Toward a methodology of chunking: applications and extensions of Linear Unit Grammar

Ray Carey
University of Helsinki
What & why

  - English as a Lingua Franca in Academic Settings
  - 1 million words of naturally occurring academic ELF

- corpus-based fluency research (cf. Götz 2013)
  - chunking as descriptive methodology
  - chunk-based patterns of ordinary dysfluency

- description of (dys)fluency features among L2 users
  - implications for language testing?
What kind of chunking?

- chunking as storage / production
  - Ellis 2002, Bybee 2010

- chunking as processing / parsing
  - Hunston & Francis 2000, Mason 2007

- Linear Unit Grammar (LUG) Sinclair & Mauranen 2006
  - oriented toward processing, with implications for storage
  - robust parsing of spoken and written text
LUG: core concepts

- chunking is intuitive & pre-theoretical:
  - prospection (suspension) completion (extension)
    - Brazil 1995, *A Grammar of Speech*
  - different analysts will have different intuitions
  - much of LUG chunking is routinized & easily learned

- labeling chunks is systematic:
  - organising (O) or incrementing message (M)
  - prospection (M-) / completion (+M)
  - linear analysis: what precedes & what likely follows?

- neither dependent on nor incompatible with traditional grammatical categories
Studies incorporating LUG

Describing the LUG model
• Sinclair & Mauranen 2006; Mauranen 2013

Organizing chunks in ELF
• Mauranen 2009, Carey 2013

LUG and metadiscourse
• Smart 2014

LUG & discourse markers
• Huang 2013

LUG-based model of turn-taking & interruption
• Carey 2011

Applying LUG to literary analysis
• Stone 2011

Tone unit boundaries & LUG chunking boundaries
• Cheng, Greaves, Warren 2008
Linear, real-time processing

whatidefendisthater(.)wecanergivetheimtheermtheknowledgerthisspecificknowledge.they.they.want
Words are chunks too

what i defend is that er (.) we can er give them the erm the knowledge this specific knowledge they they want
what i defend is | that | er | ( . ) | we can | er | give them the | erm | the knowledge | this specific knowledge | they | they want
M: message-oriented
O: organisation-oriented

what i defend is
that
er (.)
we can
er
give them the
erm
the knowledge
this specific knowledge
they
they want
M: message-oriented
O: organisation-oriented

M what i defend is
O that
O er
O (.)
M we can
O er

M give them the
O erm
M the knowledge
M this specific knowledge
M they
M they want
### OT: text organising

### OI: interaction organising

<table>
<thead>
<tr>
<th>M</th>
<th>what I defend is</th>
<th>M</th>
<th>give them the knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>OT</td>
<td>that</td>
<td>OI</td>
<td>erm</td>
</tr>
<tr>
<td>OI</td>
<td>er</td>
<td>M</td>
<td>the knowledge</td>
</tr>
<tr>
<td>OI</td>
<td>(.)</td>
<td>M</td>
<td>this specific knowledge</td>
</tr>
<tr>
<td>M</td>
<td>we can</td>
<td>M</td>
<td>they</td>
</tr>
<tr>
<td>OI</td>
<td>er</td>
<td>M</td>
<td>they want</td>
</tr>
</tbody>
</table>
5 types of M (message) chunks

| M-       | what i defend is | +MA       | give them the |
| OT       | that            | OI        | erm           |
| OI       | er              | +M        | the knowledge |
| OI       | (.)             | MR        | this specific knowledge |
| +M-      | we can          | MF        | they          |
| OI       | er              | MS        | they want     |

MA = Message Adjustment / MF = Message Fragment
MR = Message Revision / MS = Message Supplement
Toward LUG-annotated corpora

- Smart 2014: first LUG-annotated written corpus
  - 40,000 words of IMDb message board discussions

- Two major challenges:
  - consistency in judgments (systematization)
  - takes forever to do manual annotation

