Corpus Linguistics, Generative Grammar, and Database Semantics

Roland Hausser

Universität Erlangen-Nürnberg
Abteilung Computerlinguistik (CLUE)
rrh@linguistik.uni-erlangen.de

Abstract This paper begins with a comparison of the word (form) analysis in (i) *Collins COBUILD English Language Dictionary* (1987), edited by John Sinclair, (ii) the statistical CLAWS tagger for the British National Corpus, and (iii) the system of automatic word form recognition in Database Semantics. Then it turns to a question debated in Corpus Linguistics, namely whether or not "meanings are in the head," and examines the role of corpora in Generative Grammar. In conclusion, the paper discusses the distinction between *sign*-oriented and *agent*-oriented approaches to the analysis of natural language, and whether or not collocation and the context of use should be treated as part of the grammar.

1 Learner's Dictionary and Statistical Tagging

In British Corpus Linguistics (CL), two schools may be distinguished: one is associated with the University of Birmingham and its mentor John Sinclair, the other with the University of Lancaster and its mentors Roger Garside and Geoffrey Leech. The Birmingham approach has been characterized as "CL-as-theory" and "doing language", the Lancaster approach as "CL-as-method" and "doing computing" (Kirk 1998).

The difference between the two approaches is apparent in their respective analyses of a word. Take for example the word decline, analyzed in the Collins COBUILD English Language Dictionary (CCELD, Sinclair 1987) as a lexical entry with several readings:

1.1 ENTRY OF decline IN COLLINS COBUILD ELD 1987 (EXCERPT)

decline /de'klain/, **declines, declining, declined** 1 if something **declines**, it becomes less in quantity, importance, or strength. [citations]
2 If you **decline** something or **decline** to do something, you politely refuse to accept something or to do something; a fairly formal word. [citations]
3 **Decline** is the condition or process of becoming less in quantity, importance, or strength. [citations]
...

Intended for learners rather than fluent speakers of English, the forms **declines, declining**, and **declined** are explicitly listed (instead of naming the paradigm).

In a separate column (not shown in 1.1), the CCELD characterizes reading 1 as an intransitive **V**erb with the hypernym *decrease*, the cognate *diminish*, and the antonym *increase*. Reading 2 is characterized as **V, V+O, Or V+*to*-INF**, whereby **V+O** indicates a transitive verb. Reading 3 is characterized as **N UNCOUNT/COUNT:USU SING**, i.e., as a noun which is usually used in the singular. Chapter 3 of Sinclair (1991) provides a detailed discussion of this entry to explain the form and purpose of entries in the CCELD in general.

Next consider the corresponding Lancaster analysis:

1.2 Forms of decline as analyzed in the BNC 2007 XML Edition

```
3682 decline NN1                          1 declinedtocomment NN1
 451 decline VVI                        249 declines VVZ
 381 decline NN1-VVB                     26 declines VVZ-NN2
 121 decline VVB-NN1                     22 declines NN2
  38 decline VVB                          7 declines NN2-VVZ
   1 decline-and-fall AJ0-NN1           446 declining AJ0
   1 decline/withdraw VVB               284 declining VVG-AJ0
 800 declined VVN                       234 declining AJ0-VVG
 610 declined VVD                       138 declining VVG
 401 declined VVD-VVN                     1 declining-cost AJ0
 206 declined VVN-VVD                     1 declining-in AJ0
```

To evaluate the tagging, we have to look up the definitions of the relevant tag-set[1] in order to see which classifications are successful. For example, declining is assigned four different tags (ambiguity), which are defined as follows:

446	declining	AJ0	adjective (unmarked) (e.g. GOOD, OLD)
284	declining	VVG-AJ0	-ing form of lexical verb and adjective (unmarked)
234	declining	AJ0-VVG	adjective (unmarked) and -ing form of lexical verb
138	declining	VVG	-ing form of lexical verb (e.g. TAKING, LIVING)

From a linguistic point of view, it would be better to classify declining unambiguously as the progressive form of the verb, and leave the standard uses of the progressive as a predicate, a modifier, or a noun to the rules of syntax.

Critical remarks on the accuracy[2] and usefulness of statistical tagging aside, the Birmingham and the Lancaster approaches share the same methodological issues of corpus linguistics, namely sampling representativeness, size, format (and all their many sets of choices) as well as the basic techniques such as the use of frequency lists, the generation of concordances, the analysis of collocations, and the question of tagging, parsing, and other kinds of in-text annotation. And both raise the question of whether their computational analysis of machine-readable texts is just a methodology (extending the tool box) or a linguistic theory.

This question is addressed by Teubert and Krishnamurthy (2007, p.1) as follows:

1.3 Brief selection of views on corpus linguistics

Corpus linguistics is . . .

[1] The UCREL CLAWS5 tag-set is available at http://ucrel.lancs.ac.uk/claws5tags.html.

[2] Cf. FoCL'99, Sect. 15.5.

- a practice, rather than a theory
- the study of language based on evidence from large collections of computer-readable texts and aided by electronic tools
- a newly emerging empirical framework that combines a firm commitment to rigorous statistical methods with a linguistically sophisticated perspective on language structure and use
- a vital and innovative area of research

Corpus linguistics is not . . .

- a branch of linguistics, but a route into linguistics
- a distinct paradigm in linguistics but a methodology
- a linguistic theory but rather a methodology
- quite a revolt against an authoritarian ideology, it is nonetheless an argument for greater reliance on evidence
- purely observational or descriptive in its goals, but also has theoretical implications

Regarding the Birmingham "CL-as-theory" and "doing language" approach, Sinclair is quite adamant about the *authority* of real data over examples invented by linguists in the Chomsky tradition – which is a methodological issue. But when it comes to writing lexical entries, Sinclair is pragmatic, with readability for the learner as his topmost priority. For example, in his introduction to the CCELD (1987) Sinclair writes:

> Within each paragraph the different senses are grouped together as well as the words allows. Although the frequency of a sense is taken into account, the most important matter within a paragraph is the movement from one sense to another, giving as clear as possible a picture.

2 The Place of Lexical Meanings

The aim of corpus linguists and dictionary builders is to provide an accurate description of "the language" at a certain point in time or in a certain time interval. It seems to follow naturally from this perspective that a language is viewed as an object "out there in the world." As Teubert (2008) puts it:

> Language is symbolic. A sign is what has been negotiated between sign users. The meaning of a sign is not my (non-symbolic) experience of it. Meanings are not in the head, as Hilary Putnam[3] never got tired of repeating. The meaning of a sign is the way in which the members of a discourse community are using it. It is what happens in the symbolic interactions between people, not in their minds.

