Talk to me in plain English please!

Explorations in Data-driven Text Simplification

Mirella Lapata Kristian Wooodsend

School of Informatics
University of Edinburgh

ESSLLI 2011, Ljubljana

Kristian Woodsend

The Challenge

A computer that analyses and generates text the way humans can would need:

- Syntactic and semantic parsing
- Robust word sense disambiguation
- Discourse understanding and coreference resolution
- Paraphrase recognition and generation
- Text rewriting capabilities
- Make inferences about what is described and whether it is important

The Challenge

A computer that analyses and generates text the way humans can would need:

- Syntactic and semantic parsing
- Robust word sense disambiguation
- Discourse understanding and coreference resolution
- Paraphrase recognition and generation
- Text rewriting capabilities
- Make inferences about what is described and whether it is important

This is way too difficult! Find a new job!

The Challenge

A computer that analyses and generates text the way humans can would need:

- Syntactic and semantic parsing
- Robust word sense disambiguation
- Discourse understanding and coreference resolution
- Paraphrase recognition and generation
- Text rewriting capabilities
- Make inferences about what is described and whether it is important

This is way too difficult! Find a new job!

Two Owl Tales

Tale 1

Owls are the order Strigiformes, comprising 200 bird of prey species. Owls hunt mostly small mammals, insects, and other birds though some species specialize in hunting fish.

Tale 2

An owl is a bird. There are about 200 kinds of owls. Owls' prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits).

Two Owl Tales

Tale 1

Owls are the order Strigiformes, comprising 200 bird of prey species. Owls hunt mostly small mammals, insects, and other birds though some species specialize in hunting fish.

Tale 2

An owl is a bird. There are about 200 kinds of owls. Owls' prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits).

Explain unfamiliar words or concepts

Two Owl Tales

Tale 1

Owls are the order Strigiformes, comprising 200 bird of prey species. Owls hunt mostly small mammals, insects, and other birds though some species specialize in hunting fish.

Tale 2

An owl is a bird. There are about 200 kinds of owls. Owls' prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits).

Substitute rare words with more familiar words or phrases

Two Owl Tales

Tale 1

Owls are the order Strigiformes, comprising 200 bird of prey species. Owls hunt mostly small mammals, insects, and other birds though some species specialize in hunting fish.

Tale 2

An owl is a bird. There are about 200 kinds of owls. Owls' prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits).

Simplify deep syntactic structures

Two Owl Tales

Tale 1

Owls are the order Strigiformes, comprising 200 bird of prey species. Owls hunt mostly small mammals, insects, and other birds though some species specialize in hunting fish.

Tale 2

An owl is a bird. There are about 200 kinds of owls. Owls' prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits).

Remove unnecessary and complicating detail

More Examples

Source	Previous calculations show that, due to the solar wind (which drops 30% of the sun's mass), Earth could escape to a higher orbit.
Target	Previous calculations show that Earth could escape to a higher orbit. This is due to the solar wind, which drops 30% of the sun's mass.

More Examples

Source	Previous calculations show that, due to the solar wind (which drops 30% of the sun's mass), Earth could escape to a higher orbit.
Target	Previous calculations show that Earth could escape to a higher orbit. This is due to the solar wind, which drops 30% of the sun's mass.

Simplification achieved by flattening of deep syntactic structures.

More Examples

Source	John Smith, who was very tired, walked his dog to the supermarket because he was hungry but he returned to his home still hungry and even more tired because the market was closed.
Target	John Smith was very tired. Nevertheless, he walked his dog to the supermarket because he was hungry. But the market was closed. So he returned to his home still hungry and even more tired.

More Examples

Source

John Smith, who was very tired, walked his dog to the supermarket because he was hungry but he returned to his home still hungry and even more tired because the market was closed.

Target

John Smith was very tired. Nevertheless, he walked his dog to the supermarket because he was hungry. But the market was closed. So he returned to his home still hungry and even more tired.

Simplification achieved by splitting sentences.

More Examples

Source These alterations are humble, but assist in circumventing the difficulties of ascertaining the meaning of obfuscated sentences.

Target These alterations are simple, but help in getting around the difficulties of finding the meaning of confusing sentences.

More Examples

Source	These alterations are humble, but assist in circumventing the difficulties of ascertaining the meaning of obfuscated sentences.
Target	These alterations are simple, but help in getting around the difficulties of finding the meaning of confusing sentences.

Simplification achieved by lexical substitutions.

The Simplification Task

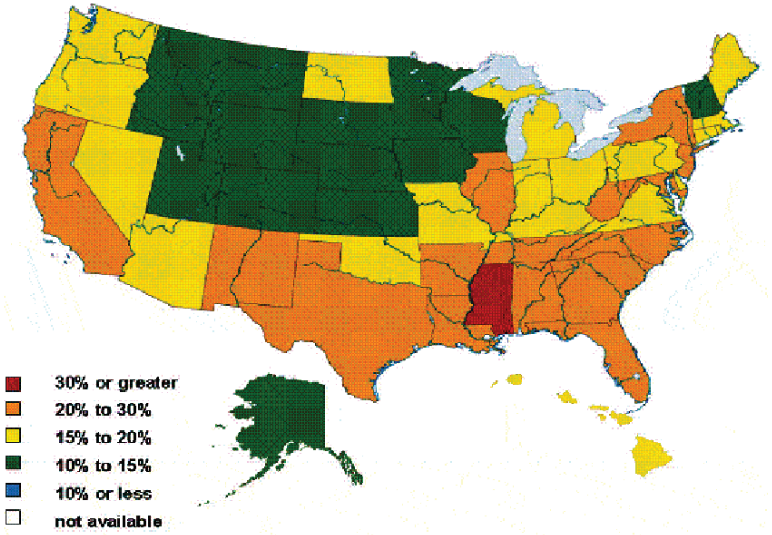
Goal: to make text easier to read and understand.

Task: involves a broad spectrum of rewrite operations including **deletion**, **substitution**, **insertion** and **reordering**.

- Simplification of **deeply-nested syntactic** elements
- Splitting clauses out into **stand-alone sentences**
- **Lexical substitution** of rare words
- **Content simplification** (e.g., removal of unimportant detail)

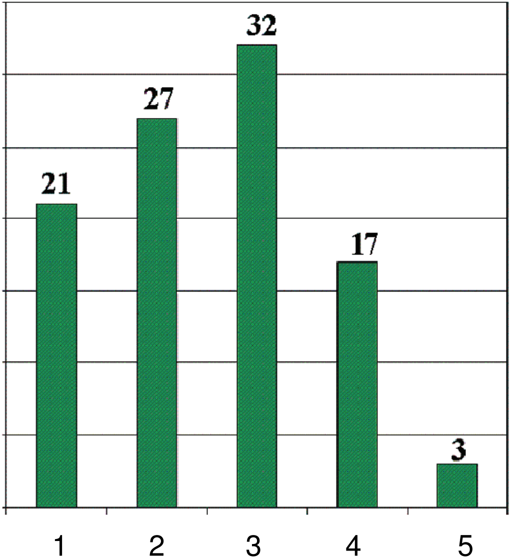
Why simplify?

1993 US National Adult Literacy Survey (grades 1-5)



Why simplify?

Percentage of the adult population for each literacy grade

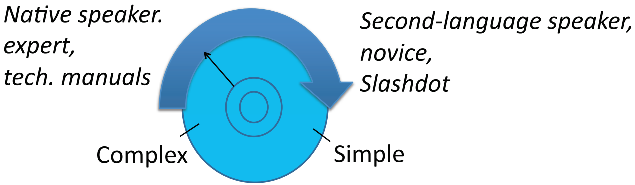


Why simplify?

1. Make more texts accessible to larger audiences.
2. Low-literacy readers (Inui et al., 2003)
3. Non-native speakers (Burstein et al., 2007)
4. Children and their teachers (Aluisio and Gasperin, 2010)
5. Individuals with language impairment (Carroll et al., 1999a)
6. Pre-processing for other NLP tasks (Chandrasekar et al., 1996; Vickrey and Koller, 2008)

Why simplify?

1. Make more texts accessible to larger audiences.
2. Low-literacy readers (Inui et al., 2003)
3. Non-native speakers (Burstein et al., 2007)
4. Children and their teachers (Aluisio and Gasperin, 2010)
5. Individuals with language impairment (Carroll et al., 1999a)
6. Pre-processing for other NLP tasks (Chandrasekar et al., 1996; Vickrey and Koller, 2008)
7. **Eventual goal:** a style dial for documents



Previous work

Rule-based methods for simplification:

- Hand-crafted syntactic rules (Chandrasekar et al., 1996; Siddharthan, 2004; Carroll et al., 1999b)
- Dictionary-based lexical simplifications (Devlin, 1999; Kaji et al., 2002; Inui et al., 2003)

Data-driven simplification (all using Simple English Wikipedia):

- Lexical substitutions from revision histories (Yatskar et al., 2010)
- Simplification as mono-lingual translation, using aligned sentences (Zhu et al., 2010; Coster and Kauchak, 2011)

This work

We want to generate simplified documents both in terms of style and content: learn **sentence simplification** and **content selection.**

✔approach should not be domain-specific
✔does not need pre-compiled resources or annotated corpora
✔can do both tasks

Generate new documents with joint model that optimizes:

1. informativeness of the selected content
2. simplicity of the rewritten text
3. overall grammaticality of the document

How to Simplify?

From Wikipedia, the free encyclopedia

WELCOME TO WIKIPEDIA

the **free encyclopedia** that anyone can change

Search

Search the 72,509 articles in the Simple English Wikipedia

How to write Simple English pages · Useful pages · Simple talk · Categories · Help

Schools Gateway (for users who want to make changes from a school)

About Wikipedia

This is the front page of the **Simple English Wikipedia**. *Wikipedias* are places where people work together to write encyclopedias in different languages. We use Simple English words and grammar here. The Simple English Wikipedia is for everyone! That includes children and adults who are learning English.

There are 72,509 articles on the Simple English Wikipedia. All of the pages are free to use. They have all been published under both the *Creative*

Selected article

Jupiter is the largest planet in the Solar System. It is the fifth planet from the Sun. Jupiter is classed as a gas giant. This is because it is very big and made up of gas. The other gas giants are Saturn,

How to Simplify?

The **Simple English Wikipedia** is an independently-maintained "spin-off" of Wikipedia.

How to Simplify?

The **Simple English Wikipedia** is an independently-maintained "spin-off" of Wikipedia.

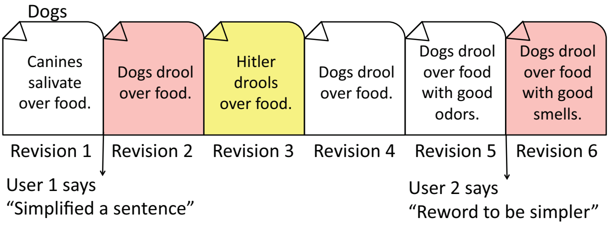
- Treat SimpleEW as translation of "complex" (regular) Wikipedia?
- But they aren't parallel: articles are written independently.
- Use alignment techniques to identify parallel sentences.

How to Simplify?

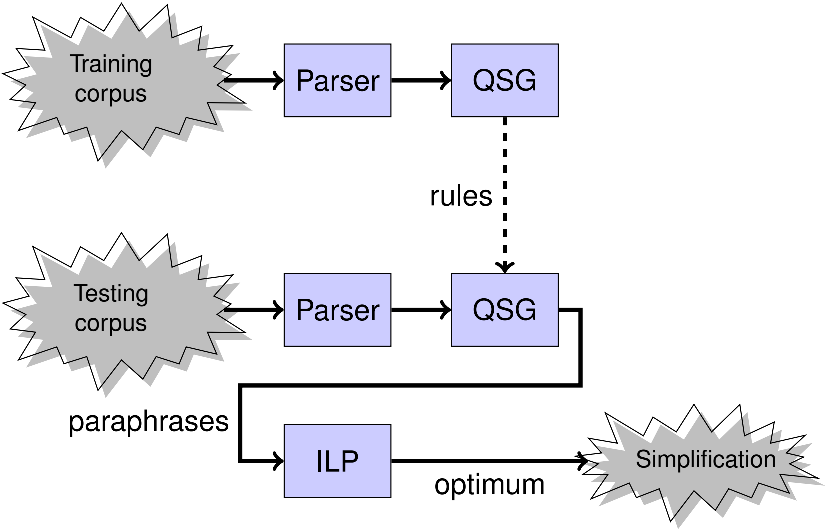
The **Simple English Wikipedia** is an independently-maintained "spin-off" of Wikipedia.

- Treat SimpleEW as translation of "complex" (regular) Wikipedia?
- But they aren't parallel: articles are written independently.
- Use alignment techniques to identify parallel sentences.

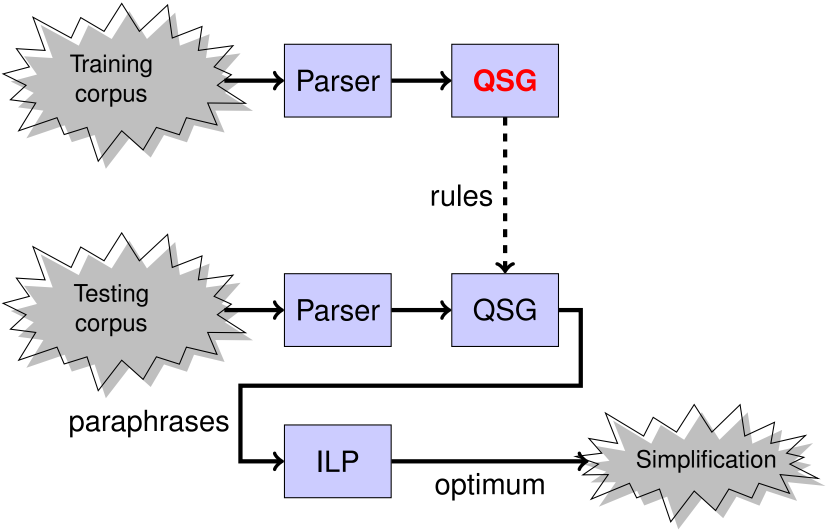
- Treat SimpleEW edits as instances of simplifications?
- But many edits aren't simplifications.
- Only consider revisions accompanied by "simpl" comments



Overview of the model



Overview of the model



Part I

Learning Simplification Paraphrases

Synchronous Grammars

Synchronous grammars are a way of simultaneously generating pairs of recursively related strings.

