Linearization in Natural Language Generation: linguistic and implementation issues

Ciprian-Virgil Gerstenberger

University of Tromsø

DGfS, Linearization Workshop, Berlin – 24th February 2010
Outline

NLG
Mereology
Linearization
GLM
Conclusion
Outline

NLG

Mereology

Linearization

GLM

Conclusion
Outline

NLG

Mereology

Linearization

GLM

Conclusion
Outline

NLG

Mereology

Linearization

GLM

Conclusion
Outline

NLG

Mereology

Linearization

GLM

Conclusion
NLG Pipeline

Content selection
- Communicative relevant information

Content organisation
- Text plan

Sentence planning
- Sentence plan

Sentence realisation
- Text
NLG Pipeline

Content selection

Communicative relevant information

Content organisation

Text plan

Sentence planning

Sentence plan

Sentence realisation

Text
Output Realization

Input: unordered tree

„Word“ Order Determination
(=Linearizing)

„Word“ Form Determination
(=Inflection)

Orthography and Punctuation Check

Output: well-formed utterance
What?

One Goal

- flexible output realization
  → getting the word sequence which fits best the communicative situation in a given context

Two subtasks

- providing flexibility (General Linearization Model)
- controlling flexibility (partly done)

One Constraint

- language-independency
Why?

Simply, because...

- the need for such a framework is increasing rapidly
- there is no such a flexible framework or system yet

Possible reasons for the lack of such a system

- statistically based quick-and-dirty solutions are much cheaper (however, not with the same results)
- surface phenomena are ‘not scientifically interesting’
- the best fitting output for a given communicative situation is hard to control
How?

Approach – language ‘mimicry’

- Look at as much and various linearization-relevant data as possible!
- Describe it for the purpose of linearization!
- Don’t explain it, model it!

Comparable to

→ developing plastics that self-clean using principles derived from study of the Lotus-leaf
Some Phenomena [1]

Speech errors/Typos
- *multimodal output* → *multimodal ouptut*
- *self-destruct instruction* → *self-instruct destruction*
- *snow flurries* → *flow snurries*
- *writing a letter to my mother* → *writing a mother to my letter*

Spoonerisms
- *our deer old queen* → *our queer old dean*

Thinking aloud
- What can be linearized in a language?
- What is the smallest linearizable entity in a language?
Some Phenomena [2]

Metathesis

- *brid* (Old English) vs. *bird* (modern English)
- *Ross* (German) vs. *horse* (English)
- *Kreuć* (German) vs. *kors* (Norwegian)
- *Roland* (French) vs. *Orlando* (Italian)

Tmesis

- *whatsoever place* → *what place soever*
- *absolutely* → *abso-bloody-lutely*

Thinking aloud

- Fan-f***ing-tastic!
Some Phenomema [3]

Particle verbs (in German)

- *Maria machte die Tür auf.*
  
  [Mary opened the door.]

- *Auf machte Maria die Tür (, nicht zu).*
  
  [Mary opened the door (, she didn’t clause it).]

- *Maria wollte die Tür aufmachen.*
  
  [Mary wanted to open the door.]

- *Maria versuchte die Tür aufzumachen.*
  
  [Mary tried to open the door.]

- *Maria hat die Tür aufgemacht.*
  
  [Mary has opened the door.]

Thinking aloud

- What is a word?
Discovering the Obvious

– What is common to all these items?

– All these items are parts of the utterance!
Mereology ≡ Partonomy

Definition
the theory of parthood relations

- part to whole
- part to part within a whole

(Greek μερος, Latin pars: ‘part’)

Thinking aloud
- What is the difference between mereology and parsing?
- What is the difference between a part and a constituent?
Phrases

“[…] headship was defined as a relation between parts and wholes; and the wholes were, in particular, ‘PHRASES’.”

Discontiguous constituents

“[…] a partonomic structure will be said to be inconsistent if its parts act as a whole by some criteria but not by others. […] Here are two examples. […] A set of words makes a phrase by joint replaceability but not by contiguity:

*The man is a friend of yours who came to see me.*”

Overlapping constituents

“Actual partonomic structures in syntax often deviate from [...] these desiderata of simplicity. Here follow some examples of complex partonomies. [...] PARTS ARE NON-UNIQUELY ASSIGNED TO WHOLES. Example: I expect him to leave town.”

Abstraction Level

Too low an abstraction level
  • speech errors, metathesis, tmesis, etc.

