

A Compound Matrix

Anders Søgaard

Center for Computational Modelling of Language

Proceedings of the HPSG04 Conference

Center for Computational Linguistics
Katholieke Universiteit Leuven

Stefan Müller (Editor)

2004

CSLI Publications

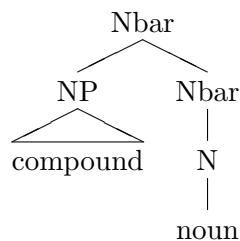
<http://csli-publications.stanford.edu/>

Abstract

This paper presents a supplement to the Grammar Matrix, namely what I call a *Compound Matrix*; in reality, it is not a matrix, since the type file includes a fully specified cross-linguistic inventory of compound types. The idea is that the grammar writer can comment out the ungrammatical types for his or her own language. The theory behind the typology is presented here in a bottom-up fashion, from the basic assumptions to the actual linguistic types.

1 Assumptions

This study deals with the semantics of *two-constituent nonargumental compound nouns*, but it does not include discussion of the syntactic nature of compounds. For clarity of exposition, we adopted the most conventional analysis of such compound nouns; see Radford (1980) for some discussion:



The inventory of the Compound Matrix is based on a typology. Two parameters were used for classifying compounds from about 30 different languages, namely:

- Each nominal constituent can refer in three ways: literally, metonymically or metaphorically. This property is called Reference.
- Independent of syntactic structure, each compound constituent is either the Pointer or Modifier of the construction, and this property is called Status.

The typological part of the study is documented in Søgaard (to appear (a)) and the properties are formally defined in Søgaard (2004). Reference is a Peircean-style trichotomy, while Status is a *functional* distinction; i.e. a Pointer “points” to the possible set of referents, whereas a Modifier modifies or restricts that set. The object of this paper is to provide a semantics for each of these constructions and to describe their implementation in the Compound Matrix. Two important assumptions relate to the translations of compound types:

- Qualia structure (Pustejovsky, 1991) with one additional quale for contour (Q_{ctr}) and Σ -roles (see below) were employed as vocabularies for talking about the meaning of compound nouns.

Table 1: The Compound Typology.

Type	Abbreviation	Example	Language
Appositional	[P(l)-P(l)]	<i>bahay-kubo</i> (house-hut; 'hut')	Tagalog
Copulative	[P(m ₁)-P(m ₁)]	<i>bassu karu</i> (bus-car; 'vehicles')	Kannada
Endocentric	[P(l)-M(m ₁)]	<i>oreh iton</i> (editor newspaper; 'newspaper editor')	Hebrew
Endocentric	[M(m ₁)-P(l)]	<i>numn numpran</i> (village-pig; 'domesticated pig')	Yimas
Endocentric	[P(l)-M(m ₂)]	<i>sundalong-kanin</i> (soldier-cooked rice; 'cowardly soldier')	Tagalog
Endocentric	[M(m ₂)-P(l)]	<i>mek'inobal</i> (mother-haze; 'rainbow')	Tzotzil
Exocentric	[P(m ₂)-M(m ₁)]	<i>panawag-pansin</i> (calling instr.-attention; 'one who wants attention')	Tagalog
Exocentric	[M(m ₁)-P(m ₂)]	<i>Romanteppich</i> (novel-tapestry; a style of prose)	German

- The translations were in (a sublanguage of) the Predicate Calculus.

This latter assumption was motivated by the wish to pass the grammar's output on to a model builder for disambiguation tasks; see Søgaard (2004) for documentation.

1.1 The Compound Typology

Logically, there are 36 possible combinations of Reference and Status. We call the compound whose left-constituent is a Modifier with metonymic Reference, and whose right-constituent is a Pointer with literal Reference, [M(m₁)-P(l)]. This corresponds to a run-of-the-mill endocentric compound in English, e.g. *lawn tennis*. Cross-linguistically, however, only 8 of these 36 types are found; see Table 1.¹

¹It is unclear whether compounds such as *hammerhead* ('shark') and *sabertooth* ('tiger') constitute a class of [M(m₂)-P(m₁)] compounds. No [P(m₁)-M(m₂)] compounds are yet attested. Or is *hammerhead* really a [M(m₂)-P(l)] compound the extension of which has been extended by metonymy? Is it suggestive that another name for sabertooths is *sabertoothed tigers*.

