Integration of SYSTRAN MT systems in an open workflow

Mats Attnäs[1] Pierre Senellart[1,2] Jean Senellart[1]

1 SYSTRAN Language Translation Technologies

2 INRIA FUTURS

MT Summit X
September 15th, 2005

SYSTRAN

Facts

- 35-year-old company
- 40+ language pairs
- 20+ different languages
- Translation service for major portals (Babelfish, Google, Yahoo!...)
- $\approx$ 35,000,000 on-line translations a day
- Large range of products (from PDA to Web servers)
- Services for corporate and institutions (Ford, Cisco, US Gov, EU...)
- Product life: v4 (2001), v5 (2004), v6 (June 2006)

SYSTRAN

Facts

- 35-year-old company
- 40+ language pairs
- 20+ different languages
- Translation service for major portals (Babelfish, Google, Yahoo!...)
- ≈ 35,000,000 on-line translations a day
- Large range of products (from PDA to Web servers)
- Services for corporate and institutions (Ford, Cisco, US Gov, EU...)
- Product life: v4 (2001), v5 (2004), v6 (June 2006)

SYSTRAN

Facts

- 35-year-old company
- 40+ language pairs
- 20+ different languages
- Translation service for major portals (Babelfish, Google, Yahoo!. . .)
- ≈ 35,000,000 on-line translations a day
- Large range of products (from PDA to Web servers)
- Services for corporate and institutions (Ford, Cisco, US Gov, EU. . .)
- Product life: v4 (2001), v5 (2004), v6 (June 2006)

SYSTRAN

Facts

- 35-year-old company
- 40+ language pairs
- 20+ different languages
- Translation service for major portals (Babelfish, Google, Yahoo!. . .)
- ≈ 35,000,000 on-line translations a day
- Large range of products (from PDA to Web servers)
- Services for corporate and institutions (Ford, Cisco, US Gov, EU. . .)
- Product life: v4 (2001), v5 (2004), v6 (June 2006)

SYSTRAN

Facts

- 35-year-old company
- 40+ language pairs
- 20+ different languages
- Translation service for major portals (Babelfish, Google, Yahoo!. . .)
- ≈ 35,000,000 on-line translations a day
- Large range of products (from PDA to Web servers)
- Services for corporate and institutions (Ford, Cisco, US Gov, EU. . .)
- Product life: v4 (2001), v5 (2004), v6 (June 2006)

SYSTRAN

Facts

- 35-year-old company
- 40+ language pairs
- 20+ different languages
- Translation service for major portals (Babelfish, Google, Yahoo!. . .)
- ≈ 35,000,000 on-line translations a day
- Large range of products (from PDA to Web servers)
- Services for corporate and institutions (Ford, Cisco, US Gov, EU. . .)
- Product life: v4 (2001), v5 (2004), v6 (June 2006)

SYSTRAN

Facts

- 35-year-old company
- 40+ language pairs
- 20+ different languages
- Translation service for major portals (Babelfish, Google, Yahoo!. . .)
- ≈ 35,000,000 on-line translations a day
- Large range of products (from PDA to Web servers)
- Services for corporate and institutions (Ford, Cisco, US Gov, EU. . .)
- Product life: v4 (2001), v5 (2004), v6 (June 2006)

SYSTRAN

Facts

- 35-year-old company
- 40+ language pairs
- 20+ different languages
- Translation service for major portals (Babelfish, Google, Yahoo!. . .)
- ≈ 35,000,000 on-line translations a day
- Large range of products (from PDA to Web servers)
- Services for corporate and institutions (Ford, Cisco, US Gov, EU. . .)
- Product life: v4 (2001), v5 (2004), v6 (June 2006)

SYSTRAN

Recent activities

- New Generation (NG) systems: Arabic → English, Swedish → English
- Tools for translators (translation memory support, bilingual terminology extraction. . .)
- New Language Pairs (FAEN, HIEN, UREN, CSEN, UKEN, SKEN, PLEN, ARFR, SREN)
- SYSTRAN Lite on PDA

SYSTRAN

Recent activities

- New Generation (NG) systems: Arabic → English, Swedish → English
- Tools for translators (translation memory support, bilingual terminology extraction...)
- New Language Pairs (FAEN, HIEN, UREN, CSEN, UKEN, SKEN, PLEN, ARFR, SREN)
- SYSTRAN Lite on PDA

SYSTRAN

Recent activities

- New Generation (NG) systems: Arabic → English, Swedish → English
- Tools for translators (translation memory support, bilingual terminology extraction...)
- New Language Pairs (FAEN, HIEN, UREN, CSEN, UKEN, SKEN, PLEN, ARFR, SREN)
- SYSTRAN Lite on PDA

SYSTRAN

Recent activities

- New Generation (NG) systems: Arabic → English, Swedish → English
- Tools for translators (translation memory support, bilingual terminology extraction...)
- New Language Pairs (FAEN, HIEN, UREN, CSEN, UKEN, SKEN, PLEN, ARFR, SREN)
- SYSTRAN Lite on PDA

MT System Complexity

- $(nb_{rules}, size_{dictionary}) \times nb_{LPs}$
- High flexibility required
- Stability
- Intrinsic complexity of language description
  - Rules and exceptions
  - Rule interaction
  - Endless dictionary completion

$\Longrightarrow$ unavoidable saturation point?

