言語と情報（２）
言語処理の技術、課題

東京大学情報学環
辻井潤一

- **知識の構造化は万人に開かれるべきであり、誰でも分かり使えるような技術が必要なのではないか（MIMAサーチや自然言語処理は難しい）**

  ユーザ・インターフェース，可視化，大量データからの知識発見
  分析ツールと質問応答システム

  既存の知識を理解する困難は常に残る
  知識の理解を促進はできるが、それ以上を望むのは知的怠惰

- **印刷技術による大きな変化ほどのものを、情報技術で起こせるかどうかは疑問だと感じた。印刷によって可能になったデータベース化を、情報技術は促進したに過ぎないのではないか。**

  印刷による一方向、非同期的なコミュニケーション
  双方向のコミュニケーション、同期的・非同期的なコミュニケーション

  蓄積と流通だけでなく，処理（加工）できる。
  テキストを素材として、情報の組織化ができる

- **情報の保存方法は、伝承→手書き文字→印刷による書籍化→電子化と移り変わってきたが、この後は何がくるだろうか。**

  長期的な保存という意味では理想的ではない、劣化、容量
  イメージではなくて、情報としての読み出しと加工ができることの重要性は残る

- **科学者が情報技術革新に邁進していった結果、それが権力者に利するように使われてしまう危険が大きいのではないか。**

  技術は中立（素朴すぎるかもしれませんが。。。）、
  権力者への情報の集中が緩和されてきたことは確かではないか？
  デマゴーグが効果的に行える危険性はあるが、その対抗手段も大きくなっていることも確か。。
  情報の世界の虚偽性とか、解釈の恣意性といったものとリアルな世界との相互関係がより重要になる。
  ただ、「リアル」とは何かという問題は残ると思う。

- **情報技術は対症療法的に発展してきているように感じる。そのため、情報量が多くなりすぎて、結局使えないものになってしまっている。**

  情報技術が対症療法的に発展してきているのかどうか、それが現在の情報過多の原因かどうか？
  ニーズが大きい新たな形態の情報流通を可能にしたことは事実だが。。。
  情報過多の中から、有効な情報を選び出す技術、あるいは、それを助ける技術が必要。判断、価値。。。
  氾濫する情報相互間の相互関係をつける技術が必要。

