Is sentence compression an NLG task?

Erwin Marsi, Emiel Krahmer
Tilburg University
The Netherlands

Iris Hendrickx, Walter Daelemans
Antwerp University
Belgium

Sentence compression

- **sentence compression** (reduction):
  summarizing a single sentence by removing information from it (Jing & McKeown, 2000)
- compressed sentence should retain most important information and remain grammatical

- applications include
  - as part of a full-blown text summarization system
  - automatic subtitling
  - displaying text on handheld devices

Compression as deletion

- **sentence compression as deletion**:
  drop any subset of words from the input sentence while retaining important information and grammaticality
  (Knight & Marcu, 2002)
- Two important properties
  - only deletions are allowed, no substitutions or insertions, and therefore no paraphrasing
  - word order is fixed
- Deletion models satisfy the **subsequence constraint**: words of the compressed sentence must be a subsequence of the input sentence

Deletion models

- Deletion models can be automatically learned from text corpora (Knight & Marcu, 2002)
  - probabilistic noisy channel model
  - shift-reduce parser + decision tree model

- Most follow up work on data-driven sentence compression adheres to the subsequence constraint
  (Minh Le & Horiguchi, 2003; Vandeghinste & Pan, 2004; Turner & Charniak, 2005; Clarke & Lapata, 2006; Zajic et al., 2007; Clarke & Lapata, 2008)

Is sentence compression an NLG task?

- Though it is a form of text-to-text generation, there is no real generation component in deletion models
- Is sentence compression therefore *not* an NLG task?

Is sentence compression an NLG task?

- Intuitively, the subsequence constraint seems a (convenient) over-simplification
- We suspect that in reality sentence compression requires:
    - transformations beyond word deletions
    - linguistic knowledge and resources typical to NLG
- To find out, we studied "real-life" sentence compression in the domain of subtitling

Overview

١. Introduction: sentence compression
٢. Material: subtitle corpus
٣. Analysis: observed compression phenomena
٤. Summary / Discussion

Material: domain

- source: *subtitles* from news broadcasts of the Dutch public television channel
- presentation space is limited:
  - 690 – 780 chars/minute
- subtitles cannot be verbatim transcription
- subtitles are often compressed form of original
- a form of *parallel text*:
  - **aut**: autocue text
  - **sub**: subtitle text