2 Σ -roles

The set of Σ -roles is defined as a (Parsons-style) vocabulary for talking about *event participants*. Since all agentive and telic qualia are eventive, compounds which get their meaning from these qualia involve Σ -roles. The collection of Σ -roles we employ, is inspired by Simon Dik’s *Semantic Function Hierarchy* (here in a slightly revised version):

- (1) Agent[?] \gg Object[?] \gg (Recipient[?]) \gg (Beneficiary[?]) \gg Instrument*
 \gg Material* \gg Location*

(I put Recipient and Beneficiary in brackets, since these roles seem almost irrelevant in the semantics of compound nouns. Though see the appendix for a few exceptions.) For illustration, the telic quale of *knife* is $\lambda x.\exists e.\mathbf{cut}(e) \wedge \mathbf{knife}(x) \wedge \Sigma_{Instr}(x, e)$. If we want to say that a bread is the object of this event, we write $\lambda x.\exists e.\exists y.\mathbf{cut}(e) \wedge \mathbf{knife}(x) \wedge \Sigma_{Instr}(x, e) \wedge \mathbf{bread}(y) \wedge \Sigma_{Obj}(y, e)$. $\Sigma^?$ is optionally expressed, but only “once per event”. (No Sigma Criterion applies here.) Σ^* is optionally expressed more than once, and SIGMA-HEAD identifies the Σ -role of α in $\exists e.\Delta_\alpha(e)$. Consequently, the value of SIGMA-HEAD in **knife1** is **sigmainstr** (a subtype of **sigma-role**).

3 The Construction Hierarchy

The hierarchy of compound constructions, i.e. with the major [S(r)-S(r)]-types as supertypes, and the different combinations of qualia and Σ -roles as subtypes, already seems monstrous and unruly. Is this necessary? There are three reasons that I think the different vocabularies *are* necessary:

- If each construction is properly restricted, ambiguity is realistic, i.e. you typically get one to three readings for each compound.
- The different properties and inventories are helpful in the semantics of adjectives, genitives, prepositions, etc.
- There is empirical evidence for the grammaticality of the specific constructions.

4 Empirical Evidence

- $[M\langle m_1, \Sigma_{Agent} \rangle - P(1)]$ is ungrammatical in English, e.g. **butcher knife* and **musician guitar*; cf. Copestake and Lascarides (1997); but $[M\langle m_2 \rangle - P(1)]$ is not, e.g. *lady snow*²

²There are two possible constraints that explain the ungrammaticality of these examples. Either a certain construction ($[M\langle m_1, \Sigma_{Agent} \rangle - P(1)]$) is ungrammatical, or Attribute

Table 2: The Translation Algorithm.

Type	Logical form
[P(1)-P(1)]	$\lambda x.\beta'(x) \wedge \alpha'(x)$
[P(m ₁)-P(m ₁)]	$\lambda z.\exists x.\exists y.x \oplus y = z.\beta'(x) \wedge \alpha'(y)$ or $\lambda x.\Delta_{F_\beta}$
[P(1)-M(m ₁)]	$\lambda x.\alpha'(x) \wedge \forall y.\exists e.\Delta_\alpha(e) \wedge \Sigma_1(x, e) \wedge \Sigma_2(y, e) \rightarrow \beta'(y)$
[M(m ₁)-P(1)]	$\lambda x.\beta'(x) \wedge \forall y.\exists e.\Delta_\beta(e) \wedge \Sigma_1(x, e) \wedge \Sigma_2(y, e) \rightarrow \alpha'(y)$
[P(1)-M(m ₂)]	$\lambda x.\alpha'(x) \wedge \exists e.\Delta_\beta(e) \wedge \Sigma_1(x, e)$
[M(m ₂)-P(1)]	$\lambda x.\beta'(x) \wedge \exists e.\Delta_\alpha(e) \wedge \Sigma_1(x, e)$
[P(m ₂)-M(m ₁)]	$\lambda x.P(x) \wedge \forall z.\exists e.\Delta_\alpha(e) \wedge \Sigma_1(x, e) \wedge \Sigma_2(z, e) \rightarrow \beta'(z)$
[M(m ₁)-P(m ₂)]	$\lambda x.P(x) \wedge \forall z.\exists e.\Delta_\beta(e) \wedge \Sigma_1(x, e) \wedge \Sigma_2(z, e) \rightarrow \alpha'(z)$