MT System Complexity

- $(nb_{rules}, size_{dictionary}) \times nb_{LPs}$
- High flexibility required
- Stability
- Intrinsic complexity of language description
  - Rules and exceptions
  - Rule interaction
  - Endless dictionary completion

$\Longrightarrow$ unavoidable saturation point?

MT System Complexity

- $(nb_{rules}, size_{dictionary}) \times nb_{LPs}$
- High flexibility required
- Stability
- Intrinsic complexity of language description
  - Rules and exceptions
  - Rule interaction
  - Endless dictionary completion

$\Longrightarrow$ unavoidable saturation point?

MT System Complexity

- $(nb_{rules}, size_{dictionary}) \times nb_{LPs}$
- High flexibility required
- Stability
- Intrinsic complexity of language description
  - Rules and exceptions
  - Rule interaction
  - Endless dictionary completion

$\Longrightarrow$ unavoidable saturation point?

MT System Complexity

- $(nb_{rules}, size_{dictionary}) \times nb_{LPs}$
- High flexibility required
- Stability
- Intrinsic complexity of language description
  - Rules and exceptions
  - Rule interaction
  - Endless dictionary completion

$\Longrightarrow$ unavoidable saturation point?

A black box system is not enough!

Need of interaction

Level	Issue	Solutions
Linguist	Why do I get this?	traceability, entry point
Translator	Where should I look?	risk area, alternatives, resource coverage
Author	How will it translate?	MT-in-mind authoring
End user	What can I change?	resource interaction

EVOLUTION REQUIRES "OPENING".

A black box system is not enough!

Need of interaction

Level	Issue	Solutions
Linguist	Why do I get this?	traceability, entry point
Translator	Where should I look?	risk area, alternatives, resource coverage
Author	How will it translate?	MT-in-mind authoring
End user	What can I change?	resource interaction

EVOLUTION REQUIRES "OPENING".

A black box system is not enough!

Need of interaction

Level	Issue	Solutions
Linguist	Why do I get this?	traceability, entry point
Translator	Where should I look?	risk area, alternatives, resource coverage
Author	How will it translate?	MT-in-mind authoring
End user	What can I change?	resource interaction

EVOLUTION REQUIRES "OPENING".

Outline

Code modernization

- From specific transliterations and encodings to standard Unicode
- From low-level paradigms to high-level object-oriented code

  $\Longrightarrow$

  - Redesign, Modularization
  - High-level linguistic data structures
- Joining parallel effort of New Generation engines (high-level, declarative)

Code modernization

- From specific transliterations and encodings to standard Unicode
- From low-level paradigms to high-level object-oriented code

$\Longrightarrow$

  - Redesign, Modularization
  - High-level linguistic data structures
- Joining parallel effort of New Generation engines (high-level, declarative)

Code modernization

- From specific transliterations and encodings to standard Unicode
- From low-level paradigms to high-level object-oriented code

  $\Longrightarrow$

  - Redesign, Modularization
  - High-level linguistic data structures
- Joining parallel effort of New Generation engines (high-level, declarative)

Code modernization

- From specific transliterations and encodings to standard Unicode
- From low-level paradigms to high-level object-oriented code

  $\Longrightarrow$

  - Redesign, Modularization
  - High-level linguistic data structures
- Joining parallel effort of New Generation engines (high-level, declarative)

