

Topics in Computational Linguistics — Grammar Engineering —

Dan Flickinger

CSLI Stanford & Saarland University

`danf@csli.stanford.edu`

Stephan Oepen

Universitetet i Oslo & CSLI Stanford

`oe@csli.stanford.edu`

<http://lingo.stanford.edu/courses/05/ge/>

The Georgetown Experiment (1954)

Russian is Turned into English by a Fast Electronic Translator

(NYT, January 8, 1954)

The switch is assured in advance by attaching the rule sign 21 to the Russian 'gyeneral' in the bilingual glossary which is stored in the machine, and by attaching the rule-sign 110 to the Russian 'mayor'. The stored instructions, along with the glossary, say whenever you read a rule sign 110 in the glossary, go back and look for a rule-sign 21. If you find 21, print the two words that follow it in reverse order.

(Journal of Franklin Institute, March 1954)

- First public MT demonstration: IBM – GU cooperation (Lèon Dostert);
- somewhat limited scale: 250 words, six 'syntax' rules → high hopes.



Transformational Grammar (1950s –)

Noam Chomsky (Associate Professor, MIT)

- *Syntactic Structures* (1957): precise formalization of hypotheses;
- *Formal Properties of Grammars* (1963): formal language theory;
- *Aspects of the Theory of Syntax* (1965): postulate ‘deep structure’;
- *Lectures on Government and Binding* (1981): universal grammar;
- *The Minimalist Program* (1995): ‘economy’ of representations.

- Contrary to original desiderata, TG theories rather hard to evaluate;
- typically many empty nodes in surface trees (few implementations);
- concerns about computational tractability (Peters & Ritchie, 1987).



Transformational vs. Generative Grammar

The term ‘generative grammar’ is sometimes also used to refer to ‘generative-transformational grammar’ [...]. There is no small irony in this usage, given the characteristic practice of transformational work of the last two decades. This body of literature, though technical in appearance, has systematically eschewed the development of consistent, broad-coverage systems whose predictions can be verified empirically. This stands in marked contrast to the formal and analytic precision that is emblematic of most of the nontransformational generative approaches [...], whose mathematical properties [...] have in many cases been explored in detail and whose empirical coverage has frequently been tested in terms of large-scale computational grammars.

(Sag, Wasow, & Bender, 2003, page 525)



Constraint-Based Lexicalism (1 of 2)

Lexical Functional Grammar (Bresnan, 1979)

- Grammatical functions as first-place concepts: functional structure;
- non-transformational: enriched lexicon plus productive lexical rules;
- projection set-up: c- and f-structure jointly determine grammaticality.

Generalized Phrase Structure Grammar (Gazdar, 1981)

- WYSIWIG: no 'deep' structure and transformations; surface-driven;
- computationally restrictive: context-free formalism (with meta-rules);
- rich use of features and feature co-occurrence principles: unification.



Constraint-Based Lexicalism (2 of 2)

Head-Driven Phrase Structure Grammar (Pollard & Sag, 1987)

