Named Entity Recognition for Danish

Eckhard Bick

Institut for Sprog og Kommunikation, Syddansk Universitet

lineb@hum.au.dk, http://visl.hum.sdu.dk

1. Corpora and annotation tools

Named entity recognition is a complex but rewarding task with a number of obvious applications - semantic corpus annotation, information retrieval, text indexing, to name but a few. As in other areas of NLP, more work has been done on English than on lesser languages, and Danish, on the brink of extinction as a language of science and research, seemed a particularly worthwhile target, and after doing independent investigations for some time, I joined the nordic *Nomen Nescio* project (Johannesen, 2002) about a year ago.

The work described in this article is based on corpus data from the Danish twin corpora *Korpus90* and *Koprus2000* (about 56 million words), which have recently been made available in grammatically annotated form. The annotation was a joint venture between the VISL project at Southern Denmark University and the Danish Society for Language and Literature (DSL), who had compiled the corpora from a variety of sources, including both literature, news and science texts.

Word based morphosyntactic information was added with a multi-level Constraint Grammar parser (*DanGram*, Bick 2001), a method that allows the annotation of running text with a high degree of robusticity and a comparatively low percentage of errors (less than 1% for word class). Tag types covered part of speech, inflexion and syntactic function (e.g. @SUBJ for subject and @ADVL for adverbial) as well as dependency markers (e.g. nominal dependent or verbal complement, pre- or postpositioned object and so on).

Nominal semantic tags were used in two different ways. (1) For common nouns, the parser added semantic prototype class tags by lexical lookup, using about 200 different prototypes covering 85% of the parser's noun lexicon of about 75.000 lexemes (for a list of categories, see http://visl.sdu.dk/visl/da/info[1]). These tags were not disambiguated themselves, but used for the disambiguation of syntactic function tags, valency instantiation etc. (2) For proper nouns, about 20 semantic subclasses were used (see discussion below), combining a variety of techniques, in sequential combination as follows:

(a) preliminary regular expression based polylexical recognition (e.g. addresses, web-urls, car names, book titles in quotes)
(b) full lexicon entries (especially major place names and Christian names, but also a number of common surnames and company names)
(c) 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).
(d) orthographical and derivational guesses (e.g. *-sen* names, abbreviated titles, *jr., sr.,* place indicators like *Sankt..., ...vej, ...gade,* ship name indicators like *M/S, H.M.S.*)
(e) a mapping Constraint Grammar capable of contextually overruling semantic class tags from the preceding modules.
(f) a disambiguation Constraint Grammar with context based rules for the selection or removal of tags, where the lexicon, heuristic guesser or mapping grammar created ambiguities.

[1] The prototype marking of the noun lexicon was carried out by Lone Hegelund in 2001 under my supervision, using a category scheme inspired by a similar system for Portuguese (Bick, 2000) and largely compatible with the core ontology semantic types of the European SIMPLE initiative.

As a matter of principle, the named entity CG module uses semantic *noun* classes as a context, more or less transferring a noun's prototype class to a proper noun, if the latter syntactically attaches to the former. Consider, for instance, the following 5 example rules covering the prototype <top> (place) and using the semantic noun type set N-TOP for contextual disambiguation:

MAP (<top>) TARGET (PROP @N<) (-1(N NOM) LINK 0 N-TOP) ;

This rule "instantiates" place-hood (<top>) for names that have been marked as postnominal dependents (@N<) by the syntactic grammar, if their head is itself a place-word (N-TOP), as in "i byen Rijnsburg". Note that this rule is simplified, as it usually has to compete with the other two name categories, that imply +LOC, i.e. <civ> (towns and countries) and <inst> (institutions).

MAP (<top>) TARGET (PROP) (1 @N<FUSE LINK 0 N-TOP) ;

This rule covers "Uppenskij katedralen", where the missing hyphen first leads to a 2-part analysis, which, however, fails to assign 'katedralen' a normal syntactic function, leaving @N<FUSE to survive all REMOVE-rules for all other functions. @N<FUSE is than used to (a) linke the two words, (b) to transfer "place-hood" to the name "Uppenskij".

MAP (<top>) TARGET (PROP) (-1 ("for" PRP)) (-2 ("syd") OR ("vest") OR ("nord") OR ("øst")) ;

This rule is an example for a place specific narrow context, looking for "syd for Odense" kind of patterns. Similar rules exist for "indgang til", "vejen til" and for the preposition "nær" in immediate left context.

ADD (<top>) TARGET (PROP @P<) (-1 ("i" PRP)) (NOT -1 @PIV) (NOT -2 <+i>) ;

This rule derives place-hood for an argument of a preposition, if that preposition is 'i' and the pp is NOT an object (@PIV) as in "deltage i" and if the there is no valency demanding nominal context left of the preposition (<+i>), as in "forelsket i". Note, that the rules run in layers, and that a later heuristic rule will not only not map <top> afte <+i> contexts, but actually *remove* "older" <top> readings in that context: REMOVE (<top>) (0 @P<) (-1 ("i" PRP)) (-2 (<+i>)), as most nouns with <+i> valency prototypically ask for non-place arguments. This is, of course, an unsafe rule, 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.