[3] Putnam attributes the same ontological status to the meanings of language as Mathematical Realism attributes to mathematical truths: they are viewed as existing eternally and independently of the human mind. In other words, according to Putnam, language meanings exist no matter whether they have been discovered by humans or not.

What may hold for mathematics is less convincing in the case of language. First of all, there are many different natural languages with their own characteristic meanings (concepts). Secondly, these meanings are constantly evolving. Thirdly, they have to be learned and using them is a skill. Treating language meanings as preexisting Platonic entities out there in the world, to be discovered by the members of the language communities, is especially doubtful in the case of new concepts such as *transitor* or *ticket machine*.

On the one hand, it is uncontroversial that language meanings should not be treated as something personal left to the whim of individuals. On the other hand, simply declaring meanings to be real external entities is an irrational method for making them "objective." The real reason why the conventionalized surface-meaning relations are shared by the speech community is that otherwise communication wouldn't work.

Even if we accept for the sake of the argument that language meanings may be viewed (metaphorically) as something out there in the world, they must also exist in the heads of the members of the language community. How else could speaker-hearers use language surfaces and the associated meanings to communicate with each other?

That successful natural language interaction between cognitive agents is a well-defined mechanism is shown by the attempt to communicate in a foreign language environment. Even if the information we want to convey is completely clear to us, we will not be understood by our hearers if we fail to use their language adequately. Conversely, we will not be able to understand our foreign communication partners who are using their language in the accustomed manner unless we have learned their language.

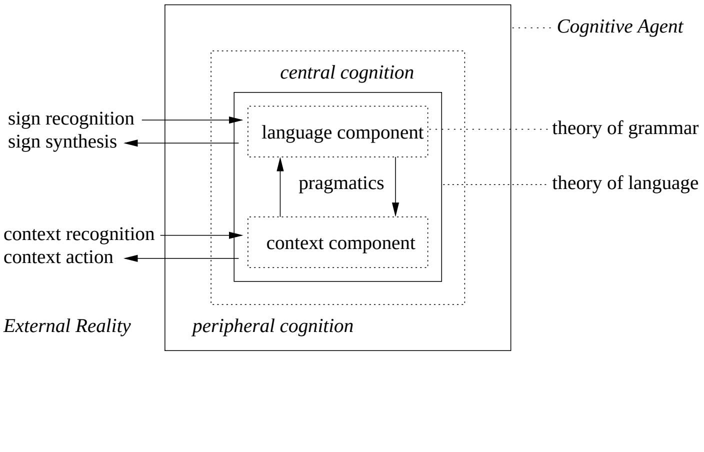
Given that natural language communication is a real and objective procedure, it is a legitimate scientific goal to model this procedure as a theory of how natural language communication works. Such a theory is not only of academic interest, but is also the foundation of free human-machine communication in natural language. The practical implications of having machines which can freely communicate in natural language are enormous: instead of having to program the machines we could simply talk to them.

3 Basic Structure of a Cognitive Agent with Language

Today, talking robots exist only in fiction, such as R2D2 in the Star Wars movies (George Lukas 1977–2005), and Roy, Rachael, etc., in the movie Blade Runner (Ridley Scott 1982). The first and so far the only effort to model the mechanism of language communication as a computational linguistic theory is Database Semantics (DBS).

DBS is developed at a level of abstraction which applies to natural agents (humans) and artificial agents (talking robots) alike. In its simplest form, the interfaces, components, and functional flow of a talking agent may be characterized schematically as follows (borrowed from NLC'06, Sect. 2.4):

3.1 STRUCTURING CENTRAL COGNITION IN AGENTS WITH LANGUAGE



According to this schema, the cognitive agent has a body out there in the world[4] with external interfaces for recognition and action. Recognition is for transporting content from the external world into the agent's cognition, action is for transporting content from the agent's cognition into the external world.[5]

In this model, the agent's immediate reference[6] with language to corresponding objects in the agent's external environment is reconstructed as a purely cognitive procedure. An example of immediate reference in the hearer mode is following a request, based on (i) language recognition, (ii) transfer of language content to the context level based on matching, and (iii) context action. An example in the speaker mode is reporting an observation, based on (i) context recognition, (ii) transfer of context content to the language level based on matching, and (iii) language production including sign synthesis.[7]

From the viewpoint of building a talking robot, the language signs existing in the external reality between communicating agents are merely acoustic perturbations (speech) or doodles on paper (writing) which are completely without any grammatical properties or meaning (cf. NLC'06, Sect. 2.2; Hausser 2009b). The latter arise via the agent's *word form recognition*, based on matching the shapes of the external surfaces with corresponding keys in a lexicon stored in the agent's memory.

This lexicon must be learned by each member of the language community. The learning procedure is self-correcting because using a surface with the wrong conventional meaning leads to communication problems. If there is anything like Teubert's and Putnam's notion of language (a position known as linguistic externalism), it is a reification of the intuitions of members of the associated language community, manifested as signs produced by speakers (or writers) in a certain intervall of time. These manifestations may then be selected, documented, and interpreted by corpus linguists.

4 Automatic Word Form Recognition

The computer may be used not only for the construction of dictionaries, e.g., by using a machine-readable corpus for improving the structure of the lexical entries, but also for their use: instead of finding the entry for a word like decline in the hardcopy of a dictionary using the alphabetical order of the lemmata, the user may type the word on a computer containing an online version of the dictionary – which then returns the corresponding entry on its screen. Especially in the case of large dictionaries with several volumes and extensive cross-referencing, the electronic version is considerably more user-friendly to the computer-literate end-user than the corresponding hardcopy.

[4] The importance of agents with a real body (instead of virtual agents) has been emphasized by emergentism (MacWhinney 2008).

[5] While language and non-language processing use the same interfaces for recognition and action, 3.1 distinguishes channels dedicated to language and to non-language interfaces for simplicity: sign recognition and sign synthesis are connected to the language component; context recognition and context action are connected to the context component.

[6] Cf. FoCL'99, Sect. 4.3; NLC'06, Sect. 2.5.

[7] For a more extensive taxonomy see the 10 SLIM states of cognition in FoCL'99, Sect. 23.5.