XML Workflow & Modules

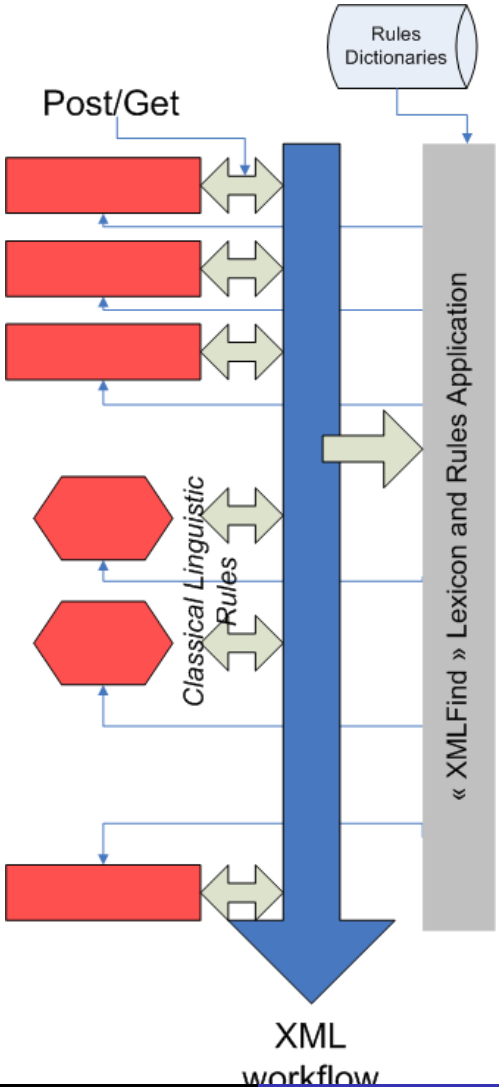
- A flow of XML data to convey more information
  - source text and history
  - intermediate status
  - document structure information
  - internal/external markup
  - confidence information

- Processed by independent, interoperable agents
  - Linguistics (Analysis/Transfer/Generation)
  - Generic (Preprocessing/Postprocessing)

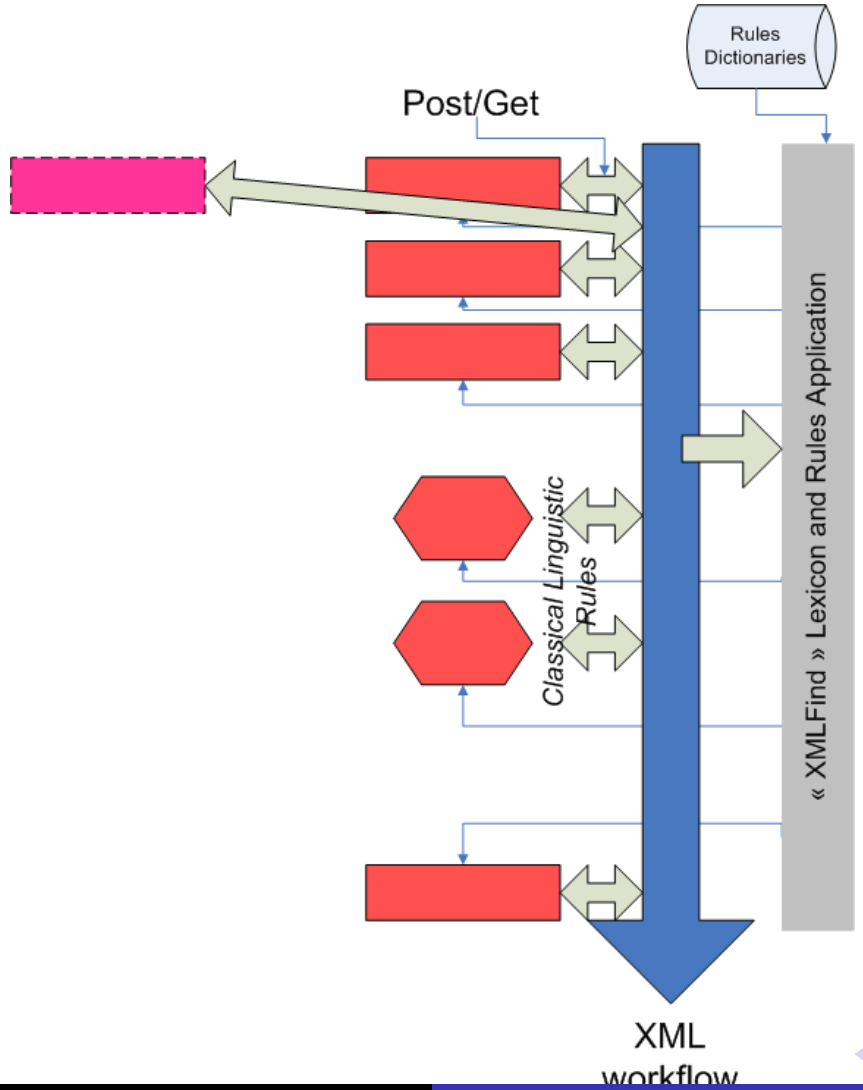
XML Workflow & Modules

- A flow of XML data to convey more information
  - source text and history
  - intermediate status
  - document structure information
  - internal/external markup
  - confidence information
- Processed by independent, interoperable agents
  - Linguistics (Analysis/Transfer/Generation)
  - Generic (Preprocessing/Postprocessing)