- [M(m₁, Σ_{Instr})-P(1, Σ_{Agent})] is ungrammatical in Danish, e.g. **knivslagter* ('knife butcher') **guitar Musiker* ('guitar musician'); but [M(m₂)-P(1)] is not, e.g. *bildækmand* ('motor car tyre man'); neither is [M(m₁, Σ_{Instr})-P(1, Σ_{Instr_2})], e.g. *guitarforstærker* ('guitar amplifier')
- [M(m₁)-P(1)] where Δ is Contour, is ungrammatical in Estonian; cf. Hiramatsu et al. (2000)
- [M(m₁, Σ_{Loc})-P(1, Σ_{Instr})] translates consistently into β per α in Italian; cf. Paggio og Ørsnes (1993). There is also a grammatical distinction between telic- and agentive-based Δ -compounds; cf. Johnston and Busa (1999)
- ...

There is similar evidence for the grammaticality of non-endocentric constructions. For example, open copulative compounds exist in Modern Greek, but not in Germanic languages. Reportedly only endocentric constructions are found in West Greenlandic (Bauer, 2001).

5 A Translation Algorithm

In our Translation Algorithm (Table 2), *bread knife* translates: 'a bread such that if it cuts anything, then it's bread'. This is of course too restrictive. A better reading is 'a bread such that if it cuts anything, then it's *typically* bread'. This is captured by introducing a Γ -operator (Chierchia, 1995). The Γ -operator is not easy to evaluate computationally. Thus, we introduce an approximation: the $\geq \frac{1}{2}$ -quantifier, a Proportional Quantifier, which denotes a subset of $Mod(\phi)$; see Sogaard (to appear (b)).³