Electronic lexical lookup is based on matching the unanalyzed surface of the word in question with the lemma of the online entry, as shown in the following schema:

4.1 MATCHING AN UNANALYZED SURFACE ONTO A KEY

unanalyzed word form surface: decline

| *matching*

lexical entry: [decline (lexical description)]

There exist several techniques for matching a given surface automatically with the proper entry in an electronic lexicon.[8]

The method indicated in 4.1 is also used for the automatic word form recognition in a computational model of natural language communication, e.g., Database Semantics. It is just that the format and the content of the lexical descriptions are different.[9] This is because the entries in a dictionary are for human users who already have natural language understanding, whereas the entries in an online lexicon are designed for building a language understanding in an artificial agent.

5 Concept Types and Concept Tokens

The basic concepts in the agent's head are provided by the external interfaces for recognition and action. Therefore, an artificial cognitive agent must have a real body interacting with the surrounding real world. The implementation of the concepts must be procedural because natural organisms as well as computers require independence from any metalanguage.[10] It follows that a truth-conditional or Tarskian semantics cannot be used.[11]

According to the procedural approach, a robot understands the concept of *shoe*, for example, if it is able to select the shoes from a set of different objects, and similarly for different colors, different kinds of locomotion like walking, running, crawling, etc. The procedures are based on concept types, defined as patterns with constants and restricted variables, and used at the context level for classifying the raw input and output data.[12]

As an example, consider the following schema showing the perception of an agent-external *square* (geometric shape) as a bitmap outline which is classified by a corresponding concept type and instantiated as a concept token at the context level:

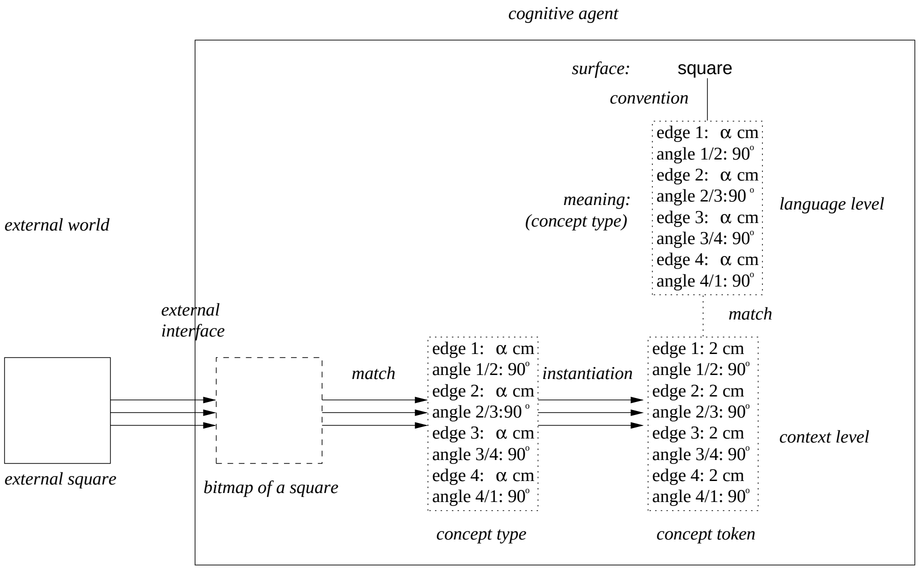
[8] See A.V. Aho & J.D. Ullman 1977, p. 336–341.

[9] Apart from their formats, a dictionary and a system of automatic word form recognition differ also in that the entries in a dictionary are for *words* (represented by their base form), whereas automatic word form recognition analyzes inflectional, derivational, and compositional *word forms* on the basis of a lexicon for allomorphs or morphemes (cf. FoCL'99, Chapt.13). Statistical tagging also classifies word forms, but uses transitional likelihoods rather than a compositional analysis based on a lexical analysis of the word form parts.

[10] Cf. FoCL'99, p. 82–83.

[11] Cf. FoCL'99, Sects. 19.3–19.5.

[12] For a more detailed discussion of the basic mechanisms of recognition and action see FoCL'99, Sects. 3.2–3.3, and NLC'06, Sects. 4.2–4.4.



The necessary properties,[13] shared by the concept type and the corresponding concept token, are represented by four attributes for edges and four attributes for angles. Furthermore, all angle attributes have the same value, namely the constant “90 degrees” in the type and the token. The edge attributes also have the same value, though it is different for the type and the token.

The accidental property of a square is the edge length, represented by the variable α in the type. In the token, all occurrences of this variable have been instantiated by a constant, here 2 cm. Because of its variable, the type of the concept *square* is compatible with infinitely many corresponding tokens, each with another edge length.

At the language level, the type is reused as the literal meaning of the English surface square, the French surface carré, and the German surface Quadrat, for example. The relation between these different surfaces and their common meaning is provided by the different conventions of these different languages. The relation between the meaning at the language level and the contextual referent at the context level is based on matching using the type-token relation.

The representation of a concept type and a concept token in 5.1 is of a preliminary holistic nature, intended for simple explanation.[14] How such concepts are exactly implemented as procedures and whether these procedures are exactly the same in every agent is not important. All that is required for successful communication is that they provide the same results (relative to a suitable granularity) in all members of a language community.

[13] Necessary as opposed to accidental (kata sumbebêkos), as used in the philosophical tradition of Aristotle.

[14] For a declarative specification of memory-based pattern recognition see L&I’05.

6 Proplets

Defining a basic meaning like *square* as a procedure for recognition and action is only the first step to make an artificial agent understand. Leaving aside questions of whether or not there is a small set of “semantic primitives” (Wierzbicka 1991) from which all other meanings can be built, and of whether or not all natural languages code content in the same way (Nichols 1992), let us turn to the form of lexical entries in DBS.

Starting from a basic meaning, the lexical entries add morpho-syntactic properties such as part of speech, tense in verbs, number in nouns, etc., needed for grammaticalized aspects of meaning, syntactic agreement, or both. These properties are coded (i) in a way suitable for computational interpretation and (ii) as a data structure fulfilling the following requirements:

First, the lexical entries of DBS are designed to provide for an easy computational method to code the semantic relations of functor-argument and coordination structure between word forms. Second, they support a computationally straightforward matching procedure, needed (i) for the application of rules to their input and (ii) for the interaction between the language and the context level inside the cognitive agent. Third, they code the semantic relations in complex expressions in an *order-free* manner, so that they can be stored in a database in accordance with the needs of storage in the hearer mode and of retrieval in the speaker mode.