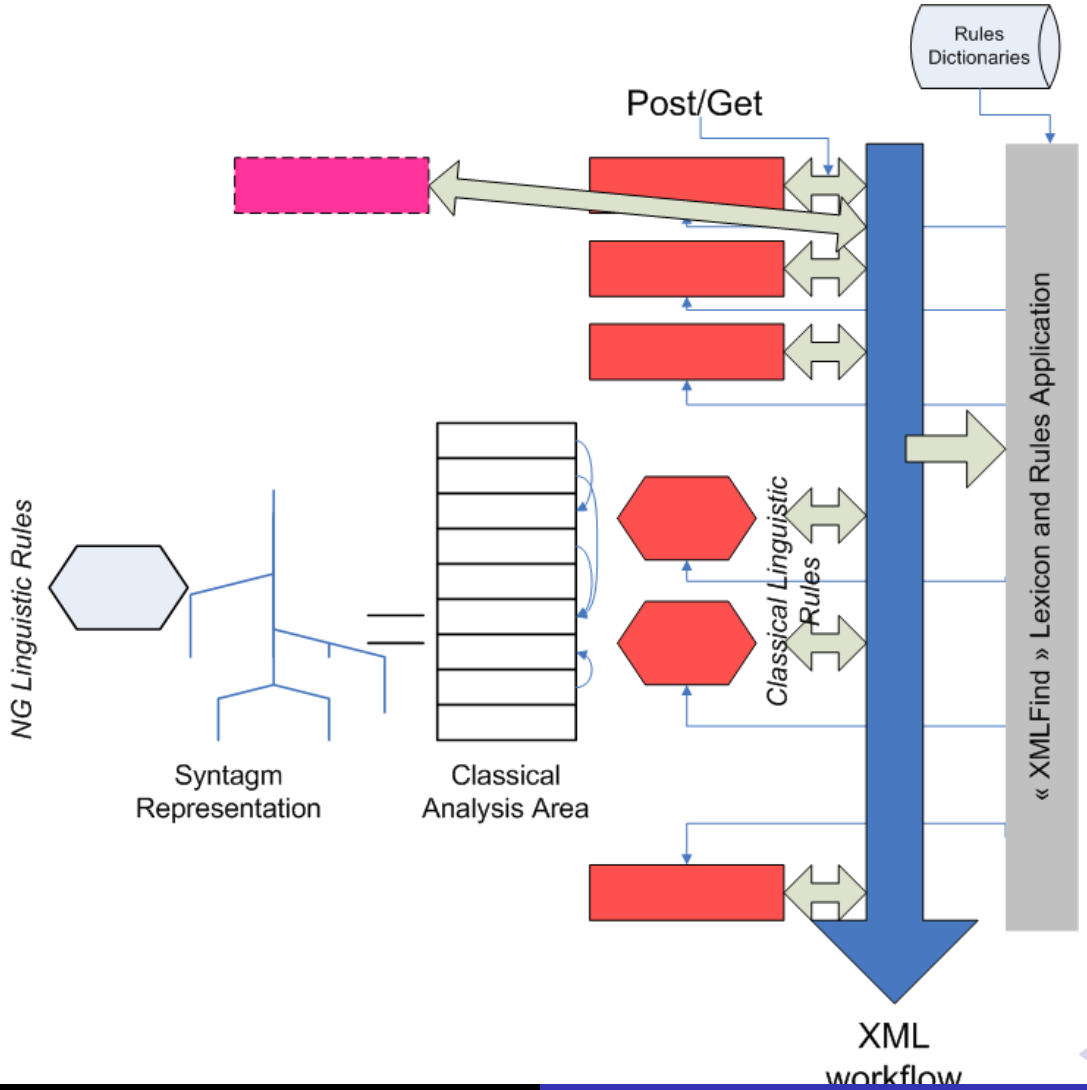
Workflow Schema



Workflow Schema



Workflow Schema

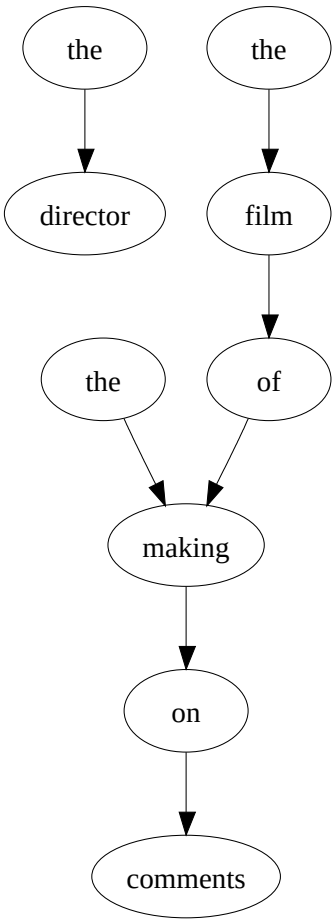


From relation graphs to syntagm trees

The director comments on the making of the film.