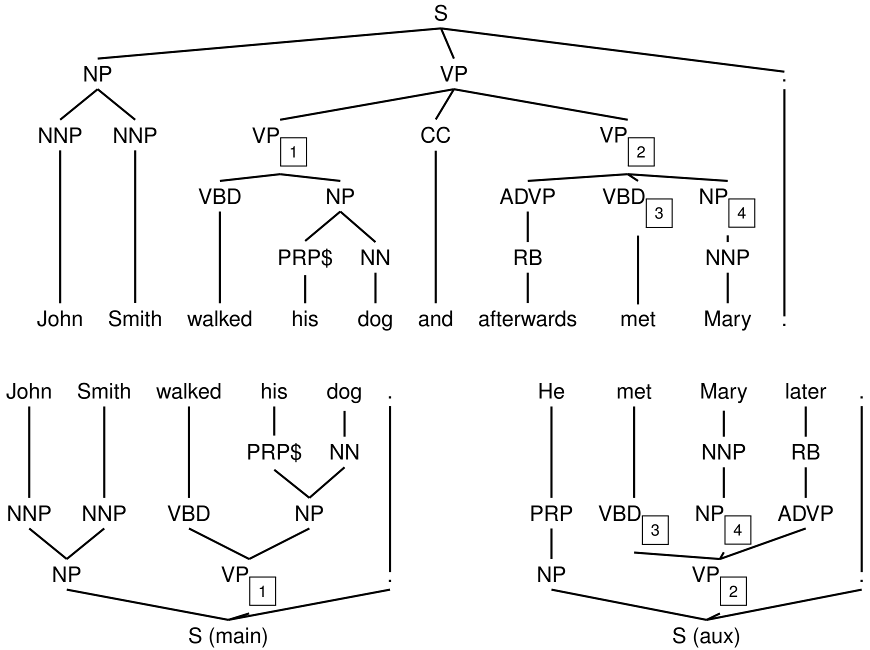
- Originally invented for programming language compilation
- Generalization of context-free grammar formalism to simultaneously produce strings in two languages.
- Have been used extensively in **syntax-based SMT:** inversion transduction grammar (ITG; Wu 1997), head transducers (Alshawi et al., 2000), hierarchical phrase-based translation (Chiang, 2007), synchronous tree substitution grammar (STSG; Eisner, 2003)

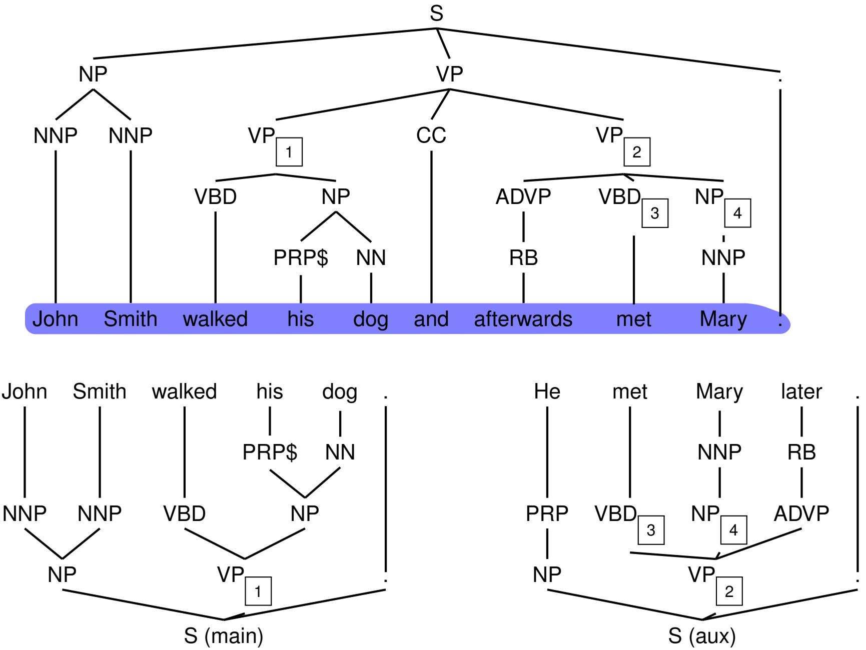
Synchronous Grammars

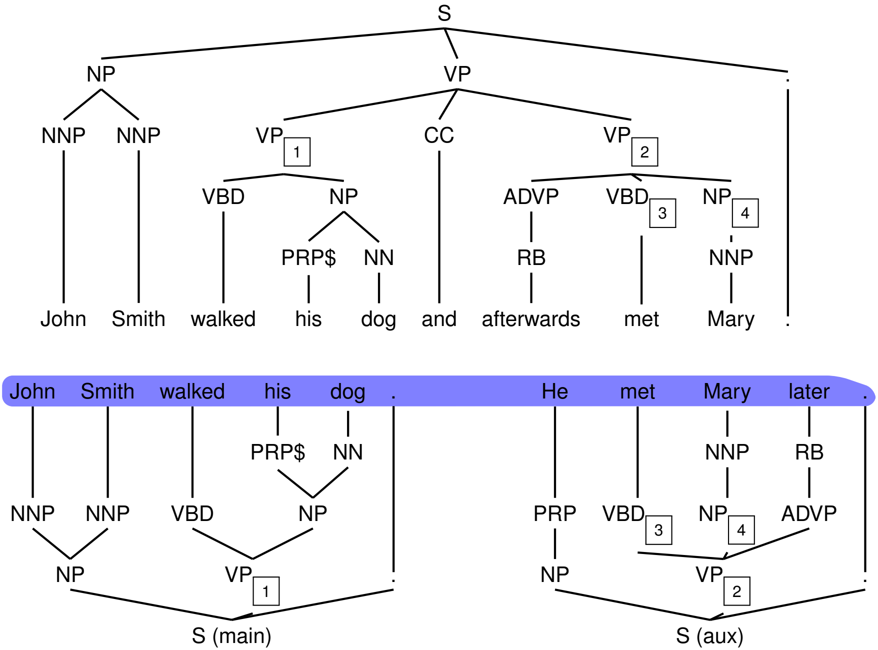
Synchronous grammars are a way of simultaneously generating pairs of recursively related strings.

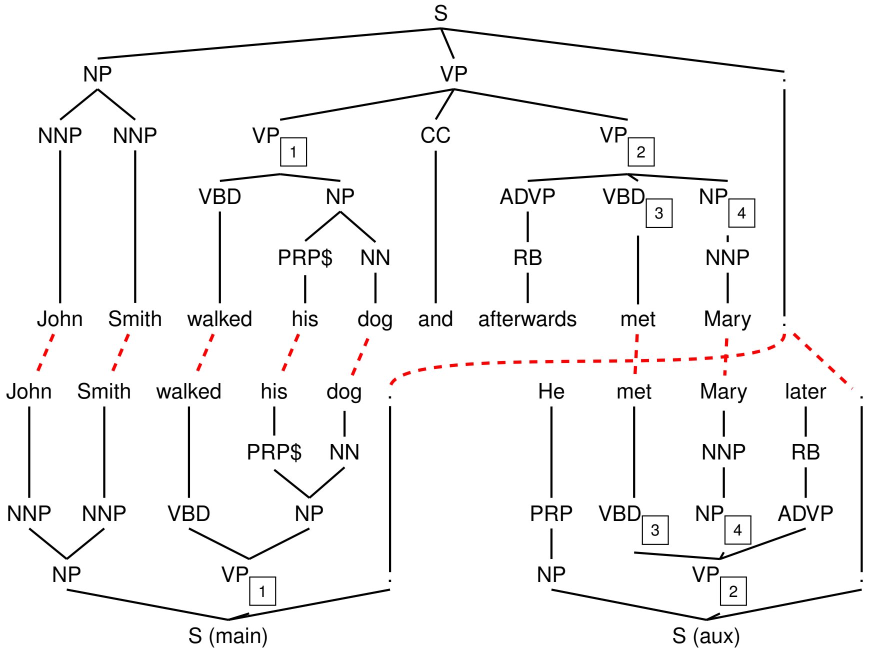
- Originally invented for programming language compilation
- Generalization of context-free grammar formalism to simultaneously produce strings in two languages.
- Have been used extensively in **syntax-based SMT:** inversion transduction grammar (ITG; Wu 1997), head transducers (Alshawi et al., 2000), hierarchical phrase-based translation (Chiang, 2007), synchronous tree substitution grammar (STSG; Eisner, 2003)

Quasi-synchronous grammar (QSG; Smith and Eisner, 2006) does not postulate strictly synchronous structure; target tree is “inspired” by source tree; allows to learn when only sub-trees align.

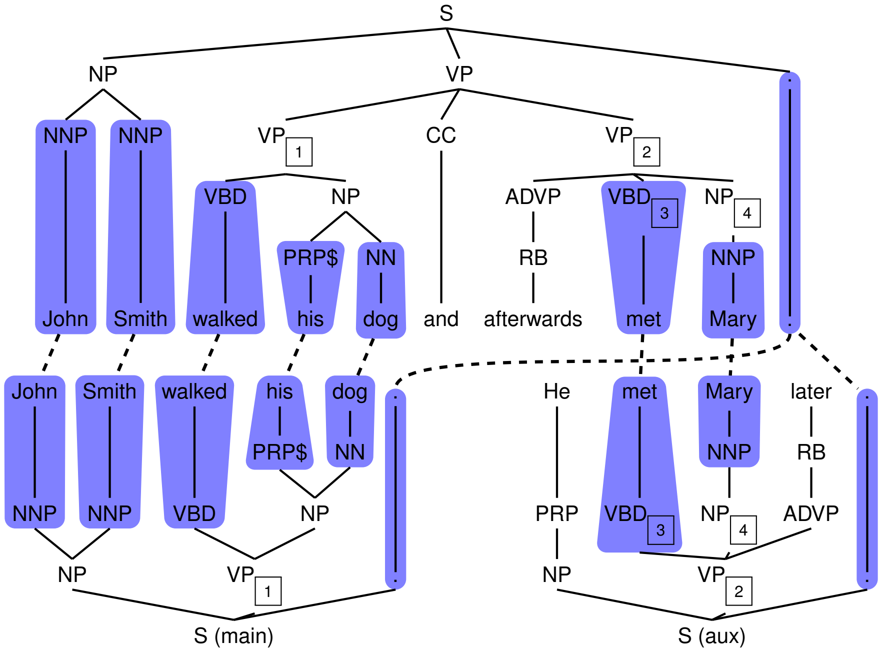

S
NP
VP
.
NNP
NNP
VP 1
CC
VP 2
VBD
NP
ADVP
VBD 3
NP 4
PRP$
NN
RB
NNP
John
Smith
walked
his
dog
and
afterwards
met
Mary
.
John
Smith
walked
his
dog
.
PRP$
NN
NNP
NNP
VBD
NP
NP
VP 1
.
S (main)
He
met
Mary
later
.
NNP
RB
PRP
VBD 3
NP 4
ADVP
NP
VP 2
.
S (aux)

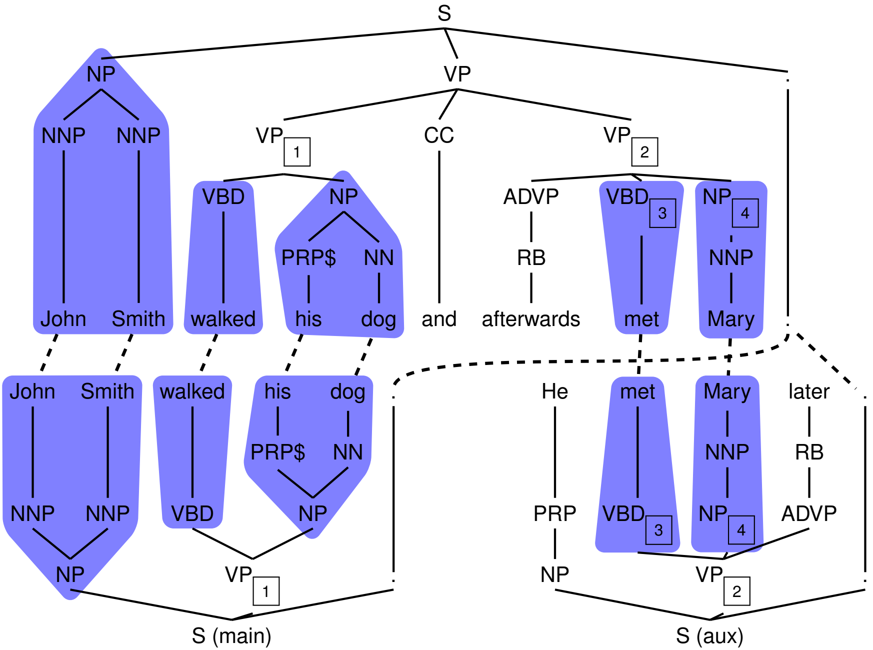

S
NP
VP
.
NNP
NNP
VP 1
CC
VP 2
VBD
NP
ADVP
VBD 3
NP 4
PRP$
NN
RB
NNP
John Smith walked his dog and afterwards met Mary .
John Smith walked his dog .
PRP$
NN
NNP
NNP
VBD
NP
NP
VP 1
.
S (main)
He met Mary later .
NNP
RB
PRP
VBD 3
NP 4
ADVP
NP
VP 2
.
S (aux)