Too high an abstraction level
  • discontiguous constituents
  • overlapping constituents

⇒ Find the best suited abstraction level for linearization!
A General Mereological Model

1. Step
→ Take the most general mereological model!

Mereological Utterance Description
- one type of entities: Linear Order Part (LOP)
- two different types of relations between LOPs
  - Part-Of relation
  - Linear Order relation
A Linear Order Part (LOP) is a language item which is (phonologically) realized as a contiguous part of an utterance.

**Complexity**
- simple (‘atomic’, elementary, primitive, input) items
- complex items
  \[ \Rightarrow \text{depending on the granularity needed!} \]

**Adequacy**
- linguistically motivated: *incredible, a red book*
- linguistically non-motivated: *a big red book*
Contiguity

Consequences of imposing contiguity

→ any contiguous part of an utterance is a LOP
→ discontiguous parts do not form a LOP

⇒ The problem of discontiguous constituents is solved!
Let $\lambda_1$ and $\lambda_2$ be two different LOPs:

**Part-Of relation**

$\lambda_1 \sqsubseteq \lambda_2$ iff $\lambda_1$ is proper part of $\lambda_2$. PO-relation is reflexive, anti-symmetric, and transitive.

$$\lambda_1 \sqsubseteq \lambda_3: \lambda_3[\lambda_1[\text{the book on the}] \lambda_1 \lambda_2[\text{table}] \lambda_2] \lambda_3$$

**Linear Order relation**

$\lambda_1 \prec \lambda_2$ iff the occurrence of $\lambda_1$ precedes the occurrence of $\lambda_2$ in the utterance. LO-relation is irreflexive, asymmetric, and transitive.

$$\lambda_1 \prec \lambda_2: \lambda_3[\lambda_1[\text{the book on the}] \lambda_1 \lambda_2[\text{table}] \lambda_2] \lambda_3$$
Property of Relations

Part-Of relation and Linear Order relation are mutually exclusive.

→ Two different LOPs can either PO-relate or LO-relate.

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

- if $\lambda_1 \subseteq \lambda_2$ then $\lambda_1 \not< \lambda_2$ and $\lambda_2 \not< \lambda_1$
- if $\lambda_1 \prec \lambda_2$ then $\lambda_1 \not\subseteq \lambda_2$ and $\lambda_2 \not\subseteq \lambda_1$
Mutual Exclusivity

\[ \lambda_1 \prec \lambda_2 \]

\[ [\text{das Buch}]_{\lambda_1} [\text{auf dem Tisch}]_{\lambda_2} \]

\[ \lambda_1 \subseteq \lambda_3 \]

\[ [[[\text{das Buch}]_{\lambda_1} [\text{auf dem Tisch}]_{\lambda_2}]]_{\lambda_3} \]
Property of Entities

Two different LOPs can not proper-overlap.

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

- then either $\lambda_1 \subseteq \lambda_3$ or $\lambda_3 \subseteq \lambda_1$
Non-proper-overlapping

Consequences of imposing non-proper-overlapping

→ parts (simpler LOPs) are uniquely assigned to wholes
  (more complex LOPs)

⇒ The problem of overlapping constituents is solved!
2. Step

→ Constrain the model to the relevant part for linearization in NLG!

From analysis/description to generation/prescription

• exclude possible, yet undocumented data
  → useless

• exclude ungrammatical data
  → uninteresting

⇒ Consider only grammatically well-formed utterances and their linguistically interesting partitions!
3. Step

→ Find the best suited level(s) of abstraction for linearization in NLG!

Main issues

- ‘linearization primitives’
- temporal order of the different levels of abstraction
- forming complex items
- linearizing elementary and complex items
The Way

— Look at the LO-relation!
— Look at data!
LO-relation

Binary relation

→ absolutely relative positions: preceding vs. subsequent

Special cases

• relatively absolut positions: first vs. last part of the whole they belong to
• immediate precedence
Intermezzo

Why not relatively absolut positions for all parts?
→ too high an abstraction level
→ danger of linearizing emptyness

Example:

Topological Field Model in German

<table>
<thead>
<tr>
<th>Vorfeld</th>
<th>Linke Klammer</th>
<th>Mittelfeld</th>
<th>Rechte Klammer</th>
<th>Nachfeld</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kannst</td>
<td>du Mozart oder Bach</td>
<td>spielen?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spiele</td>
<td>Mozart!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Die Kinder</td>
<td>spielen</td>
<td>Mozart.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wer</td>
<td>spielt</td>
<td>Mozart?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wenn du willst,</td>
<td>kannst</td>
<td>Du Mozart</td>
<td>spielen.</td>
<td>das von Mozart komponiert wurde.</td>
</tr>
<tr>
<td>Maria</td>
<td>spielt</td>
<td>ein Stück,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wenn</td>
<td>du</td>
<td></td>
<td>willst, . . .</td>
<td></td>
</tr>
<tr>
<td>... das</td>
<td>von Mozart</td>
<td>komponiert wurde</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Polish markers for person and number in past (so-called ‘floating affixes’)

(1) Nie widzieliśmy tego.  [We didn’t see this.]
not see-pst-m-pl-1pl this

(2) Tegośmy nie widzieli.
this-1pl not see-pst-m-pl

(3) Myśmy tego nie widzieli.
we-1pl this not see-pst-m-pl

Thanks to Magdalena Wolska for this example!

Thinking aloud

– ‘Floating affixes’, such a great misnomer!
Romanian weak pronouns

(4) Să îl faceți! [Do it!]
that it do-conj-pl

(5) Să-l faceți!
that it do-conj-pl

(6) Faceți-l!
do-imp-pl it
Separable particle verbs in German

(7) Sie will das Fenster aufmachen.  
She wants the window to open

[She wants to open the window.]

(8) Sie macht das Fenster auf.  
She makes the window open

[She opens the window.]
Phrasal verbs

(9) They call up John.
(10) They call John up.

Thanks to Edith A. Moravcsik for this example!
Counterexample

German suffix

(11) Sie soll\text{te} \text{kochen}. [\text{She should cook.}]
(12) *Sie \text{kochente}.

\rightarrow \text{irrelevant for linearization}
\rightarrow \text{treatable in the inflection component}

Thanks to Trond Trosterud for this example!
Linearization test

In search of the suited granularity . . .

Given two items $\alpha$ and $\beta$ at morpho-syntactic level in a specific language: if the language allows for both $\alpha \prec \beta$ and $\beta \prec \alpha$ then these items are linearization primitives.
Help wanted!

The author of this presentation is scared of the simplicity of the linearization test. Therefore, he kindly asks you to send him any kind of weird examples in any language to refine or refute it.

Thanks in advance!
Temporal Order

Syntax vs. inflectional morphology

Tentative temporal ordering
- clear temporal sequential order
  → all operations in a module follow all operations in another module
- fuzzy temporal sequential order
  → too fuzzy to describe
Temporal Order

Clear Temporal Order

- syntax comes **before** morphology
  (e.g., Minimalist Program)
- syntax comes **after** morphology
  (e.g., Lexical Function Grammar,
  Head-Driven Phrase Structure Grammar)
  → keyword *Lexical Integrity Principle*
Temporal Order

Fuzzy Temporal Order

- syntax and morphology are **synchronous**\(^1\) (e.g., Functional Unification Grammar, Radical Construction Grammar)
- syntax and morphology are **parallel** (e.g., Parallel Morphology (?))
- syntax and morphology are **alternating**\(^2\) (e.g., ???)
- morphology is **distributed** among syntax, semantics and phonology (e.g., Distributed Morphology)

\(^1\)Thanks to William Croft for pointing out this category to me!
\(^2\)Thanks to Edith A. Moravcsik for pointing out this possible category to me!
Romanian

Romanian *this-NP*

(13) acest om  
this man

(14) *acest a omul*  
this-def man-def

(15) *acest a om*  
this-def man

(16) *acest omul*  
this man-def

(17) omul acesta  
man-def this-def

(18) *om acest*  
man this

(19) *omul acest*  
man-def this

(20) *om acesta*  
man this-def

Thinking aloud

— **Thanks** to Ciprian-Virgil Gerstenberger for this example! :-)

---

This example demonstrates the use of Romanian pronouns and definite articles in the context of natural language generation (NLG). The examples show how the pronoun 'aceste' (this) and the definite article 'om' (the man) can be used to form different phrases. The use of 'aceste' without a definite article is marked with an asterisk to indicate that it is a non-standard or non-native use. The examples also show the correct use of 'omul' (the man) with a definite article, which is more common in Romanian. Thinking aloud with examples like these can help in understanding the nuances of language and its practical applications.
Solution

If inflection before linearization:

- Which value for the Romanian *this*-NP: `def+` or `def-`?
- How to “backtrack” to obtain all grammatical variants?
- Which consequences for NLG systems based on, e.g., LFG or HPSG?

