A General Linearization Model (GLM) for NLG

Ciprian Gerstenberger

Department of Computational Linguistics, Saarland University

19th Dec 2005
Outline

Phenomena

Linearization

Mereology

Model

Worked Examples

Further ideas
Outline

Phenomena

Linearization

Mereology

Model

Worked Examples

Further ideas
Outline

Phenomena

Linearization

Mereology

Model

Worked Examples

Further ideas
Outline

Phenomena
Linearization
Mereology
Model
Worked Examples
Further ideas
Outline

Phenomena

Linearization

Mereology

Model

Worked Examples

Further ideas
Outline

Phenomena

Linearization

Mereology

Model

Worked Examples

Further ideas
Psycholinguistics

1. speech errors
   - *multimodal output* → *mutlimodal ouptut*
   - *rote Türen* → *tote Rüren*
   - *snow flurries* → *flow snurries*
   - *self-destruct instruction* → *self-instruct destruction*
   - *writing a letter to my mother* → *writing a mother to my letter*

2. spoonerisms
   - *our deer old queen* → *our queer old dean*
   - *Kentucky Fried Chicken* → *Mirco Nontschew läßt grüßen!*

3. children’ secret languages
   - *Was ist denn das?* → *Wabasibistdebenndabas?*
Comparative Linguistics

- **metathesis**
  - *brid* (Old English) vs. *bird* (modern English)
  - *brennen* (German) vs. *burn* (English)
  - *Born* (German) vs. *Brunnen* (German)/ *bron* (Dutch)
  - *Ross* (German) vs. *horse* (English)
  - *Warze* (German) vs. *wrat* (Dutch)
  - *Presse* (German)/ *press* (English) vs. *pers* (Dutch)
  - *Kreuz* (German) vs. *kors* (Norwegian)
  - *Roland* (French) vs. *Orlando* (Italian)

- **tmesis**
  - *whatsoever place* → *what place soever*
  - *absolutely* → *abso-bloody-lutely*
  - *fantastic* → *fan-f***ing-tastic*
Clitics (in Romanian [1])

The indefinite article
- un sat frumos [a nice village]
- un frumos sat

The definite article
- satul frumos [the nice village]
- frumosul sat
Clitics (in Romanian [1])

The definite article

- satul frumos [the nice village]
- frumosul sat
- satul meu frumos [my nice village]
- frumosul meu sat
- satul frumos al meu
- frumosul sat al meu
- al meu sat frumos
- al meu frumos sat
Clitics (Romanian [2])

Weak pronominal forms

- *te rog să îmi dai server-ul/serverul*
  [(I) beg you to give me the server. Please give me the server.]
- *te rog să-mi dai server-ul/serverul*
- *te rog să mi-l dai*
  *te rog să-mi-l dai*
  *te rog să-mi îl dai*
  *te rog să îmi îl dai*
- *dă-mi server-ul/serverul*
  [*dă îmi server-ul]*
- *dă-mi-l*
  [*dă mi-l |
  *dă-mi îl |
  *dă îmi îl]*
- *mi l-ai dat*
  [*mi-l ai dat |
  *mi îl ai dat |
  *îmi îl ai dat*]
Clitics (Romanian [3])

Clitic clustering

- *mi l-ai dat*
- *dǎ-mi-l*
- *vedea-le-aş mari şi sănătoase*
- *vedea-ţi-le-aş mari şi sănătoase*
- *vedea-ţi-i-aş mari şi sănătoşi*

A: *Copiii fetei mele sînt încă mici.*
B: *Vedea-i-i-ai mari şi sănătoşi!*
Clitics (Polish [1])

The particle śmy [1st.Pl]

Q: Zrobiliście to?
[Did you do this?]

A1: Nie zrobiłśmy tego. (canonical)
[We didn’t do this.]

A2: (My) tego nie zrobiłśmy. (canonical)

A3: Tego (my) nie zrobiłśmy.

A4: Tegośmy nie zrobił.

A5: Myśmy tego nie zrobił.

A6: Tego myśmy nie zrobił.
The particle śmy [1st.Pl]

**Q:** Byliście tam pierwszego maja?  
[Were you there on the 1st of May?]

**A1:** Pierwszego tam nie byliśmy.  
[We were not there on the 1st of May.]

**A2:** Pierwszego nie byliśmy tam.  

**A3:** Tam pierwszego nie byliśmy.  

**A4:** Pierwszegośmy tam nie byli.  

**A5:** Pierwszego żeśmy tam nie byli. (coll.)
The particle śmy [1st.Pl]

Q: *Widzieliście białego królika?*  
[Did you see the white rabbit?]

A1: *Białego królika śmy nie widzieli.*  
[We didn’t see the white rabbit.]

A2: *Białego śmy królika nie widzieli.*
Phenomena in German

- **compound coordination in German**
  - be- und entladen
  - **Schweins- und Kalbsbraten**
  - *Das ist nicht nur eine Geld-, sondern auch eine Platzfrage.*
  - Originaltexte und -melodien sind nicht erhalten.