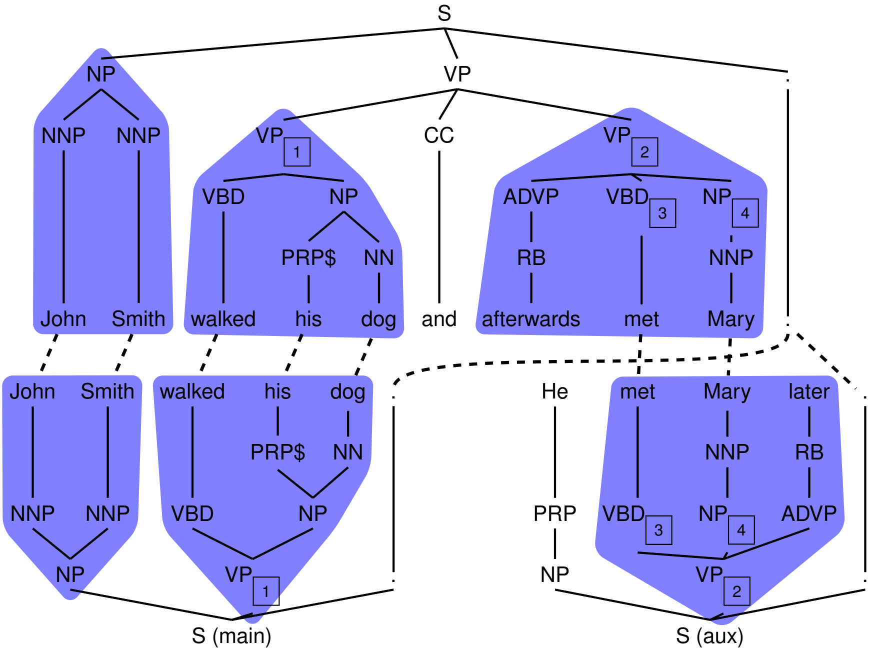

S
NP
VP
.
NNP
NNP
VP 1
CC
VP 2
VBD
NP
ADVP
VBD 3
NP 4
PRP$
NN
RB
NNP
John
Smith
walked
his
dog
and
afterwards
met
Mary
.
John Smith walked his dog . He met Mary later .
PRP$
NN
NNP
RB
NNP
NNP
VBD
NP
PRP
VBD 3
NP 4
ADVP
NP
VP 1
.
NP
VP 2
.
S (main)
S (aux)

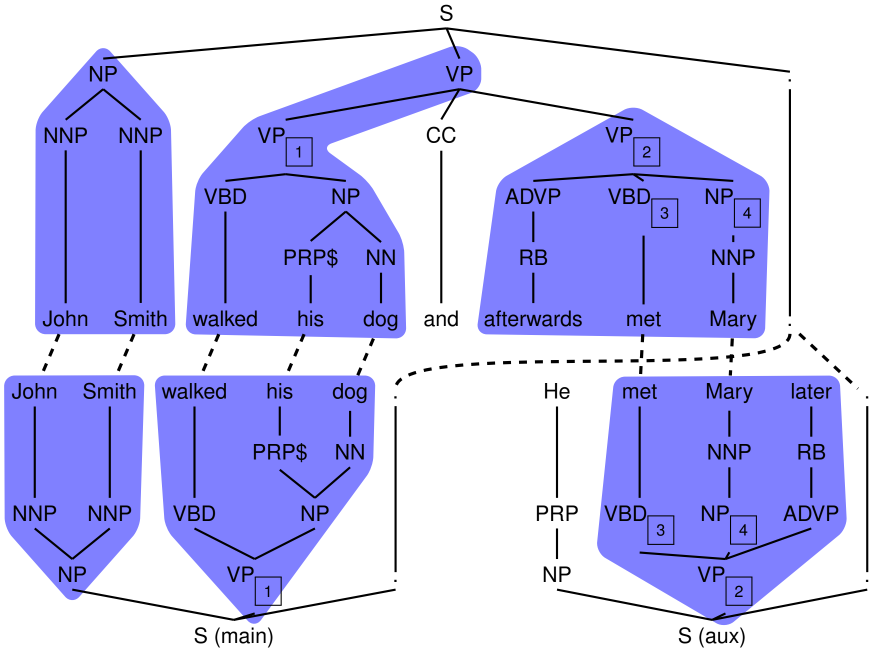


S
NP
VP
.
NNP
NNP
VP 1
CC
VP 2
VBD
NP
ADVP
VBD 3
NP 4
PRP$
NN
RB
NNP
John
Smith
walked
his
dog
and
afterwards
met
Mary
.
John
Smith
walked
his
dog
.
He
met
Mary
later
.
PRP$
NN
NNP
RB
NNP
NNP
VBD
NP
PRP
VBD 3
NP 4
ADVP
NP
VP 1
.
NP
VP 2
.
S (main)
S (aux)

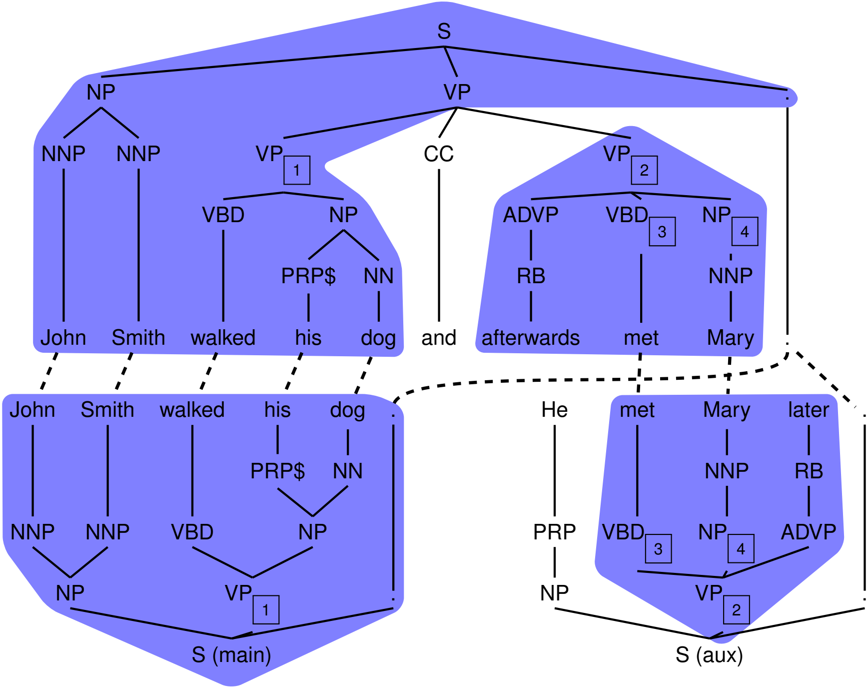

S
NP
VP
.
NNP
NNP
VP 1
CC
VP 2
VBD
NP
ADVP
VBD 3
NP 4
PRP$
NN
RB
NNP
John
Smith
walked
his
dog
and
afterwards
met
Mary
.
John
Smith
walked
his
dog
.
He
met
Mary
later
.
NNP
NNP
VBD
PRP$
NN
RB
NNP
NP
PRP
VBD 3
NP 4
ADVP
NP
VP 1
NP
VP 2
S (main)
S (aux)

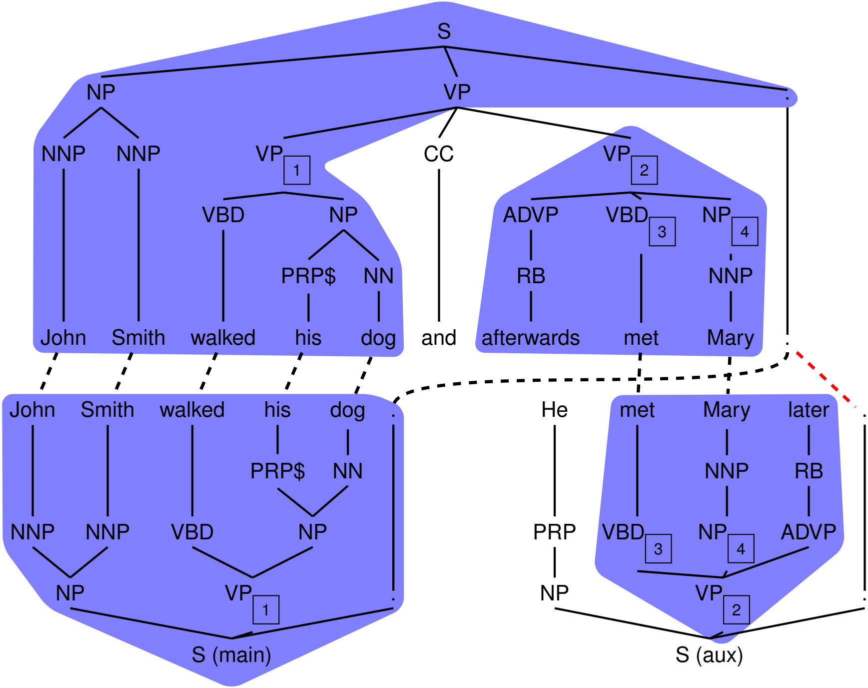

S
NP
VP
.
NNP NNP
VP 1
CC
VP 2
VBD
NP
ADVP
VBD 3
NP 4
PRP$ NN
RB
NNP
John Smith walked his dog and afterwards met Mary .
John Smith walked his dog
PRP$ NN
NNP NNP VBD NP
NP VP 1 .
S (main)
He met Mary later .
RB
PRP VBD 3 NP 4 ADVP
NNP
NP VP 2
S (aux)

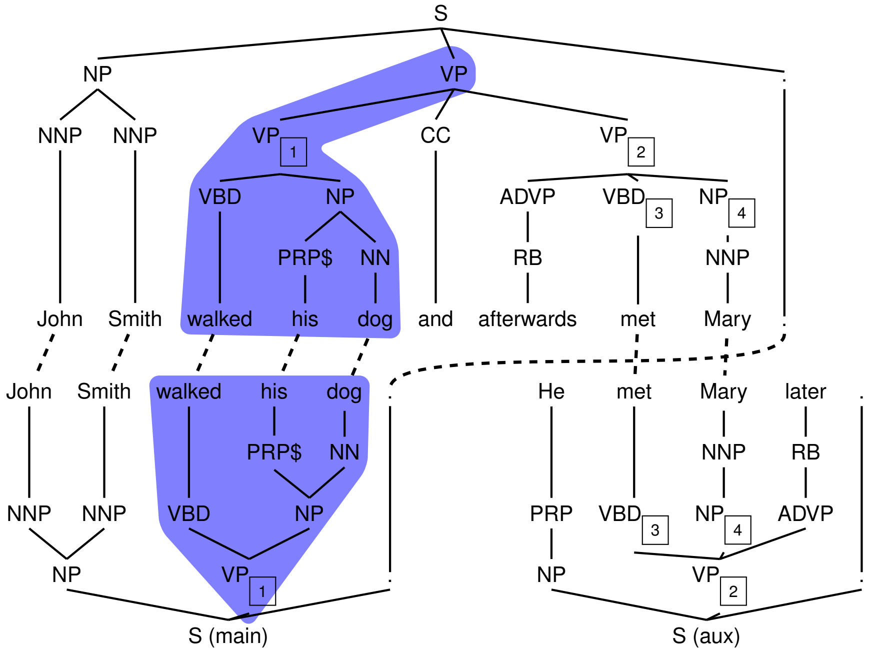

S
NP
VP
NNP
NNP
VP 1
CC
VP 2
VBD
NP
ADVP
VBD 3
NP 4
PRP$
NN
RB
NNP
John
Smith
walked
his
dog
and
afterwards
met
Mary
John
Smith
walked
his
dog
He
met
Mary
later
PRP$
NN
NNP
RB
NNP
NNP
VBD
NP
PRP
VBD 3
NP 4
ADVP
NP
VP 1
NP
VP 2
S (main)
S (aux)



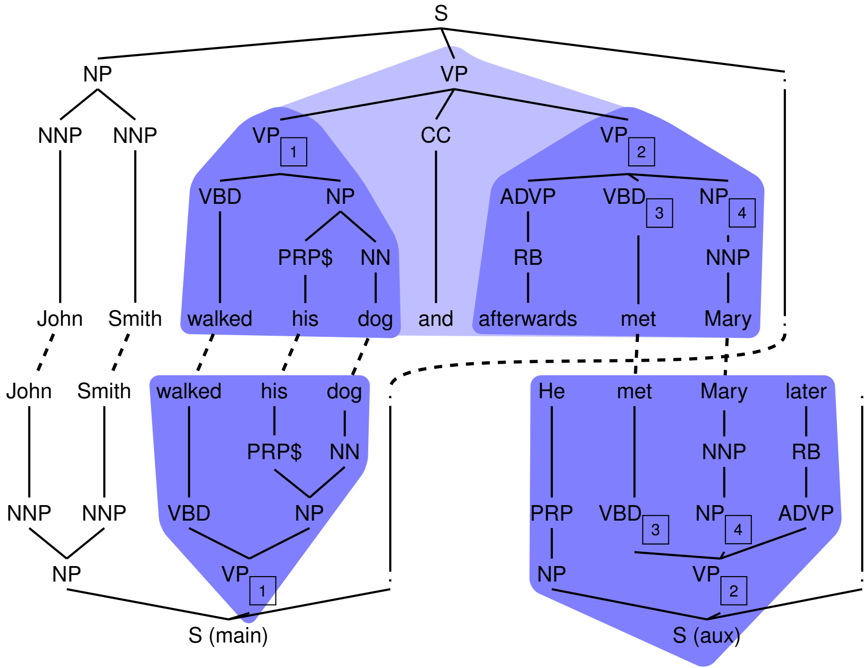
S
NP
VP
NNP NNP
VP 1
CC
VP 2
VBD
NP
ADVP
VBD 3
NP 4
PRP$ NN
RB
NNP
John Smith walked his dog and afterwards met Mary
John Smith walked his dog
He met Mary later
PRP$ NN
NNP RB
NNP NNP VBD NP
PRP VBD 3 NP 4 ADVP
NP VP 1
NP VP 2
S (main)
S (aux)


S
NP
VP
.
NNP
NNP
VP 1
CC
VP 2
VBD
NP
ADVP
VBD 3
NP 4
PRP$
NN
RB
NNP
John
Smith
walked
his
dog
and
afterwards
met
Mary
.
John
Smith
walked
his
dog
.
He
met
Mary
later
.
PRP$
NN
NNP
RB
NNP
NNP
VBD
NP
PRP
VBD 3
NP 4
ADVP
NP
VP 1
NP
VP 2
S (main)
S (aux)