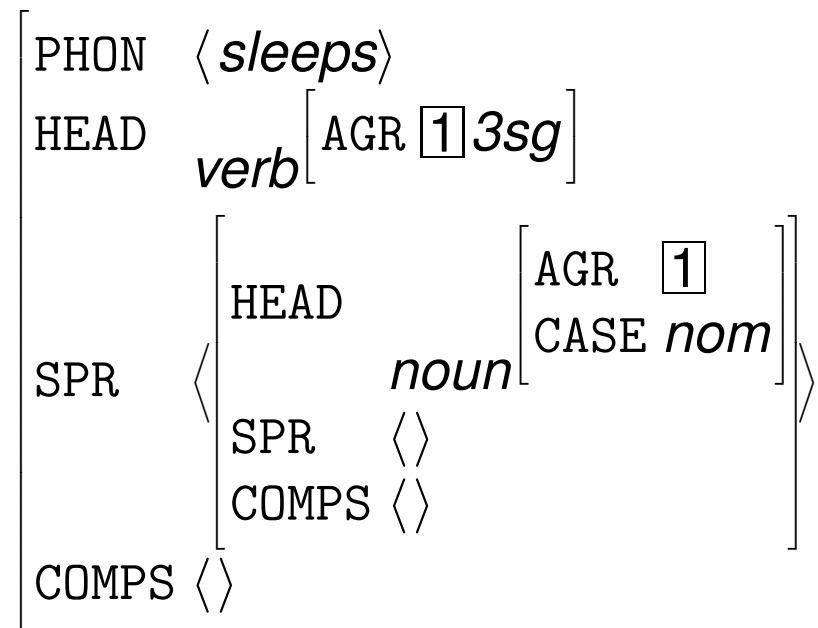
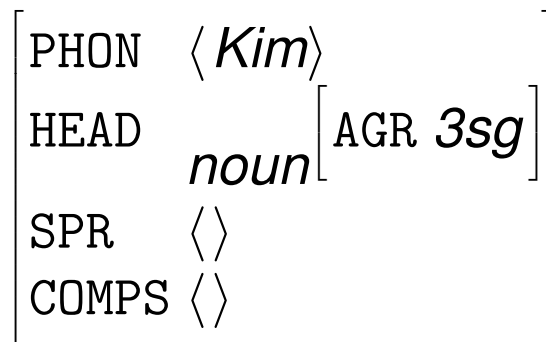
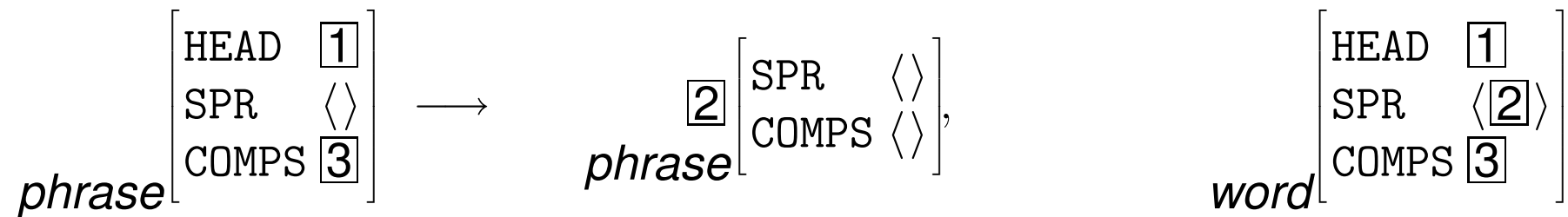
- Immediate successor to GPSG (Natural Language Project at HPL);
- headedness: importance of (lexical) head in most constructions;
- rich, hierarchically-organized lexicon (as typed feature structures);
- small inventory of schematic phrase structure rules for combination;
- *sign*-based approach: tight integration of syntax and semantics.

Grammar Writing

- Since early 1980s, grammar engineering as a task in its own right;
- HP Labs, (Xerox) PARC, SRI (Menlo Park), CSLI among centers.



Interaction of Lexicon and Phrase Structure Schemata



Third Generation of Broad-Coverage Grammars

Characteristics

- Multi-dimensional, strongly hierarchical encoding of knowledge;
- declarativity and reversibility: support both parsing and generation;
- focus on engineering methodologies and processing efficiency;
- multi-lingual development, often distributed accross several sites.

Examples

- ParGram (PARC, XRCE Grenoble, Stuttgart, Fuji Xerox, Bergen);
- DELPH-IN (CSLI, Saarbrücken, Tokyo, Cambridge, Sussex, UiO);
- F|XTAG (University of Pennsylvania, Université Paris 7, Korea).



HPSG Parsing — Then and ‘Now’

Version	Platform	Test Set	filter %	etasks ϕ	pedges ϕ	tcpu ϕ (s)	space ϕ (kb)
October 1996	PAGE	<i>'tsnlp'</i>	49.9	656	44	4.77	19,016
		<i>'aged'</i>	51.3	1763	97	36.69	79,093
August 1999	PET	<i>'tsnlp'</i>	93.9	170	55	0.03	333
		<i>'aged'</i>	95.1	753	292	0.14	1,435
		<i>'fuse'</i>	95.5	3084	1140	0.65	10,589

(generated by [incr tsdb()] at 5-nov-1999 (21:23 h))

Cumulative Break-Through in Parsing Efficiency

- Oldest comparable profiles: net speed-up of around 260 (excluding gc);
- grammar evolution: problem size (in edges) increased by factor of three;
- additional factors (hardware, packing): above four orders of magnitude.



Unification-Based Processing Underway to Dot Com

Tutorial at the 38th Annual Meeting of
the ACL (Hong Kong, October 2000)

Dan Flickinger

CSLI, Stanford University
and YY Software Corporation
dan@csli.stanford.edu

Stephan Oepen

Saarland University
and YY Software Corporation
oe@coli.uni-sb.de

Deep Linguistic Processing with HPSG: DELPH-IN

Set-Up

- Loosely organized group of institutions and interested individuals;
- rooted in 'linguistic' NLP but geared towards practical applications;
- DELPH-IN resources in wide use: research, education, applications.

History

- Originally CSLI Stanford – Saarland University (VerbMobil, 1994);
- Tokyo University (Tsuji Laboratory): logic compilation techniques;
- Cambridge and Sussex Universities: efficient and accurate NLP;
- Trondheim and Oslo Universities, NTT: additional languages, MT.