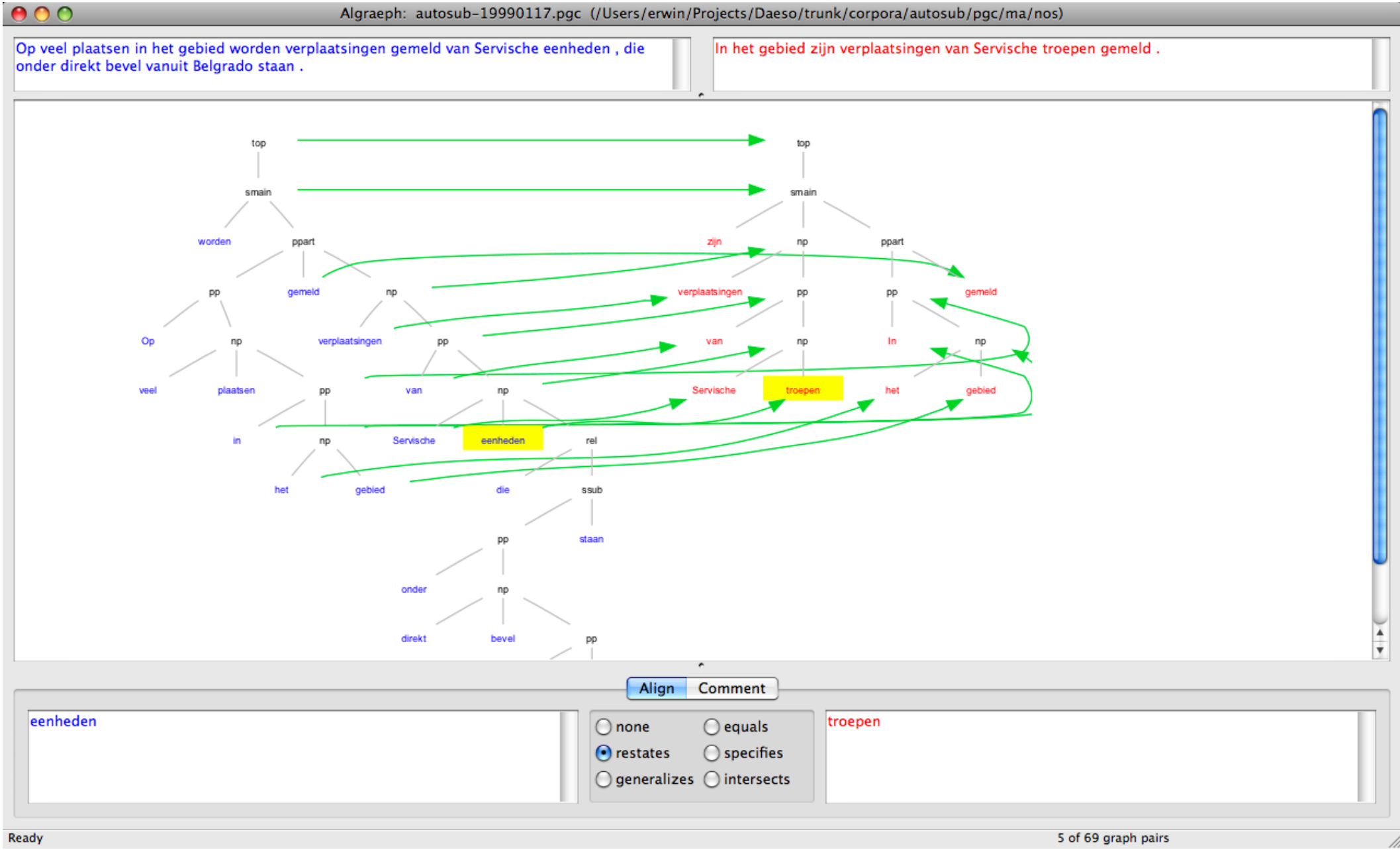
Material: preprocessing

- Subtitle corpus originally collected for studying automatic subtitling (Vandeghinste & Tsjong Kim Sang, 2004)
- automatically tokenized
- automatically aligned at sentence level
- sentence alignments manually checked

Material: further processing

-
- subtitle corpus has become part of DAESO corpus
  - monolingual treebank of parallel/comparable Dutch text (Marsi & Krahmer, 2007)
- all sentences syntactically parsed
- syntax trees manually aligned
    - alignment of similar syntactic nodes
    - labeled with semantic similarity relations
- current work only uses the *word alignments*

Material: aligned trees



Material: alignment degree

Degree	Autocue	(%)	Subtitle	(%)
0	3607	20.74	12542	46.75
1	12382	71.19	13340	49.72
2	1313	7.55	901	3.36
3	83	0.48	41	0.15
4	8	0.05	6	0.02

- *alignment degree*: number of other sentences that a sentence is aligned to

Material: alignment degree

Degree	Autocue	(%)	Subtitle	(%)
0	3607	20.74	12542	46.75
1	12382	71.19	13340	49.72
2	1313	7.55	901	3.36
3	83	0.48	41	0.15
4	8	0.05	6	0.02

- almost half of the subtitles has no corresponding autocue because
  - in a foreign language
  - live interviews

Material: alignment degree

Degree	Autocue	(%)	Subtitle	(%)
0	3607	20.74	12542	46.75
1	12382	71.19	13340	49.72
2	1313	7.55	901	3.36
3	83	0.48	41	0.15
4	8	0.05	6	0.02

