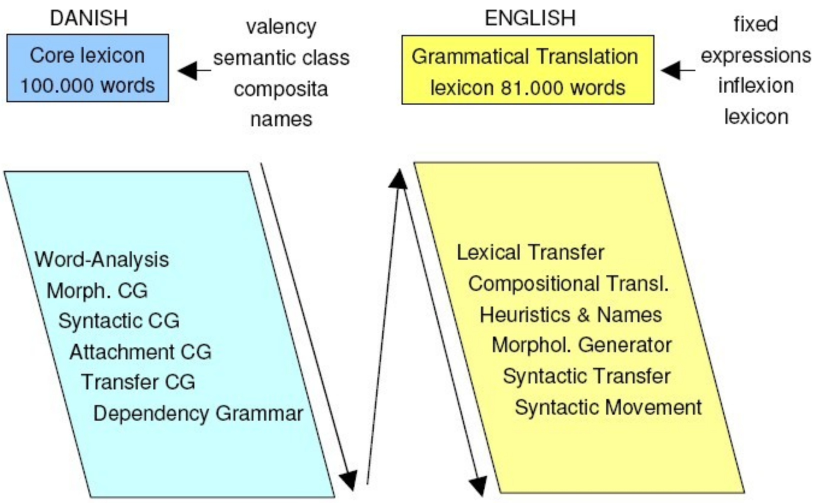

Fig 1: The Dan2eng system

For more details on the Dan2eng MT system, see http://beta.visl.sdu.dk/ (demo, documentation, NLP papers).

[1] The ordering of differentiator-translation pairs is important - defaults, with fewer restrictions, have to come last. For the numerical value of a given translation, 1/rank is used.

[2] 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 ...)*