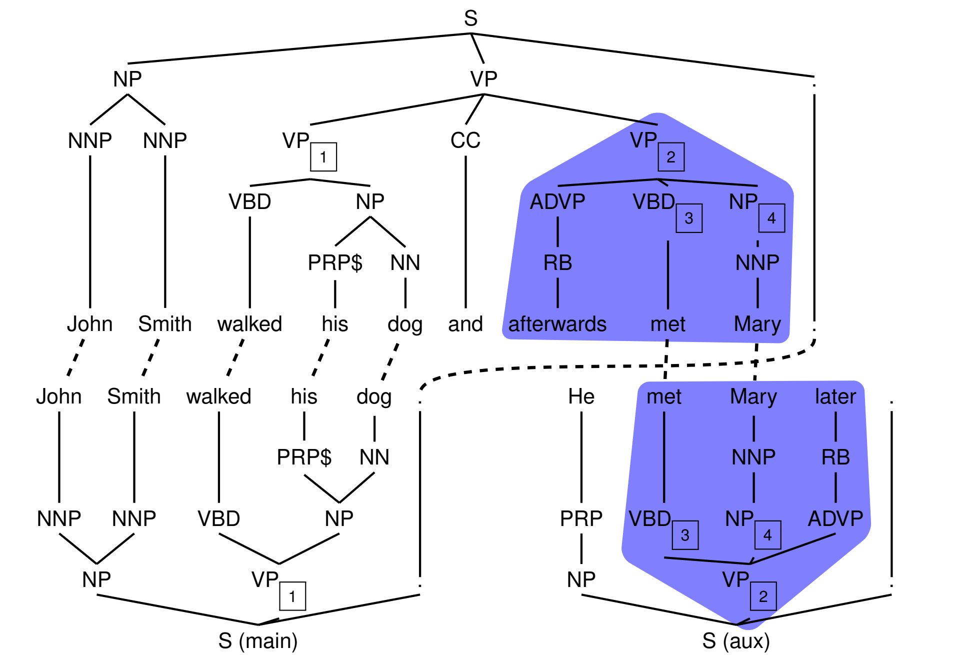

S
NP
VP
.
NNP
NNP
VP 1
CC
VP 2
VBD
NP
ADVP
VBD 3
NP 4
PRP$
NN
RB
NNP
John
Smith
walked
his
dog
and
afterwards
met
Mary
.
John
Smith
walked
his
dog
.
He
met
Mary
later
.
PRP$
NN
NNP
RB
NNP
NNP
VBD
NP
PRP
VBD 3
NP 4
ADVP
NP
VP 1
.
NP
VP 2
.
S (main)
S (aux)



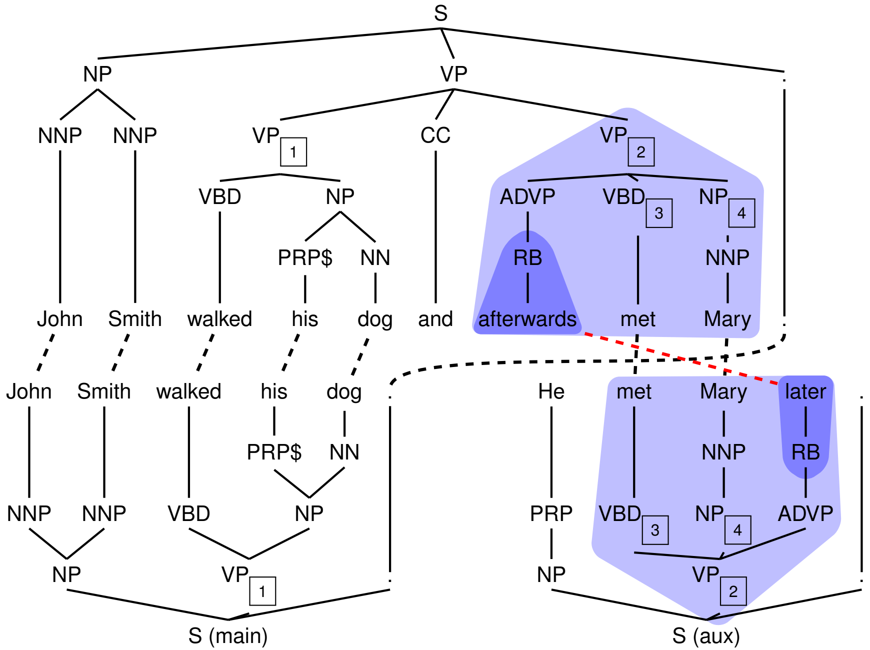
$$\langle \textbf{VP, VP} \rangle \rightarrow \langle \textbf{[VP}_{\boxed{1}} \textbf{ CC VP], [VP}_{\boxed{1}}\textbf{]} \rangle$$



⟨VP, ⟨VP, S⟩ ⟩ → ⟨[VP$_{\boxed{1}}$ *and* VP$_{\boxed{2}}$], ⟨[VP$_{\boxed{1}}$], [NP [PRP *He*] VP$_{\boxed{2}}$.]⟩ ⟩

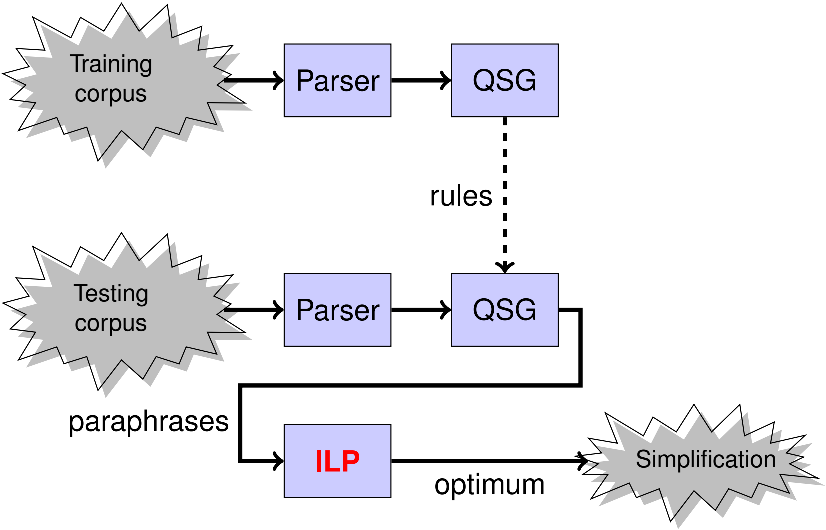


⟨VP, VP⟩ → ⟨[ADVP [RB *afterwards*] VBD$_{\boxed{3}}$ NP$_{\boxed{4}}$], [VBD$_{\boxed{3}}$ NP$_{\boxed{4}}$ ADVP [RB *later*]]⟩



$\langle$**RB, RB**$\rangle \rightarrow \langle$**[*afterwards*], [*later*]**$\rangle$

Overview of the Model



Part II

A Brief Introduction into ILP

What is Integer Linear Programming?

- Optimisation Technique.
- Find minimum or maximum value of a linear objective function.
- With respect to a set of constraints.
- ILP is an extension of Linear Programming; every LP has:
  - decision variables
  - a linear objective function
  - constraints on the variables

Linear Programming: Telfa Example

- Telfa Corporation manufactures tables and chairs.
- A table requires 1 hour of labour and 9 square board feet of wood.
- A chair requires 1 hour of labour and 5 square board feet of wood.
- They have 6 hours of labour and 45 square board feet of wood.
- Each table generates $8 of profit and each chair $5.
- **Goal:** Maximise profit.

(from Winston and Venkataramanan, 2003)

Telfa Example: LP Model

Decision Variables

$$\begin{aligned} x_1 &= \text{tables manufactured} \\ x_2 &= \text{chairs manufactured} \end{aligned}$$

Telfa Example: LP Model

Decision Variables

$$\begin{aligned} x_1 &= \text{tables manufactured} \\ x_2 &= \text{chairs manufactured} \end{aligned}$$

Objective function

$$\text{Profit} = 8x_1 + 5x_2$$

Telfa Example: LP Model

Decision Variables

$$\begin{aligned} x_1 &= \text{tables manufactured} \\ x_2 &= \text{chairs manufactured} \end{aligned}$$

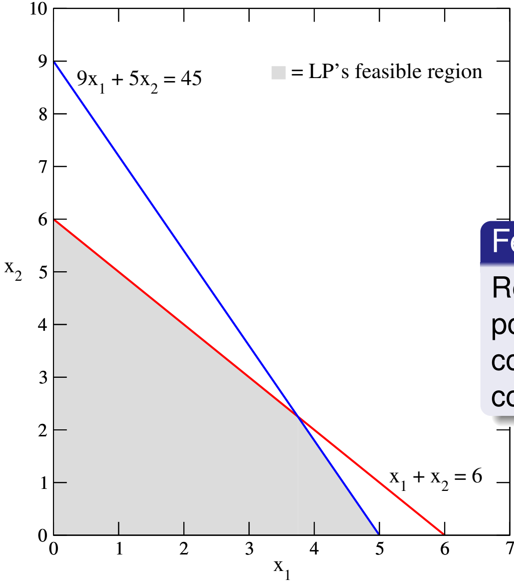
Objective function

$$\text{Profit} = 8x_1 + 5x_2$$

Constraints

Labour constraint	x_1	$+$	x_2	$\leq$	6
Wood constraint	$9x_1$	$+$	$5x_2$	$\leq$	45
Variable constraints			x_1	$\geq$	0
			x_2	$\geq$	0

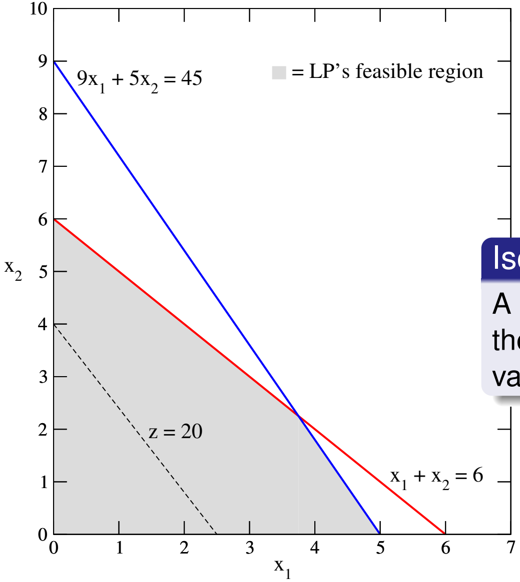
Solving LP Models



Feasible Region

Region that contains all the points that satisfy the LP constraints. A polyhedral convex set.

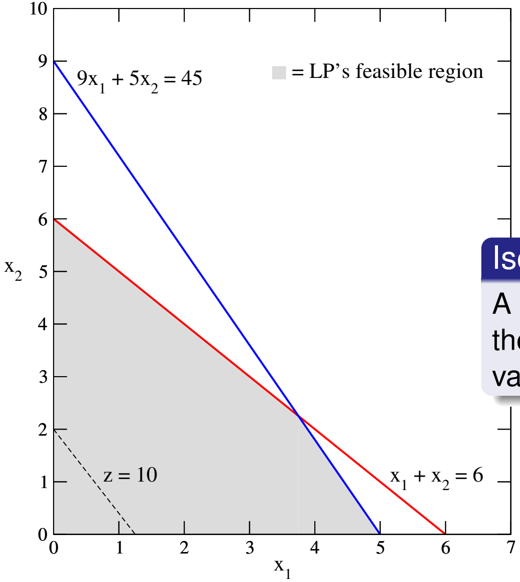
Solving LP Models



Isoprofit Line

A line on which all points have the same objective function value.

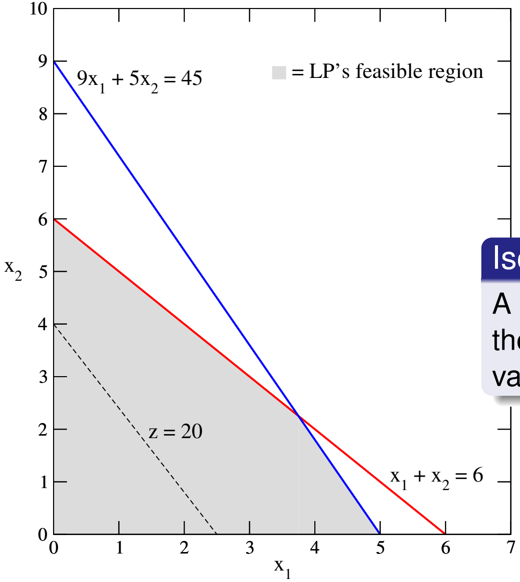
Solving LP Models



Isoprofit Line

A line on which all points have the same objective function value.

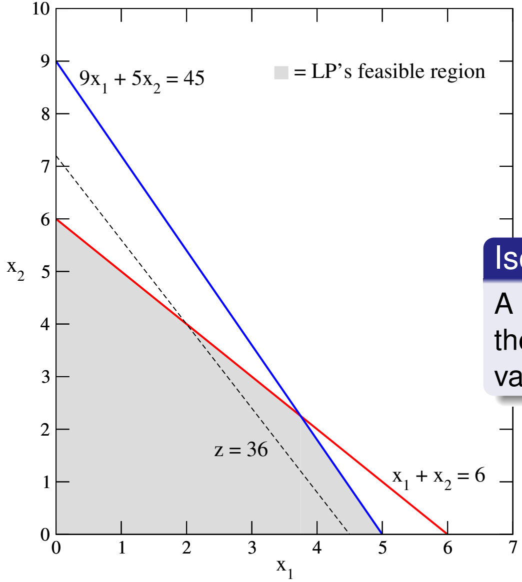
Solving LP Models



Isoprofit Line

A line on which all points have the same objective function value.

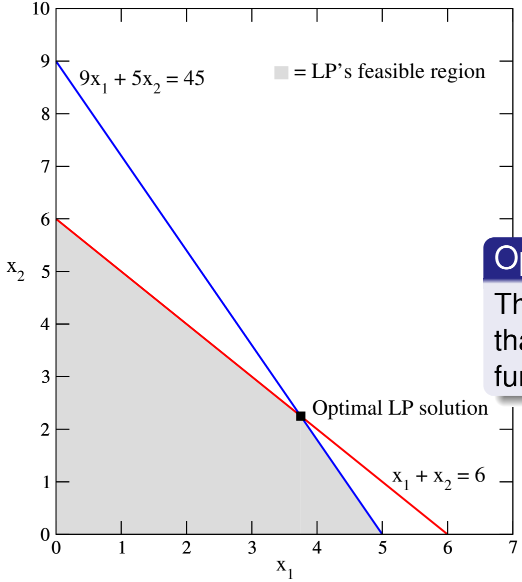
Solving LP Models



Isoprofit Line

A line on which all points have the same objective function value.