SELECT (<top>) (0 @SUBJ>) (*1 @MV LINK 0 <vk> LINK *1 @<SC LINK 0 N-TOP) ;

This rule matches subject names with subject complements that are safe place nouns, as in "Moskva er en by i Rusland", and could be run in the reverse, as in "Den største by i Rusland er Moskva". More complicated versions of this rule cover, for instance, "Moskva er en af de største byer i Rusland".

SELECT (<top>) (1 KC) (*2C <top> BARRIER @NON->N) ;

This rule handles coordination with another named entity that has already been recognized as a topological, or is unambiguously known from the lexicon.

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);

This rule, the most complex that will be given here, checks for place subject complements (@SC N-TOP) in relative clauses (@FS-N<) with a relative pronoun (<rel> INDP) at most one komma away to the right of the named entity in question.

2. Discussion of name categories

Prior to classifying names, there should be a stable definition of the parent category, *proper noun.* Strictly speaking, this is not a part in the present project, since the existing parser already made a distinction between ordinary nouns and proper nouns. However, a number of changes was made both in the preprocessor and the disambiguation grammar, the latter especially with regard to sentence initial words. The basic definition of proper noun (PROP) was a morphological one: Upper case words or word chains with upper case initials in the first and last parts[2], non-inflecting in number and definiteness, but (for Danish) inflecting inc case (NOM - GEN). Consequently, simplex names in lower case (pharmaceuticals, biological scientific names) are treated as generic nouns (N), and inflected names are regarded as nominalized nouns *(Mac'en, PC'en, han har spillet i 3 VM'er, vore 3 bedste Pro'e r).* Upper case names with articles/modifiers to the left are still treated as PROP, as long as they do not show (allow?) enclitic definiteness (assumed to be barred by the PROP's inherent definiteness): *et Europa i rødt, *Europa'en, *Europa'et.* Nouns with upper case initial in mid-sentence are still word class marked as nouns (N), but may be marked as <prop> with a secondary tag[3]: *han ringede til Kommunen.* Composita of names and common nouns are treated as common nouns: *Europaminister, EU-kontingent, Martinusbevægelsen, Danmarks-tourné.* Danish does not appear to have a clear strategy with regard to hyphenization in these cases, so these are treated like ordinary composita. Romance numerals and arabic ordinals are added to names as dependents, without influence on the name category in question: *Christian II <hum>, "Terminator 2" <tit>*

In agreement with general Nomen Nescio strategy, 6 core categories were used: Human names, places, organisations, events, titles and brands/objects, the latter also being used for "others". 10 further categories are operational in the lexicon and grammar, while 4 categories remain experimental. All secondary classes can be expressed as combinations or subcategories of the 6 core categories.

2.1. Human names <hum>

People's names can be lexically recognized as a chain of registered simplex personal names: Christian, middle and surnames. Morphologically, certain structural lower case words may precede a surname: *von, van, de, de la, di, du, da, do, das, dos, y, ten, zu, bin, ibn.* These elements may either head the name chain or break the regular flow of upper case words, creating a preprocessor lumping task. In a few cases structural markers become part of a surname as such, as in gaelic *Mac (MacMillan, McNamara),* the Scandinavian patronym endings *-sen/-son* and *-dottir (Jensen, Axelsson, Kristindottir)* or the Slavic endings *-ajev* and *-owa.*

In Danish, only a few Christian names, most notably *Hans* and *Otte,* are - in sentence initial position - co-ambiguous with non-propria classes. Surnames, however, are frequently co-ambiguous with place names, like in *Sprogø, Togeby, Svendstrup* etc., and are sometimes used for creating names for certain cultural concepts or creations,:

a) diseases: *M. Crohn, M. Parkinson, M. Cushing*, used with either *M., Morbus* or as a naked name
b) prizes: *en Bodil, Nobel-prisen, Pulitzer-prisen*
c) paintings: *en Picasso, en ægte van Gogh*
d) stipends: *Betty Jensen Legat*

Note, that certain of these classes, unlike personal names as such, allow articles, an indication that the categories in question (b, c, d) belong in the category for brand or object names, which generally

[2] In the case of book titles, upper case in the first element of a polylexical was regarded as a sufficient criterion.,
[3] Upper case nouns that are not generic and do not allow articles or modifiers, could still sensibly be treated as real PROP names in the lexicon: *Hjemmeværnet, Kommunekontoret.*

admits an oscillation between names and nouns (e.g. car brands). <disease> has been used as an experimental category of its own.