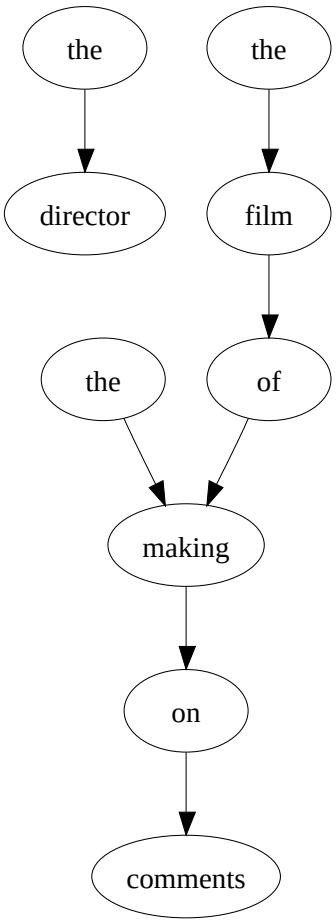
From relation graphs to syntagm trees

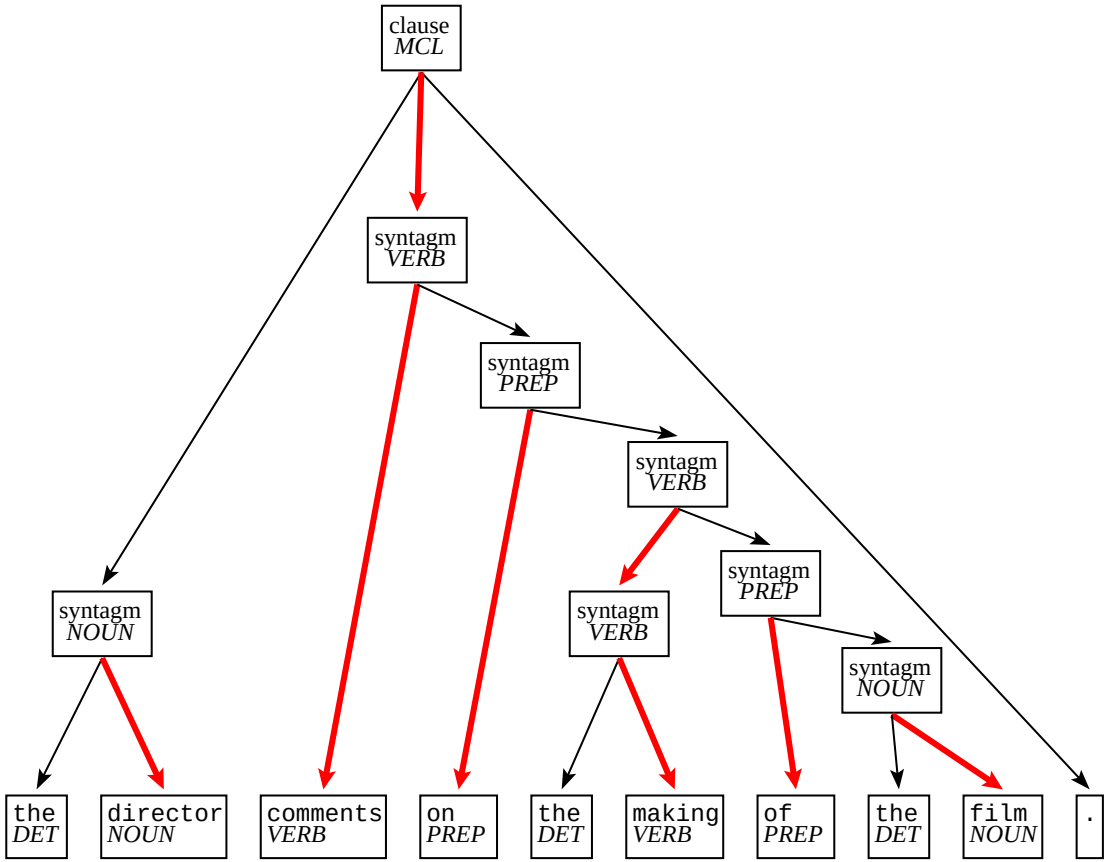
The director comments on the making of the film.