The format for satisfying these linguistic and computational requirements are flat (non-recursive) feature structures called proplets. As an example consider the lexical analysis of the English word surface square as a noun (as in Anna drew a square), as a verb (as in Lorenz squared his account), and as an adjective (as in Jacob has a square napkin).

6.1 Different proplet structures with the same core value

noun, singular	*noun, plural*
sur: square	sur: squares
noun: *square*	noun: *square*
cat: sn	cat: pn
sem: sg	sem: pl
mdr:	mdr:
fnc:	fnc:
prn:	prn:

transitive verb *3rd pers. sg present*	*non-3rd-pers.sg present*	*past tense/past participle*	*progressive*
sur: squares	sur: square	sur: squared	sur: squaring
verb: *square*	verb: *square*	verb: *square*	verb: *square*
cat: ns3′ a′ v	cat: n-s3′ a′ v	cat: n′ a′ v	cat: a′ be
sem: pres	sem: pres	sem: past/perf	sem: prog
mdr:	mdr:	mdr:	mdr:
arg:	arg:	arg:	arg:
prn:	prn:	prn:	prn:

adjective

$$\begin{bmatrix}\text{sur: square}\\ \text{adj: }\textit{square}\\ \text{cat: adn}\\ \text{sem:}\\ \text{mdr: B}\\ \text{mdd:}\\ \text{prn:}\end{bmatrix}$$

These proplets contain the same concept type *square* (illustrated in 5.1) as the value of their respective core attributes, i.e., noun, verb, and adj, providing the part of speech. Different surface forms are specified as values of the surface attribute and different morpho-syntactic properties[15] are specified as values of the category and semantics attributes. For example, the verb forms are differentiated by the combinatorially relevant cat values ns3′ a′ v, n-s3′ a′ v, n′ a′ v, and a′ be, whereby ns3′ indicates a valency position (Herbst et al. 2004, Herbst and Schüller 2008) for a nominative 3rd person singular noun, n-s3′ for a nominative non-3rd person singular noun, n′ for a nominative of any person or number, and a′ for a noun serving as an accusative. They are further differentiated by the sem values pres, past/perf, and prog for tense and aspect.

This method of characterizing variations in lexical meaning by inserting the same concept as a core value into different proplet structures applies also to the word decline:

6.2 LEXICAL ANALYSIS OF decline IN DBS

$$\begin{array}{lll}\textit{intransitive verb} & \textit{transitive verb} & \textit{noun}\\ \begin{bmatrix}\text{sur: decline}\\ \text{verb: decline}\\ \text{cat: n-s3}'\text{ v}\\ \text{sem: pres}\\ \text{mdr:}\\ \text{arg:}\\ \text{prn:}\end{bmatrix} & \begin{bmatrix}\text{sur: decline}\\ \text{verb: decline}\\ \text{cat: n-s3}'\text{ a}'\text{ v}\\ \text{sem: pres}\\ \text{mdr:}\\ \text{arg:}\\ \text{prn:}\end{bmatrix} & \begin{bmatrix}\text{sur: decline}\\ \text{noun: decline}\\ \text{cat: sg}\\ \text{sem: count/mass}\\ \text{mdr:}\\ \text{fnc:}\\ \text{prn:}\end{bmatrix}\end{array}$$

The intransitive and the transitive verb variants are distinguished by the absence versus presence of the a′ valency position in the respective cat values. The verbs and the noun are distinguished by their respective core attributes verb and noun as well as by their cat and sem values. The possible variations of the base form surfaces correspond to those in 6.1.

7 Grammatical Analysis in the Hearer Mode of DBS

Compared to the CCELD 1987 dictionary entries for decline (cf. 1.1), the corresponding DBS proplets 6.2 may seem rather meagre. However, in contrast to dictionary

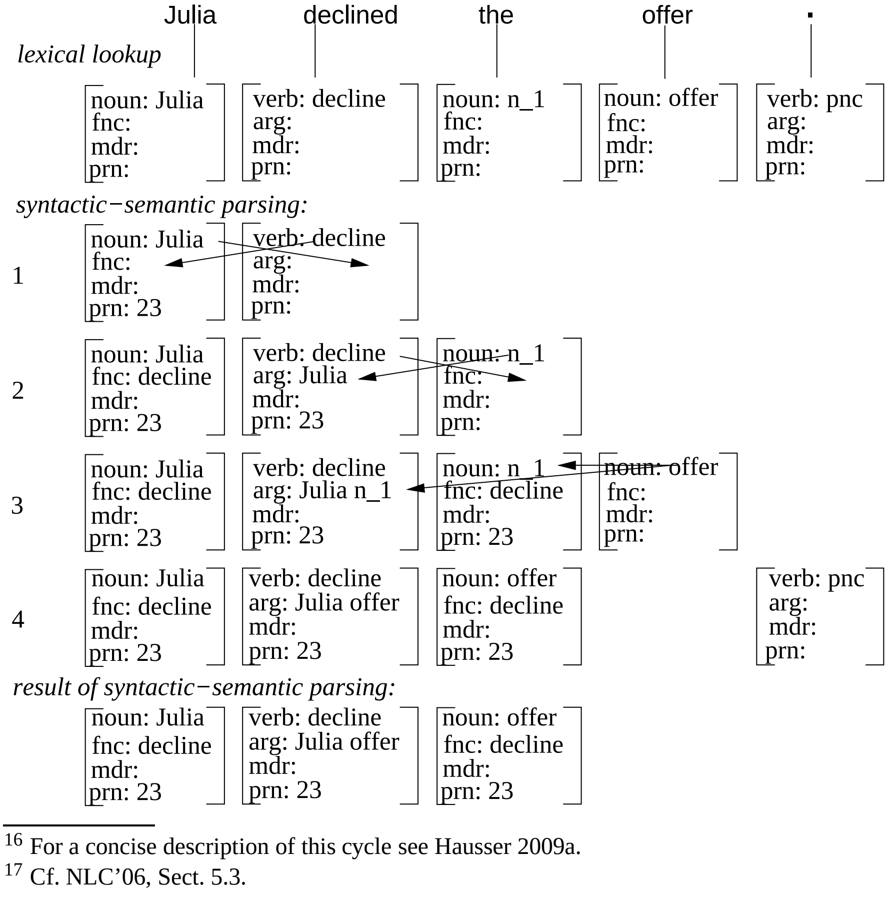
[15] For simplicity, proplets for the genitive singular and plural forms of the noun and any comparative and superlative forms of the adjective are omitted. Also, the attributes nc (next conjunct) and pc (previous conjunct) for the coordination of nouns, verbs, and adjectives have been left out. For a detailed explanation of the lexical analysis in Database Semantics see NLC'06, Sects. 4.1 and 13.1.

