

Hybridlogik og natursprogsgenkendelse

Anders Søgaard
Center for Language Technology
Njalsgade 80
DK-2300 Copenhagen
anders@cst.dk

April 4, 2006

- Lidt om det universelle genkendelsesproblem og grammatikkers extension.
- En simpel unifikationsbaseret grammatik, og hvordan den forholder sig til de lingvistiske teorier, der anvendes i dag.
- Tre forskellige slags logikker, eksemplificeret v. $H(\Downarrow)$, $HDL_r(\Downarrow)$ og PDL^+ .
- Om afgørlighed og $NPTIME$.

0.1 Introduktion til matematisk lingvistik

Mål: En grammatikformalisme, der beskriver mængden af mulige natursprogsgrammatikker. Mængden begrænses "nedefra" af ekspressivitet (trad. i forh. til Chomskyhierarkiet), og "ovenfra" af kompleksitet og lærbarhed. Her tales om universel genkendelse og problemets kompleksitet. Målet her er traditionelt PTIME eller NPTIME.

- (1) He is **both** a painter **and** a linguist.
- (2) mer em Hans.DAT es huus.ACC hälfed.DAT aastriiche.ACC.
- (3) govel-i.NOM igi.NOM sisxl-i.NOM saxl-isa-j.GEN.NOM m-is.GEN
Sail-is-isa-j.GEN.GEN.NOM ('all (the) blood.NOM (of the) house.GEN
(of) Saul.GEN)
- (4) dat Jan Piet Marie Fred^k (horde leren^k uitnodigen)⁺ en zag leren^k
omhelzen.

0.2 Natursprogsenkendelse

Sprog $L(G)$ defineret i forhold til en grammatik G :

$$\begin{aligned} L(G) &= \{\sigma \mid G \vdash \sigma\} \\ \text{eller } L(G) &= \{\sigma \mid \sigma \models G\} \end{aligned}$$

Eller:

$$L(G) = \{\sigma \mid \exists M. \models_M \sigma \wedge G\}$$

Definition 0.1 (Universel genkendelse). En grammatisk *teori* beskriver en mængde grammatikker, defineret over et vokabularium A . Givet en grammatik i den mængde G , og en streng $\sigma \in A^*$: $\sigma \in L(G)$?

Example 0.2 (Kontekstfri grammatik). En kontekstfri grammatik består af et vokabularium, en mængde non-terminale symboler, genskrivningsregler og et designeret startsymbol. Her reglerne:

- $A \rightarrow a, A$
- $A \rightarrow a, b$

Lad A være startsymbolet. $L(G) = \{a^+b\}$. Universel genkendelse for kontekstfri grammatikker kan afgøres i PTIME ($\mathcal{O}(n^3)$). Altså: Uanset hvor kompleks en grammatik og hvor kompleks en streng, kan det, hvis grammatikken er kontekstfri, afgøres, om strengen tilhører grammatikkens ekstension (sproget), i tiden $k \times n^3$, hvor k er en konstant, og n er strengens længde.

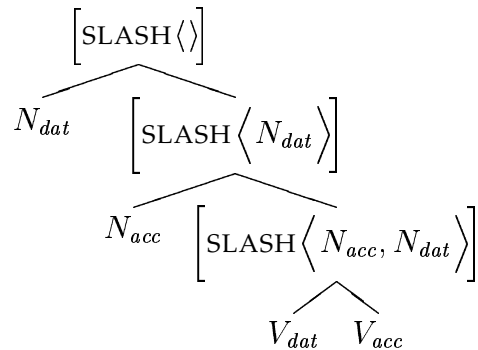
0.3 Unifikationsbaseret grammatik

I unifikationsbaseret grammatik repræsenteres syntaktisk og semantisk information som trækstrukturer, og den centrale operation er unifikation. I simple teorier erstatter trækstrukturer blot syntaktiske kategorier.