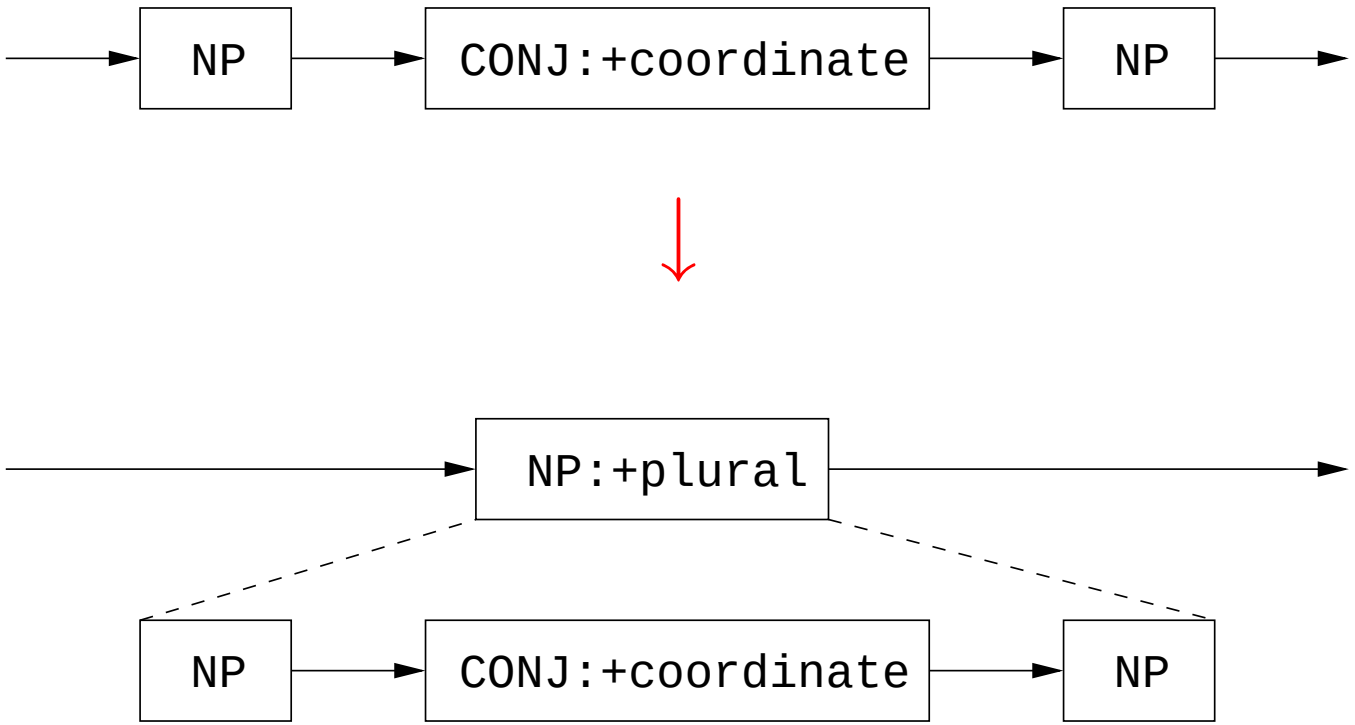
From relation graphs to syntagm trees

The director comments on the making of the film.





Declarative linguistic rules



Matching as a Fundamental

- Dictionary lookup
- Rule triggering
- Sentence fuzzy matching

ALL OF THIS REQUIRES "SMART MATCHING".

- XML structured matching
- Low-level lookup operators (numbers, dates, chemical formulas. . .)
- On-the-fly spellcheck, normalization. . .

Matching as a Fundamental

- Dictionary lookup
- Rule triggering
- Sentence fuzzy matching

ALL OF THIS REQUIRES "SMART MATCHING".

- XML structured matching
- Low-level lookup operators (numbers, dates, chemical formulas. . .)
- On-the-fly spellcheck, normalization. . .

Matching as a Fundamental

- Dictionary lookup
- Rule triggering
- Sentence fuzzy matching

ALL OF THIS REQUIRES "SMART MATCHING".

- XML structured matching
- Low-level lookup operators (numbers, dates, chemical formulas. . .)
- On-the-fly spellcheck, normalization. . .

Matching as a Fundamental

- Dictionary lookup
- Rule triggering
- Sentence fuzzy matching

ALL OF THIS REQUIRES "SMART MATCHING".

- XML structured matching
- Low-level lookup operators (numbers, dates, chemical formulas. . .)
- On-the-fly spellcheck, normalization. . .

Matching as a Fundamental

- Dictionary lookup
- Rule triggering
- Sentence fuzzy matching

ALL OF THIS REQUIRES "SMART MATCHING".

- XML structured matching
- Low-level lookup operators (numbers, dates, chemical formulas. . .)
- On-the-fly spellcheck, normalization. . .

Matching as a Fundamental

- Dictionary lookup
- Rule triggering
- Sentence fuzzy matching

ALL OF THIS REQUIRES "SMART MATCHING".

- XML structured matching
- Low-level lookup operators (numbers, dates, chemical formulas...)
- On-the-fly spellcheck, normalization...