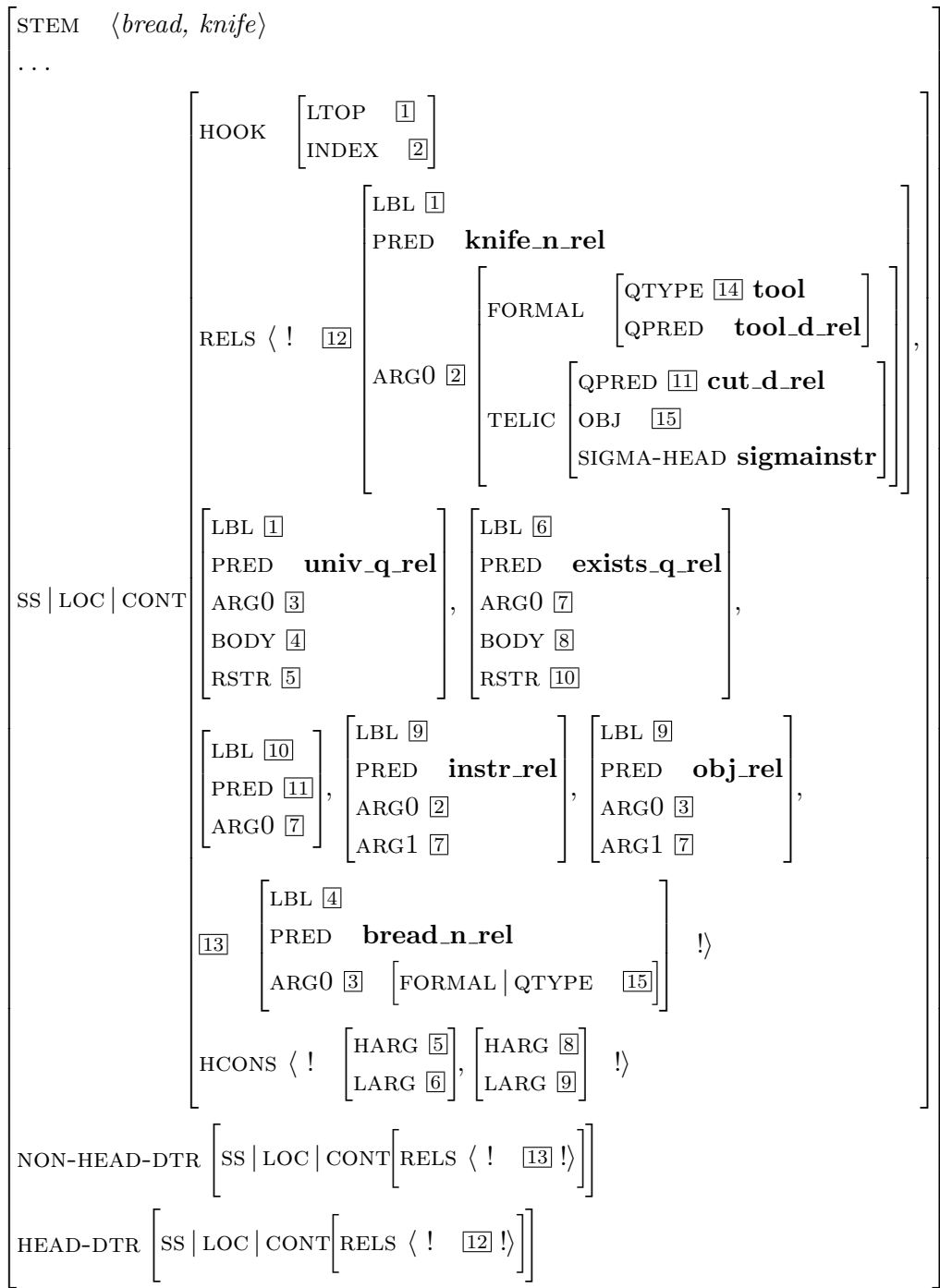
6 The Compound Matrix

The feature structure below represents the semantics of the compound *bread knife* at the \bar{N} -level in the Compound Matrix. It corresponds to the λ -formula above. The reading - 'a bread such that if it cuts anything, then it's bread' - is licensed by the fact that the formal qualia (type) of the non-head daughter unifies with the restriction on Σ_{Obj} of the telic quale of the head daughter.

nouns are restricted in their formal quale to be non-human. The predictions differ. Our first theory wrongly claim that *dog food* is ungrammatical, while the second theory wrongly claims that *child bed* is ungrammatical. Things seem to be fuzzy here. There is a tendency that $[M\langle m_1, \Sigma_{Agent} \rangle - P(1)]$ is not expressed in English, while the corresponding genitive is; on the other hand, $[M\langle m_1, \Sigma_{Obj} \rangle - P(1)]$ is expressed, even with animate or human modifiers. Consider the following examples: *dogtag*, *dog's tongue*, *dog Latin*, *dogwatch*.

³Another simplification is the translation of the exocentric compounds. In our translation, there is only room for one dependent type relation between α and β and the referent. Sometimes this is reasonable, as in the analysis of *dust bowl* ('anything which contains dust'), while in some cases there seem to be more than one relation; e.g. *iron horse* would be analyzed as 'anything made out of iron', which is obviously too unrestrictive. A better analysis would involve both agentive and telic qualia; namely 'anything made out of iron, which is used to transport human beings'. (In Danish, the equivalent of *iron horse*, i.e. *jernhest*, refers to both trains and bicycles.) Exocentric compounds are always underspecified (our analysis still allows plains and boats in $[[iron\ horse]]$). The exact reading of *iron horse* may be due to the historical origin of the word and it's rapid lexicalization.

