# A Named Entity Recognizer for Danish

**Eckhard Bick**

University of Southern Denmark
Rugbjergvej 98, DK-8260 Viby J
eckhard.bick@mail.dk

**Abstract**

This paper describes how a preexisting Constraint Grammar based parser for Danish (DanGram, Bick 2002) has been adapted and semantically enhanced in order to accommodate for named entity recognition (NER), using rule based and lexical, rather than probabilistic methodology. The project is part of a multi-lingual Nordic initiative, Nomen Nescio, which targets 6 primary name types (human, organisation, place, event, title/semantic product and brand/object). Training data, examples and statistical text data specifics were taken from the Korpus90/2000 annotation initiative (Bick 2003-1).

The NER task is addressed following the progressive multi-level parsing architecture of DanGram, delegating different NER-subtasks to different specialised levels. Thus named entities are successively treated as first strings, words, types, and then as contextual units at the morphological, syntactic and semantic levels, consecutively. While lower levels mainly use pattern matching tools, the higher levels make increasing use of context based Constraint Grammar rules on the one hand, and lexical information, both morphological and semantic, on the other hand. Levels are implemented as a sequential chain of Perl-programs and CG-grammars.

Two evaluation runs on Korpus90/2000 data showed about 2% chunking errors and false positive/false negative proper noun readings (originating at the lower levels), while the NER-typer as such had a 5% error rate with 0.1 - 0.5% remaining ambiguity, if measured only for correctly chunked proper nouns.

## 1. Introduction

This paper describes how a preexisting Constraint Grammar (CG) based parser for Danish (DanGram, Bick 2001) has been adapted and semantically enhanced in order to accommodate for named entity recognition (NER), using rule based and lexical, rather than probabilistic methodology. The project was part of a recently concluded multi-lingual 2-year initiative, Nomen Nescio, which was supported by, and documented for, the Nordic Research Academy NorFA (Bick 2003). Nomen Nescio targets 6 primary name types (human, organisation, place, event, title/semantic product and brand/object), that for Danish were subdivided into 20 subtypes, allowing for systematic ambiguity and the treatment of atomic semantic features. Training data, examples and statistical text data specifics were taken from the Korpus90/2000 annotation initiative, a joint venture involving the Danish Society of Litterature and Language (corpus compilation, http://www.dsl.dk) and the VISL project at the University of Southern Denmark (grammatical tagging, http://beta.visl.sdu.dk). Current applicative uses of DanGrams NER-system include a commercial Question-Answering (QA) system and enrichment of annotated corpora (about 60 million words, http://corp.hum.sdu.dk). Part of the corpus material has been manually revised, and because DanGram also annotates syntactic function, the corpus can be used to extract syntactic patterns for individual name types (for instance, subject-object preference for name classes in conjunction with key verbs).

## 2. NER as a distributed parsing task

Since DanGram in its basic structure is based on progressive multi-level parsing, it was a natural step to delegate NER-tasks to different specialised levels, too, thus treating named entities successively as first strings, words, types, and then as contextual units at the morphological, syntactic and semantic levels, consecutively.