entries, proplets are not intended for being read by humans. Instead, proplets are a data structure designed for processing by an artificial agent. The computational processing is of three kinds, (i) the hearer mode, (ii) the think mode, and (iii) the speaker mode. Together, they model the cycle of natural language communication.[16]

In the hearer mode, the processing establishes (i) the semantic relations of functor-argument and coordination structure between proplets (horizontal relations) and (ii) the pragmatic relation of reference between the language and the context level (vertical relations, cf. 3.1). In the think mode, the processing is a selective activation of content in the agent's memory (Word Bank) based on navigating along the semantic relations between proplets and deriving new content by means of inferences.[17] In the speaker mode, the navigation is used as the conceptualization for language production.

Establishing semantic relations in the hearer mode is based solely on (i) the time-linear order of the word form surfaces (ii) and a lexical lookup provided by automatic word form recognition. As an example, consider the syntactic-semantic parsing of Julia declined the offer., based on the DBS algorithm of LA-grammar.

7.1 TIME-LINEAR DERIVATION ESTABLISHING SEMANTIC RELATIONS


Julia
declined
the
offer
.
lexical lookup
noun: Julia fnc: mdr: prn:
verb: decline arg: mdr: prn:
noun: n_1 fnc: mdr: prn:
noun: offer fnc: mdr: prn:
verb: pnc arg: mdr: prn:
syntactic−semantic parsing:
1
noun: Julia fnc: mdr: prn: 23
verb: decline arg: mdr: prn:
2
noun: Julia fnc: decline mdr: prn: 23
verb: decline arg: Julia mdr: prn: 23
noun: n_1 fnc: mdr: prn:
3
noun: Julia fnc: decline mdr: prn: 23
verb: decline arg: Julia n_1 mdr: prn: 23
noun: n_1 fnc: decline mdr: prn: 23
noun: offer fnc: mdr: prn:
4
noun: Julia fnc: decline mdr: prn: 23
verb: decline arg: Julia offer mdr: prn: 23
noun: offer fnc: decline mdr: prn: 23
verb: pnc arg: mdr: prn:
result of syntactic−semantic parsing:
noun: Julia fnc: decline mdr: prn: 23
verb: decline arg: Julia offer mdr: prn: 23
noun: offer fnc: decline mdr: prn: 23

[16] For a concise description of this cycle see Hausser 2009a.

[17] Cf. NLC'06, Sect. 5.3.

The analysis is surface compositional in that each word form is analyzed as a lexical proplet (cf. lexical lookup, here using simplified proplets). The derivation is time-linear, as shown by the stair-like addition of a lexical proplet in each new line. Each line represents a derivation step, based on a rule application. The semantic relations are established by no more and no less than copying values, as indicated by diagonal arrows.[18]

The result of this derivation is a representation of *content* as an order-free set of proplets. Given that the written representation of an order-free set requires some order, though arbitrary, the following example uses the alphabetical order of the core values:

7.2 CONTENT OF Julia declined the offer.

sur:	sur:	sur:
verb: decline	noun: Julia	noun: offer
cat: decl	cat: nm	cat: def sn
sem: past	sem: sg	sem: sg count
arg: Julia offer	fnc: decline	fnc: decline
mdr:	mdr:	mdr:
prn: 23	prn: 23	prn: 23

The proplets are order-free because the grammatical relations between them are coded solely by attribute-value pairs (for example, [arg: Julia offer] in the *decline* proplet and [fnc: decline] in the *Julia* proplet) – and not in terms of dominance and precedence in a hierarchy. As a representation of content, the language-dependent surfaces are omitted. Compared to 7.1, the proplets are shown with additional cat and sem features.

8 Abstract Coding of Semantic Relations

Linguistically, the DBS derivation 7.1 and the result 7.2 are traditional in that they are based on explicitly coding functor-argument (or valency) structure[19] as well as morphosyntactic properties. Given that other formal grammar systems, even within Chomsky's Nativism, have been showing an increasing tendency to incorporate traditional notions of grammar, there arises the question of whether DBS is really different from them. After all, Phrase Structure Grammar, Categorial Grammar, Dependency Grammar, and their many subschools[20] have arrived at a curious state of peaceful coexistence[21] in

[18] For more detailed explanations, especially the function word absorptions in line 3 and 4, see NLC'06 and Hausser (2009a, 2009b).

[19] In addition, the DBS method is well-suited for handling extrapropositional functor-argument structure (subclauses) and intra- and extrapropositional coordination including gapping, as shown in NLC'06, Chapts. 7–9.

[20] Known by acronyms such as TG (with its different manifestations ST, EST, REST, and GB), LFG, GPSG, HPSG, CG, CCG, CUG, FUG, UCG, etc.

[21] This state is being justified by a whole industry of translating between the different grammar systems and proposing conjectures of equivalence. An early, pre-statistical instance is Sells 1985, who highlights the common core of GB, GPSG, and LFG. More recent examples are Andersen, Nioche, Briscoe and Carroll 2008, who propose a treebank based on Dependency Grammar for the BNC, and Liu and Huang 2006 for Chinese; Hockenmaier and Steedman 2007 describe CCGbank as a translation of the Penn Treebank (Marcus, Santorini, and Marcinkiewicz 1993) into a corpus of Combinatory Categorial Grammar derivations; etc., etc.

which the choice between them is more a matter of local tradition and convenience than a deliberate research decision.

DBS is essentially different from the current main stream grammars mentioned above because DBS hearer mode derivations map the lexical analysis of a language surface directly into an order-free set of proplets which is suitable (i) for storage in and retrieval from a database and thus (ii) suitable for modeling the cycle of natural language communication.[22] This would be impossible without satisfying the following requirements:

8.1 Requirements for modeling the cycle of communication

1. The derivation order must be strictly time-linear.
2. Semantic relations must be coded solely by means of attribute-value pairs.
3. Proplets must be defined as non-recursive feature structures.

These DBS requirements are incompatible with the other grammars for the following reasons: (1) and (2) preclude the use of grammatically meaningful tree structures, and as a consequence of (3) there is no place for unification. Behind the technical differences of method there is a more general distinction: the current main stream grammars are *sign*-oriented, whereas DBS is *agent*-oriented.