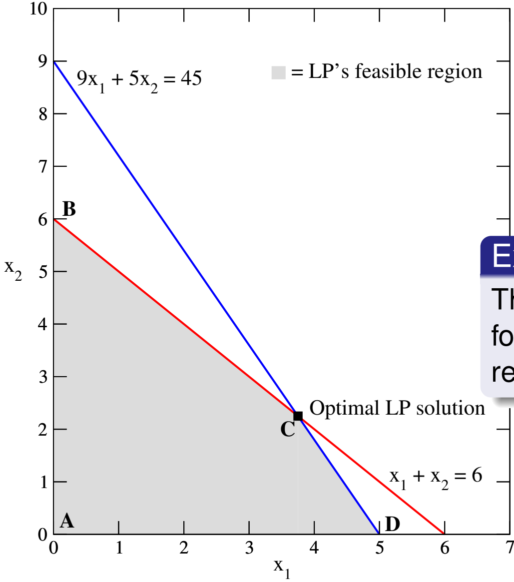
Solving LP Models



Optimal Solution

The point within feasible region that has maximum objective function value.

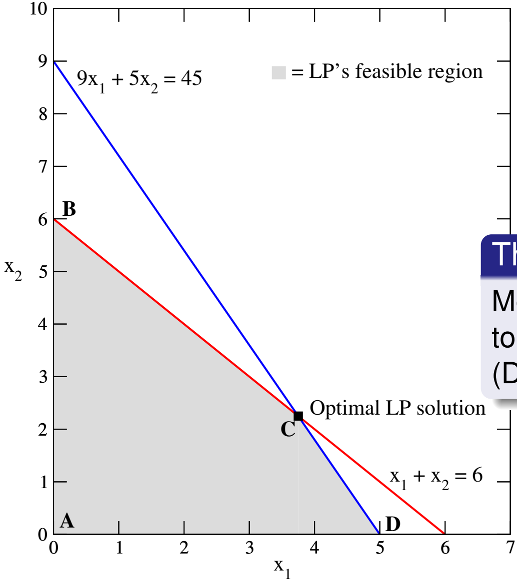
Solving LP Models



Extreme Point

The intersections of lines that form boundaries of feasible region.

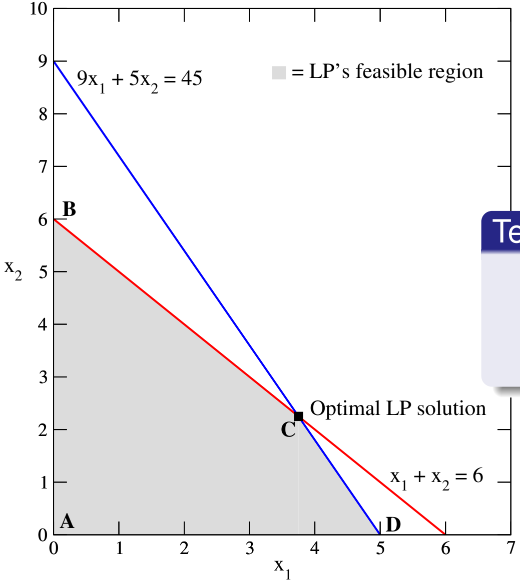
Solving LP Models



The simplex algorithm

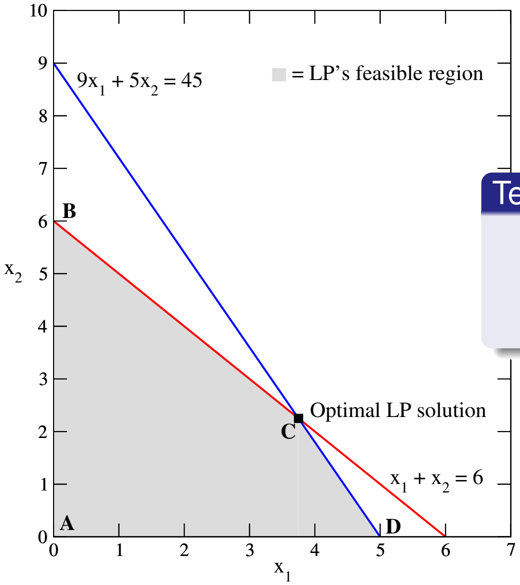
Moves from one extreme point to an adjacent extreme point (Dantzig, 1963).

Solving LP Models



Telfa Problem Solution

- $z = 41.25$
- $x_1 = 3.75$
- $x_2 = 2.25$



Telfa Problem Solution

- $z = 41.25$
- $x_1 = 3.75$
- $x_2 = 2.25$

We cannot build a fraction of a chair or table!

Integer Linear Programming

Integer linear programs are LP problems in which some or all of the variables must be non-negative integers.

Integer Linear Programming

Integer linear programs are LP problems in which some or all of the variables must be non-negative integers.

Telfa LP model

$$\max z = 8x_1 + 5x_2 \quad \text{(Objective function)}$$

subject to (s.t.)

$$\begin{array}{rcrcll} x_1 & + & x_2 & \leq & 6 & \text{(Labour constraint)} \\ 9x_1 & + & 5x_2 & \leq & 45 & \text{(Wood constraint)} \\ & & x_1 & \geq & 0; & \\ & & x_2 & \geq & 0; & \end{array}$$

Integer Linear Programming

Integer linear programs are LP problems in which some or all of the variables must be non-negative integers.

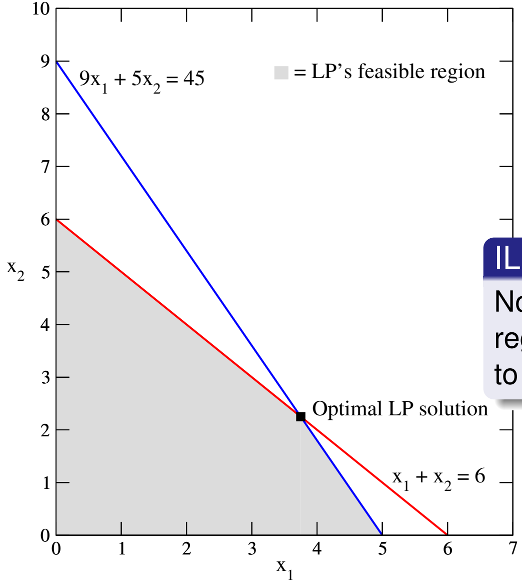
Telfa ILP model

$$\max z = 8x_1 + 5x_2 \quad \text{(Objective function)}$$

subject to (s.t.)

$$\begin{array}{rcrcll} x_1 & + & x_2 & \leq & 6 & \text{(Labour constraint)} \\ 9x_1 & + & 5x_2 & \leq & 45 & \text{(Wood constraint)} \\ & & x_1 & \geq & 0; & x_1 \text{ integer} \\ & & x_2 & \geq & 0; & x_2 \text{ integer} \end{array}$$

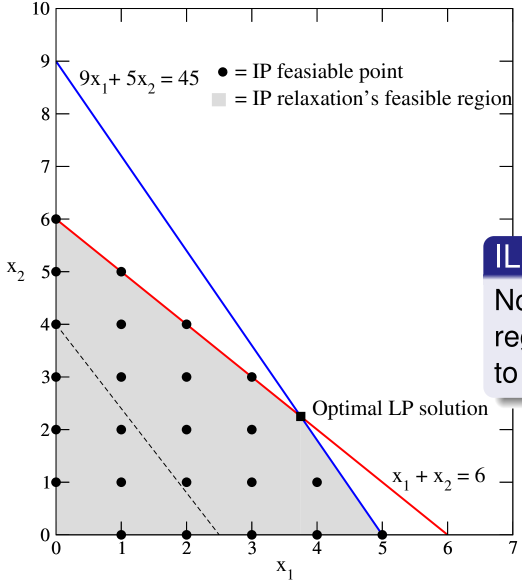
Solving ILP Models



ILP Solutions

Not all points within feasible region of an LP will be solutions to ILP problem.

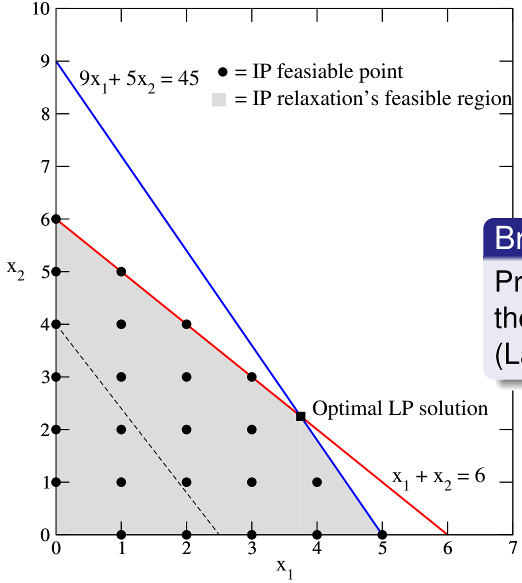
Solving ILP Models



ILP Solutions

Not all points within feasible region of an LP will be solutions to ILP problem.

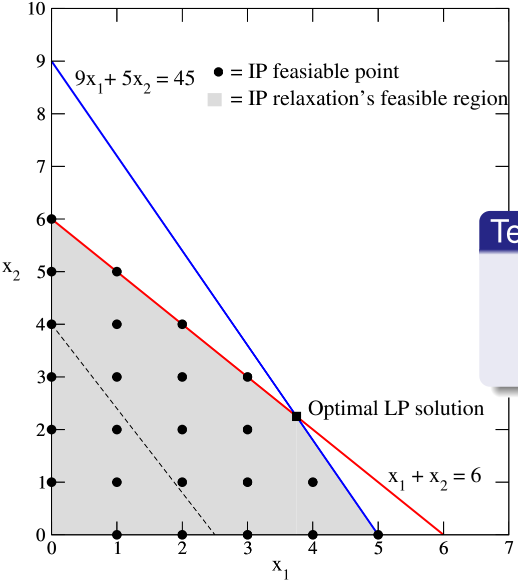
Solving ILP Models



Branch and Bound

Prunes sub-optimal sections of the feasibility region (Land and Doig, 1960).

Solving ILP Models



Telfa ILP Solution

- $z = 40$
- $x_1 = 5$
- $x_2 = 0$

Part III

Learning to Simplify Sentences

ILP for Sentence Simplification

$$
\begin{aligned}
\max_x \quad & \textstyle\sum_{i \in \mathcal{P}} g_i x_i + h_w + h_{sy} \\
\text{s.t.} \quad & x_j \rightarrow x_i \qquad \forall i \in \mathcal{P}, j \in \mathcal{D}_i \\
& x_i \rightarrow y_s \qquad \forall i \in \mathcal{P}, s \in \mathcal{A}_i \\
& \sum_{j \in \mathcal{C}_i} x_j = x_i \quad \forall i \in \mathcal{C}, j \in \mathcal{C}_i \\
& x_i \rightarrow y_s \qquad \forall s \in \mathcal{S}, i \in \mathcal{P}_s \\
& \sum_{s \in \mathcal{S}} y_i \geq 1 \quad x_i \in \{0, 1\} \forall i \in \mathcal{P} \\
& \qquad\qquad y_s \in \{0, 1\} \forall s \in \mathcal{S}.
\end{aligned}
$$

- Parse tree nodes x, Sentences y
- Rewrite probabilities g_i
- Readability indices h_w and h_{sy}
- Build tree
- Sentence splitting
- Ensure single QSG choice
- Ensure logical consistency

ILP for Sentence Simplification

$$
\begin{aligned}
\max_x \quad & \sum_{i \in \mathcal{P}} g_i x_i + h_w + h_{sy} \\
\text{s.t.} \quad & x_j \rightarrow x_i && \forall i \in \mathcal{P}, j \in \mathcal{D}_i \\
& x_i \rightarrow y_s && \forall i \in \mathcal{P}, s \in \mathcal{A}_i \\
& \sum_{j \in \mathcal{C}_i} x_j = x_i && \forall i \in \mathcal{C}, j \in \mathcal{C}_i \\
& x_i \rightarrow y_s && \forall s \in \mathcal{S}, i \in \mathcal{P}_s \\
& \sum_{s \in \mathcal{S}} y_i \geq 1 && x_i \in \{0,1\} \forall i \in \mathcal{P} \\
& && y_s \in \{0,1\} \forall s \in \mathcal{S}.
\end{aligned}
$$

- Parse tree nodes x, Sentences y
- Rewrite probabilities g_i
- Readability indices h_w and h_{sy}
- Build tree
- Sentence splitting
- Ensure single QSG choice
- Ensure logical consistency

ILP for Sentence Simplification

$$
\begin{aligned}
\max_x \quad & \textstyle\sum_{i \in \mathcal{P}} g_i x_i + h_w + h_{sy} \\
\text{s.t.} \quad & x_j \rightarrow x_i \qquad \forall i \in \mathcal{P}, j \in \mathcal{D}_i \\
& x_i \rightarrow y_s \qquad \forall i \in \mathcal{P}, s \in \mathcal{A}_i \\
& \sum_{j \in \mathcal{C}_i} x_j = x_i \quad \forall i \in \mathcal{C}, j \in \mathcal{C}_i \\
& x_i \rightarrow y_s \qquad \forall s \in \mathcal{S}, i \in \mathcal{P}_s \\
& \sum_{s \in \mathcal{S}} y_i \geq 1 \quad x_i \in \{0,1\} \forall i \in \mathcal{P} \\
& \qquad\qquad y_s \in \{0,1\} \forall s \in \mathcal{S}.
\end{aligned}
$$

- Parse tree nodes x, Sentences y
- Rewrite probabilities g_i
- Readability indices h_w and h_{sy}
- Build tree
- Sentence splitting
- Ensure single QSG choice
- Ensure logical consistency