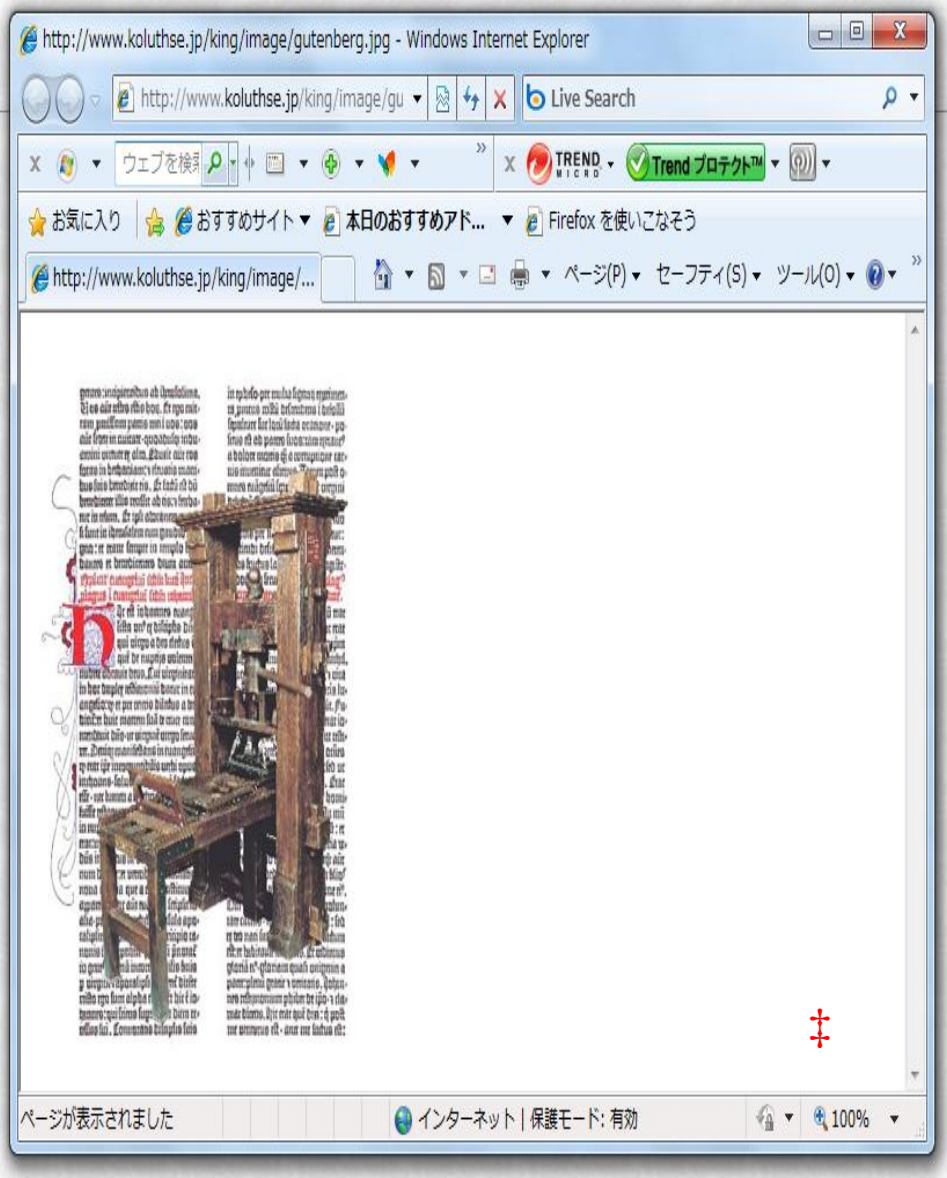
言語と（意味・文脈・記憶・構造・解釈）

- 幼児、絵本
- Bilingualの人
- 公正＝to be fair
- Freedom Fighters , Terrorists
- 遊び、甘えの構造

16世紀になると、聖書はグーテンベルグの発明した印刷機によって大量に、しかも正確に印刷され、さらに世界各国の言葉で印刷されるようになりました。グーテンベルグの印刷機で大量に印刷されたドイツ語やフランス語の聖書は、それまでそれぞれの地域でつかわれていた「方言」を標準化し、現代のドイツ語、フランス語の基礎ができました。こうしてグーテンベルグの印刷機が、ルネサンスなど文化の基盤を作り、近代への扉を開くことになったのです。

展示品

グーテンベルグ印刷機 (復刻機



‡日本聖書協会

辻井の講義へのコメント

- ネットの中では独特の日本語ができつつあるが、それによってアイデンティィが分裂してしまうことがあるだろうか。
- 言語の平準化は個々の文体や個性を奪うことになるのではないか。

ネットの世界と現実の世界で、性質が大きく変る人がいることは確か
その場合に2つの世界で言葉の使い方まで変わることはあり得ると思う、
これはネットのコミュニティでの言葉が人格を変えたのではなく、因果関係が逆？
ただ、言葉が人格を変えているように見える場合もあるので、別人格の確立を助長することはあるでしょう

ネットや電子的なコミュニケーションの進展が「言語を平準化する」という前提があるようだが、本当か？
むしろ、印刷技術が平準化を生み出したのとは逆の、言語の多様化、コミュニケーションのFragmentationが起こっているように見えるが。。。

- ことばの解釈はその人のバックグラウンドによって左右される。コンピュータに言語処理させる際にはどのような下地をもたせればよいのか。

人工知能の最初の頃のTuringテストの話，　Turingの予言とレーブナー賞

言語理解の研究：言語とその解釈、知識の表現（70-80年代）

むしろ、この問題を迂回することで技術的な発展があったと思う

- 印刷技術の発達で言語が統一されたように、言語処理技術の発達によって、人間の解釈の仕方も統一されるのだろうか。

言語とその解釈の結びつきを標準化していくことで、コミュニケーションの効率化を図る試みはある。
Semantic Web の試み：興味のある人は、調べてみるとよい。

- 言語が意味や情報、知識などの人間の文化と密接につながっているならば、言語による情報を自動的に収集、蓄積するにあたって、その言語の背景の文化はどう処理されるのだろうか。

