The General Linearisation Model:

A Mereological Description of the Natural Language Utterance

Ciprian Gerstenberger
University of Saarland

IGK Summer School September 2004 Edinburgh
Outline

1. Flexible Output Realisation
2. Analysis versus Generation
3. The General Linearisation Model
4. Worked Example
5. Conclusions
Flexible Output Realisation

- Planning dialogue output (TALK Project)
  - multi-modal (speech, text, graphics)
  - multi-lingual
  - domain-adaptable

  - great flexibility in output realisation

  - great flexibility in getting the word sequence which fits best the communicative situation in a given context
Natural Language Analysis (1)

Maria singt ein Lied.

continuous speech signal
(a whole)

tokens
(parts)
+ linear order

continuous string
(a whole)
Natural Language Analysis (2)

\[
\begin{align*}
Maria & \quad \text{NP} & \quad Maria \\
\text{singt} & \quad \text{VVFIN} & \quad \text{singen} \\
\text{ein} & \quad \text{ART} & \quad \text{ein} \\
\text{Lied} & \quad \text{NN} & \quad \text{Lied} \\
. & \quad . \\
\end{align*}
\]

- tokens + linear order
- +pos tags+lemmata
- syntactic structure (tree)
- tokens+linear order
- +pos tags+lemmata+
- grammatical function+
- case+tense+...
Natural Language Generation (1)

\[
\begin{align*}
\text{singen} & \quad \text{subj} & \quad \text{dobj} \\
\text{Maria} & & \text{Lied} \\
\text{ein} & & \text{det} \\
\end{align*}
\]

\[
\text{Maria singt ein Lied.} \\
\text{Ein Lied singt Maria.}
\]

syntactic structure (tree) +
pos tags + lemmata +
grammatical function +
case + tense + ...

(all) grammatically
well-formed variants
(continuous strings)
Natural Language Generation (2)

Content selection

Communicative relevant information

Content organisation

Text plan

Sentence planning

Sentence plan

Sentence realisation

Text
Sentence Realisation

Input: unordered tree

Word Order Determination
(=Linearising)

Word Form Determination
(=Inflection)

Orthography and Punctuation Check

Output: well-formed utterance
Analysis versus Generation

- analysis assumes complex structures (e.g. trees) for language description
- the result of generation is just a chain of words (a string, i.e., no tree!)
- language models postulate some elements which are actually not existent (i.e. not perceivable) in the string of words (traces, empty topological fields, etc.)
- the linearisation step does not involve any empty elements
Approaches to Linear Precedence


• the common idea of all these approaches is the dissociation of syntactic structure from surface structure

Criticism:

• syntactic and surface structures are not dissociated radically enough

• all these theories fail to get the right perspective on linearisation as a step of text generation
  – the use of empty elements (traces, topological fields, etc.)
  – the use of (non-unary) trees as result of linearisation
  – the use of unnecessarily complicated theoretical models for linearisation
Bridging the Linearisation Gap

Maria singt ein Lied.

Ein Lied singt Maria.
Idea

- use utterance parts that always stay together
  - no matter which size they have
  - no matter which syntactic structure they belong to
The General Linearisation Model

- a sole type of entities: Linear Order Part (LOP)

- two different types of relations holding between LOPs
  - Part-Of Relation
  - Linear Order Relation

- two different types of rules
  - PO-relating rules (mereological rules)
  - LO-relating rules (linear rules)
GLM - Definitions

• **Linear Order Part:**
  A Linear Order Part is a phonologically realisable language item which has to be linearised as a continuous part of a grammatically well-formed utterance.

• **Part-Of Relation:**
  A Part-Of relation holding between two different LOPs \( \lambda_1 \) and \( \lambda_2 \) \((\lambda_1 \sqsubseteq \lambda_2)\) states that \( \lambda_1 \) is part of \( \lambda_2 \). The Part-Of relation is reflexive, anti-symmetric, and transitiv.

• **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 transitiv.
GLM - Examples of LOPs

- a phoneme is a LOP (the smallest!)
- a syllable is a LOP
- a morpheme is a LOP
- a word is a LOP
- different group of words are LOPs:
  - non-discontinuous parts of non-discontinuous constituents (*der rote Apfel*
  - non-discontinuous constituents (*der rote Apfel*
  - non-empty topological fields are LOPs (*Maria gab Hans einen roten Apfel; (dass) Maria Hans einen roten Apfel gab*)
  - whole (main/subordinate) clauses are LOPs (*Peter glaubte, dass Maria Hans einen roten Apfel gab; Dass Maria Hans einen roten Apfel gab, glaubte Peter*)
  - whole sentences are LOPs
  - ...
GLM - Properties (1)

- 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 $\lambda_1 \not\sqsupseteq \lambda_2$
2. if $\lambda_1 \sqsubseteq \lambda_2$ then $\lambda_2 \not\sqsupseteq \lambda_1$
3. if $\lambda_1 \not\preceq \lambda_2$ then $\lambda_1 \not\preceq \lambda_2$
4. if $\lambda_1 \not\preceq \lambda_2$ then $\lambda_2 \not\preceq \lambda_1$
GLM - Properties (2)

\[ \lambda_1 \prec \lambda_2 \]

\[ [(\text{das Buch})]^{\lambda_1} \quad [(\text{auf dem Tisch})]^{\lambda_2} \]

\[ \lambda_1 \subseteq \lambda_3 \]

\[ [[[\text{das Buch}]]^{\lambda_1} \quad [(\text{auf dem Tisch})]^{\lambda_2}]^{\lambda_3} \]
GLM - Properties (3)

- **No Overlapping:**
  Two different LOPs can not overlap.