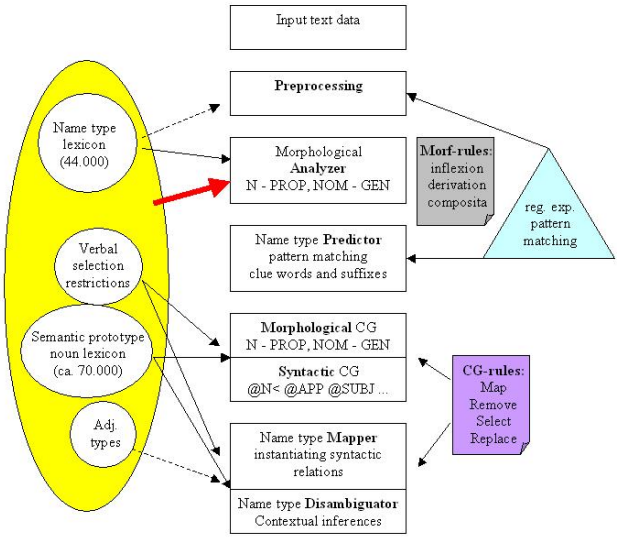


**Figure 1**

Levels are implemented as a sequential chain of Perl-programs and CG-grammars. While lower levels mainly use pattern matching tools, the higher levels make increasing use of context based Constraint Grammar rules on the one hand, and lexical information, both morphological and semantic, on the other hand. The morphological anlayzer uses a regular name lexicon of ca. 45.000 entries, while the pattern-matching levels use extensive affix and key-word lists for the classification of complex name chains (including addresses, web-urls, version numbers, book titles). Dangram's ordinary CG modules, adding word based morphosyntactic information, interact with the NER CG-modules in two ways: First, the inherent robustness of the Constraint Grammar method will supply later name-rules with reliably analyzed context. Thus, for newspaper text, Dangram's F-scores approach 99% for word class (PoS categories), and 95% for syntactic function/dependency (e.g. subject, direct object, relative clause). Second, since

Dangram's noun lexicon (ca. 75.000 lexemes) contains semantic prototype information and semantic feature typing (e.g <Hprof> for human professionals, <Aorn> for birds, <Lh> for human-created places), higher-level CG-rules can instantiate NER-classes from noun classes, once syntactic functions like postnominal, apposition, subject complement etc. are known. Semantic class can even be derived from information embedded in relative clauses, or be projected from verbs' selection restrictions onto name subjects and objects.

## 2.1. NE string recognition at raw text level

A first preprocessing task in name recognition is the distinction between names and sentence initial capitalisation. For instance, upper-case highlighting has to be distinguished from name abbreviations (UNPROFOR, USA). The NE string recognition module itself, a part of the general tokenizer, uses regular expressions to identify polylexicals, creating "words" by substituting in-word blanks with equal signs. Here, one guiding NER principle is, of course, to fuse consecutive uppercase-initial strings into name units (Nyrup=Rasmussen, Australian=Open).

For personal names, the pattern recognizer checks for a list of international in-name prepositions, articles and co-ordinators (*Maria dos Santos, Ras al Kafj).* Similarly, certain international chaining particles are accepted in other name types, covering e.g. organisation names (Dansk Selskab for Akupunktur) or place names (Place de la Concorde). Once recognised, name chains are allowed to grow to the right. Thus, 'Den/Det' in mid-sentence is a safe name-initialiser, even with an intervening lower case word: *Det ny Lademann Aps.*

A special task is classifying punctuation characters as either token separators or parts of name chains, - as in initials, web-urls, abbreviations or book titles, and sometimes, creative brand names (Carl Th. Philipsen, cand. polit., H.M.S. Polaris, Nr.=Nissum). Similarly, numerals appear, for instance, in version numbers and car names (Honda Civic 1,4 GL sedan).

Not all NE chain fusion can be done with pattern matching alone, and lexical context information may be necessary. Thus, 'for' is a name part only in conjunction with 'selskab', 'organisation' etc. Here the recognizer needs lexical knowledge at the simplest level, in the form of an "allowed word list", optionally with an inflexional addition like (en|et|ern?e?)?. At a higher level, the full name lexicon is used, and name chain candidates are subjected to lexicon sensitive splitting rules. Thus, in the case of genitive candidates, split halves are lexicon checked individually, and splitting is allowed if the first part is a known person, organisation, institution or political unit (i.e. humanoids): Sonofons <org> # GSM 900-net, Richard Strauss' <hum> # Zarathustra, New Yorks <civ> # Empire State Building.

## 2.2. NE recognition at the morphological level

The morphological analyzer, *dantag*, uses – besides compositional rules - lexicon name data in two ways:

(a) full lexicon entries (especially major place names and Christian names, but also a number of common surnames and company names, ca. 45.000 entries), semantically subclassified into 20 classes.