- about 1 in 5 autocue sentences is completely dropped

Material: alignment degree

Degree	Autocue	(%)	Subtitle	(%)
0	3607	20.74	12542	46.75
1	12382	71.19	13340	49.72
2	1313	7.55	901	3.36
3	83	0.48	41	0.15
4	8	0.05	6	0.02

- *sentence merging*
  - about 8% of the (short) autocue sentences are merged into a single subtitle
  - cf. sentence aggregation

Material: alignment degree

Degree	Autocue	(%)	Subtitle	(%)
0	3607	20.74	12542	46.75
1	12382	71.19	13340	49.72
2	1313	7.55	901	3.36
3	83	0.48	41	0.15
4	8	0.05	6	0.02

- *sentence splitting*
  - about 4% of the (long) autocue sentences are split into multiple subtitles
  - cf. sentence simplification

Material: alignment degree

Degree	Autocue	(%)	Subtitle	(%)
0	3607	20.74	12542	46.75
1	12382	71.19	13340	49.72
2	1313	7.55	901	3.36
3	83	0.48	41	0.15
4	8	0.05	6	0.02

- sentence deletion, splitting and merging are important for automatic subtitling
- however, not part of sentence compression proper
  - rather compression at the text level
- so we focus on one-to-one aligned sentences only

Material: word compression

- compression is partly obtained by word compression
  - *seven* → *7*
  - *United States* → *US*
  - *Euro* → *€*
- word compression is important for automatic subtitling
- however, not part of sentence compression proper
  - rather compression at the lexical level

Material: compression ratio

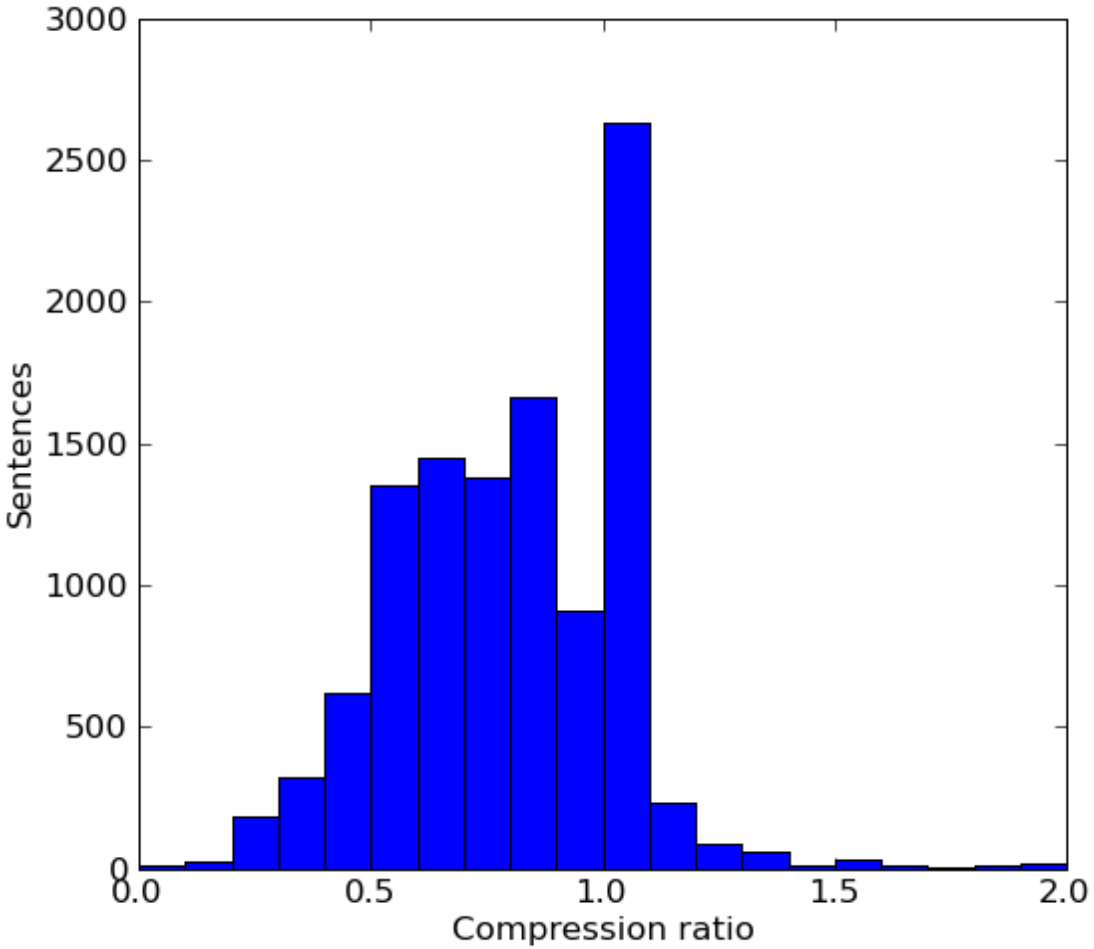
- so we measure compression in terms of *tokens* rather than *characters*