For someone working in sign-oriented linguistics, the idea of an agent-oried approach may take some getting used to.[23] However, an agent-oriented approach is essential for a scientific understanding of natural language, because the general structure of language is determined by its function,[24] and the function of natural language is communication.

Like any scientific theory, the DBS mechanism of natural language communication must be *verified*. For this, the single most straightforward method is implementing the theory computationally as a talking robot. This method of verification is distinct from the repeatability of experiments in the natural sciences, and may serve as a unifying standard for the social sciences.

Furthermore, once the overall structure of a talking robot (i.e., interfaces, components, and functional flow, cf. 3.1, 5.1, Hausser 2009a) has been determined, partial solutions may be developed without the danger of impeding the future construction of more complete systems.[25] For example, given that the procedural realization of recognition and action is still in its infancy in robotics, DBS currently makes due with English

[22] A formal difference is that LA-grammar is the first and so far the only algorithm with a complexity hierarchy orthogonal to the Chomsky hierarchy (TCS'92).

[23] Also, there seems to be an irrational fear of creating artificial beings resembling humans. Such *homunculi*, which occur in the earliest of mythologies, are widely regarded as violating the tabu of *doppelganger* similarity (Girard 1974). Another matter is the potential for misuse – which is a possibility in any basic science with practical ramifications. Misuse of DBS (in some advanced future state) must be curtailed by developing responsible guidelines for clearly defined laws to protect privacy and intellectual property while maintaining academic liberty, access to information, and freedom of discourse.

[24] This is in concord with Darwin's theory of evolution in which anatomy, for example, will be structured according to functions associated with use.

[25] The recent history of linguistics contains numerous examples of naively treating morphological as well as semantic phenomena in the syntax, pragmatic phenomena in the semantics, etc. These are serious mistakes, some of which have derailed scientific progress for decades.

words as places holders for core values. As an example, consider the following lexical proplets, which are alike except for the values of their sur and noun attributes:

8.2 DIFFERENT CORE VALUES IN THE SAME PROPLET STRUCTURE

$$\begin{bmatrix}\text{sur: squares}\\ \text{noun: }\textit{square}\\ \text{cat: pn}\\ \text{sem: pl}\\ \text{mdr:}\\ \text{fnc:}\\ \text{prn:}\end{bmatrix} \begin{bmatrix}\text{sur: triangles}\\ \text{noun: }\textit{triangle}\\ \text{cat: pn}\\ \text{sem: pl}\\ \text{mdr:}\\ \text{fnc:}\\ \text{prn:}\end{bmatrix} \begin{bmatrix}\text{sur: circles}\\ \text{noun: }\textit{circle}\\ \text{cat: pn}\\ \text{sem: pl}\\ \text{mdr:}\\ \text{fnc:}\\ \text{prn:}\end{bmatrix} \begin{bmatrix}\text{sur: tables}\\ \text{noun: }\textit{table}\\ \text{cat: pn}\\ \text{sem: pl}\\ \text{mdr:}\\ \text{fnc:}\\ \text{prn:}\end{bmatrix} \begin{bmatrix}\text{sur: chairs}\\ \text{noun: }\textit{chair}\\ \text{cat: pn}\\ \text{sem: pl}\\ \text{mdr:}\\ \text{fnc:}\\ \text{prn:}\end{bmatrix} \begin{bmatrix}\text{sur: trees}\\ \text{noun: }\textit{tree}\\ \text{cat: pn}\\ \text{sem: pl}\\ \text{mdr:}\\ \text{fnc:}\\ \text{prn:}\end{bmatrix}$$

These proplets represent a class of word forms with the same morpho-syntactic properties. This class may be represented more abstractly as a proplet pattern.[26]

8.3 REPRESENTING A CLASS OF WORD FORMS AS A PROPLET PATTERN

$$\begin{bmatrix}\text{sur: }\alpha\text{+s}\\ \text{noun: }\alpha\\ \text{cat: pn}\\ \text{sem: pl}\\ \text{mdr:}\\ \text{fnc:}\\ \text{prn:}\end{bmatrix} \text{ where } \alpha \,\varepsilon\, \{\textit{square, triangle, circle, table, chair, tree, ...}\}$$

By restricting the variable α to the core values used in 8.2, the representation 8.3 as a proplet pattern is equivalent to the explicit representation of the proplets class 8.2. Proplet patterns with restricted variables are used for the base form lexicon of DBS, making it more transparent and saving a considerable amount of space.

In concatenated (non-lexical) proplets, the (i) core meaning and (ii) the compositional semantics (based on the coding of morpho-syntactic properties) are clearly separated. This becomes apparent when the core values of a content are replaced by suitably restricted variables, as shown by the following variant of 7.2:

8.4 COMPOSITIONAL SEMANTICS AS A SET OF PROPLET PATTERNS

$$\begin{bmatrix}\text{sur:}\\ \text{verb: }\alpha\\ \text{cat: decl}\\ \text{sem: past}\\ \text{arg: }\beta\ \gamma\\ \text{mdr:}\\ \text{prn: k}\end{bmatrix} \begin{bmatrix}\text{sur:}\\ \text{noun: }\beta\\ \text{cat: nm}\\ \text{sem: sg}\\ \text{fnc: }\alpha\\ \text{mdr:}\\ \text{prn: k}\end{bmatrix} \begin{bmatrix}\text{sur:}\\ \text{noun: }\gamma\\ \text{cat: def sn}\\ \text{sem: sg count}\\ \text{fnc: }\alpha\\ \text{mdr:}\\ \text{prn: k}\end{bmatrix}$$

By restricting the variable α to the values *decline, buy, eat,* or any other transitive verb, β to the values *Julia, Susanne, John, Mary* or any other proper name, and γ to the values *offer, proposal, invitation*, etc., this combinatorial pattern may be used to represent the compositional semantics of a whole set of English sentences, including 7.2.

[26] MacWhinney (2005) describes "feature-based patterns" arising from "item-based patterns," which resembles our abstraction of proplet patterns from classes of corresponding proplets.

9 Collocation

At first glance, 8.4 may seem open to the objection that it does not prevent meaningless or at least unlikely combinations like Susanne ate the invitation, i.e., that it fails to handle collocation (which has been one of Sinclair's main concerns). This would not be justified, however, because the hearer mode of DBS is a recognition system taking time-linear sequences of unanalyzed surfaces as input and producing a content, represented by an order-free set of proplets, as output. In short, in DBS the collocations are in the language, not in the grammar.