Human name chains can be complements of title nouns, as in *Baronesse Blixen, Jomfru Ane, Hr Jensen, Frøken Mathilde, Mrs Smith, Broder Jakob, Dr. Schnelling*. Though the noun's <+prop> valency does suggest internal structure in these cases, I have in the Danish parser opted for a non-analytic 1-token name reading integrating the titles. One distinction justifying this solution as opposed to the one chosen for profession noun + name *(gravøren Peter Jensen, danserinden Mia Maertens),* is the fact that titles cannot be definiteness inflected before names, while profession nouns can, suggesting that the former are under the inherent definiteness scope of the (larger) name chain.

Given this distinction, both preprocessing and name recognition use a list of title nouns and abbreviations in several languages.

Mythical names denoting humanoids, literary heroes or gods *(Snehvide, Odin, Herkules, Mumrik)* are also treated as <hum>, but suffer a fair deal of ambiguity due to loan usage in astronomy, titles and other fields.

Articles and demonstratives may interfere with name recognition, as in *den Peter jeg husker,* where the article implies added definiteness, and thus a generic reading for Peter in isolation. While this supports a noun reading, upper case morphology favours a name reading, as does the fact that generic names have been allowed in other cases (scientific biological names, cars, brands)

Attributive modifiers *(lille Ida)* are also unusual for names, and one might consider fusing *lille* onto the name token. However, the fact that no article is provided already nicely distinguishes this construction from ordinary noun-np's.

2.2. Place names <top>

The prototypical Danish place name is written as one token without an article, though international exceptions do exist *(Den Haag, O Porto, La Santa, Trinidad y Tobago, San Francisco).* In particular, *Sankt, San, Santo, Santa, São, Saint* etc. are very productive first parts of topologica in Christian countries. Morphologically, place names can often be recognized by geographical elements: *-vig, -havn, -bjerg,* or the town-specific *-rup, -strup, -by, -lev, -sted.*

Syntactically, place names often have a characteristical left verbal context (<va+LOC>, <vta+LOC>, <va+DIR>, <vta+DIR>, MOVE-verbs) or left prepositional context. In particular, *ved* and *i* , to a certain degree *fra* and *til,* are suggestive of topologicals, though a lot of contextual restrictions apply.

However, place names for human settlements (countries, towns, villages etc.) in particular may also function as +HUM subjects of cognitive verbs, as genitive-marked "owners" or with the case role of AGENT (constructing, sending), blurring the distinctional line between <hum> and <top>. For semantic reasons, and to allow for CG rules using +HUM and -HUM syntactic contexts, I have introduced the **civitas <civ>** category for these names: *Danmark, Ikast, USA, London, Folkerepublikken_Kina.* The new category also nicely covers terms like *Det Tredje Rige, Romerriget, Sovjetunionen,* which are neither typical topologicals nor typical organisations.

Buildings with a geographical value, like churches *(St. Peters Katedralen, Ribe Domkirke, Vor Frue Kirke, Bedsted Kirke),* are treated as <top>, if all parts are in upper case, or as a name + noun unit, if the last element i in lower case *(Uppenskij katedralen).* Many buildings, however, have an institutional value that allows them, if only metaphorically, to assume +HUM traits in many sentence contexts. Thus, places like *Mønsterbageriet, diskoteket SiSi, Louvre, Legoland, Kommunekontoret, Det Hvide Hus* can *offer, invite, earn* like names from the person or organisation classes, while at the same time allowing for locational prepositions and the somewhat less human, ergative, opening and closing actions. For these cases, I use the category **institution <inst>**, a kind of hybrid between <org> and <top>, a topologically defined multi-human unit. A characteristical, but unsafe, prepositional context is non-literal *på:* Compare: *på SiSi, på Station Nord* versus *på Ribe Domkirke.* Individual hotels, restaurants and supermarkets *(Illum, London Hilton,*

Rådhuskælderen) are treated as <inst> , while chains are <org> *(Dansk Supermarked, MacDonalds, Best Western).*

A good introspective test for both <top>, <civ> and<inst> is subject-hood for *ligge + LOC.* And even though people and things can *ligge*, too (i seng, på bordet), the semantic difference can be tested with the question: *Hvor ligger X henne?*

Place names can form composita, often with a hyphen, marking either suburbs, motorway access points and postal districts *(Århus-Syd, Odense-NØ)*, geographical ambiguity *(Nykøbing Mors)* or "fused" locations *(Köln-Wahn).* These cases are difficult to distinguish from route denominations like *Århus-Kalundborg, Dover-Calais*, or *trekanten Berlin-Wien-Rom*, which therefore also are treated as <top> (though so far the distinction as *not* <civ> is hard to make).