$$Compression\ Ratio\,(CR) = \frac{\#\text{tokens}_{\text{sub}}}{\#\text{tokens}_{\text{aut}}}$$

- this way we abstract from word compression

Material: compression ratio

histogram of CR distribution for 1-to-1 aligned sentences



- many autocue sentences not compressed (CR=0)
- some autocue sentences are in fact expanded (CR>0)
- we keep only sentences with CR<1

Material: parsing failures

- 0.2% sentences failed to pass the parser
- no parse tree, therefore no tree alignment, therefore no word alignment...
- so we skipped pairs containing a parsing failure

Material: disregarded

To sum up, we:

- disregard autocue-subtitle pairs *not* 1-to-1 aligned (because text compression)
- measure CR in terms of tokens
- disregard pairs with CR>=0
- disregard pairs with parsing errors

Material: remaining

- we kept 5233 out of original 15289 pairs

	Min	Max	Mean	SD
Aut-tokens	2	43	15.41	5.48
Sub-tokens	1	29	10.26	3.72
CR	0.07	0.96	0.69	0.17

Overview

١. Introduction: sentence compression
٢. Material: subtitle corpus
٣. Analysis: observed compression phenomena
٤. Summary / Discussion

Analysis: edit operations

- Sentence compression can be regarded as a string transformation involving word deletion, substitution and insertion
- These edit operation can be deduced from the alignment of the syntax trees:
  - ➔ if an autocue word is *not* aligned (to a subtitle word), then it was deleted
  - ➔ if a subtitle word is *not* aligned (to an autocue word), then it was inserted
  - ➔ if different autocue and subtitle words are aligned, then substitution occurred
  - ➔ if alignments cross each other, then the word order was changed

Analysis: edit operations

- Several advantages over calculating conventional string edit distance
    - e.g. clearly distinguishes word order changes

Analysis: deletions

	Min	Max	Sum	Mean	SD
Del	1	34	34728	6.64	4.57
Sub	0	6	4116	0.79	0.94
Ins	0	17	7768	1.48	1.78
Reorder			1688	0.32	

- deletion is by far most frequent operation
- on average 7 words per sentence

Analysis: substitutions & insertions

	Min	Max	Sum	Mean	SD
Del	1	34	34728	6.64	4.57
Sub	0	6	4116	0.79	0.94
Ins	0	17	7768	1.48	1.78
Reorder			1688	0.32	

- perhaps surprising, insertions are more frequent then substitutions

Analysis: reordering

	Min	Max	Sum	Mean	SD
Del	1	34	34728	6.64	4.57
Sub	0	6	4116	0.79	0.94
Ins	0	17	7768	1.48	1.78
Reorder			1688	0.32	

- word reordering is a binary variable
- about 1 in 3 sentences is reordered

Analysis: subsequences

- the subtitle is a *subsequence* of the autocue if there are only deletions, i.e.
  - no substitutions
  - no insertions
  - no word order changes
- only 16% of all autocue sentences are proper subsequences!
- does this imply that a deletion model can *not* account for 84% of the observed data?