(b) partial lexicon entries, used by the lexical analyser prior to disambiguation, in order to heuristically assess unknown multi-word names with a known first, second or third part (e.g. known Christian names with unknown surnames, or known company names with geographical extensions: *Toshiba Denmark*).

This lexicon module introduces both cross word class ambiguity, e.g. *Otte, Hans* in sentence initial position, and name class internal ambiguity, e.g. *Lund* (place/person), *Audi* (company/vehicle) etc. The latter is not necessarily lexicon-registered, if it is systematic, like in the use of town names for sports teams. The category <media> expresses a systematic ambiguity title/organisation *(Jyllandsposten, TV2).*

Upper case words with no entry in the name lexicon will first be checked as lower case against the ordinary lexicon, using full inflexional and derivational analysis. Second, a compositional reading is tried, cutting the word in parts and matching the first against the name lexicon and the rest against the ordinary and suffix lexicon. The resulting analysis draws its word class tag from the word class of the final part of the composition, eg. nouns (ANC-kontor, G8-mødet, Marsåret) or adjectives (EU-godkendt, Heisenberg'ske). Similarly, names may be treated as nouns in inflected disguises (EMSen, EMS'en). Heuristic name readings are assigned to upper case words, if all else fails, or if all non-name readings are "unsafe" (compounds, sentence initial).

## 2.3. Semi-lexical NE type prediction (patterns)

This semantic type predictor uses compositional heuristics. The program respects safe, i.e. fully lexicon based subtype readings from the morphological analyser and tries to confirm lexicon based half-guesses (e.g. known name type as first part of a name chain). In all other cases, the type predictor tries to instantiate morphological (derivational) and otherwise pattern based type clues for the different categories, or even to assign negative markers, as <non-hum> for name strings incorporating non-alphabetic characters of function words. Sometimes, 2 readings will be assigned for later disambiguation (e.g. <civ> [civitas] and <hum> for certain place suffixes).

As a general example, the <org> class uses patterns like the following: in-word capitals *(MediaSoft),* "suffixes": *Amba, GmbH, A/S, AG, Bros., & Co ...*, type indicators: *=Holding, =Organisation, =Society, =Network, Banco=d[eiao], K/S, I/S, Klub=, Fonden=,* morphological indicators: *-com, -ex, -rama, -tech, -soft*

One problem are interferences between the different sections of the type predictor. For instance, place names can be part of club names, and human names can be part of prize titles. This is why ordering the sections is important, or even running parts of a certain type predictor disjunctly or iteratively. Also, some of the pattern-lists have NOT-conditions quoting partial or overlapping patterns that would indicate other semantic name classes.

## 2.4. NE word class and case disambiguation (CG)

This module is a full-fledged cg-grammar with 3.300 sentence-wide context sensitive rules. It has access to word class, inflexion, verbal and nominal valency potential, semantic class etc. The idea is to disambiguate the morphological readings suggested by the analyzer at an earlier stage. Names are only a minor part of this task. Nevertheless it is much safer to contextually

disambiguate, say sentence initial imperatives from heuristic proper nouns, than at the pattern matching stages. Also, some names have to be morphologically disambiguated as to nominative (NOM) and genitive (GEN). Sentence-initially, names are discarded in favor of verbs and function words, if followed by an np, and non-compound "lexical" nouns win over heuristic names.

### 2.5. NE chunking repair module

This module's general task is to instantiate word fusing and word separating choices too hard or too ambiguous for the preprocessors to make, and therefore left to the CG level for contextual disambiguation. In the case of name chains, faulty NE chunking from earlier modules can be undone. The programme fuses *Hans Jensen* og *Otte Nielsen*, but keeps *Hans Porsche* [his Porsche] and *Otte PC'er* [eight PCs] separate, drawing on the CG's word class and subtype disambiguation of *Jensen <hum>*, *Nielsen <hum>*, *Porsche <V>* and *PC'er <cc-h>*.