  難問、今後の課題でしょう。 言語の背後にあるものが、言語のデータに反映するので、それを処理するモデルというのは徐々に研究されている。あるいは、このデータから評価・感情表現を処理する技術も研究されつつあるので、将来は。。。

- 多言語間で、一つの概念を表す言葉どうしも厳密には1対1で対応していないと思うが、言語処理技術ではどのレベルまで結びつけて考えているのか。

  これは、本体の講義の中で。。。。

- 日常的に使用する言葉が、言語処理技術に適応していく可能性はあるか。

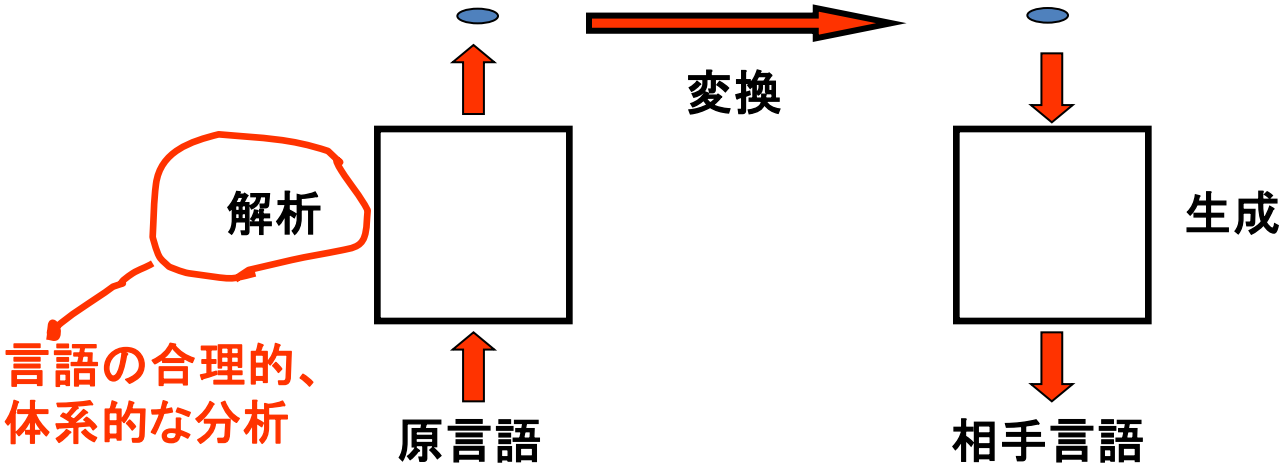
- 世界統一言語があった方がよいのではないか。

  うまく行かないのでは？ 言語によって表現される側の多様性、その動的な性質、新語の問題、専門用語・カタカナ語の問題

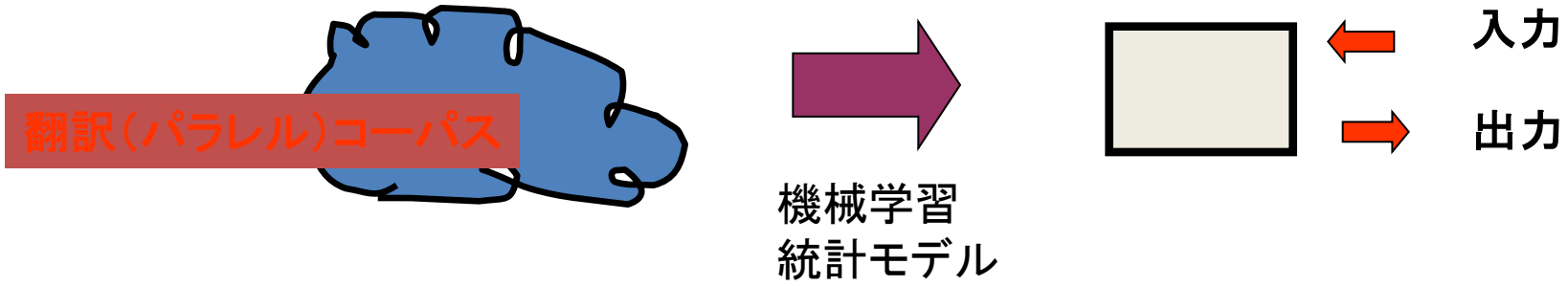
言語処理の技術：
機械翻訳を例にして

Minority Languageの機械翻訳への関心