As demonstrated by the <civ> and <inst> categories, ystematic polysemy allows some names to oscillate between +PLACE and +HUM. Similarly, place names can metaphorically move into the event category <occ>, as in soccer or other sports "pairings": *Danmark-Norge 2:1, semifinalerne Rusland-Spanien (tirsdag) og Frankrig-Italien (onsdag)*, or sporting or political events: *han vandt Wimbledon, Paris-Dakar startede i går*, *efter Maastricht, siden Watergate*

2.3. Organisations <org>

This category covers companies, organisations, ideological movements and the like. Though in principle lexicographically accessible, this category is very productive, unlike place names who form a more stable inventory, and more in the style of personal names. Morphologically, full <org> names are often polylexicals, involving "safe" company markers like *A/S, K/S, I/S, ApS, Ltd, Gmbh, & Co., & bros.,* nominal compound parts like *...selskab, ...forening, Selskab for ...,* or classifier second parts like ... *Foods, ... Electronics, ... Industries, ... Airlines.* Complex <org> names may involve articles (marked as upper case) as in *Den Danske Bank, Den Danske Forening,* or prepositions, in particular *for* and *af* (usually not in upper case): *Sammenslutningen af Alternative Behandlere, Medicinsk Forening for Akupunktur.* Fairly safe graphical markers, used in simplex names, are upper case letters in mid-word: *GrammarSoft, ItauTech,* and - in little proud and umbilical Denmark - names initiating in *Dan...* or *Scan...*

Typical of movements and organisations, and to a certain degree of companies, are abbreviations of 2-3 or more capital letters *(NATO, SF, WHO, AUDI, AEG, AGF),* though there is a considerable ambiguity with chemical formula *(CO, NOX)*, events *(OL, VM)* and other types of abbreviations, calling for maximal lexicographical treatment of abbreviations. A well-defined sub-category of <org> are **political parties <party>**, whose names are often abbreviated, and quite common in newpaper corpora.

Especially in (sports) club names, <org> names can consist of an abbreviation followed by a place name: *FC København, MC Herning.*

Semantically, the <org> category shares a lot of its functional distribution with both person names and institutions (agent case role, +HUM subject-hood with cognitive and speech verbs), but can be distinguished from the former by allowing the preposition *i* to the left, and from the latter by being –PLACE, not allowing narrow place prepositions like *ved.* An introspective semantic test is that <org> can be founded and joined, but not situated and touched. <org> is an abstract entity, while <inst> is a physical entity.

A special case are names of newspapers, radio channels and tv stations, most of which are used as both <org> names and title names <tit>. They can be read in, watched or listened to (like books, films and songs), but at the same time allow for a degree of +HUM agenthood and subjecthood for cognitive verbs: *Jeg har læst i Jyllandsposten, at .... Jyllandsposten har ansat 20 nye medarbejdere, BT tør hvor andre tier ...* In order to capture this double potential, I have introduced the name category of **media <media>**: *Miljø Nyt, Bådnyt, Ekstrabladet, Aftonbladet, Le Figaro, Sunday Times, Frankfurter Allgemeine.*

2.4. Events and occasions <occ>

This category is used for both natural and organized events, periods and time names in general. A good distributional test are time-prepositions, in particular *efter, indtil* and *siden.* As subjects, most members of this category allow time verbs like *vare, stå på* (duration feature), *foregå, forløbe* (process, activity)*, ske* (event)*, begynde, slutte* etc. (not *starte, stoppe,* which also work for moving things).

In Danish, many lexically fixed "time names", like the names of days, months and Christian holidays *(tirsdag, januar, jul, påske)*, are written in lower case, thus, if not contradicting, then at least discouraging proper name analysis in terms of morphological word class, - though most of these do not inflect in number or definiteness, which might be considered criteria i favour of a proper noun reading. There are, however, many upper case polylexicals where either the first or last part are nouns denoting events or occasions, thus allowing to classifiy the name in question as <occ>: *Operation Barbarossa, US Open, Tour de France, Tønder Festival, Første Verdenskrig, Golfkrigen, Europamesterskaberne.*

The last two examples are morphologically different, since they are inflecting, and could also be read as composita consisting of a proper noun *(Golf, Europa)* and a common noun. Such an analysis has, in fact, proven to be very robust. Being analytical, the latter method makes optimal use of the parser's existing lexicon, without the need for a more heuristical name recognizer, or individual lexicon additions.

As mentioned above, there is a metaphoric transfer from sites to events *(Siden Kranskaja Gora 1988).* Exposition and conference names are treated in a similar way. Such names are often marked by including a date *(Expo 98)* or an ordinal *(den 5. internationale Arkitekturudstilling),* but may be formally unrecognizable, as in the title-derived exposition name*"Keltiske fyrster".* Projects *(PaNoLa, Cordial Syn , Hjerteugen)* have a systematic polysemy with the title category <tit>, but allow +TIME dependents and verbs *(vare, starte, lanceres etc.).*

2.5. Book, film and music titles <tit>

Titles can be simplex words *("Skyggen", Iliaden, Genesis, Eddaen),* but will often consist of more than one word, and display a syntactic structure of their own, while still functioning as constituent units in higher order syntactic structures. The strategy of the Danish CG parser is to assign internal structure only where titles appear in isolation, and to lump titles as one-name units in all other cases (titles as subject, object, argument of preposition):
"Frøken_Smillas_fornemmelse_for_sne" blev filmatiseret på engelsk.