Example 0.3 (Unifikationsbaseret grammatik I). Her en udvidet kontekstfri grammatik:

- $\left[\begin{array}{l} \text{TYPE } non\text{-terminal} \\ \text{CAT } \#1A \end{array} \right] \rightarrow \left[\begin{array}{l} \text{TYPE } terminal \\ \text{PHON } a \end{array} \right], \left[\begin{array}{l} \text{TYPE } non\text{-terminal} \\ \text{CAT } \#1 \end{array} \right]$
- $\left[\begin{array}{l} \text{TYPE } non\text{-terminal} \\ \text{CAT } A \end{array} \right] \rightarrow \left[\begin{array}{l} \text{TYPE } terminal \\ \text{PHON } a \end{array} \right], \left[\begin{array}{l} \text{TYPE } terminal \\ \text{PHON } b \end{array} \right]$

“#1” betyder, at de to værdier er identiske. Hvis man for eksempel vil sørge for kongruens, kan et CASE-træk tilføjes, og trækværdierne kan kræves at være identiske. Unifikationsbaserede grammatikker er mere ekpressive end kontekstfri grammatikker; se f.eks.



De er også rigere end mildt kontekst-sensitive formalismer. *Suffixaufnahme* (Oldgeorgisk) og hollandsk koordination beskrives let vha. tokenidentitet (reentrancy).

Example 0.4 (Unifikationsbaseret grammatik II). Her en model-teoretisk udgave af samme grammatik:

$$\left[\text{TYPE } \textit{non-terminal} \right] \Rightarrow \left[\begin{array}{l} \text{CAT A } \mathbb{I} \\ \text{LEFT } \left[\begin{array}{l} \text{TYPE } \textit{terminal} \\ \text{PHON a} \end{array} \right] \\ \text{RIGHT } \left[\begin{array}{l} \text{TYPE } \textit{non-terminal} \\ \text{CAT } \mathbb{I} \end{array} \right] \end{array} \right] \vee$$

$$\left[\begin{array}{l} \text{CAT A} \\ \text{LEFT } \left[\begin{array}{l} \text{TYPE } \textit{terminal} \\ \text{PHON a} \end{array} \right] \\ \text{RIGHT } \left[\begin{array}{l} \text{TYPE } \textit{terminal} \\ \text{PHON b} \end{array} \right] \end{array} \right]$$

Her er intet egentlig startsymbol, så hvad kræver natursprogsgenkendelse? Svaret er: (i) en logisk beskrivelse af σ , (ii) et eksistentielt postulat, at der findes en *non-terminal* rod, og (iii) en logisk beskrivelse af velformede trækstrukturer, herunder et krav, at modellen (trækstrukturen) er forbundet.

0.4 Samtidige model-teoretiske formalismer

En *trækstrukturgrammatik med nedrivning* (TN) består af et hierarki af træstrukturer med token-identiteter og aksiomer med global kvantifikation. Kategorial unifikationsgrammatik og konstruktionsgrammatik er TNER. HPSG findes i forskellige udgaver. Uden leksikalske regler, mængder, og hvis relationelle og funktionelle afhængigheder simuleres, er HPSG en TN. LFG, PATR-II og GPSG er *generative* unifikationsbaserede formalismer.

0.5 Grammatik og logik

Model-teoretiske grammatikker er altså logiske teorier. En trækstruktur, som f.eks.

$$\begin{bmatrix} A a \\ B \perp \\ C [B \perp] \end{bmatrix}$$

kan beskrives i førsteordenslogik

$$\exists x, y, z, v R_A(x, y) \wedge a(y) \wedge R_B(x, z) \wedge R_C(x, v) \wedge R_B(v, z)$$

Theorem 0.5. *Model-checking for førsteordenslogik kan afgøres i PSPACE, men ikke i PTIME. En PTIME-algoritme kræver, at antallet af assignments begrænses til n .*

0.6 Egenskaber

En passende logik har globalt udsyn (regler, principper, binding) og en måde at påtvinge token-identitet (agreement, linking). Det betyder, at logikken, vi leder efter, *ikke* er invariant under "disjoint unions" eller "generated substructures"; og det betyder også, at logikken ikke har "tree model property." Det er altså klart, at der er brug for en rigere logik end, for eksempel, almindelig modallogik.