ILP for Sentence Simplification

$$
\begin{aligned}
\max_x \quad & \textstyle\sum_{i \in \mathcal{P}} g_i x_i + h_w + h_{sy} \\
\text{s.t.} \quad & x_j \rightarrow x_i \qquad \forall i \in \mathcal{P}, j \in \mathcal{D}_i \\
& x_i \rightarrow y_s \qquad \forall i \in \mathcal{P}, s \in \mathcal{A}_i \\
& \sum_{j \in \mathcal{C}_i} x_j = x_i \qquad \forall i \in \mathcal{C}, j \in \mathcal{C}_i \\
& x_i \rightarrow y_s \qquad \forall s \in \mathcal{S}, i \in \mathcal{P}_s \\
& \sum_{s \in \mathcal{S}} y_i \geq 1 \qquad x_i \in \{0,1\} \forall i \in \mathcal{P} \\
& \qquad\qquad\qquad y_s \in \{0,1\} \forall s \in \mathcal{S}.
\end{aligned}
$$

- Parse tree nodes x, Sentences y
- Rewrite probabilities g_i
- Readability indices h_w and h_{sy}
- Build tree
- Sentence splitting
- Ensure single QSG choice
- Ensure logical consistency

ILP for Sentence Simplification

$$\max_x \quad \sum_{i \in \mathcal{P}} g_i x_i + h_w + h_{sy}$$

$$\text{s.t.} \quad x_j \rightarrow x_i \qquad \forall i \in \mathcal{P}, j \in \mathcal{D}_i$$

$$x_i \rightarrow y_s \qquad \forall i \in \mathcal{P}, s \in \mathcal{A}_i$$

$$\sum_{j \in \mathcal{C}_i} x_j = x_i \qquad \forall i \in \mathcal{C}, j \in \mathcal{C}_i$$

$$x_i \rightarrow y_s \qquad \forall s \in \mathcal{S}, i \in \mathcal{P}_s$$

$$\sum_{s \in \mathcal{S}} y_i \geq 1 \qquad x_i \in \{0, 1\} \forall i \in \mathcal{P}$$

$$y_s \in \{0, 1\} \forall s \in \mathcal{S}.$$

- Parse tree nodes x, Sentences y
- Rewrite probabilities g_i
- Readability indices h_w and h_{sy}
- Build tree
- Sentence splitting
- Ensure single QSG choice
- Ensure logical consistency

ILP for Sentence Simplification

$$\max_x \quad \sum_{i \in \mathcal{P}} g_i x_i + h_w + h_{sy}$$

$$\text{s.t.} \quad x_j \rightarrow x_i \qquad \forall i \in \mathcal{P}, j \in \mathcal{D}_i$$

$$x_i \rightarrow y_s \qquad \forall i \in \mathcal{P}, s \in \mathcal{A}_i$$

$$\sum_{j \in \mathcal{C}_i} x_j = x_i \qquad \forall i \in \mathcal{C}, j \in \mathcal{C}_i$$

$$x_i \rightarrow y_s \qquad \forall s \in \mathcal{S}, i \in \mathcal{P}_s$$

$$\sum_{s \in \mathcal{S}} y_i \geq 1 \qquad x_i \in \{0, 1\} \forall i \in \mathcal{P}$$

$$y_s \in \{0, 1\} \forall s \in \mathcal{S}.$$

- Parse tree nodes x, Sentences y
- Rewrite probabilities g_i
- Readability indices h_w and h_{sy}
- Build tree
- Sentence splitting
- Ensure single QSG choice
- Ensure logical consistency

ILP for Sentence Simplification

$$
\begin{aligned}
\max_x \quad & \textstyle\sum_{i \in \mathcal{P}} g_i x_i + h_w + h_{sy} \\
\text{s.t.} \quad & x_j \rightarrow x_i \quad && \forall i \in \mathcal{P}, j \in \mathcal{D}_i \\
& x_i \rightarrow y_s \quad && \forall i \in \mathcal{P}, s \in \mathcal{A}_i \\
& \sum_{j \in \mathcal{C}_i} x_j = x_i \quad && \forall i \in \mathcal{C}, j \in \mathcal{C}_i \\
& x_i \rightarrow y_s \quad && \forall s \in \mathcal{S}, i \in \mathcal{P}_s \\
& \sum_{s \in \mathcal{S}} y_i \geq 1 \quad && x_i \in \{0,1\} \forall i \in \mathcal{P} \\
& && y_s \in \{0,1\} \forall s \in \mathcal{S}.
\end{aligned}
$$

- Parse tree nodes x, Sentences y
- Rewrite probabilities g_i
- Readability indices h_w and h_{sy}
- Build tree
- Sentence splitting
- Ensure single QSG choice
- Ensure logical consistency

Objective of the Model

$$\max_{x} \quad \sum_{i \in \mathcal{P}} g_i x_i + h_w + h_{sy}$$

Objective of the Model

$$\max_{x} \quad \sum_{i \in \mathcal{P}} g_i x_i + h_w + h_{sy}$$

- Log-probability for rewriting: $g_i = \log\left(\frac{n_r}{N_r}\right)$.

Objective of the Model

$$\boxed{\max_{x} \quad \sum_{i \in \mathcal{P}} g_i x_i + {\color{red}h_w} + h_{sy}}$$

- Log-probability for rewriting: $g_i = \log\left(\frac{n_r}{N_r}\right)$.
- **Number of words** against **target words per sentence**:

$$h_w(x, y) = \text{wps} \times \sum_{i \in \mathcal{S}} y_i - \sum_{i \in \mathcal{P}} l_i^{(w)} x_i.$$

Objective of the Model

$$\max_{x} \quad \sum_{i \in \mathcal{P}} g_i x_i + h_w + h_{sy}$$

- Log-probability for rewriting: $g_i = \log\left(\frac{n_r}{N_r}\right)$.
- **Number of words** against **target words per sentence**:

$$h_w(x, y) = \text{wps} \times \sum_{i \in \mathcal{S}} y_i - \sum_{i \in \mathcal{P}} l_i^{(w)} x_i.$$

- **Number of syllables** against **target syllables per word**:

$$h_{sy}(x) = \text{spw} \times \sum_{i \in \mathcal{P}} l_i^{(w)} x_i - \sum_{i \in \mathcal{P}} l_i^{(sy)} x_i.$$

Objective of the Model

$$\boxed{\max_{x} \quad \sum_{i \in \mathcal{P}} g_i x_i + h_w + h_{sy}}$$

- Log-probability for rewriting: $g_i = \log\left(\frac{n_r}{N_r}\right)$.
- **Number of words** against **target words per sentence**:

$$h_w(x, y) = \text{wps} \times \sum_{i \in \mathcal{S}} y_i - \sum_{i \in \mathcal{P}} l_i^{(w)} x_i.$$

- **Number of syllables** against **target syllables per word**:

$$h_{sy}(x) = \text{spw} \times \sum_{i \in \mathcal{P}} l_i^{(w)} x_i - \sum_{i \in \mathcal{P}} l_i^{(sy)} x_i.$$

- Linear approximation of Flesch-Kincaid Grade Level

Experimental Setup

Data sets:

1. Train model on MainEW–SimpleEW aligned sentences
2. And aligned sentences from revision histories
3. Use same test set as Zhu et al. (2010)

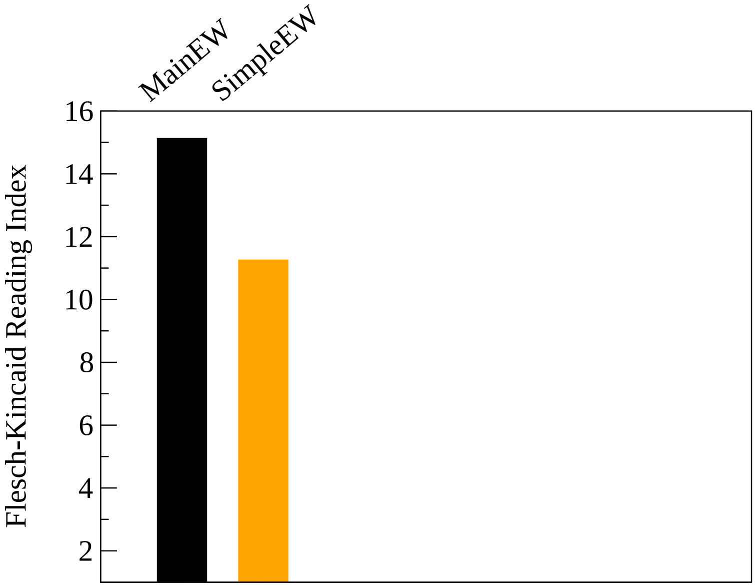
Comparison systems:

1. Zhu et al.'s (2010) system (based on Yamada and Knight 2001)
2. Joshua tree-based SMT system (Li et al., 2010)
3. SimpleEW's editor SpencerK's lexical substitution system

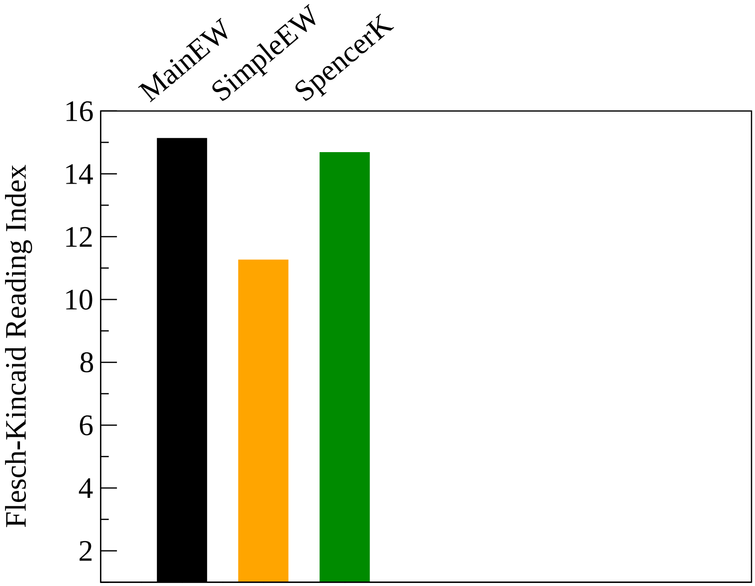
Evaluation:

1. Flesch Kincaid reading index
2. Simplicity Is the target sentence simpler than the source?
3. Grammaticality Is the target sentence grammatical?
4. Meaning Does the target preserve the meaning of the source?