Graphical criteria for recognizing titles are quotes and - where surviving in electronic corpora - italics. Furthermore, the 1. word in a title, as well as most content words, will be in upper case, though there seems to be a great deal of variation as to this point, not least because many title names in Danish texts are in English, German, French or other languages. Recognizing title names is thus largely a preprocessing task, and once recognized as a token, most titles are fairly unambiguous.

Semantically, a distinction can be made between literary "running text" titles on the one hand *("The Quick and the Dead", "Det lille hus på prærien", "Mit liv som hund", "Cosi fan tutte", "På loftet sidder nissefar"),* and "classifying" titles on the other. This latter category is rarely quote marked, does not exceed np-structure, and can usually be recognized by a classifying key nominal element: *Den Danske Ordbog, Bill of Rights, Lov om ..., Grundloven,Warszawa-pagten.* Again, a number of these could also be read as name+noun composita.

Problematic cases are prizes, stipends and collections, whose names often sound like titles *("De Gyldne Palmer", Kollektionen "Purple Pal"),* but will here be treated as object/brand names.

Another special case are "name titles":*Der står stadig "Pavarotti" på dem. - Filmen "Oscar".* Here a name of another category, <hum>, fills all of the title name. My current view is that <hum> would be the "analytical" internal reading, while <tit> should be maintained in sentence context, quotes and syntax overruling internal name recognition.

A related category is that of **<genre>**, which subsumes the names of literary or philosophical traditions *(Science Fiction, Islam),* areas of study *(Anatomi, Fysiologi)*, games *(Backgammon, "4-på stribe")* and dances *(Cha cha cha, Square Dance, Togtur til Vejle, Paso Doble)*. These cases are somewhat reminiscent of the <title> category, since they are names for cognitive creations, and sometimes use quotes *("4-på-stribe")* or some internal structure *(Togtur til Vejle).* Dance names in particular, are often co-extensive with corresponding music names. On the other hand, members of the <genre> category are more generic, less individual names, and thus closer to the realm of common nouns. In Danish, there is some corresponding telltale fluctuation in the usage of upper and lower case, where the individual writer can express a varying degree of name-hood by using either one or the other.

An introspective test for the <genre> class is direct object-hood for *dyrke, lære, undervise, forkynde.*

2.6. Object names, and in particular, brand names <brand>

Brand names cover a wide range of products, for instance foods *(Corn flakes),* drinks *(Coca Cola),* operating systems *(Linux),* sanitary products *(Reponse Shampoo)* etc., defying a homogeneous semantic description and turning the category into the the naming system's obvious waste bin category, not least for object names, with the notable exception of the clearly defined class of vehicles (to be discussed below). A morphological feature that nevertheless helps characterizing this category is their members' strong tendency to turn into "real" inflecting common nouns: *Volvo'en, hans gamle Macintosh, at drikke 4 Tuborg, spise en After Eight, tage 3 kodimagnyler/Kodimagnyler, at smøre Kærgården på brødet.* Unless listet in the lexicon in inflected form, the Danish CG parser will treat such inflected words as common nouns in spite of their name etymology and upper case first letter, assigning gender, numer, case and definitness features. In fact, there is a tendency for very common brand names to shed the upper case initial and become ordinary nouns (which the would deserve to be listed in the lexicon as such): *betale med dankort, drikke cola, købe en pc'er.* Semantically, brand names denote things, concrete movable physical entities, that can be *had* and *brought* (not to mention *bought*).

Brand names are often derived from company names, allowing for a certain ambiguity *(Tuborg, Volvo).* In other cases, the company name enters as first part of a polylexical brand name *(Tuborg Gold, VW 1300, Peugeot 603 Cashmere, Apple II, Konica E240 Super SR),* allowing the name recognizer to assign a <brand> tag on the grounds of a recognized company name and a variable second part, in many cases consisting of a well-patterned combination of Arabic or Roman numerals, capital letters and brand type specific key words, for instance *coupé, sedan* or *station car* for cars. Typical for brand names in a more general sense are superlative markers like *Ultra, Super, Extra, de Luxe.*

Wine names are usually derived from regions or other place denominations, and are thus ambiguous between <brand> and <top>, unless a year number or type specific extension *(Apellation, Cru, Sec, Blanco)* force the distinction.