Common Background Assumptions and Goals

Linguistic Framework and Descriptive Formalism

- Established frameworks: HPSG and Minimal Recursion Semantics;
- mutual feedback between theory development and implementation;
- convergence on typed feature structure formalism (details below).

Joint Goals

- Establish joint repository of grammars, processors, reference data;
- advance theory building and implementation through synergy;
- reusability: make as much as possible available in open-source;
- devise efficient and flexible technology for practical applications.



Collaborative Research on HPSG Processing

Common Reference Formalism

- Strongly typed, conjunctive, closed world typed feature structure logic;
- blend of [Carpenter, 1992], [Copestake, 1992], and [Krieger, 1995].

Engineering and Processing Environments

- LKB: grammar development environment (Lisp) [Copestake, 2002];
- PET: efficient, industry quality runtime engine (C⁺⁺) [Callmeier, 2000];
- [incr tsdb()]: competence and performance profiler [Oepen, 2000].

Common Grammars on Multiple Platforms

- English (CSLI): LinGO English Resource Grammar; Dan Flickinger et al.
- Japanese (DFKI, NTT Research): Melanie Siegel & Emily M. Bender.
- Norwegian, Italian, Korean, Greek, Portuguese, Spanish are underway.



Choice of Descriptive Formalism

Desiderata

- **Declarativity** clear separation of declarative and procedural knowledge: (exact) same grammar is used for parsing and generation;
- **Adequacy** non-redundant account of linguistic generalizations;
- **Efficiency** build on known techniques, aiming for near real-time.

Design Decisions

- Unification of typed feature structures is the *one* central operation;
- multiple inheritance, strict appropriateness, complex constraints;
- closed world; no disjunctive, implicational, or relational constraints



The LinGO English Resource Grammar (ERG)

Development Background (1993 – today)

- General-purpose, wide-coverage, computational English grammar;
- mainly Dan Flickinger, with Rob Malouf, Emily M. Bender, Jeff Smith;
- supported in multiple HPSG processing environments (LKB & PET);
- range of projects: speech, email, UseNet (no WSJ); one company.

Design

- HPSG [Pollard & Sag 1994]: constraint-based, strongly lexicalized;
- MRS [Copestake et al., 1999]: flat, Davidsonian, underspecified;
- type hierarchies defining principles, lexical classes, constructions;
- strict grammaticality assumption: generator using same grammar.



LinGO ERG: Coverage and Size

Linguistic Coverage

- 85 % of 12,000 transcribed dialogue turns from VerbMobil domains;
- 80⁺ % of customer emails in financial and ecommerce domains;
- both fairly short utterances: average 9 words, ranging from 1 – 40;
- 88 % of phenomena-based examples in Hewlett Packard test suite.;
- more recently, 95 % on excerpts from tourism brochures (13 words).

Size of Grammar (as of October 2004)

- some 2,600 types for fundamentals, lexicon, rules, and semantics;
- 12,358 lexical entry stems (around 3,800 verbs and 6,300 nouns);
- 27 lexical (15 inflectional) rules and 96 phrase structure schemata.



Sample Data (LOGON Domain) Analyzed by LinGO English Grammar

- 1 *Be considerate of game, farm animals and other hikers.*
- 109 *Kjeragveggen has interested climbers since the 1970s.*
- 304 *But there are things to do for those with knickers and anoraks too.*
- 39 *Follow the road past NUTEC and continue up along Kvarvenveien, past the recreation area.*
- 248 *The first part of the trip goes with the Hurtigruta to Torvik, with a bicycle ride at night into the sunrise out to Runde, and a hike to Norway's southernmost bird mountain.*
- 326 *If there is one thing Swedes are concerned with, it is preparing delicious dishes.*



Grammatical Coverage on Tourism Excerpts

'lingo/08-nov-03/hike/03-11-14/pet'						
Aggregate	total items #	word string ϕ	lexical items ϕ	parser analyses ϕ	overall coverage %	time (s) ϕ
$35 \leq i\text{-length} < 40$	1	35.00	109.00	2372.00	100.0	2.00
$30 \leq i\text{-length} < 35$	2	32.50	109.00	1768.00	100.0	1.12
$25 \leq i\text{-length} < 30$	7	26.71	100.57	1393.14	100.0	1.32
$20 \leq i\text{-length} < 25$	28	21.68	78.36	931.93	100.0	0.52
$15 \leq i\text{-length} < 20$	72	16.89	54.08	136.18	93.1	0.26
$10 \leq i\text{-length} < 15$	119	11.77	39.85	35.87	95.0	0.10
$5 \leq i\text{-length} < 10$	95	7.47	23.49	5.79	93.7	0.04
$0 \leq i\text{-length} < 5$	6	4.00	7.67	1.33	100.0	0.01
Total	330	12.86	42.85	177.17	94.8	0.19

(generated by [incr tsdb()] at 14-nov-2003 (22:49 h))