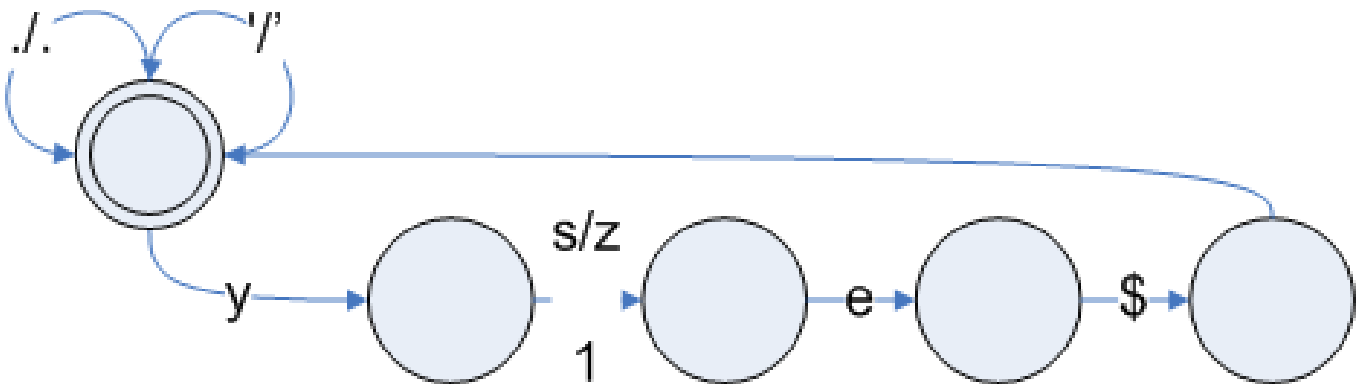
Matching as a Fundamental

- Dictionary lookup
- Rule triggering
- Sentence fuzzy matching

ALL OF THIS REQUIRES "SMART MATCHING".

- XML structured matching
- Low-level lookup operators (numbers, dates, chemical formulas...)
- On-the-fly spellcheck, normalization...

Spellcheck



Outline

Linguistic Improvement by Input Simplification

LET SPECIALIZED AGENTS BRING THEIR EXPERTISE.

- Focus on local grammars
- Entity recognition
- Third-party components
- Extra-linguistics decision making

Linguistic Improvement by Input Simplification

LET SPECIALIZED AGENTS BRING THEIR EXPERTISE.

- Focus on local grammars
- Entity recognition
- Third-party components
- Extra-linguistics decision making

Linguistic Improvement by Input Simplification

LET SPECIALIZED AGENTS BRING THEIR EXPERTISE.

- Focus on local grammars
- Entity recognition
- Third-party components
- Extra-linguistics decision making

Linguistic Improvement by Input Simplification

LET SPECIALIZED AGENTS BRING THEIR EXPERTISE.

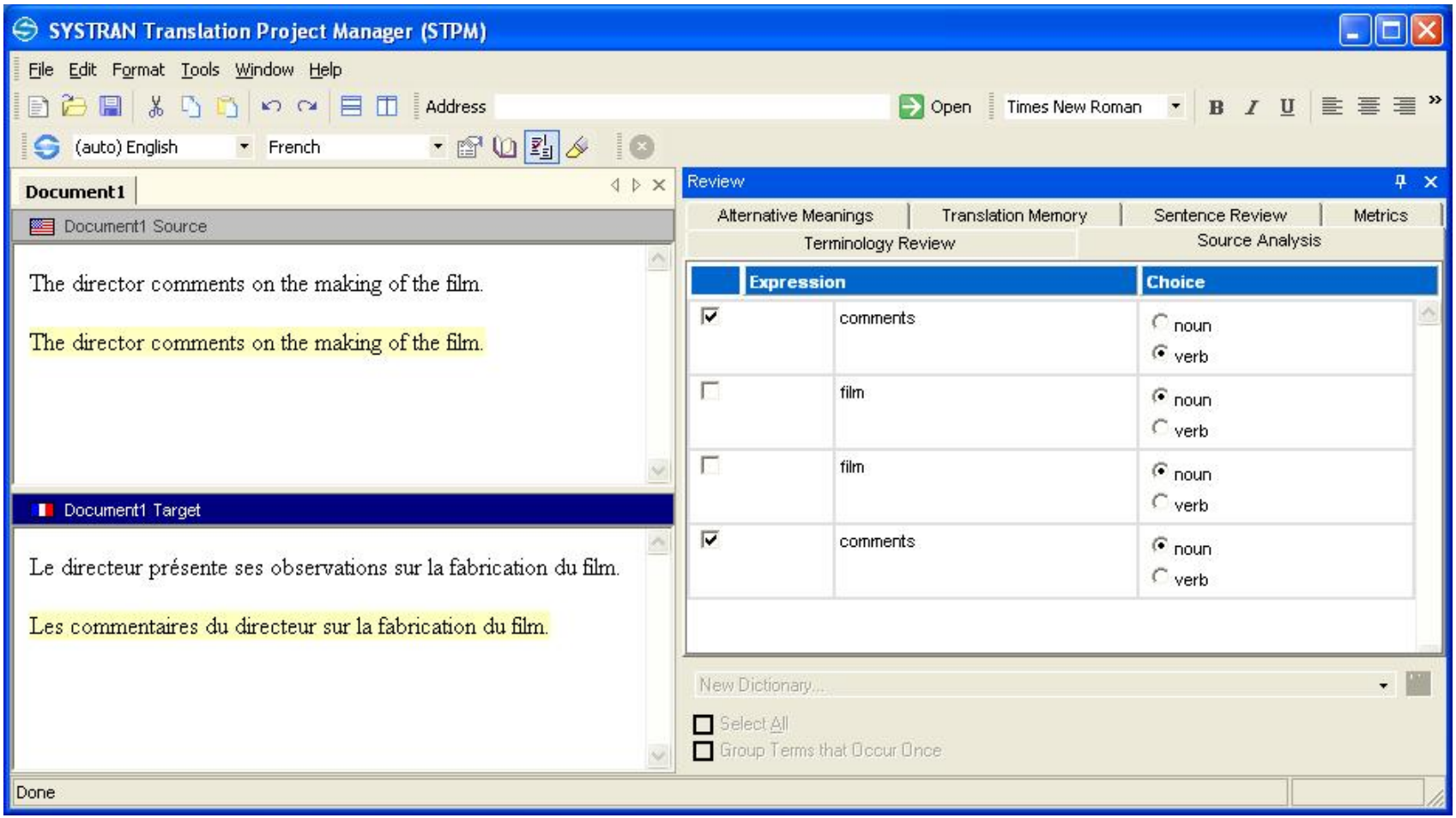
- Focus on local grammars
- Entity recognition
- Third-party components
- Extra-linguistics decision making