Ship names can often be recognized by a systematic first part *(USS, HMS, S/S, M/S),* followed by a variable name part. In the present tagging system, following the semantic prototyping of ordinary nouns, cars, ships, planes and space shuttles are lumped together as **vehicles <V>**, since they differ from other brand names in that they allow movement verbs, like members of the +ANIM classes <hum> and <A>. A distinction not made explicit at present, is that between generic and individual vehicle names, the ship names above being examples of the latter, car names an example of the former. A similar distinction holds for biological names vs. pet animal names. In the lexicon, but not yet in the parser, <v> and <a> are used for generic names, <V> and <A> for individual names.

As mentioned above, the names of collections, stipends and prizes often "borrow" the shape of another name category, in particular, human names and titles. However, what is denoted, can stil be described as named object (the things collected, the money or trophey awarded), and these categories will thus be included in the object/brand category.

Collections: *Skagens Samlingen, Top 10, Top 50, Kollektionen "Purple Pal"*
Stipends etc.: *H.V.Jensens Legat, Betty Jensen Fondats (<org?>)*
Prizes: *Nobelprisen, De Gyldne Palmer*

2.7. Substance and material names <mat>

Apart from the experimental categories, this is the name category that most systematically contradicts upper case usage, though there is still a great deal of orthographical variation in case spelling in the corpora. Thus, it is not clear whether upper case / lower case as a token feature should force the N/PROP distinction, or whether the lexicon should decide.

The semantics of the category is parallel to the corresponding common noun class *(træ, gummi, klister, salt),* the difference being the degree of "arteficiality" and "scientific specifity". The largest group are pharmaceuticals and household substances: *salvarsan, kodymagnyl, agiolax.* To a certain degree, these can be heuristically captured, via endings like *-am, -cid, -lax,* or additions like *retard* or *forte.*

Another <mat> group consists of chemical abbreviations: *NaCl, NO2, H2O etc.,* though their lower case long forms *(natriumklorid etc.) ,* somewhat illogically, are still treated as nouns.

2.8. Animal and plant names <A>, <B>

This is the category used for scientific biological species names of the type *Mus musculus, Bacillus subtilis, Arenomya arenaria, Quercus robur,* where the first, higher order part of a typically polylexical name is in upper case, the rest in lower case. Heuristic morphological recognition can be attempted using certain characteristical Latin inflexion and derivation endings.

Other sets of names tagged <A> are pet names like *Hundi, Fido, Rex, Blacky* etc., and mythical beasts like *Pegasus, Cerberus* etc., both denoting individuals rather than a species. Like with vehicles, the parser does not make this distinction explicit, whereas the lexicon distinguishes between <A> for pet names and <a> for species names. Moving pet names into the <hum> domain would also solve this ambiguity, but might compromise matching qualitiy with regard to semanticosyntactic selection restrictions (e.g. speech and thinking verbs).

2.9. Astronomical and astrological names <astro>

This category subsumes names of planets, stars, moons and other celestial bodies, as well as constellations and human space installations that are not vehicles (i.e. space stations): *Mars, Merkur, Io, Ganymed, Halley, Mir, Karlsvognen, Orion, Sirius* etc.

Most of these names were inspired by mythical names from the Greek and Roman classics tradition, and are thus ambiguous between <astro> and <hum>, demanding contextual disambiguation, relying, among other things, on the instantiation of ±HUM selection restrictions by CG rules.

2.10. Other, experimental categories

A number of minor name categories can be defined, but not easily subsumed under any of the larger categories. Current candidates are **<disease>, <ling>, <race>** and **<wea>.**

Disease names <disease> are sometimes based on surnames *(Parkinson, M. Parkinson, Morbus Parkinson, Hodgin, Menière),* but there is also a Latin scientific classification *(Diabetes mellitus),* and in some cases ordinary Danish disease nouns are spelled in upper case, suggesting a certain name consciousness on the part of the writer *(Mæslinger, Tuberkulose, Leddegigt).*

Language names <ling> are usually treated as nouns, but (increasing?) upper case usage, or the lack of the usual nationality adjective analogue can suggest a name reading *(latin, esperanto, pidgin, volapyk).*

Race or ethnicity names <race> are also commonly treated as nouns (in fact often language names at the same time), which is supported by the fact that many instances are inflected in the

definite plural *(Yanomamierne, Irokeserne, Maori, Xhosa, Yoruba).* However, words from this category are often spelled in upper case in Danish, mimicking English usage.

Weather phenomena <wea> are sometimes assigned names too, most notably storms. A fashionable example is *El Niño.* Weather names are spelled with upper case initials, and could be subsumed under the event category <occ> .

2.11. Semantic features

In its CG rules files, the Palavras parser uses both semantic prototype tags and atomic semantic features (Bick 2000, pp. 298-327), and established proper noun classes can be integrated into the disaguatio system on par with common nouns. Thus, both prototypes and features will provide context for valency and selection restriction based syntactic disambiguation rules. The table below shows which semantic features are linked to the individual prototype categories discussed above.