The Generative Grammars of Nativism, in contrast, generate tree structures of possible sentences by means of substitutions, starting with the S node. Originally a description of syntactic wellformedness, Generative Grammar was soon extended to include world knowledge governing *lexical selection*. For example, according to Katz and Fodor (1963), the grammar must characterize ball in the man hit the colorful ball as a round object rather than a festive social event. In this sense, Nativism treats collocations as part of the Generative Grammar and Sinclair is correct in his frequent protests against Nativist linguists' modeling their own intuitions instead of looking at "real" language.

In response, Generative grammarians have turned to annotating corpora by hand or statistically (treebanks) for the purpose of obtaining broader data coverage. For example, the U. of Edinburgh and various other universities are known to have syntactically parsed versions of the BNC. The parsers used are the RASP, the Minipar, the Charniak and the IMS parser. Unfortunately, the resulting analyses are not freely available. Yet even if one of them succeeded to achieve complete data coverage (according to some still to be determined standard of wider acceptance) there remains the fact that constituent-structure-based Generative Grammars and their tree structures were never intended to model communication and are accordingly unsuitable for it.

In DBS, the understanding of collocations by natural and artificial agents is based on interpreting (i) the core values and (ii) the functor-argument and coordination structure of the compositional semantics (as in 7.1) – plus the embedding into the appropiate context of use and the associated inferencing. This is no different from the understanding of newly coined phrases (syntactic-semantic neologisms), which are as much a fact of life as are collocations.

The members of a language community utilize the *productivity* of natural language in word formation and compositional semantics to constantly coin new phrases. For example, Republican US Senator Tom Coburn called the stimulus package the largest generational theft bill on record, which was bounced around on CNN for a few days. Or take the creative use of navigate in President Obama has to navigate varying advice on Afghanistan.

Another matter are idioms, such as a blessing in disguise or a drop in the bucket. As frozen non-literal uses, they are either interpretable by the same inferencing as spontaneous non-literal uses (e.g., metaphor, cf. NLC'06, Sect. 5.4) or they must be learned. For example, an ax(e) to grind (German: ein Hühnchen rupfen) may be viewed as similarly opaque (non-compositional or non-Fregean) in syntax-semantics as cupboard is in morphology. Just as cupboard must be equated with kitchen cabinet in the agent's cognition, an axe to grind (attributed to Benjamin Franklin) must be equated with expressing a serious complaint.

10 Context

The attempt of Generative Grammar to describe the tacit knowledge of the speaker-hearer without the explicit reconstruction of a cognitive agent has led not only to incorporating lexical selection into the grammar, but also the *context of use*. Pollard and Sag (1994), for example, propose a treatment of context in HPSG which consists in adding an attribute to lexical entries (see also Green 1997). The values of this attribute are called constraints, and have the form of such definitions[27] as

(a) "*the use of the name* John *is legitimate only if the intended referent is named* John."

(b) "*the complement of the verb* regret *is presupposed to be true.*"

For a meaningful computational implementation this is sadly inadequate, though for a self-declared "sign-based" approach it is probably the best it can do.

Instead of cramming more and more phenomena of language use into the Generative Grammar, Database Semantics clearly distinguishes between the agent-external real world and the agent-internal cognition. The goal is to model the agent, not the external world.[28] Whether the model is successful or not can be verified, i.e., determined objectively, (i) by evaluating the artificial agent's behavior in its interaction with its environment and with other agents and (ii) by directly observing the agent's cognitive operations via the service channel (cf. NLC'06, Sect. 1.4).

In the agent's cognition, DBS clearly separates the language and the context component (cf. 3.1), and defines their interaction via a computationally viable matching procedure based on the data structure of proplets (cf. NLC'06, Sect. 3.2). In addition, DBS implements three computational mechanisms of reference for the sign kinds *symbol, indexical*, and *name*.[29] This is the basis for handling the HPSG context definition (a), cited above, as part of a general theory of signs, whereas definition (b) is treated as an inference by the agent.

For systematic reasons, DBS develops the context component first, in concord with ontogeny and phylogeny (cf. NLC'06, Sect. 2.1). To enable easy testing and upscaling, the context component is reconstructed as an autonomous agent without language. The advantage of this strategy is that practically all constructs of the context component can be reused when the language component is added. The reuse, in turn, is crucial for ensuring the functional compatibility between the two levels.

For example, the procedural definition of basic concepts, pointers, and markers provided by the external interfaces of the context component are reused by the language

[27] These definitions are reminiscent of Montague's (1974) *meaning postulates* for constraining a model structure of *possible worlds*, defined purely in terms of set theory. Supposed to represent spatio-temporal stages of the actual world plus counterfactual worlds with unicorns, etc., a realistic definition or programming of the model structure is practically impossible. Therefore, it is always defined "in principle" only. Cf. FoCL'99, Sect. 20.2.

[28] This is in contrast to the assumptions of truth-conditional semantics, including Montague Grammar, Situation Semantics, Discourse Semantics, or any other metalanguage-based approach. Cf. FoCL'99, Chapts. 19–21; NLC'06, Sect. 2.3.

[29] Cf. FoCL'99, Sect. 6.1; NLC'06, Sect. 2.6. The type-token relation between corresponding concepts at the language and the context level illustrated in 5.1 happens to be the reference mechanism of symbols.

component as the core meanings of symbols, indexicals, and names, respectively. The context component also provides for the coding of content and its storage in the agent's memory, for inferencing on the content, and for the derivation of adequate actions, including language production.

In human-machine communication, the context component is essential for reconstructing two of the most basic forms of natural language interaction. One is telling the artificial cognitive agent what to do, which involves contextual action. The other is the artificial cognitive agent's telling what it has perceived, which involves contextual recognition.

Conclusion

From the linguists' perspective, the learner is for an English learner's dictionary what the artificial cognitive agent is for Database Semantics: each raises the question of what language skills the learner/artificial agent should have.

However, the learner already knows how to communicate in a natural language. Therefore, the goal is to provide her or him with information of how to speak English well, which requires the compilation of an easy to use, accurate representation of contemporary English.

Database Semantics, in contrast, has to get the artificial agent to communicate with natural language in the first place. This requires the reconstruction of what evolution has produced in millions of years as an abstract theory which applies to natural and artificial agents alike.