0.7 Hybridlogik

Definition 0.6 (H(\Downarrow)). Over Kripke modeller $\langle W, R, V \rangle$. $\text{PROP} \cap \text{NOM} = \emptyset$.
Syntaks:

$$\phi \doteq i | p | \phi \wedge \psi | \neg \phi | \langle \alpha \rangle \phi | @_i \phi | \Downarrow x. \phi$$

Semantik:

$$\begin{array}{lll} M, w, g \models i & \text{hviss} & V(i) = \{w\} \\ M, w, g \models p & \text{hviss} & w \in V(p) \\ M, w, g \models \phi \wedge \psi & \text{hviss} & M, w, g \models \phi \& M, w, g \models \psi \\ M, w, g \models \neg \phi & \text{hviss} & M, w, g \not\models \phi \\ M, w, g \models \langle \alpha \rangle \phi & \text{hviss} & \exists w'. R_\alpha(w, w') \& M, w', g \models \phi \ (R_\alpha \in R) \\ M, w, g \models @_i \phi & \text{hviss} & \exists w'. M, w', g \models \phi \& V(i) = \{w'\} \\ M, w, g \models \Downarrow x. \phi & \text{hviss} & \exists w', g'. g' \stackrel{x}{=} g \& g'(x) = w \& M, w', g' \models \phi \end{array}$$

$$ST_x(\Downarrow y. \phi) = \exists y, z. (x = y \wedge ST_z(\phi))$$

Andre nyttige operatorer:

$$\begin{array}{ll} \exists x. \phi & \doteq \Downarrow z. \Downarrow x. (z \wedge \phi) \\ E(\phi) & \doteq \Sigma x. \phi \text{ hvor } x \text{ ikke optræder i } \phi \\ \Sigma x. \phi & \doteq \Downarrow z. \Downarrow x. (x \wedge \phi) \end{array}$$

Definition 0.7 ($\text{HDL}(\Downarrow)_r$). Ligesom Definition 0.6, men kvantorindlejring er begrænset til r . Derudover tilføjes non-deterministisk Kleene-star, i.e. R_α^* forkorter $\bigcup_n (R_\alpha)^n$, hvor n er non-deterministisk valgt.

Definition 0.8 (PDL^+). Ligesom Definition 0.7, men NOM fjernes og i stedet tilføjes $\langle \alpha \cap \beta \rangle \phi$, hvor

$$M, w \models \langle \alpha \cap \beta \rangle \phi \quad \text{hvis} \quad \exists w' R_\alpha(w, w') \wedge R_\beta(w, w') \& M, w' \models \phi$$

PDL^+ svarer til PDL udvidet med E og \cap .

0.8 Grammatik i hybridlogik

Husk:

“... (i) en logisk beskrivelse af σ , (ii) et eksistentielt postulat, at der findes en *non-terminal*, og (iii) en logisk beskrivelse af velformede trækstrukturer, herunder et krav, at modellen (trækstrukturen) er forbundet.”

Her i $\text{HDL}_r(\Downarrow)$:

- (i) $\phi \wedge \langle \prec \rangle (\psi \wedge \langle \prec \rangle \dots)$
- (ii-iii) (a) $\neg \Sigma x. \langle + \rangle x$ (acyklisk)
- (b) $\exists x. A(\langle * \rangle^{-1} x) \wedge A(x \rightarrow \neg \langle \alpha \rangle^{-1} \top)$ (sammenhængende og med rod/start)
- (c) $\neg \Sigma x. \exists y, z. \langle \alpha \rangle y \wedge \langle \alpha \rangle z \rightarrow \neg @_y z$ (deterministiske træk)

I $\text{H}(\Downarrow)$ må defineres en transitiv og irrefleksiv relation R_{path} , der subsumerer alle andre relationer, så f.eks. $\neg \Sigma x. \langle path \rangle x$ (acyklisk), eftersom mastermodaliteten ikke kan defineres. I PDL^+ kan disse egenskaber blandt andet kodes således:

$$\begin{array}{ll}
 \text{Det.} & \langle \alpha \rangle \phi \wedge \langle \alpha \rangle \psi \rightarrow \langle \alpha \rangle (\phi \wedge \psi) \\
 \text{Acykl.} & \neg \langle \alpha \cap (\alpha; \beta; *) \rangle \top \\
 \text{Samm.} & E(\text{root} \wedge \neg \langle \alpha \rangle^{-1} \top) \wedge A(\langle * \rangle^{-1} \text{root}) \wedge \\
 & A(\text{root} \rightarrow [+]\neg \text{root})
 \end{array}$$

Der er også andre ting, der skal kodes, f.eks. typer (e.g. $\text{sign} \rightarrow \langle \text{CAT} \rangle \top$) og typehierarkiet (Boolesk).