Also, the \forall -based analysis of endocentric compounds may be redundant in some cases. While a straight-forward \exists -analysis is far to weak for the non-deictic and non-lexicalized use of a compound like *salmon knife*, it suffices for the agentive reading of *Eskimo's knife* (which is a true compound in many Germanic languages). It seems foolish to say that an Eskimo's knife is a 'knife which, whenever (or, worse, at least half of the times) it is made, is made by Eskimos'.



7 The Implementation of a Danish Compound Grammar

The Implementation Algorithm:

- extract qualia information from a SIMPLE dictionary (Lenci et al., 2000) - if there's one for your particular language, that is
- modify the matrix file
- comment out compound types which are ungrammatical in that language, and restrict the grammatical types with appropriate **semantics**
- load the matrix file, the language specific grammar, and the "UG-ish" compound grammar

A simple Perl program was written for extraction. Not all relevant information (i.e. Contour and SIGMAHEAD) are contained in the SIMPLE dictionaries, so we restricted the Perl output in various ways:⁴

- `bil1 := nom-lxm & [STEM <"bil">,SYNSEM.LOCAL.CONT [HOOK.INDEX [FORMAL [QTYPE vehicle, QPRED "vehicle.d_rel"], AGENTIVE [QPRED "fremstille.cre.d_rel", SIGMAHEAD sigmaagentive], TELIC [QPRED "transportere.d_rel", SIGMAHEAD sigmarole], CONTOUR [QTYPE contoursort, QPRED "bil-shaped.d_rel"]], RELS <![PRED ".bil_n_rel"!>]]].`
- `bil1 := nom-lxm & [STEM <"bil">,SYNSEM.LOCAL.CONT [HOOK.INDEX [FORMAL [QTYPE vehicle, QPRED "vehicle.d_rel"], AGENTIVE [QPRED "fremstille.cre.d_rel", SIGMAHEAD sigmaagentive], TELIC [QPRED "transportere.d_rel", SIGMAHEAD sigmainstr], CONTOUR [QTYPE cubic, QPRED "bil-shaped.d_rel"]], RELS <![PRED ".bil_n_rel"!>]]].`

The necessary modifications of the matrix file are:

- remove the constraint that quantifiers only quantify over **ref-ind**, i.e. include events
- add a **semantics** ontology (e.g. one based on the SIMPLE dictionary) and a **contoursort** ontology
- add types for qualia

8 Interpretation hierarchy

(Parallels complexity in processing, i.e. economy.)

- Words, i.e. lexicalized compounds

⁴For those who don't speak Danish: `bil` is *car*, `fremstille` is *manufacture*, and `transportere` is *transport*.

- Endocentric compounds (incl. appositional and copulative compounds; is there any internal ranking?)
- Exocentric compounds
- Pragmatic interpretations (incl. deictic compounds?)

9 Differences between Danish and English

Using the Compound Matrix types, we constructed a Danish Compound Grammar, following the algorithm above. The grammar is about 2000 words and it blocks 4 types out of about 70 compound constructions. We also constructed an English test grammar; see Søgaard (2004). One of the immediate advantages of compatible grammars is that one can easily describe the differences between compound components of different grammars. Some of the major differences between Danish and English are mentioned here:

- $[M\langle m_1, \Sigma_{Instr} \rangle - P\langle l, \Sigma_{Agent} \rangle]$ is blocked in Danish
- $[M\langle m_1, \Sigma_{Agent} \rangle - P(l)]$ is blocked in English
- While $[P(l) - M(m_1)]$ works in the domain of law in English (e.g., *Code Napoleon*), it doesn't in Danish
- ...

10 Conclusion and Applications

The Grammar Matrix as such adopts a pragmatic approach to compound semantics (Flickinger and Bender, 2003). Such an approach is theoretically inadequate for a variety of reasons. This is evidenced by some of the data presented here, i.e. the ungrammaticality of certain compound types in certain languages, but see Liberman and Sproat (1992), Copestake and Lascarides (1997) and Søgaard (2004) for some more detailed discussion. The supplement presented here addresses this problem in the current design of the Grammar Matrix. The supplement provides exactly the kind of analyses which are necessary for applications such as machine translation and knowledge-based disambiguation.