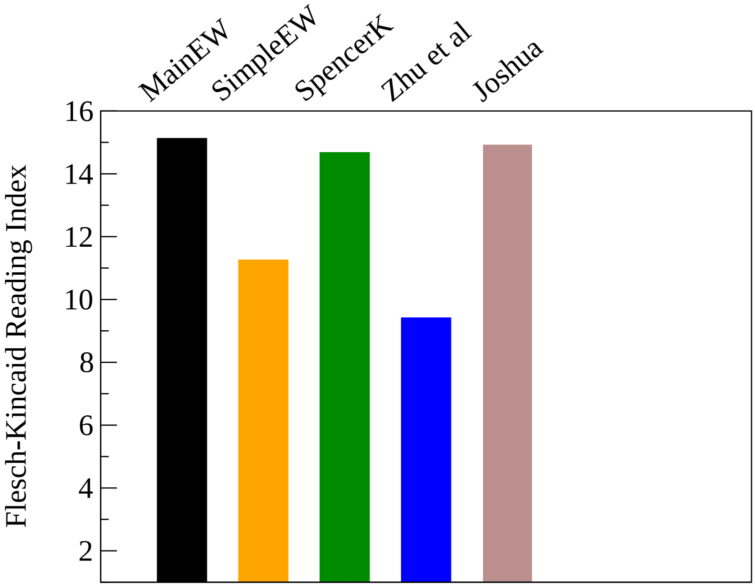
System Readability



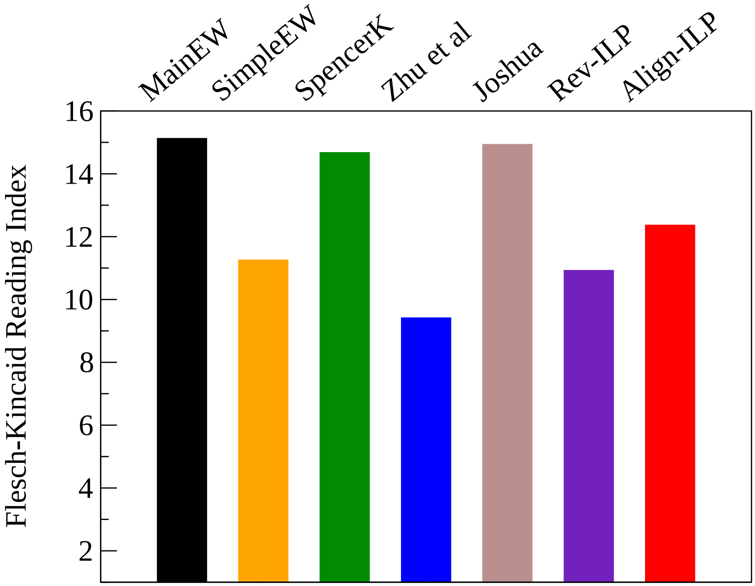
Readability and accuracy measures



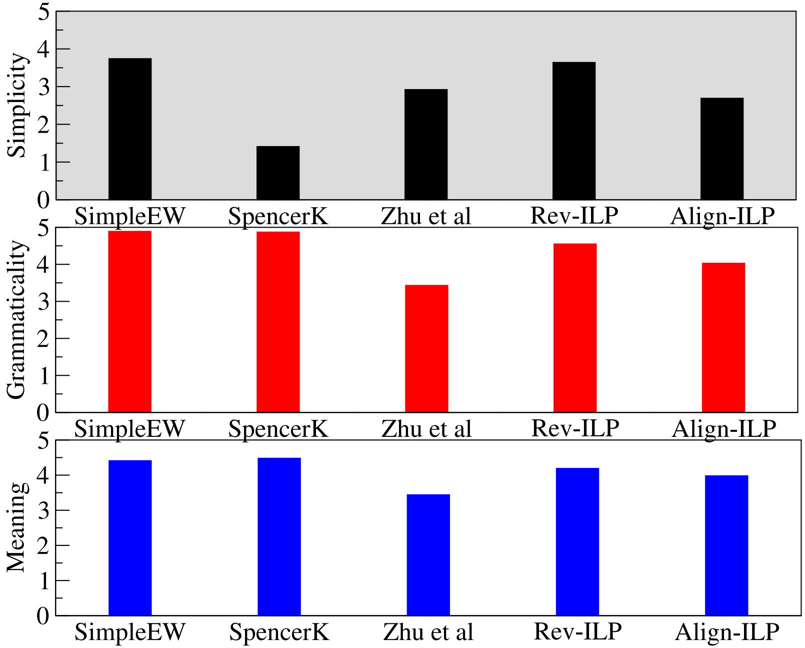
Readability and accuracy measures



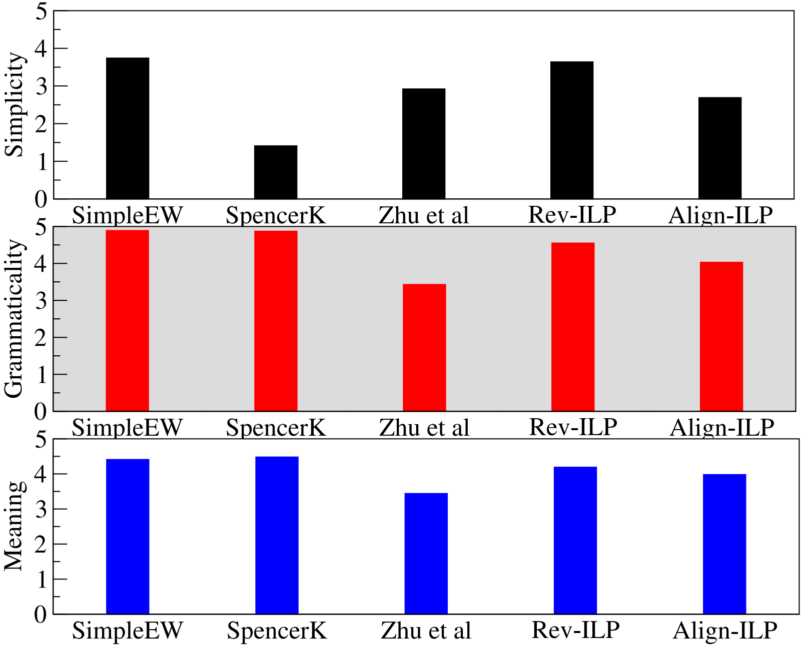
Readability and accuracy measures



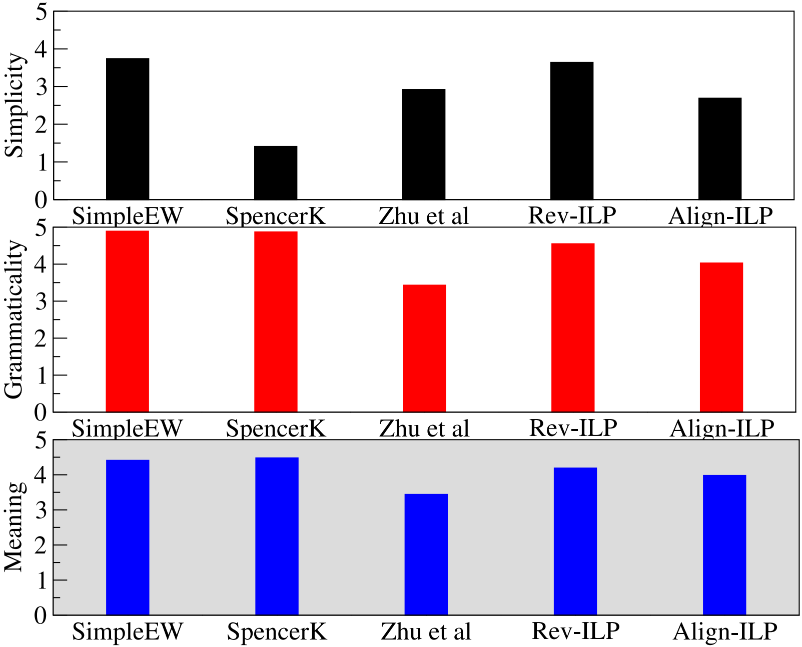
Human Evaluation



Human evaluation



Human evaluation



Out-of-domain Simplification: Little Red Riding Hood

Gutenberg Source	There was once a sweet little maid who lived with her father and mother in a pretty little cottage at the edge of the village. At the further end of the wood was another pretty cottage and in it lived her grandmother. Everybody loved this little girl, her grandmother perhaps loved her most of all and gave her a great many pretty things. Once she gave her a red cloak with a hood which she always wore, so people called her Little Red Riding Hood.
Rev-ILP Output	There was once a sweet little maid. She lived with her father and mother in a pretty little cottage at the edge of the village. At the further end of the wood it lived her grandmother. Everybody loved this little girl. Her grandmother perhaps loved her most of all. She gave her a great many pretty things. Once she gave her a red cloak with a hood, so persons called her Little Red Riding Hood.

The mean FKGL on simplified stories was 3.78 (7.04 for source).

Part IV

Learning to Simplify Documents

ILP for Document Simplification

$$
\begin{array}{lll}
\max_x & \sum_{i \in \mathcal{P}} (f_i + g_i) x_i + h_w + h_{sy} & \\
\text{s.t.} & \sum_{i \in \mathcal{P}} l_i^{(w)} x_i \leq L_{\max} & \\
& x_j \rightarrow x_i & \forall i \in \mathcal{P}, j \in \mathcal{D}_i \\
& x_i \rightarrow y_s & \forall i \in \mathcal{P}, s \in \mathcal{A}_i \\
& \sum_{j \in \mathcal{C}_i} x_j = x_i & \forall i \in \mathcal{C}, j \in \mathcal{C}_i \\
& x_i \rightarrow y_s & \forall s \in \mathcal{S}, i \in \mathcal{P}_s \\
& \sum_{s \in \mathcal{S}} y_i \geq 1 & x_i \in \{0, 1\} \forall i \in \mathcal{P} \\
& & y_s \in \{0, 1\} \forall \{s \in \mathcal{S}\}.
\end{array}
$$

ILP for Document Simplification

$$\max_x \quad \sum_{i\in\mathcal{P}}(f_i + g_i)x_i + h_w + h_{sy}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{P}} l_i^{(w)} x_i \leq L_{\max}$$

$$x_j \rightarrow x_i \qquad \forall i \in \mathcal{P}, j \in \mathcal{D}_i$$

$$x_i \rightarrow y_s \qquad \forall i \in \mathcal{P}, s \in \mathcal{A}_i$$

$$\sum_{j\in\mathcal{C}_i} x_j = x_i \quad \forall i \in \mathcal{C}, j \in \mathcal{C}_i$$

$$x_i \rightarrow y_s \qquad \forall s \in \mathcal{S}, i \in \mathcal{P}_s$$

$$\sum_{s\in\mathcal{S}} y_i \geq 1 \quad x_i \in \{0,1\} \forall i \in \mathcal{P}$$

$$y_s \in \{0,1\} \forall \{s \in \mathcal{S}\}.$$

- Salience scores f_i

ILP for Document Simplification

$$
\begin{aligned}
\max_x \quad & \sum_{i\in\mathcal{P}}(f_i + g_i)x_i + h_w + h_{sy} \\
\text{s.t.} \quad & \sum_{i\in\mathcal{P}} l_i^{(w)} x_i \le L_{\max} \\
& x_j \to x_i \qquad \forall i \in \mathcal{P}, j \in \mathcal{D}_i \\
& x_i \to y_s \qquad \forall i \in \mathcal{P}, s \in \mathcal{A}_i \\
& \sum_{j\in\mathcal{C}_i} x_j = x_i \quad \forall i \in \mathcal{C}, j \in \mathcal{C}_i \\
& x_i \to y_s \qquad \forall s \in \mathcal{S}, i \in \mathcal{P}_s \\
& \sum_{s\in\mathcal{S}} y_i \ge 1 \quad x_i \in \{0,1\} \forall i \in \mathcal{P} \\
& \qquad\qquad y_s \in \{0,1\} \forall \{s \in \mathcal{S}\}.
\end{aligned}
$$

- Salience scores f_i

ILP for Document Simplification

$$
\begin{aligned}
\max_x \quad & \textstyle\sum_{i \in \mathcal{P}} (f_i + g_i) x_i + h_w + h_{sy} \\
\text{s.t.} \quad & \textstyle\sum_{i \in \mathcal{P}} l_i^{(w)} x_i \leq L_{\max} \\
& x_j \rightarrow x_i \qquad \forall i \in \mathcal{P}, j \in \mathcal{D}_i \\
& x_i \rightarrow y_s \qquad \forall i \in \mathcal{P}, s \in \mathcal{A}_i \\
& \sum_{j \in \mathcal{C}_i} x_j = x_i \quad \forall i \in \mathcal{C}, j \in \mathcal{C}_i \\
& x_i \rightarrow y_s \qquad \forall s \in \mathcal{S}, i \in \mathcal{P}_s \\
& \sum_{s \in \mathcal{S}} y_i \geq 1 \quad x_i \in \{0, 1\} \forall i \in \mathcal{P} \\
& \qquad\qquad y_s \in \{0, 1\} \forall \{s \in \mathcal{S}\}.
\end{aligned}
$$

- Salience scores f_i

ILP for Document Simplification

$$\max_x \quad \sum_{i \in \mathcal{P}} (f_i + g_i) x_i + h_w + h_{sy}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{P}} l_i^{(w)} x_i \leq L_{\max}$$

$$x_j \rightarrow x_i \qquad \forall i \in \mathcal{P}, j \in \mathcal{D}_i$$

$$x_i \rightarrow y_s \qquad \forall i \in \mathcal{P}, s \in \mathcal{A}_i$$

$$\sum_{j \in \mathcal{C}_i} x_j = x_i \quad \forall i \in \mathcal{C}, j \in \mathcal{C}_i$$

$$x_i \rightarrow y_s \qquad \forall s \in \mathcal{S}, i \in \mathcal{P}_s$$

$$\sum_{s \in \mathcal{S}} y_i \geq 1 \quad x_i \in \{0,1\} \forall i \in \mathcal{P}$$

$$y_s \in \{0,1\} \forall \{s \in \mathcal{S}\}.$$

- Salience scores f_i
- Overall length budget

ILP for Document Simplification

$$
\begin{aligned}
\max_x \quad & \textstyle\sum_{i \in \mathcal{P}} (f_i + g_i) x_i + h_w + h_{sy} \\
\text{s.t.} \quad & \textstyle\sum_{i \in \mathcal{P}} l_i^{(w)} x_i \leq L_{\max} \\
& x_j \rightarrow x_i \qquad \forall i \in \mathcal{P}, j \in \mathcal{D}_i \\
& x_i \rightarrow y_s \qquad \forall i \in \mathcal{P}, s \in \mathcal{A}_i \\
& \sum_{j \in \mathcal{C}_i} x_j = x_i \quad \forall i \in \mathcal{C}, j \in \mathcal{C}_i \\
& x_i \rightarrow y_s \qquad \forall s \in \mathcal{S}, i \in \mathcal{P}_s \\
& \sum_{s \in \mathcal{S}} y_i \geq 1 \quad x_i \in \{0,1\} \forall i \in \mathcal{P} \\
& \qquad\qquad\quad y_s \in \{0,1\} \forall \{s \in \mathcal{S}\}.
\end{aligned}
$$

- Salience scores f_i
- Overall length budget
- Where do salience scores come from?

ILP for Document Simplification

$$\max_x \quad \sum_{i \in \mathcal{P}} (f_i + g_i) x_i + h_w + h_{sy}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{P}} l_i^{(w)} x_i \leq L_{\max}$$

$$x_j \rightarrow x_i \qquad \forall i \in \mathcal{P}, j \in \mathcal{D}_i$$

$$x_i \rightarrow y_s \qquad \forall i \in \mathcal{P}, s \in \mathcal{A}_i$$

$$\sum_{j \in \mathcal{C}_i} x_j = x_i \quad \forall i \in \mathcal{C}, j \in \mathcal{C}_i$$

$$x_i \rightarrow y_s \qquad \forall s \in \mathcal{S}, i \in \mathcal{P}_s$$

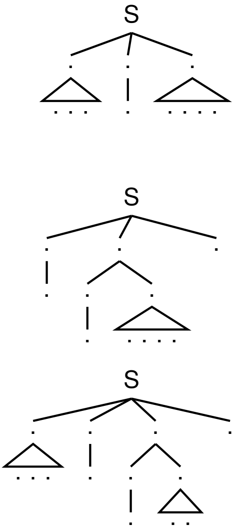
$$\sum_{s \in \mathcal{S}} y_i \geq 1 \quad x_i \in \{0, 1\} \forall i \in \mathcal{P}$$

$$y_s \in \{0, 1\} \forall \{s \in \mathcal{S}\}.$$

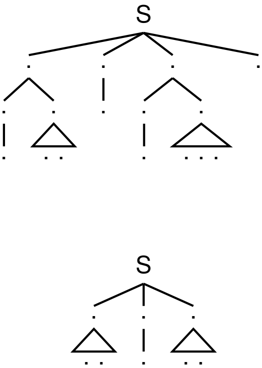
- Salience scores f_i
- Overall length budget
- Where do salience scores come from?
- Use SVM to learn f_i scores from features ϕ