Linguistic Improvement by Input Simplification

LET SPECIALIZED AGENTS BRING THEIR EXPERTISE.

- Focus on local grammars
- Entity recognition
- Third-party components
- Extra-linguistics decision making

Interacting with the user



Post-edition

1 With more than 1.5 million installed, the Cisco 2500 series is one of the most popular solutions for a wide range of cost-effective configurations, including dual LAN, integrated router/hub, and integrated access server models.

Avec plus de 1.5 million étant installé, les 2500 séries de Cisco sont l'une des solutions les plus populaires pour un éventail de configurations, y compris le LAN duel, de routeur/hub intégrés, et de modèles intégrés de serveur d'accès.

9 For example, integrated call switching and call handling features enable small or branch offices to use their Cisco access solution for call handling and remote access instead of having to invest in a PBX system.

Par exemple, la commutation d'appel et les dispositifs intégrés de manipulation d'appel permettent succursales de petites ou d'employer leur solution d'accès de Cisco pour la manipulation d'appel et l'accès à distance au lieu de devoir investir dans un système de PBX.

Developing new language pairs

Analysis, Transfer, Generation:

- Independence
- Genericity
- Maturity

Challenge

40 new cross LPs in one year?

Developing new language pairs

Analysis, Transfer, Generation:

- Independence
- Genericity
- Maturity

Challenge

40 new cross LPs in one year?

Developing new language pairs

Analysis, Transfer, Generation:

- Independence
- Genericity
- Maturity

Challenge

40 new cross LPs in one year?

Developing new language pairs

Analysis, Transfer, Generation:

- Independence
- Genericity
- Maturity

Challenge

40 new cross LPs in one year?

Developing new language pairs

Analysis, Transfer, Generation:

- Independence
- Genericity
- Maturity

Challenge

40 new cross LPs in one year?

Summary

- Still a complex system
- But clear entry points
- Open to external approaches
- Strengthened interactivity with linguist/translator/author/user

Summary

- Still a complex system
- But clear entry points
- Open to external approaches
- Strengthened interactivity with linguist/translator/author/user

Summary

- Still a complex system
- But clear entry points
- Open to external approaches
- Strengthened interactivity with linguist/translator/author/user

Summary

- Still a complex system
- But clear entry points
- Open to external approaches
- Strengthened interactivity with linguist/translator/author/user

Perspectives

- Effective use of alternative approaches in v6
- To know the system's hesitations and help it decide (rules weighting)
- Smaller agents
- Improved quality, New LPs!

Perspectives

- Effective use of alternative approaches in v6
- To know the system's hesitations and help it decide (rules weighting)
- Smaller agents
- Improved quality, New LPs!

Perspectives

- Effective use of alternative approaches in v6
- To know the system's hesitations and help it decide (rules weighting)
- Smaller agents
- Improved quality, New LPs!

Perspectives

- Effective use of alternative approaches in v6
- To know the system's hesitations and help it decide (rules weighting)
- Smaller agents
- Improved quality, New LPs!