The Compound Matrix is a flexible module that could easily be fit into other packages, e.g. ERG or the Matrix of Mainland Scandinavian which is currently being developed by the Scandinavian HPSG community. Since semantics is first-order axiomatizable, knowledge-based disambiguation is possible with theorem provers and model builders; cf. Søgaard (2004). The Compound Matrix is for now available upon request.

11 Appendix 1: Σ -roles in $[M(m_1)-P(l)]$ compounds

Where $\alpha = Agent$:

- (2) (a) eskimomusik ('Eskimo's music') ($\beta = Object$)
- (b) eskimokniv ('Eskimo's knife') ($\beta = Instrument$)
- (c) kunstnerværksted ('artists' workplace') ($\beta = Location$)
- (d) børneler ('children's clay') ($\beta = Material$)

Where $\alpha = Object$:

- (3) (a) romanforfatter ('novel writer') ($\beta = Agent$)
- (b) tomatkniv ('tomato knife') ($\beta = Instrument$)
- (c) grøntsagstorv ('vegetable marketplace') ($\beta = Location$)
- (d) skulpturler ('sculpture clay') ($\beta = Material$)

Where $\alpha = Recipient$ or *Beneficiary* (seldom):

- (4) (a) børneforfatter ('children's writer') ($\beta = Agent$)
- (b) kirkeskat ('church tax') ($\beta = Object$)
- (c) børnepenge (children-money; 'financial support to parents')
($\beta = Object$)
- (d) næsedråber ('nose drops') ($\beta = Instrument$)

Where $\alpha = Instrument$:

- (5) (a) * ($\beta = Agent$)
- (b) sværddans ('sword dance') ($\beta = Object$)
- (c) elsav ('power saw') ($\beta = Instrument$)
- (d) knivkøkken ('knife kitchen') ($\beta = Location$)
- (e) støbeformsjern ('mold/cast iron') ($\beta = Material$)

Where $\alpha = Location$:

- (6) (a) koncertmusiker ('concert musician') ($\beta = Agent$)
- (b) skolemad ('school food') ($\beta = Object$)
- (c) køkkenkniv ('kitchen knife') ($\beta = Instrument$)
- (d) skolebod ('school booth/shop') ($\beta = Location$)
- (e) skoleler ('school clay') ($\beta = Material$)

Where $\alpha = Material$:

- (7) (a) stålsmed ('steel smith') ($\beta = Agent$)
- (b) oliemaleri ('oil painting') ($\beta = Object$)
- (c) ?boghvedeovn ('buckwheat oven') ($\beta = Instrument$)
- (d) træværksted ('wood workshop') ($\beta = Location$)

12 Appendix 2: Σ -roles in $[M(m_1)-P(m_2)]$ compounds

Where $\alpha = Agent$:

- (8) (a) børnechampagne (children's champagne; 'juice') ($\beta = Object$)
- (b) hundeskål ('dogs' bowl') ($\beta = Instrument$)
- (c) hundehus ('dogs' house') ($\beta = Location$)
- (d) hippietobak (hippies' tobacco; marihuana) ($\beta = Material$)

Where $\alpha = Object$:

- (9) (a) fodbolddommer (football judge; 'referee') ($\beta = Agent$)
- (b) fiskelomme (fish pocket; 'place with many fish') ($\beta = Instrument$)
- (c) auraværksted (aura workshop/repair office; 'psychologists' office') ($\beta = Location$)
- (d) fjeldhyttebeton (Norwegian cabin beton/concrete; 'wood') ($\beta = Material$)

Where $\alpha = Instrument$:

- (10) (a) * ($\beta = Agent$)
- (b) bogstavrim (letter rhyme; alliteration) ($\beta = Object$)
- (c) flaskepost (bottle mail; 'bottle message') ($\beta = Instrument$)
- (d) vindmølle ('wind mill') ($\beta = Location$)
- (e) printer lead ('ink, cartridge') ($\beta = Material$)