Feature bundling in the major name categories (synopsis)

	<vq> +COGN X *siger, tilbyder*	+LOC (place) *være dér* ved/i X	<cc> (concrete movable object) *bring X*	made, built, invented (HUM-cause)	+TIME X *vare, begynde, slutte* siden X	+LIFE	+MOVE
<hum>	+ (1)	-	-	-	-	+	+
<top>	-	+	-	-	-	-	-
<inst><civ>	+	+	-	built	-	-	-
<org><media> <party>	+ (group)	-	-	constituted	-	metaph.	metaph.
<tit><media>	+	-	metaph.	authored	-	-	-
<genre>	+	-	-	taught	-	-	-
<brand><mat>	-	-	+	produced	-	-	-
<V> (<v>)	-	-	+	produced	-	-	+
<A> (<a>)	metaph.	-	+	-	-	+	+
<B> (<b>)	-	(-)	+	-	-	+	-
<astro>	-	+	-	-?	-	-	+
<occ>	-	metaph.	-	(held)	+	-	-

In the table above, certain feature bundling structures become evident. The yellow cell block shows, how the features +COGNITION and +LOCATION can be used to distinguish <hum>/<org> from <top> and <inst>/<civ>, respectively, while <brand> and <occ> have neither of these features. The blue cells lump together discrete physical entities, which may possess both generic and individual names, with the former exhibiting a certain tendency towards inflexion and lower case. The green cell block makes the necessary distinctions within the entity block, using the 4 possible permutations of ±LIFE and ±MOVE. The red block shows the verbal-semantic tests necessary to classify names within the "human-made" categories.

3. Statistics

Name type incidences in running text

	Korpus2000 (ca. 28.630.000 words)				Korpus90[4] (ca. 26.000.000 words)			
<hum>	601.487	46,1%	**601.487**	**46,1%**	517.436	49,9%	**517.436**	**49.9%**
<top>	79.100	6,8%	**424.306**	**32,5%**	89.050	8,6%	**295.677**	**28,5%**
<astro>	883	0.1%			892	0,1%		
<inst>	71.082	5,4%			44.461	4,3%		
<civ>	273.241	20,1%			161.274	15,6%		
<org>	178.931	13,7%	**216.221**	**16,6%**	140.269	13,5%	**164.778**	**15,9%**
<media>	22.008	1,7%			18.640	1,8%		
<party>	15.282	1,2%			5.869	0,6%		
<tit>	31.586	2,4%	**31.586**	**2,4%**	36.537	3,5%	**36.537**	**3,5%**
<occ>	17.120	1,3%	**17.120**	**1,3%**	10.090	1,0%	**10.090**	**1,0%**
<brand>	3.816	0,1%	**14.620**	**1,1%**	4.058	0,4%	**11.790**	**1,1%**
<V>	8.462	0,6%			4.332	0,4%		
<genre>	946	0,1%			953	0,1%		
<mat>	1220	0,1%			1.694	0,2%		
<A>	152	0,0%			627	0,1%		
<B>	24	0,0%			126	0,0%		
	1.305.340		**1.305.340**		1.036.308		**1.036.308**	
All PROP	1.303.358			**4,6%**	1.033.323			**4,0%**
Precision	**99,8%**				**99,7%**			

The statistics was done on the above mentioned "quote corpora" of mixed running text (DSL's Korpus90 and Korpus2000). As the table indicates, no major differences in name distribution were found between the two corpora, and both had an overall 4-5% proper noun incidence. Almost half of all names were found to be personal names <hum>, a third were place names <top> (of these 3/4 "humanoid") and a sixth were "place-less" human non-physical entities <org>. Title frequency <tit> was around 3%, while events <occ> and brands/others <brand> hover around the 1% mark..

In roughly two thirds of all tokens a primary proper noun subclass reading was assigned from the lexicon, in one third a primary reading was assigned by morphological, contextual or heuristic means. In both cases, ambiguity was resolved by a Constraint Grammar module. This module also contains context sensitive mapping rules which can force a name subclass reading other than the primary reading, or increase ambiguity by adding further readings before the disambiguation rules are run.

Once a corpus is annotated, it can of course be used for all kinds of *lexical* statistics, which is often more transparent and more interesting for ordinary users. Using the major name categories discussed in chapter (2), it can be shown, for instance, which person, place, brand or organisation was most renowned - in statistical terms - in Danish texts around the year 2000.

[4] The numbers are for an early version of Korpus90 - the present version amounts to 27.990.000 words.