Analysis: subsequences

- No, because sentence compression is not a problem with a unique solution
    - just like NLG, MT, ...
- There may very well exist semantically equivalant compressions which *do* satisfy the subsequence constraint
- So how many of the observed non-subsequences have subsequence alternatives?

Analysis: subsequences

- manual exercise:
  - for a random sample of 200 non-subsequences
  - try to find a proper subsequence with the same meaning and the CR
  - performed by one author; checked by second

Aut: *in zijn residentie is het een chaos*
in his residence is it a chaos

Sub: *chaos heerst in de residentie*
chaos rules in the residence

Seq: *zijn residentie is een chaos*
his residence is a chaos

Analysis: subsequences

Difference in tokens between original and rewritten subtitle

Token-diff	Count	%
-2	4	2.0
-1	18	9.0
0	73	36.5
1	42	21.0
2	32	16.0
3	11	5.5
4	9	4.5
5	5	2.5
7	2	1.0
8	2	1.0
9	1	0.5
11	1	0.5

- 95 out of 200 (47%) can be rewritten as a subsequence with same CR (or smaller)
- 16% of original data was already subsequence
- so 55% (16% + 47% of 84%) is compatible with a deletion model

Analysis: remaining problems

- even though the subsequence constraint is not as problematic as it seemed, about 45% of the observed data is still violates a deletion model
- our exercise reveals examples where insertion, substitution and word order changes are essential for obtaining the targeted CR
- found three main categories:

    1) obligatory word reordering
    2) referring expressions
    3) paraphrasing

Analysis: obligatory reordering

- after deletion of a constituent, word reordering is often obligatory to preserve meaning and/or grammaticality
- observed in 24 out 200 sentences

Aut: *in PLAATS heeft IEMAND IETS besloten*
in location has somebody something decided

Sub: **heeft IEMAND IETS besloten*
has somebody something decided

IEMAND heeft IETS besloten
someone has something decided

Analysis: referring expressions

- referring expressions are often replaced by
  - a shorter, less precise expression
  - a real anaphor
- requires context modeling: transcends the per-sentence paradigm of sentence compression
- shows that generating referring expressions is relevant for an application like automatic subtitling

Aut: Many of them are deported by he Serbs in crammed trains

Sub: *Refugees are deported by train*

Analysis: paraphrasing

- fixed lexical paraphrases
  - *since a few years* → *nowadays/recently/now*
- paraphrases with slots