Where $\alpha = Location$:

- (11) (a) køkkenmusiker ('kitchen musician') ($\beta = Agent$)
- (b) bordtennis ('table tennis') ($\beta = Object$)
- (c) rumskib ('space ship') ($\beta = Instrument$)
- (d) kloakrestaurant (underground restaurant; 'foot chamber for rats') ($\beta = Location$)
- (e) Grønlandsmursten (Greenland brick; 'ice block (for iglos)') ($\beta = Material$)

Where $\alpha = Material$:

- (12) (a) betontømrer ('beton carpenter') ($\beta = Agent$)

- (b) lufttennis ('air tennis') ($\beta = \textit{Object}$)
- (c) luftguitar ('air guitar') ($\beta = \textit{Instrument}$)
- (d) genværksted ('genetics repair office') ($\beta = \textit{Location}$)

I realize that *fodboldommer* and *vindmølle* could also be analyzed as endocentric compounds, accepting very underspecified lexical semantics of *ommer* and *mølle*. These examples are just for illustration, so this problem is ignored. There are plenty of examples in both these categories, e.g., respectively, *bogorm* ('bookworm') and *cykelmotorvej* (bicycle highway; 'broad path for bicycles').

13 References

- Bauer, Laurie. 2001. Compounds. In Martin Haspelmath et al. (eds.), *Language typology and language universals*. Berlin: de Gruyter.
- Chierchia, Gennaro. 1995. Individual-level predicates as inherent generics. In Gregory Carlson and Francis Pelletier (eds.), *The generic book*. Chicago: University of Chicago Press.
- Copestake, Ann; Lascarides, Alex. 1997. Integrating symbolic and statistical representations: the lexicon-pragmatics interface. *Proceedings of ACL-EACL 1997*. Madrid, Spain.
- Hiramatsu, Kazuko; Snyder, William; Roeper, Thomas; Storrs, Stephanie; Saccomon, Mathew. 2000. On musical hand chairs and linguistic swing. *Proceedings of the 24th Boston University Conference on Language Development*. Somerville, USA.
- Johnston, Michael; Busa, Federica. 1999. Qualia structure and the compositional interpretation of compounds. In Evelyne Viegas (ed.), *Breadth and depth of semantic lexicons*. Dordrecht: Kluwer.
- Lenci, Alessandro; Busa, Federica; Ruimy, Nilda; Gola, Elisabetta; Monacchini, Monica; Galzolari, Nicoletta; Zampolli, Antonio; Guimier, Emilie; Recourcé, Gaëlle; Humphreys, Lee; von Rekovsky, Ursula; Ogonowski, Antoine; McGauley, Clare; Peters, Wim; Peters, Ivonne; Gaizauskas, Robert; Villegas, Marta. 2000. *Linguistic specifications*. SIMPLE Work Package 2.
- Lieberman, Mark; Sproat, Richard. 1992. The stress and structure of modified noun phrases in English. In Anna Szabolcsi and Ivan Sag (ed.), *Lexical matters*. Stanford: CSLI.
- Paggio, Patrizia; Ørnsnes, Bjarne. 1993. Automatic translation of nominal compounds. *Rivista di Linguistica* 5: 129-156.
- Pustejovsky, James. 1991. The generative lexicon. *Computational Linguistics* 17: 409-441.
- Radford, Andrew. 1980. *Transformational grammar*. Cambridge: Cambridge University Press.
- Søgaard, Anders. (2004). *Compound semantics: interpretation, representation and implementation*. Copenhagen: Dpt. of Computational Linguistics, Copenhagen Business School. (MA-thesis)
- Søgaard, Anders. (to appear (a)). Compounding theories and linguistic diversity. In Zygmunt Frajzyngier (ed.), *Language theories and linguistic diversity*. Amsterdam: John Benjamins.
- Søgaard, Anders. (to appear (b)). Proportional quantifiers in first-order logic. Submitted to the Workshop on Computational Semantics 2005.