$$f_i = \sum_j w_j \phi_j + w_0.$$

MainEW article

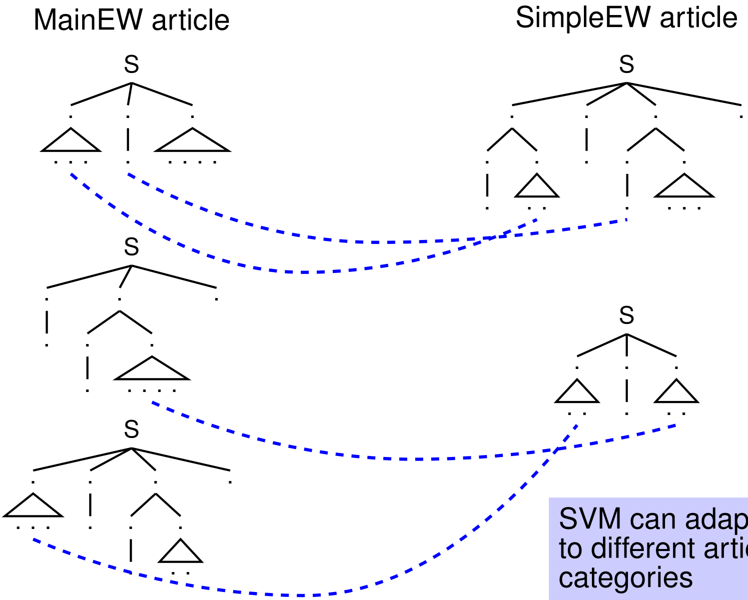


SimpleEW article



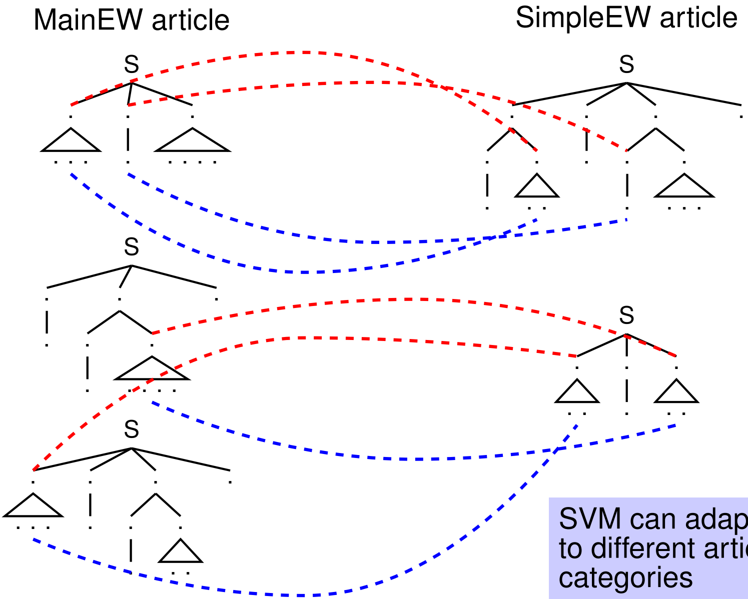
SVM can adapt
to different article
categories

Unsupervised labelling of training data



SVM can adapt
to different article
categories

Unsupervised labelling of training data



SVM can adapt to different article categories

Experimental Setup

Data sets:

- QSG rules obtained from 14,831 sentence pairs
- 3 Wikipedia categories: Animals, Celebrities and Cities
- Generated 5 articles in each category

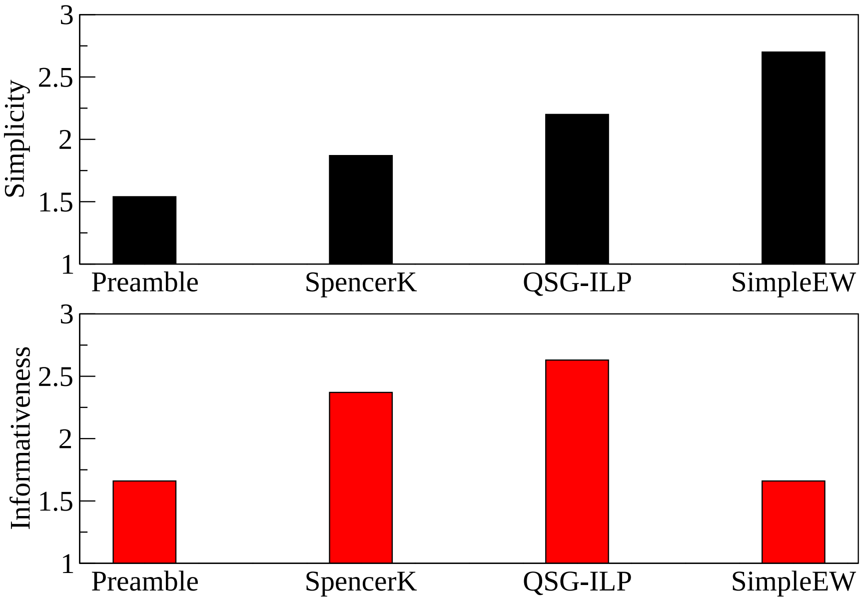
Comparison systems:

1. Preamble: Introductory sentences of original article
2. Extract-SK: Sentence extraction plus Spencer Kelly's lexical substitution dictionary
3. SimpleEW: Simple Wikipedia articles as gold standard

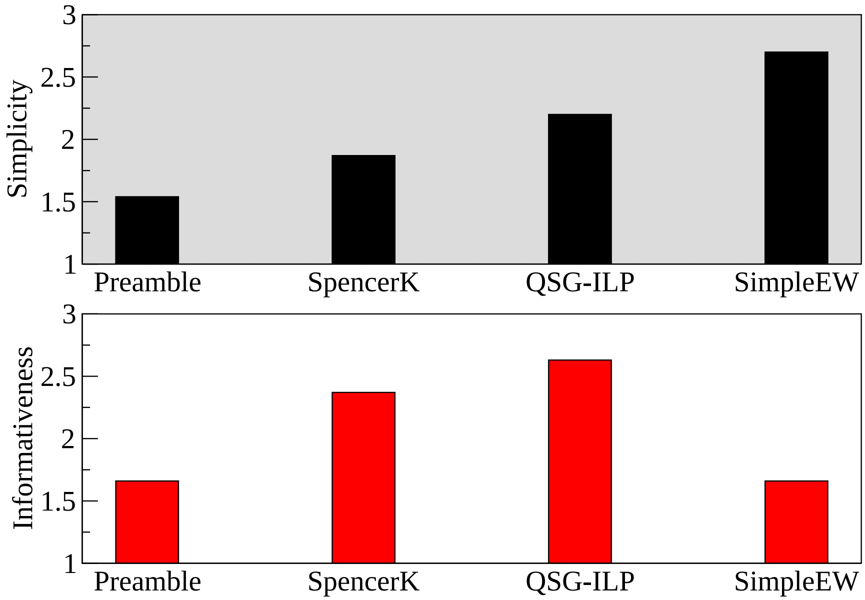
Evaluation:

- Human evaluation using non-native English speakers
- Simplicity: is the text simple or complicated?
- Informativeness: does article capture most important information?

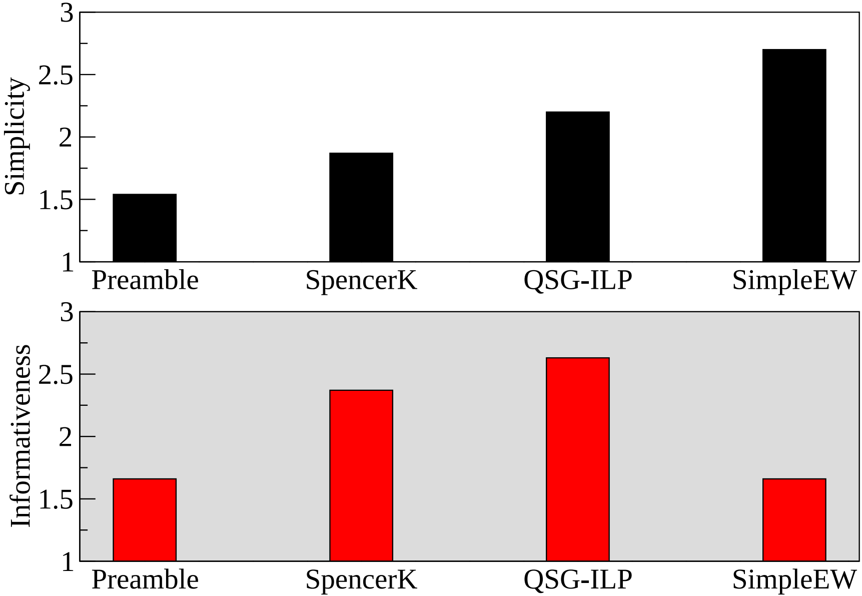
Human Evaluation



Human Evaluation



Human Evaluation



Example Output: Owls

Source

Owls are the order Strigiformes, comprising 200 bird of prey species. Owls hunt mostly small mammals, insects, and other birds though some species specialize in hunting fish.

Output

Owls are the order Strigiformes, making up 200 bird of prey species. Owls hunt mostly small mammals, insects, and other birds.

Example Output: Owls

Source	Owls are the order Strigiformes, comprising 200 bird of prey species. Owls hunt mostly small mammals, insects, and other birds though some species specialize in hunting fish.

Output	Owls are the order Strigiformes, making up 200 bird of prey species. Owls hunt mostly small mammals, insects, and other birds.

Rare words substituted with more familiar phrase

Example Output: Owls

Source

Owls are the order Strigiformes, comprising 200 bird of prey species. Owls hunt mostly small mammals, insects, and other birds though some species specialize in hunting fish.

Output

Owls are the order Strigiformes, making up 200 bird of prey species. Owls hunt mostly small mammals, insects, and other birds.

Removed unnecessary detail

Example Output: Senegal bushbaby

Simple English Wikipedia

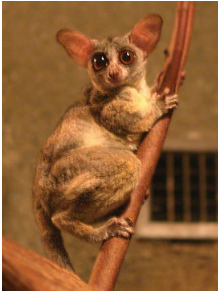
Getting around

- Main Page
- Simple start
- Simple talk
- New changes
- Show any page
- Help
- Give to Wikipedia

Page Talk Read Change

Senegal bushbaby

From Wikipedia, the free encyclopedia



The **Senegal bushbaby** is also known as the **Senegal galago**, or the **lesser bush baby**. It is a small, nocturnal primate. The word **bush baby** may come from the animals' cries.

They are small primates (130mm and 95-300 grams) with woolly thick fur that ranges from silvery gray to dark brown. They are agile leapers. They have 1-2 babies per litter, with gestation period being 110—120 days.

Conclusions

- Framework for modeling simplification.
- System for simplifying Wikipedia articles.
- Jointly selects content and rewrites text.
- Output is informative, and simpler than baselines.
- Learns from Wikipedia content and revision process.

Future work:

- Enrich the model with discourse-level document structure.
- User-specific and genre-specific objectives.
- On-line text simplification, extend to other languages.

Questions

Objective of the model

$$\max_{x} \quad \sum_{i \in \mathcal{P}} (f_i + g_i) x_i + h_w + h_{sy}$$

- Raw SVM salience score, from features ϕ: $f_i = \sum_j w_j \phi_j + w_0$.
- Log-probability for rewriting: $g_i = \log\left(\frac{n_r}{N_r}\right)$.
- **Number of words** against **target words per sentence**:

$$h_w(x, y) = \text{wps} \times \sum_{i \in \mathcal{S}} y_i - \sum_{i \in \mathcal{P}} l_i^{(w)} x_i.$$

- **Number of syllables** against **target syllables per word**:

$$h_{sy}(x) = \text{spw} \times \sum_{i \in \mathcal{P}} l_i^{(w)} x_i - \sum_{i \in \mathcal{P}} l_i^{(sy)} x_i.$$

- Linear approximation of Flesch-Kincaid Grade Level:

$$\text{FKGL} = 0.39 \left(\frac{\text{total words}}{\text{total sentences}}\right) + \left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

Statistics and readability measures

System	Token count	FKGL Index
MainEW		10.48 ± 2.08
SimpleEW	196 ± 111	8.81 ± 2.65
Preamble	203 ± 149	11.23 ± 2.76
Extract-SK	238 ± 52	9.79 ± 2.13
QG-ILP	165 ± 53	7.34 ± 1.79

Results of human evaluation

System	Simplicity	Informativeness
SimpleEW	**2.70**	1.66
Preamble	1.54	1.66
Extract-SK	1.87	2.37
QG-ILP	2.20	**2.63**

Simplicity: Is the text simple or complicated?
Informativeness: Does the article capture the most important information?

Various statistics on experiments

Models	Articles	Data	Rules	FKGL	BLEU
MainEW				15.12	0.50
SimpleEW				11.25	—
SpencerK			2,855	14.67	0.47
Zhu et al.	65,133	108,016	?	9.41	0.38
C&K	10,000	137,000	?	14.93	0.48
Rev-ILP	14,831	84,769	769	10.92	0.42
Align-ILP	15,000	141,872	622	12.36	0.34
Joshua	15,000	141,872	365,633	14.93	0.48

Aluisio, Sandra and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: The porsimples project for simplification of portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*. Association for Computational Linguistics, Los Angeles, California, pages 46–53.

Burstein, Jill, Jane Shore, John Sabatini, Yong-Won Lee, and Matthew Ventura. 2007. The automated text adaptation tool. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. Association for Computational Linguistics, Rochester, New York, USA, pages 3–4.

Carroll, John, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999a. Simplifying text for language-impaired readers. In *Proceedings of the 9th EACL*. Bergen, Norway, pages 269–270.

Carroll, John, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999b. Simplifying text for language impaired readers. In *Proceedings of the 9th EACL*. Bergen, Norway, pages 269–270.

Chandrasekar, Raman, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th COLING*. Copenhagen, Denmark, pages 1041–1044.

Coster, William and David Kauchak. 2011. Simple english wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 665–669.

Devlin, Siobhan. 1999. *Simplifying Natural Language for Aphasic Readers*. Ph.D. thesis, University of Sunderland.

Inui, Kentaro, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: A project note. In *Proceedings of Workshop on Paraphrasing*. Sapporo, Japan, pages 9–16.

Kaji, Nobuhiro, Daisuke Kawahara, Sadao Kurohashi, and Satoshi Sato. 2002. Verb paraphrase based on case frame alignment. In *Proceedings of the 40th ACL*. Philadelphia, PA, pages 215–222.

Siddharthan, Advaith. 2004. Syntactic simplification and text cohesion. *Research on Language and Computation* 4(1):77–109.

Vickrey, David and Daphne Koller. 2008. Sentence simplification for semantic role labeling. In *Proceedings of ACL-08: HLT*. Columbus, OH, pages 344–352.

Yatskar, Mark, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Proceedings of NAACL*. pages 365–368.

Zhu, Zhemin, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing, China, pages 1353–1361.