Also, upper case nouns are chained to preceding names, e.g. creating an <org> name from a proper noun and a following human group common noun (Betty Nansen Foreningen), now drawing on the full semantic noun lexicon, not only on the patterns and lists of the name type predictor (2.3.).

### 2.6. NE function classes assignment (CG)

Names can, of course, exercise most of the syntactic functions of ordinary np's (subject, object etc.), and these are handled by a 4000 rule syntactic CG. However, some functions are either more limited and specific in use for proper nouns, or, on the contrary, more elaborate. Thus, name predicatives appear with a more limited set of verbs (ad-subject with *være, hedde*, ad-object with *kalde, døbe*, not, e.g., *blive, gøre*), and cannot be free predicatives. On the other hand, especially in a news text corpus, there are name specific apposition types and other postmodifiers with a high frequency for names (*filmen "The Matrix" - John Andersen, distrikschef, Billund, 60 år).*

All of these syntactic relations, once established, can at a later level be used by the system to derive semantic name types from lexical semantic information residing in the corresponding noun heads.

### 2.7. NE semantic type grammar (CG)

This level works with the same 20 NE types used by the lexicon (3) and type-predictor (4), but has the decisive edge of being able to draw on syntactic relations and sentence context. Thus, rules can exploit lexical semantic information pertaining to *other* word classes, especially noun prototypes and verbal subcategorization, to a lesser degree adjective types. Like for syntax, there is both a mapping CG and a disambiguation CG. The former is capable of adding to or even overriding semantic class types from the preceding levels, while the latter removes or selects semantic type tags where the lexion, heuristic type predictor or mapping grammar created ambiguities.

#### 2.7.1 Cross-nominal prototype transfer

As a matter of principle, the NE CG module uses semantic *noun* classes as a context, more or less transferring a head noun's prototype class to a proper noun dependent, where the latter syntactically attaches to the former, as in postnominals (<top> in "byen Rijnsburg" [the town of Rijnsburg]) or subject complements ("Moscow is a town in Russia"). Note that the latter rule can be run in reverse ("The largest town in Russia is Moscow"). More complex rules can check, for instance, for place subject complements (@SC N-TOP) in relative clauses (@FS-N<) with a relative pronoun (<rel> INDP) at most one comma away to the right of the named entity in question:

SELECT (<top>) (0 NOM) (*1 (<rel> INDP @SUBJ>) BARRIER NON-KOMMA LINK *1 VFIN LINK 0 @FS-N< LINK -1 ALL LINK *1 @MV LINK 0 <vk> LINK *1 @<SC LINK 0 N-TOP);

#### 2.7.2 Coordination based type inference

Drawing on and matching syntactic tags from the syntactic CG-module, the name type-mapper correlates same-function conjuncts, e.g. to discard the set of non-human name tags (and thus arriving at an organisation reading) for "Palæstinas Selvstyre" using the coordination with a (safe) person name in the subject group "Arafat og [and] hans [his] Palæstinas Selvstyre [home-rule]".

#### 2.7.3 PP-contexts

Rules can derive place-hood for an argument of certain "place"-prepositions, e.g. 'i' [in], but such rules must be conditioned by the preposition phrase NOT being an object (as in "deltage i" [participate in]) and there NOT being valency demanding nominal context left of the preposition (<+i>), as in "forelsket i" [enamoured in]. Note, that all rules are run in consecutive layers, and that a later, heuristic rule will not only not map <top> after <+i> contexts, but actually *remove* "older" <top> readings in that context, as most nouns with <+i> valency prototypically ask for non-place arguments[1].