- **particle verbs in German**
  - *Maria machte die Tür auf.*
  - *Auf machte Maria die Tür.*
  - *Maria wollte die Tür aufmachen.*
  - *Maria versuchte die Tür aufzumachen.*

What about the orthography of particle verbs?

- *Maria möchte gerne radfahren vs. gerne Rad fahren*
- *Maria ist radgefahren vs. Rad gefahren*
Discontinuous Constituents

- partial fronting:
  - *Ein Buch hat Maria gelesen.*

- extraposition:
  - *Ein Buch hat Maria gelesen, das von ihm war.*
  - *Ich möchte eine Playlist erstellen mit drei Liedern.*

- split NPs:
  - *Rote Äpfel habe ich drei gegessen.*
Topological Field Models

Macro-structure of Germanic languages

- capturing the generalization of TFM to restrict linearization when necessary
- coping with special phenomena
  - partial fronting
  - scrambling in the Mittelfeld
  - extraposition
  - multiple fronting
Polysynthetic Languages

An example from Western Greenlandic, a polysynthetic and agglutinating (but not incorporating) language:

*Aliikusersuillammassuaanerartassagaluarpaalli.*

aliiku-sersu-i-llammas-sua-a-nerar-ta-ssa-galuar-paal-li

entertainment-provide-SEMITRANS-one.good.at-COP-say.that-REP-FUT-sure.but-3plSUBJ/3sgOBJ-but

‘However, they will say that he is a great entertainer, but . . .’

(12:1 morpheme-to-word ratio)

Source: http://encyclopedia.thefreedictionary.com/polysynthetic+language
What is linearization?

Linearization ≠ Ordering Words

1. What about “sublexical” phenomena?
2. What about “supralexical” phenomena (discontinuous constituents, topological fields)?
Was is a word?
End product of NLG: natural language output

- provide the flexibility for generation of natural language output
- to deliberately control its production in order to pick up the most appropriate utterance in a specific context
Analysis vs. Generation

Crucial difference between analysis and generation

- analysis is description/interpretation, generation is prescription
- totally different premises
- totally different goals
- analysis assumes sophisticated structures for language description and interpretation, the result of generation is just a chain of words
Observation

The goal of NLG is natural language output

- just a string (be it phonological or graphical)
- we do NOT convey syntactic structure when speaking/writing
- we do NOT use especially designated graphical/acoustic signs for empty nodes/topological fields, ellipsis, traces, etc.
The Place of Linearization

**NLU**

tokenizing → tagging → parsing → semantic interpretation

**NLG**

text polishing ← inflection ← linearization ← [lexicalization, generating referring expressions, aggregation, etc.] ← text planning ← content selection

**NLG**

black = CGR  red = HPR
CGR

1. linearization
2. inflection
3. text polishing

- relying heavily on the whole linguistic knowledge from analysis
- relying heavily on the decisions previously made during the generation process
- no need of empty structures nor metaphores (traces, liberation, emancipation, climbing, landing)
- syntactic structures, constituency: obstacles for linearization

What else?
Why not merelogy?

Ideas:

- primordiality of speech over written language
- primordiality of handwriting over electronic forms
- difference between type and token
- difference between langue and parole
- the quality of automatic speech recognition / speech synthesis
- what about trying to make out the parts of an utterance of a language you don’t know or are about to learn?

Still not convinced? Ciprian, use you most cruel argument!
Other models?

Why not using lattices to model the linearization (cf. G. Link’s modeling of plurals?)
⇒ lattices are still sets (posets)

Why not using temporal logic for linearization?
⇒ Ciprian, sing again!
Speech has other different parameters than the temporal component!
A General Linearization Model

1. only one type of entities: Linear Order Part (LOP)
2. two different types of relations holding between LOPs
   - Part-Of relation
   - Linear Order relation
3. two different types of rules
   - PO-relating rules (mereological rules)
   - LO-relating rules (linear rules)
     - horizontal
     - vertical
     - diagonal
Definitions

[Linear Order Part] A Linear Order Part (LOP) is a phonologically realizable language item which has to be linearised as a contiguous part of a grammatically well-formed utterance.

[Part-Of Relation] A Part-Of relation holding between two different LOPs $\lambda_1$ and $\lambda_2$ ($\lambda_1 \subseteq \lambda_2$) states that $\lambda_1$ is part of $\lambda_2$. The Part-Of relation is reflexive, anti-symmetric, and transitive.

[Linear Order Relation] A Linear Order relation holding between two different LOPs $\lambda_1$ and $\lambda_2$ ($\lambda_1 \prec \lambda_2$) states that $\lambda_1$ precedes $\lambda_2$. The Linear Order relation is irreflexive, asymmetric, and transitive.
Exclusivity

The Part-Of relation and the Linear Order relation are mutually exclusive, i.e., two different LOPs can either PO-relate or LO-relate but not both.