In other words, Database Semantics must start from a much more basic level than a learner's dictionary. For DBS, any given natural language requires

- automatic word form recognition for the expressions to be analyzed,
- syntactic-semantic interpretation in the hearer mode, resulting in
- content which is stored in a database and
- selectively activated and processed in the think mode, and
- appropriately realized in natural language in the speaker mode.

On the one hand, each of these requirements constitutes a sizeable research and software project. On the other hand, the basic principles of how natural language communication works is the same for different languages. Therefore, once the software components for automatic word form recognition, syntactic-semantic parsing, etc., have been developed in principle, they may be applied to different languages with comparatively little effort.[30]

Because the theoretical framework of DBS is more comprehensive than that of a learner's dictionary, DBS can provide answers to some basic questions. For example, DBS allows to treat basic meanings in terms of recognition and action procedures, phenomena of language use with the help of an explicitly defined context component, and

[30] For example, given (i) an on-line dictionary of a new language to be handled and (ii) a properly trained computational linguist, an initial system of automatic word form recognition can be completed in less than six month. It will provide accurate, highly detailed analyses of about 90% of the word form types in a corpus.

collocations produced in the speaker mode in terms of what the agent was exposed to in the hearer mode. Conversely, a learner's dictionary as a representation of a language is much more comprehensive than current DBS and thus provides a high standard of what DBS must accomplish eventually.

Acknowledgments

This paper benefited from comments by Thomas Proisl, Besim Kabashi, Johannes Handl, and Carsten Weber (CLUE, Erlangen), Haitao Liu (Communication Univ. of China, Beijing), Kiyong Lee (Korea Univ., Seoul), and Brian MacWhinney (Carnegie Mellon Univ., Pittsburgh).

Bibliography

Aho, A.V. and J.D. Ullman (1977) *Principles of Compiler Design*, Reading, MA: Addison-Wesley

Andersen, O., J. Nioche, E. Briscoe and J. Carroll (2008) "The BNC parsed with RASP4UIMA," in *Proceedings of the Sixth Language Resources and Evaluation Conference* (LREC), Marrakech, Morocco.

BNC XML Edition (2007) Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

FoCL'99 = Hausser, R. (1999) *Foundations of Computational Linguistics, Human–Computer Communication in Natural Language, 2nd ed. 2001*, Berlin Heidelberg New York: Springer

Girard, R. (1972) *La violence et le sacré*, Paris: Bernard Grasset

Green, G. (1997) "The structure of CONTEXT: The representation of pragmatic restrictions in HPSG." Proceedings of the 5th annual meeting of the Formal Linguistics Society of the Midwest, edited by James Yoon. Studies in the Linguistic Sciences

Hausser, R. (2009a) "Modeling Natural Language Communication in Database Semantics," in M. Kirchberg and S. Link (eds.), Proceedings of the APCCM 2009, Australian Computer Science Inc., CIPRIT, Vol. 96

Hausser, R. (2009b) "From Word Form Surfaces to Communication," in H. Kangassalo, Y. Kiyoki, and T. Welzer (eds.), *Information Modelling and Knowledge Bases XXI*. Amsterdam: IOS Press Ohmsha

Herbst, T., D. Heath, I. Roe, and D. Goetz (2004) *A Valency Dictionary of English: A Corpus-based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives*, Berlin: Mouton de Gruyter

Herbst, T., and S. Schüller (2008) *Introduction to Syntactic Analysis, A Valency Approach*, Tübingen: Gunter Narr

Hockenmaier, J., and M. Steedman (2007) "CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank" *Computational Linguistics*, Vol. 33.3:355-396

Katz, J.J., and J.A. Fodor (1963) “The Structure of a Semantic Theory,” *Language*, Vol.39:170-210

Kirk, J.M. (1998) “Review of T. McEnery and A. Wilson 1996, and of G. Barnbrook 1996,” *Computational Linguistics*, Vol. 24.2:333-335

L&I’05 = Hausser, R. (2005) “Memory-Based Pattern Completion in Database Semantics,” *Language and Information*, 9.1:69–92, Seoul: Korean Society for Language and Information

Liu, Haitao, and Wei Huang (2006) “A Chinese Dependency Syntax for Treebanking,” *Proceedings of the 20th Pacific Asia Conference on Language, Information, Computation*, p. 126–133, Beijing: Tsinghua Univ. Press

MacWhinney, B. (2008) “How Mental Models Encode Embodied Linguistic Perspective,” in L. Klatzky et al. (eds.) *Embodiment, Ego-Space, and Action*, New York: Psychology Press

MacWhinney, B. (2005) “Item-based constructions and the logical problem,” *Association for Computational Linguistics (ACL)*, 46-54

Marcus, M., B. Santorini, and M. Marcinkiewicz (1993) “Building a large annotated corpus of English: the Penn Treebank,” *Computational Linguistics*, Vol. 19

Montague, R. (1974) *Formal Philosophy*, New Haven: Yale Univ. Press

NLC’06 = Hausser, R. (2006) *A Computational Model of Natural Language Communication: Interpretation, Inference, and Production in Database Semantics*, Berlin Heidelberg New York: Springer

Nichols, J. (1992) *Linguistic Diversity in Space and Time*, Chicago: Univ. of Chicago Press

Pollard, C., and I. Sag (1994) *Head-Driven Phrase Structure Grammar*. Stanford: CSLI

Putnam, H. (1975) “The meaning of “meaning”,” in *Mind, Language and Reality. Philosophical Papers*, Vol. 2:215-271, Cambridge: Cambridge University Press

Sinclair, J. (1987) *Collins COBUILD English Language Dictionary* (Editor in Chief), London and Glasgow: Collins

Sinclair, J. (1991) *Corpus Concordance Collocation*, Oxford: Oxford Univ. Press

Sells, P. (1985) *Lectures on Contemporary Syntactic Theories: An introduction to GB Theory, GPSG, and LFG*, Stanford: CSLI

TCS’92 = Hausser, R. (1992) “Complexity in Left-Associative Grammar,” *Theoretical Computer Science*, 106.2:283-308

Teubert, W. and Krishnamurthy, R. (eds.) (2007). *Corpus Linguistics*. London: Routledge

Teubert, W. (2008) Corpus List, Sat 16/08/2008 12:46

Wierzbicka, A. (1991) *Cross-Cultural Pragmatics: The Semantics of Human Interaction*. Berlin: Mouton de Gruyter