	<hum>	<top> <civ>	<org> <inst> <party> <media>	<occ>	<brand>
1.	Gud	Danmark	Venstre	VM	Windows
2.	Poul Nyrup Rasmussen	København	Folketinget	Anden Verdenskrig	Linux
3.	Clinton	USA	Politiken	DM	Dannebrog
4.	Ligulf	EU	Jyllands-Posten	Tour de France	Explorer
5.	Nyrup	Europa	NATO	OL	Deep Blue
6.	Jesus	Århus	Socialdemokratiet	EM	Wap
7.	Sara	Tyskland	SF	Wimbledon	Pentium
8.	Bush	Frankrig	FN	French Open	HF
9.	Bill Clinton	Sverige	Dansk Folkeparti	Roskilde Festival	Java
10.	Ritt Bjerregaard	Rusland	Tele Danmark	Den Kolde Krig	Ny_Løn
11.	Marianne Jelved	Kina	DR	Første Verdenskrig	Roundup
12.	Peter	Norge	AGF	Golfkrigen	Colgate
13.	Milosevic	England	DSB	Grand Prix	Bordeaux
14.	Washington	Odense	TV 2	World Cup	Danablu
15.	Tue	Italien	Enhedslisten	Giro d' Italia	Word
16.	Svend Auken	Israel	CD	Australian Open	WordPerfect
17.	Bo Johansson	London	Microsoft	US Open	PlayStation
18.	Mogens Lykketoft	Paris	LO	Melodi Grand Prix	Outlook
19.	Jeltsin	Brøndby	Københavns Universitet	Post Danmark Rundt	HTX
20.	Teodor	Spanien	Den Danske Bank	Europa Cup	Cipramil

The table proves, against all political sense, that God still outranked Poul Nyrup, and Jesus beat Bush. Venstre's latest election victory could have been predicted, and the merger of Politiken and Jyllandsposten as brothers in spirit could have been announced earlier, had this data been made public ... The <occ> list shows, that there's no event like sports and wars, and though Microsoft still ranked below public or semipublic companies like DSB, DR and Tele Danmark, the IT revolution put Bill Gates' products - Windows, Explorer and Word - high in the hit list of brands. He should worry, though, Linux was a close second ... And Danes are still addicted to wine and cheese, with Bordeau and Danablu as a natural pack leader. At least when leaving out cars (<V>), as done in this table.

4. Evaluation

Two 100.000 word subcorpora from the Korpus90 quote corpus were inspected for name tagging errors, evaluating all name readings in one case, and only heuristically based readings in the other. Error types were 'wrong major class' (6 name classes) and 'wrong minor class' (but correct major class), as well as PoS errors concerning the PROP tag itself, i.e. false positive and false negative proper noun readings. The latter also include tokenisation errors, i.e. fusing to much into a PROP token (false positive), or too little (false negative). The latter figure (false negative) is somewhat unsafe, since evaluation focused on PROP contexts only. A special source of errors was the corpus type as such: In spite of a tailor made automatic preprocessing module, there were a fair number of sentence chunking errors (irremediable due to the mixed sentence order of the quote corpus), as well as many upper case first and second words, presumably used as a kind of bold facing, but

creating false positive PROP readings. In all, such problems accounted for an additional error percentage of 1.3% of all PROP readings, and were not included in the table below. Naturally, most of these cases involved non-lexicon readings.

Name type errors in running text

	chunk 1: lexicon-PROP		chunk 2: heuristic PROP	
	instances (100.000 words)	percentage of all PROP readings	instances (100.000 words)	percentage of non-lexicon readings
wrong major class (6 classes)	266	5.0 %	151	9.2 %
wrong subclass, same major class	31	0.6 %	3	0.2 %
false positive PROP reading (incl. "overchunking")	56	1.2 %	27	1.6 %
false negative (missing) PROP (incl. "underchunking")	22	0.4 %	13	0.8 %
cross-class ambiguity (major classes)	7	0.1 %	0	0.0 %
all proper nouns	**5330**		4793	
of these: not in lexicon	1833 (34.4%)		**1641** (34.2%)	

As can be seen from the table, major semantic class errors are nearly twice as frequent for names without a lexicon entry, but since many names ask for ambiguous lexicon entries, these are of course no guarantee against errors. Subclass errors within the same major class are fairly rare, possibly due to the absolute rareness of certain subclasses (cf. previous table). Word class (PoS) and tokenisation (chunking) errors concerning the PROP category run at an error rate of about 1.6 %, which is slightly worse than a Constraint Grammar parser's average PoS error rate, a probable reason being the fact that proper noun recognition is much more dependent on tokenisation and preprocessing than the recognition of other word classes, which to a higher degree can be based on grammatical features and context conditions alone.

References:

Bick, Eckhard, *The Parsing System 'Palavras' - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Århus: Aarhus Universitetsforlag, 2000

Bick, Eckhard, "En Constraint Grammar Parser for Dansk", in Peter Widell & Mette Kunøe (eds.): *8. Møde om Udforskningen af Dansk Sprog, 12.-13. oktober 2000,* Århus Universitet, 2001

Johannesen, Janne Bondi, "Nomen Nescio-prosjekt, Bakgrunn og aktiviteter", i Henrik Holmboe (ed.), *Nordisk Sprogteknologi, Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004,* Museum Tusculanums Forlag, København 2002