0.9 Egenskaber (II)

Theorem 0.9. $H(\Downarrow)$ og $HDL(\Downarrow)_r$ er invariante under total bisimulering-med-navne, og PDL^+ er invariant under power bisimulering.

Proof. $H(E)$ er invariant under total bisimulering-med-navne [tC05]. \exists er “safe” for total bisimulering-med-navne (?), og $*$ er (endda) “safe” for bisimulering. For PDL^+ , see [vB03]. \square

Theorem 0.10. $H(\Downarrow)$, $HDL(\Downarrow)_r$ og PDL^+ er uafgørlige.

Proof. Se [BS95] for $H(\Downarrow)$. Heraf følger $HDL(\Downarrow)_r$'s uafgørlighed (lad r være det antal indlejringer, det kræver at definere et uafgørligt problem i $H(\Downarrow)$). PDL^+ kan vises ved reduktion til recurrent tiling (høj uafgørlighed). \square

Theorem 0.11. $H(\Downarrow)$, $HDL(\Downarrow)_r$ og PDL^+ har model-checking-problemer, der kan løses i PTIME.

Proof. Se [FdR06] og [Lan06]. \square

0.10 Polysize model property

Definition 0.12 (Polysize model property). En logik har pmp, hvis $M \models \phi$ betyder, at der findes en model M' , så $|M'| \leq (k \times |\phi|^c)$ og $M' \models \phi$.

Lemma 0.13. Hvis Λ (en konsistent, normal modallogik) har pmp, og hvis model-checking kan afgøres i PTIME, er Λ afgørlig i NPTIME, i.e. Λ er NP-komplet.

Theorem 0.14. Unifikationsbaserede grammatikker har, hvis alle unære projektioner er acykliske, pmp.

Proof. Hvis $M \models \phi$, findes der en M' , så $|M'| \leq (2|\sigma| - 1) \times (u + 1) \times \mathbf{paths}$. $(2|\sigma| - 1)$ er den maksimale størrelse på et træ uden unære projektioner, u er antallet af unære regler/fraser/typer, og $\mathbf{paths} = \{\pi \in \mathbb{L} \mid \pi \text{ indeholder intet label to gange}\}$. \square

Corollary 0.15. Unifikationsbaserede grammatikker er, hvis alle unære projektioner er acykliske, NP-komplette.

0.11 Udsigt

NP-komplethedensbeviset er kun interessant, fordi de kendte unifikations-baserede grammatikker, der har tilsvarende kompleksitet, f.eks. [Tra95], er mindre ekspressive. F.eks. kan der i $\text{HDL}(\Downarrow)_r$ og PDL^+ skrives grammatikker med fri ordfølge, scrambling og "discontinuous constituency". Ortogonal nedrivning, "functional uncertainty", og subsumption defineres let. NP-komplethedensresultatet er opnået under et model-teoretisk perspektiv, der også bærer nogle fordele med sig, e.g. et mere fleksibelt grammatikalitetsbegreb, et åbent leksikon og fuld deklarativitet.

Ekspressivitet	[our result]	[Tra95]
fri WO	✓	
scrambling	✓	
disc.const.	✓	
detours		
funct.unc.	✓	
subsump.	✓	
Egenskaber	[our result]	[Tra95]
åbent leksikon	✓	
deklarativitet	✓	

References

- [BS95] Patrick Blackburn and Jerry Seligman. Hybrid languages. *Journal of Logic, Language and Information*, 4:251–272, 1995.
- [FdR06] Massimo Franceschet and Maarten de Rijke. Model checking for hybrid logics. *Journal of Applied Logic*, 2006. In press.
- [Lan06] Martin Lange. Model checking propositional dynamic logic with all extras. *Journal of Applied Logic*, 4:39–49, 2006.
- [tC05] Balder ten Cate. *Model theory for extended modal languages*. PhD thesis, University of Amsterdam, Amsterdam, the Netherlands, 2005. ILLC Dissertation Series DS-2005-01.
- [Tra95] Marten Trautwein. *Computational pitfalls in tractable grammar formalisms*. PhD thesis, University of Amsterdam, Amsterdam, the Netherlands, 1995. ILLC Dissertation Series DS-1995-15.
- [vB03] Johan van Benthem. Logic and game theory. In G. Mints and R. Muskens, editors, *Games, logic, and constructive sets*, volume 161 of *CSLI Lecture Notes*, pages 3–22. CSLI Publications, Stanford, California, 2003.