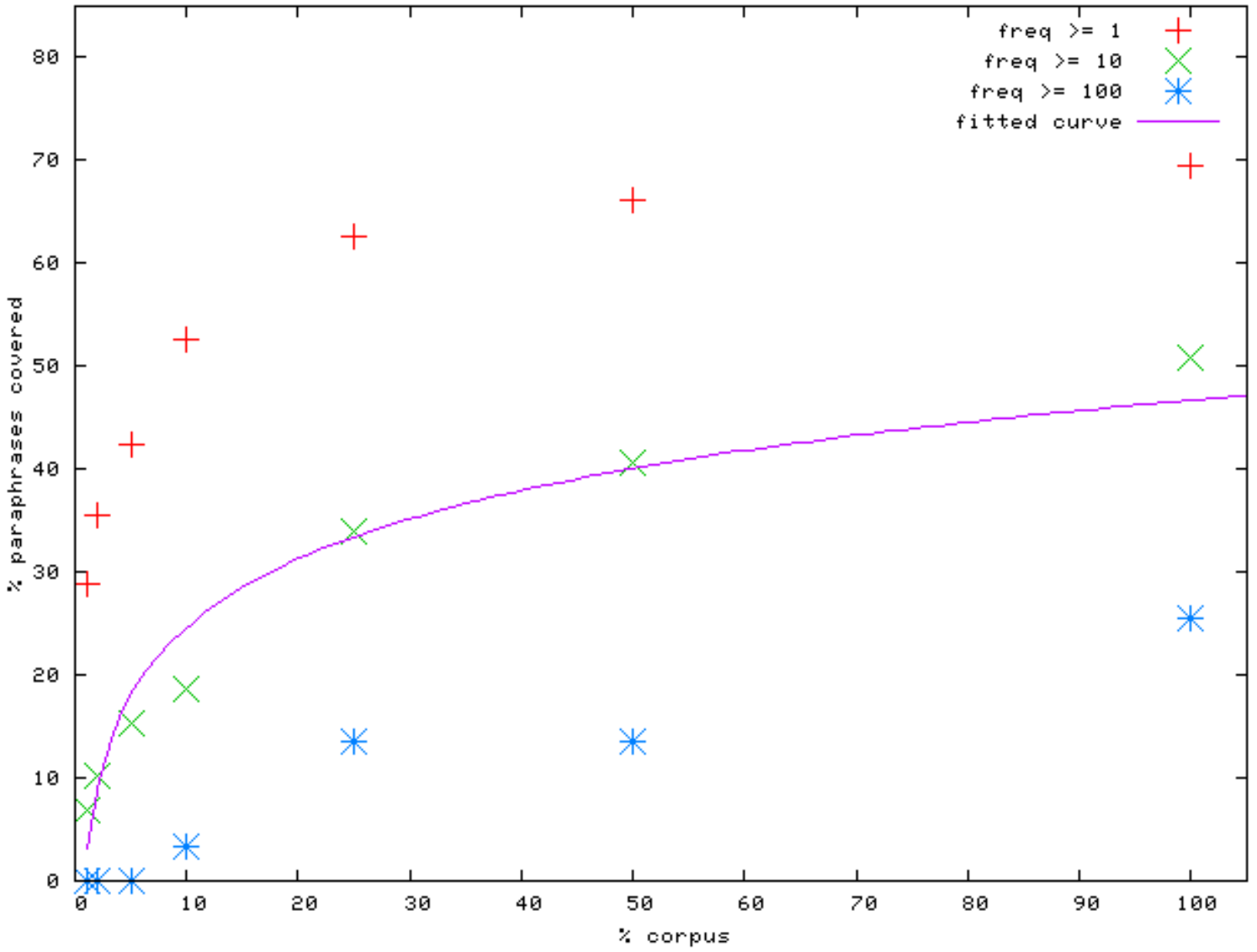
Aut: *X neemt het initiatief tot oprichting van Y*
X takes the initiative to raising of Y

Sub: *X zet Y op*
X sets Y up ("X raises Y")

Automatic paraphrase extraction

- there is more and more work on automatic paraphrase extraction(Lin & Pantel, 2001; Barzilay & Lee, 2003; Dolan et al; 2004; ...)
- how many of the paraphrases encountered in our sample can be automatically extracted from a text corpus?
- assuming a “perfect learner”, paraphrases must at least occur with a sufficient frequency in the text corpus
- Twente News Corpus: 325M words

Automatic paraphrase extraction



Overview

١. Introduction: sentence compression
٢. Material: subtitle corpus
٣. Analysis: observed compression phenomena
٤. Summary / Discussion

Summary

- deletion model of sentence compression:
  - delete any subset of words from the input sentence
  - while retaining important information and grammaticality
- can account for only 16% of observed compressions in the subtitle domain
- rewriting to proper subsequences suggests it can account for about 55%
- for the remaining 45%, substitution, insertions (and word order changes) are crucial
- issues: fix word order, referring expressions, paraphrasing

Discussion

- Is sentence compression an NLG task?
  - **no**, because for my application X I am happy with a simple deletion model which accounts for roughly 55% of the cases
  - **yes**, because I need more than deletion to account for the remaining 45% of the cases
- Sentence compression as part of NLG should include:
  - text revision / grammar-based transformation
  - generating (shorter) paraphrases
  - generating (shorter) referring expressions
  - sentence splitting & merging (aggregation)
  - ...