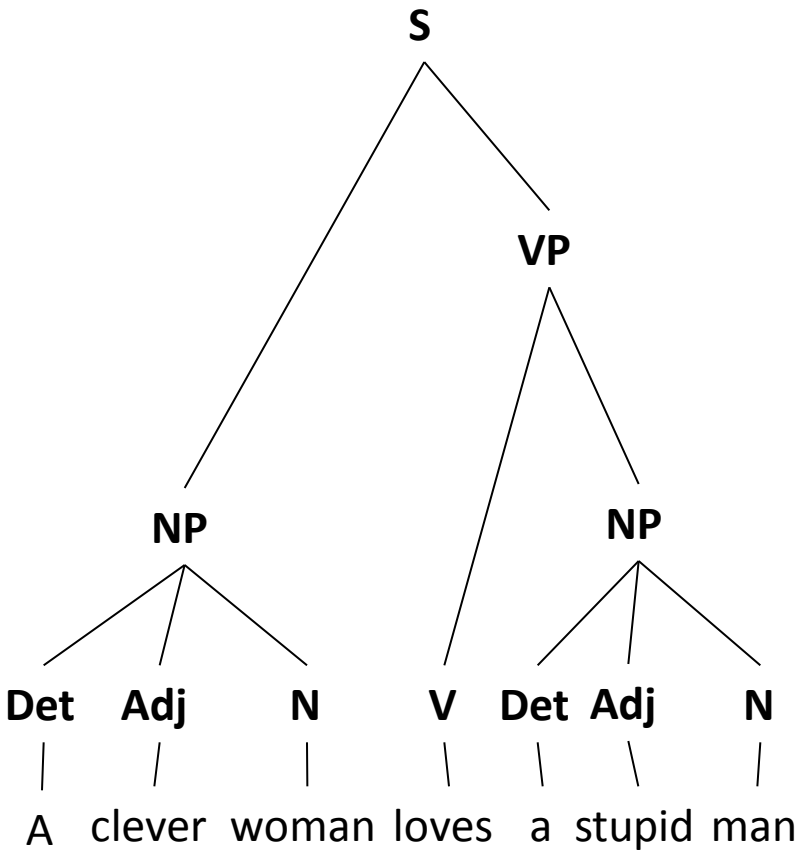
（1） 合理主義の機械翻訳： 規則に基づく翻訳

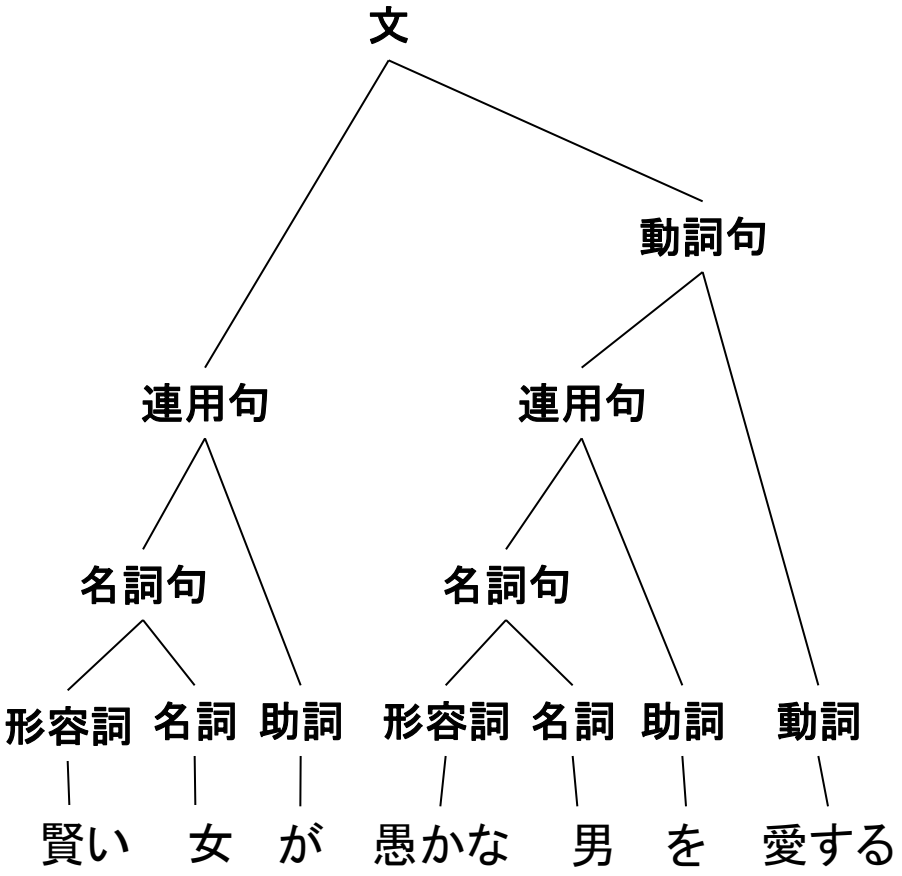


（2） 経験主義の機械翻訳： データに基づく翻訳

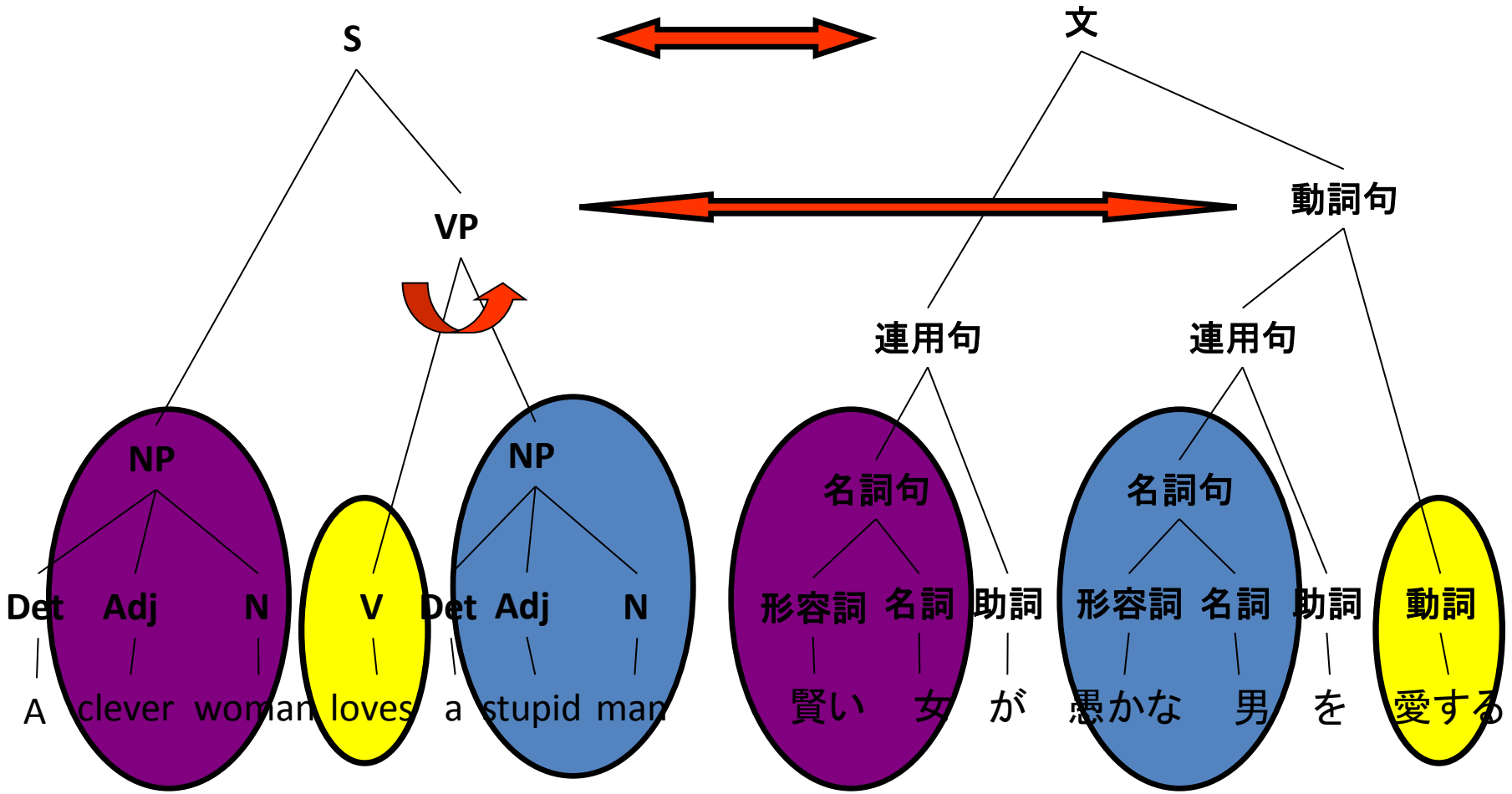


構成的な翻訳


S
VP
NP
NP
Det
Adj
N
V
Det
Adj