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

If $\lambda_2 \sqsubseteq \lambda_1$ and $\lambda_2 \sqsubseteq \lambda_3$ then either $\lambda_1 \sqsubset \lambda_3$ or $\lambda_3 \sqsubset \lambda_1$. 

![Diagram showing the relationships between different LOPs](image)
GLM - Properties (3)

- **No Overlapping:**
  Two different LOPs can not overlap.

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

\[
\text{if } \lambda_2 \sqsubseteq \lambda_1 \text{ and } \lambda_2 \sqsubseteq \lambda_3 \text{ then either } \lambda_1 \sqsubseteq \text{ or } \lambda_3 \sqsubseteq \text{ }
\]
GLM - Properties (3)

- **No Overlapping:**
  Two different LOPs can not overlap.

Let \( \lambda_1, \lambda_2 \) and \( \lambda_3 \) be different LOPs:

if \( \lambda_2 \sqsubseteq \lambda_1 \) and \( \lambda_2 \sqsubseteq \lambda_3 \) then either \( \lambda_1 \sqsubseteq \lambda_3 \) or \( \lambda_3 \sqsubseteq \lambda_1 \)
GLM - Properties (3)

- **No Overlapping:**
  Two different LOPs can not overlap.

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

if $\lambda_2 \sqsubseteq \lambda_1$ and $\lambda_2 \sqsubseteq \lambda_3$ then either $\lambda_1 \sqsubset$ or $\lambda_3 \sqsubset$

![Diagram showing overlapping LOPs with German words: "das Buch" and "dem Tisch" under $\lambda_1$ and $\lambda_3$, and "auf" under $\lambda_2$.]
No Overlapping:
Two different LOPs can not overlap.

Let \( \lambda_1, \lambda_2 \) and \( \lambda_3 \) be different LOPs:

if \( \lambda_2 \sqsubseteq \lambda_1 \) and \( \lambda_2 \sqsubseteq \lambda_3 \) then either \( \lambda_1 \sqsubset \lambda_3 \) or \( \lambda_3 \sqsubset \lambda_1 \)
GLM - Corrolaries

- Let $\lambda_1, \lambda_2, \lambda_3$ and $\lambda_4$ be different LOPs, and $\lambda_3 \prec!$: 
  
  1. if $\lambda_1 \sqsubseteq \lambda_3$ then $\lambda_1 \prec!$
  2. if $\lambda_2 \sqsubseteq \lambda_4$ then $\lambda_3 \prec!$
  3. if $\lambda_1 \sqsubseteq \lambda_3$ and $\lambda_2 \sqsubseteq \lambda_4$ then $\lambda_1 \prec$

- Let $\lambda_1, \lambda_2, \lambda_3$ and $\lambda_4$ be different LOPs, and $\lambda_1 \prec!$: 
  
  1. if $\lambda_1 \sqsubseteq \lambda_3$ and $\lambda_2 \nsubseteq \lambda_3$ then $\lambda_3 \prec$
  2. if $\lambda_2 \sqsubseteq \lambda_4$ and $\lambda_1 \nsubseteq \lambda_4$ then $\lambda_1 \prec$
  3. if $\lambda_1 \sqsubseteq \lambda_3$, $\lambda_2 \sqsubseteq \lambda_4$, $\lambda_3 \nsubseteq \lambda_4$, and $\lambda_4 \nsubseteq \lambda_3$ then $\lambda \sim$
GLM - Forming LOPs (1)
GLM - Forming LOPs (2)

- LOP-forming rules

Rule Name: $AD[J|V]Modification_lop$

Condition Slot

\[ X \text{ mod-> Y; } \]
\[ X\.pos = \text{ADJA | ADV;} \]

Action Slot

\[
[X; Y] :::- lop007;
\]
GLM - Linearising LOPs (1)
GLM - Linearising LOPs (2)

- horizontal Linear Order rules

<table>
<thead>
<tr>
<th>Rule Name:</th>
<th>det_H</th>
</tr>
</thead>
</table>
| Condition Slot | X det-> Y;  
                 | X α-> Z;   |
| Action Slot   | Y prec-> Z; |


GLM - Linearising LOPs (3)
GLM - Linearising LOPs (4)

- vertical Linear Order rules

<table>
<thead>
<tr>
<th>Rule Name: \textit{det}_V</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Condition Slot</strong></td>
</tr>
<tr>
<td>\textbf{X det-&gt; Y;}</td>
</tr>
<tr>
<td><strong>Action Slot</strong></td>
</tr>
<tr>
<td>\textbf{Y prec-&gt; X;}</td>
</tr>
</tbody>
</table>
GLM: Linearising LOPs (5)
GLM - Linearising LOPs (6)

- diagonal Linear Order rules

**Rule Name:** *relClauseVerb_D*

**Condition Slot**

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

*OUTPUT:* \{X\_lop prec\rightarrow Q\_lop;

\quad Q\_lop.gender \neq V.gender; \ldots \}

**Action Slot**

\[ X \mathrm{prec}\rightarrow V; \]
GLM - Worked Example (1)
Paris gibt der sehr schönen Helena einen roten Apfel.

Paris gibt einen roten Apfel der sehr schönen Helena.
GLM - Future Work

- using the model to write flexible, parametrizable linearisation grammars for German

- incorporating Information Structure into the rules

- testing the model for several „exotic“ languages

- using annotated corpora to learn LOP-forming and linearisation rules
GLM - Conclusions

- GLM as utterance surface oriented model:
  - describing utterances as mereological structures
  - reflecting the way humans learn their mother tongue
  - treating different linguistic entities in a uniform way
  - accounting for context in an easy, flexible way
  - allowing for extending and parametrizing the grammar without changing the theory
  - neutral with respect to syntactic theories
  - language-independent