- Partial automation
  - no substitute for human intuition
  - advantage of provisional annotations
Step 1: find the O chunks

- Following Sinclair & Mauranen 2006:
  - O chunks are fixed & recurring, easy to find
  - good place to start placing boundaries

- Data-driven lists of O chunks
  - ELFA corpus n-grams: 2-6 words
    - manually examining concordance lines
  - single units: *er, and, but, yeah, erm, mhm*
  - bigrams: high-frequency chunk boundaries
Step 2: guess the M chunks

- Message Fragments (MF)
  - preliminary boundaries at repeats, false starts (wha-)

- Shallow noun phrase chunking:
  - remaining data is part-of-speech tagged by TreeTagger
  - Natural Language Tool Kit: regular expression chunker
    - Bird, Loper & Klein 2009
  - data-driven regex patterns (based on 24 min. of data)

- Shallow NP combined with surrounding words, rough guesses at M chunk labels
Human analysis in XML

- Preliminary chunker outputs data in XML format
  - data is “overchunked” by design: easier to merge than create new annotations

- XML is formatted using cascading stylesheet (CSS)
  - analyzed & revised in Oxygen XML editor
  - chunk boundaries are merged, altered
  - chunk labels are assigned through drop-down menus
Working with XML through CSS

```xml
<e type="M-">
  <w pos="WP" lemma="what">what</w>
  <w pos="PP" lemma="i">i</w>
  <w pos="VVP" lemma="defend">defend</w>
  <w pos="VBZ" lemma="be">is</w>
</e>
<e type="OT">that</e>
<e type="OI">er</e>
<e type="OI">
  <pause type="P2-3sec"/>
</e>
<e type="+M-">
  <w pos="PP" lemma="we">we</w>
  <w pos="MD" lemma="can">can</w>
</e>
<e type="OI">er</e>

M- what I defend is
OT that
OI er
OI P: 2-3 sec
+M- we can
OI er
+MA give them the
OI erm
+M the knowledge
MR this specific kn
MF they
MS they want
```

Element: e

Name: type
Value: +MA

More...
Measuring chunker accuracy

- Each token treated as an observation
  - ends with chunk boundary or no boundary
  - chunker output compared to gold standard
  - true/false positives, true/false negatives calculated

- After 32,000 words (3.5 hours of data)
  - accuracy: 87% (% of observations in agreement)
  - recall: 92% (% of gold boundaries predicted)
  - precision: 80% (% of predicted boundaries correct)
Is LUG reproducible among different analysts?

- inter-rater reliability test
  - trained research assistant using 1400 words of text
  - developed set of chunking guidelines (<2 pages)
  - independently chunked 5210 words of spoken text
    - 2k words from ELFA corpus, 3k words from Hong Kong Corpus of Spoken English (prosodic)

- Cohen’s Kappa = 0.885
  - >.81 “almost perfect” (Landis & Koch 1977)
  - Smart 2014: test with 288 words, Kappa=0.92
Relation between chunks and tone units?

- Hong Kong Corpus of Spoken English (prosodic)
  - HKCSE: Cheng, Greaves, Warren 2008
  - 900,000 words of spoken English (L1/L2), annotated for tone unit boundaries
  - 3,135 words for inter-rater chunking (1,108 TUBs)

- How well does LUG predict tone unit boundaries?
  - accuracy: 80% (% of observations in agreement)
  - recall: 70% (% tone unit boundaries predicted)
  - precision: 73% (% of predicted boundaries correct)
Conclusions

• LUG is a robust descriptive model of language
  • can handle transcriptions of challenging spoken data
  • judgments can be regularised and systematised
  • good potential for inter-rater reproducibility
  • supported by intonation patterns
  • automatisation is within reach

• Ideal for corpus-based studies of (dys)fluency
  • Message Fragments (MF), Message Adjustments (MA)
  • inter-speaker distribution of (dys)fluency features