N
A
clever
woman
loves
a
stupid
man

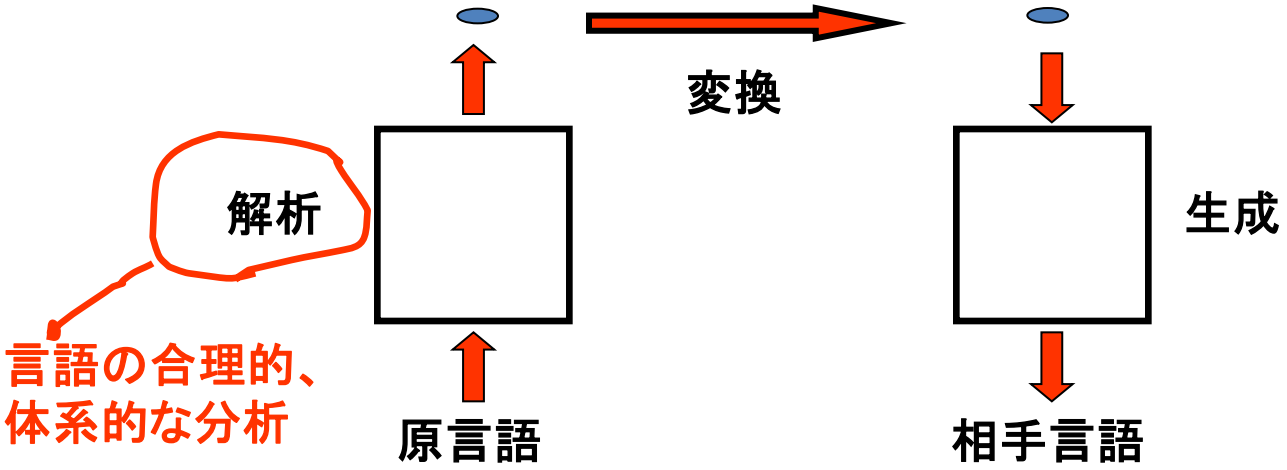

文
動詞句
連用句
連用句
名詞句
名詞句
形容詞
名詞
助詞
形容詞
名詞
助詞
動詞
賢い
女
が
愚かな
男
を
愛する

構成的な翻訳

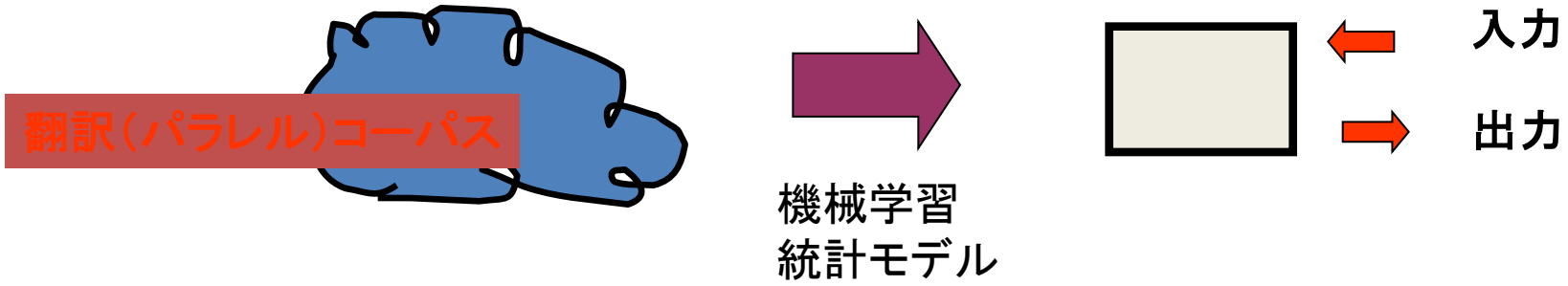

S
VP
NP
Det
Adj
N
V
NP
Det
Adj