⇒ Linearization before inflection!
Complex Items

German NP

(21) Das Buch ist schön.
    the book is nice
    The book is nice.

(22) Schön ist das Buch.
    nice is the book
    The book is nice.

(23) Ist das Buch schön?
    is the book nice
    Is the book nice?
Complex items

Hypothetical language

(24) SUBJ ADV\textsubscript{temp} VERB ADV\textsubscript{caus} DOBJ.

(25) VERB SUBJ ADV\textsubscript{temp} ADV\textsubscript{caus} DOBJ.

(26) VERB ADV\textsubscript{caus} SUBJ ADV\textsubscript{temp} DOBJ.

(27) VERB ADV\textsubscript{caus} DOBJ SUBJ ADV\textsubscript{temp}.
Total Permutation Constraint (TPC)
If two or more linearization primitives permute always as a unit they form a complex Linear Order Part.

Observations
- independent of them forming a constituent in traditional sense
- contiguity constraint applied to complex units
Some Similarity Noticeable?

Mobile. (William B. Wade)
Complex items

(28) Peter hat gestern ein Buch, das schön ist, gekauft.

Yesterday, Peter bought a nice book.

<table>
<thead>
<tr>
<th>Vorfeld</th>
<th>Linke Klammer</th>
<th>Mittelfeld</th>
<th>Rechte Klammer</th>
<th>Nachfeld</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>hat</td>
<td>gestern ein Buch, das schön ist, gekauft.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peter</td>
<td>hat</td>
<td>ein Buch, das schön ist, gestern gekauft.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>*Peter</td>
<td>hat</td>
<td>ein Buch gestern, das schön ist, gekauft.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peter</td>
<td>hat</td>
<td>gestern ein Buch gekauft, das schön ist.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peter</td>
<td>hat</td>
<td>ein Buch gestern gekauft, das schön ist.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Partial Permutation Constraint (TPC)
If two or more linearization primitives permute sometimes as a unit they form a complex Linear Order Part in the given contexts.

⇒ expressing and modelling contextual linearization factors
Ingredients

- entities: Symbolic Linear Order Parts (SLOPs) ⇔ LOPs in MUD
  
- two different types of rules ⇔ relations in MUD
  → PO-rules: forming complex SLOPs
    – from any possible combination of dependency tree nodes (!)
  → LO-rules: linearizing SLOPs
    – horizontal
    – vertical
    – diagonal (modelling discontinuous constituents)
A General Linearization Model

Input

- dependency tree
- nodes are elementary SLOPs + any piece of information stemming from the previous NLG steps
  → building a sound interface between surface and the rest of NLG pipeline
ID/LP Mapping

Task

→ map Immediate Dominance (ID) structures (SLOPs!) into their grammatical Linear Precedence pendants (i.e., all possible grammatical output sequences)
Forming SLOPs

The diagram illustrates the structure of a sentence in German, with the verb 'geben' (to give) as the subject and the noun 'Paris' as the direct object. The sentence 'Helena schön sehr Apfel' is parsed into its constituent parts, indicated by labels such as 'iobj', 'dobj', 'subj', and 'mod'. The diagram shows the syntactic relationships between the words, with 'Helena' as the direct object of 'geben', 'schön' as an adverb modifying 'geben', and 'sehr' as an adverb modifying 'schön'. The phrase 'Apfel' is a direct object of 'geben', with 'rot' serving as an adverb modifying 'Apfel'.
Forming SLOPs
Forming SLOPs
Forming SLOPs
Forming SLOPs
Forming SLOPs
Horizontal LO-rules
Vertical LO-rules

[Diagram showing a tree structure with nodes labeled 'Helena', 'Paris', 'Apfel', 'geben', 'd', 'schön', 'sehr', 'ein', 'rot', 'det', 'mod', 'iobj', 'subj', 'dobj']
Diagonal LO-rules
Conclusion

- development of a linearization module for NLG
- mereological description of natural language utterances
- flexible wrt. linearization phenomena
- language-independent
- flexibility (Allow all!)
  ⇔ grammaticality (Constrain gradually!)
- temporality: system design ⇔ linguistic data

Thinking aloud
  - Stop talking and implement the module!
Thank you!

lichtung

manche meinen
lechts und rinks
kann man nicht velwechseln
werch ein illtum!

Elnst Jandr (1925–2000)