#### 2.7.4 Prenominal context: Using adjective classes

Like nouns, adjectives have been semantic type classified in DanGram's Lexicon. Though not extensively used yet, this information is used by a few rules in the name type cg, the most important drawing on the class of human adjectives (lexicon typed as e.g. <psych> or <soc>, or set defined as word lists, e.g. "adfærdsvanskelig", "alfaderlig"), which can be exploited for mapping <hum> tags onto hitherto untyped (<heur>) names, if a dependency relation can be established syntactically.

## 3. Evaluation and perspective

A preliminary evaluation of the current NER-system was carried out on 2 random text chunks from Korpus90 (100.000 words) and Korpus2000 (43.000 words, table 1). There were about 2% chunking errors[2]. and false positive/false negative proper noun readings (originating at the lower levels), while the NER-typer as such had a 5% error rate with 0.1 – 0.5% remaining ambiguity, if measured only for correctly chunked proper nouns.

[1] This method is, of course, imperfect, as there are metaphorical and other exceptions ("forelsket i Venedig"), but if other, earlier and safer, rules have already disambiguated the place/human ambiguity, no harm will be done even in the exception cases.

[2] In fact, in 36 cases, probably due to scanning or excerpting techniques, the corpus contains (partly) uppercase headlines, author headings etc. fused onto subsequent sentences in a way escaping automatic sentence boundary recognition. "Name" candidates resulting from such fusion were left out of this evaluation, since they do not reflect regular running input.

| **Korpus2000** (jan. 2003) | **all PROP** | **heuristic only** |
|---|---|---|
| ca. 43.000 words, ca. 2000 names (ca. 3100 text tokens), about half with lexicon entries | % of all PROP readings | % of non-lexicon readings |
| wrong major class (6 classes) | 5.0 % | 7.3% |
| wrong subclass, same major class | 0.8 % | 0.9 % |
| false positive PROP reading (incl. "overchunking") | 1.4 % | 2.2 % |
| false negative (missing) PROP (incl. "underchunking") | 0.8 % | 1.7 % |
| cross-class ambiguity (major classes) | 0.5 % | 0.1 % |

**Table 1**

These numbers translate into an F-score (accuracy) of just under 93%, slightly better than the 91.85% achieved by a similar system for Portuguese (Bick 2003). This also compares favourably with published data for automated learning systems for English. Thus, at the MUC-7 conference (1998), the best performing system (LTG, Mikheev et. al. 1998) achieved an overall F-measure of 93.39, using hybrid techniques involving both probabilistics/HMM, name/suffix lists and sgml-manipulating rules. Borthwick et. al. (1998), report in-domain/same-topic F-scores of up to 92.20, for maximum-entropy training (MENE). But though *same-topic* performance went up to 97.12% for a hybrid system, when integrating other MUC-7-systems, a possible weakness of trained systems is indicated by the fact that in MUC's *cross-topic* formal test, F-scores dropped to 84.22 and 92 for pure and hybrid MENE, respectively. In this context, it must be stressed that Korpus90/2000 texts are highly cross-domain/cross-topic, since sentence order has been randomized for copyright reasons.

### 3.1. Subtypes

| *tagged as:* / *should be:* | **<hum>** | **<org>** | **<top>** | **<occ>** | **<tit>** | **<brand>** | sum cross-cat. |
|---|---|---|---|---|---|---|---|
| **hum**, A, B | 4 | 2 | 7 | 1 | 1 | 2 | **13** |
| **org**,party,media | 17 | 1 | 5 | 1 | 1 | 2 | **26** |
| **top**, civ, inst | 19 | 7 | 11 | 0 | 0 | 2 | **28** |
| **occ** (events) | 1 | 0 | 0 | 0 | 0 | 0 | **1** |
| **tit**, genre, ling | 14 | 4 | 1 | 1 | 0 | 1 | **21** |
| **brand**, objects | 4 | 4 | 2 | 0 | 0 | 0 | **10** |
| sum cross-categ. | **55** | **17** | **15** | **3** | **3** | **7** | **99** |
| sum all | 59 | 18 | 26 | 3 | 2 | 7 | |
| error-bias = n(tag-error) / n(cat-error) | **4.2** | 0.7 | 0.5 | 3 | **0.14** | 0.7 | |
| tag frequency | 925 | 358 | 607 | 22 | 61 | 30 | 2005 |
| tag frequency % | 46.1 % | 17.9 % | 30.3 % | 1.1 % | 3.0 % | 1.5 % | |
| error-incidence= n(tag-error) / n(tag) | 5.9 % | 4.7 % | **2.5 %** | 13.6 % | 4.9 % | **23.3 %** | |