Let $\lambda_1$ and $\lambda_2$ be different LOPs:

1. if $\lambda_1 \sqsubseteq \lambda_2$, then $\lambda_1 \not\approx \lambda_2$
2. if $\lambda_1 \sqsubseteq \lambda_2$, then $\lambda_2 \not\approx \lambda_1$
3. if $\lambda_1 \prec \lambda_2$, then $\lambda_1 \not\sqsubseteq \lambda_2$
4. if $\lambda_1 \prec \lambda_2$, then $\lambda_2 \not\sqsubseteq \lambda_1$

Examples:

$\lambda_1 \prec \lambda_2$: [das Buch]$_{\lambda_1}$ [auf dem Tisch]$_{\lambda_2}$

$\lambda_1 \sqsubseteq \lambda_2$: [[das Buch]$_{\lambda_3}$ [auf dem Tisch]$_{\lambda_1}$]$_{\lambda_2}$
Two different LOPs can not overlap.

Let $\lambda_1$, $\lambda_2$, and $\lambda_3$ be different LPUs:

1. if $\lambda_2 \subseteq \lambda_1$ and $\lambda_2 \subseteq \lambda_3$, then either $\lambda_1 \subseteq \lambda_3$ or $\lambda_3 \subseteq \lambda_1$

Examples:

$[\lambda_1 \text{ der rote }] \lambda_1 \ [\lambda_2 \text{ Apfel }] \lambda_2$

$[\lambda_1 \text{ der }] \lambda_1 \ [\lambda_2 \text{ rote Apfel }] \lambda_2$

$*[\lambda_1 \text{ der } [\lambda_2 \text{ rote }] \lambda_1 \text{ Apfel }] \lambda_2$
Examples of LOPs [1]

- a phoneme is a LOP (the smallest!)
- a phoneme cluster (not necessarily a syllable) is a LOP
- a syllable is a LOP
- a morpheme is a LOP
- a word is a LOP
- different groups of words are LOPs:
  
  ...
Examples of LOPs [2]

- contiguous subparts of constituents
  + [der rote] Apfel
  + der [rote Apfel]

- contiguous constituents
  + [der rote Apfel] ist schön
  + [der rote Apfel, den Maria Hans gab], ist schön
  + *[der rote Apfel] ist schön, [den Maria Hans gab]

**NB:** Contiguous constituents are LOP not by virtue of being constituents but due to the fact that they have to be realized as a contiguous part of an utterance!
Examples of LOPs [3]

- non-empty topological fields
  + Maria gab Hans einen roten Apfel
  + (dass) Maria Hans einen roten Apfel gab
- whole (main/subordinate) clauses
  + Peter glaubte, dass Maria Hans einen roten Apfel gab
  + Dass Maria Hans einen roten Apfel gab, glaubte Peter
  + Peter glaubte, dass Maria Hans einen roten Apfel gab
  + Dass Maria Hans einen roten Apfel gab, glaubte Peter
- whole sentences are LOPs
- ...
Forming LOPs [1]
Forming LOPs [2]
Forming LOPs [3]
Forming LOPs [4]
Forming LOPs [5]
Forming LOPs [6]

Rule Name: $AD[J]V$Modification

Condition Slot

$X \text{mod} \rightarrow Y$
$X.\text{pos} = \text{ADJA} \mid \text{ADV}$

Action Slot

$[X; Y] ::= \text{lop007}$
Linearizing LOPs (Horizontal) [1]
Linearizing LOPs (Horizontal) [2]

Rule Name: \textit{det}_H

Condition Slot

\begin{center}
\begin{tabular}{l}
X \textit{det-} \rightarrow Y; \\
X \textit{α-} \rightarrow Z;
\end{tabular}
\end{center}

Action Slot

\begin{center}
\begin{tabular}{l}
Y \textit{prec-} \rightarrow Z;
\end{tabular}
\end{center}
Linearizing LOPs (Vertical) [1]
Linearizing LOPs (Vertical) [2]

Rule Name: \textit{det}_V

\textbf{Condition Slot}

\[ \text{X \textit{det} \rightarrow Y; } \]

\textbf{Action Slot}

\[ \text{Y \textit{prec} \rightarrow X; } \]
Linearizing LOPs (Diagonal) [1]
Linearizing LOPs (Diagonal) [2]

Rule Name: relClauseVerb_D

Condition Slot

\[ X \alpha \rightarrow Y; \quad Y \text{mod} \rightarrow Z; \quad Z \beta \rightarrow V; \quad V.\text{pos} = \text{PRELS}; \]

\textit{OUTPUT}: \{X_\text{lop} \text{prec} \rightarrow Q_\text{lop};

\quad \quad \quad Q_\text{lop}.\text{gender} \neq V.\text{gender}; \ldots\}\]

Action Slot

\[ X \text{prec} \rightarrow V; \]
Example of Linearization [1]
Example of Linearization [2]

Paris → geben → d → sehr → schön → Helena → ein → rot → Apfel

Paris gibt der sehr schönen Helena einen roten Apfel.
Paris gibt einen roten Apfel der sehr schönen Helena.
Summary

- General Linearization Model: simple mereological model for linearization
- providing a new view of the natural language utterance
- GLM is NO linguistic theory
- extremely useful for this CGR-specific tasks of NLG, but of no use for the HPR-tasks
- (almost) of no use for NLU
- no use of unnecessary constructs
- language-independent
- as general as to regard both a phoneme and a whole text belonging to the the same type (Linear Order Part)
What next?

In construction!