**Table 2**

### 3.2. Genre variation

Table 2 breaks down error frequency according to name type. Obviously, the big categories <hum> (almost half), <top> (a third) and <org> (a sixth) contribute correspondingly large error shares. In relative terms, however, <top> stands out as particularly "safe" (2.5 %), while the rare event <occ> and brand categories impress as "unsafe" (13% and 23% errors, respectively).

Another interesting detail becomes visible when computing an "error bias", here defined as the ratio between erroneous usage of a given tag-category ("over-usage") and recall-failures for a given category ("under-usage"). For instance, person names <hum> are over-used by the system, usurping proper nouns that should have been read as something else, while titles <tit> are under-used, i.e. often tagged as something else (usually when unquoted). Category-internal errors occur almost exclusively in the <top> category, where rare <civitas> names may be underspecified as ordinary place names[3].

In order to investigate the effect of genre-variation as opposed to topic-variation, three test texts were taken from (a) transcribed parliamentary discussions (46.400 words, 2.3% names), (b) an archaeological journal (43.000 words, 3.0% names), (c) a novel (15.200 words, 5.2% names), all contained in the Danish dfk-corpus (http://corp.hum.sdu.dk). Novel-results (F-score 92.7) resembled Korpus2000 (mostly news-) text, while the parliament transcripts performed better (F-score 94.8), and the archaeological journal performed worse (F-score 89.7), especially in the light of roughly constant token-recognition errors. A possible explanation is that, while mixed (news-) text has a 3:2:1 ratio for person-place-organisation, (b) had 50% place names and (c) 40% organisations. Only (a) resembled the mixed-text ratio, though with an even higher person proportion (70%). Future research should determine whether these differences are due to textual/distributional features alone, or caused by a (remediable?) bias in the lexicon or rule body. Balancing strengths and weaknesses, future work should also examine to which degree probabilistic systems can interface with or supplement CG based NER systems.

## References


Bick, Eckhard (2001). En Constraint Grammar Parser for Dansk. In: Widell, Peter & Kunøe, Mette (eds.): 8. Møde om Udforskningen af Dansk Sprog (pp. 40-50). Århus: Århus University

Bick, Eckhard (2003): Named Entity Recognition for Danish. I: Henrik Holmobe (ed.): Årbog 2002 for Nordisk Sprogteknologisk Forskningsprogram 2000-2004 (pp. 331-350). Copenhaguen: Museum Tusculanum.

Bick, Eckhard (2003), Multi-Level NER for Portuguese in a CG Framework, in: Mamede, Nuno et. al. (eds.) Computational Processing of the Portuguese Language (pp. 118-125). Faro: Springer

Borthwick, Andrew & Sterling, John & Agichtein, Eugene & Grishman, Ralph (1998). NYU: Description of the MENE Named Entity System as Used in MUC-7. In: Proc. of the 7th Message Understanding Conf. (MUC7), April 29th - May 1st, 1998, Fairfax

Mikheev, Andrei & Grover, Claire & Moens, Marc (1998). Description of the LTG System used for MUC-7. In: Proc. of the 7th Message Understanding Conf. (MUC7), April 29th - May 1st, 1998, Fairfax


[3] This error may in some cases even stem from the lexicon, since place names were partly heuristically compiled from "safe" contexts, and only later partly moved into the newly introduced <civitas> category.