N
A
clever
woman
loves
a
stupid
man
文
動詞句
連用句
連用句
名詞句
名詞句
形容詞
名詞
助詞
形容詞
名詞
助詞
動詞
賢い
女
が
愚かな
男
を
愛する

Minority Languageの機械翻訳への関心

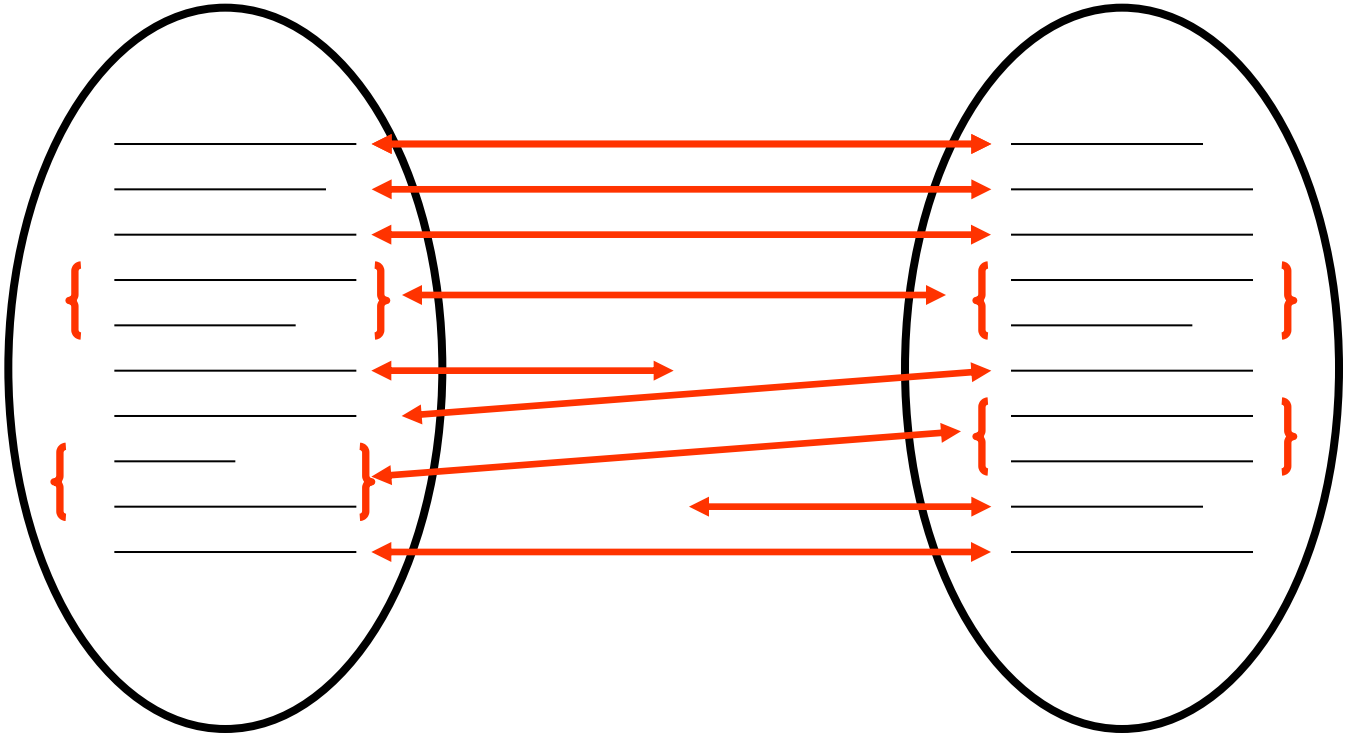
（1） 合理主義の機械翻訳： 規則に基づく翻訳



（2） 経験主義の機械翻訳： データに基づく翻訳



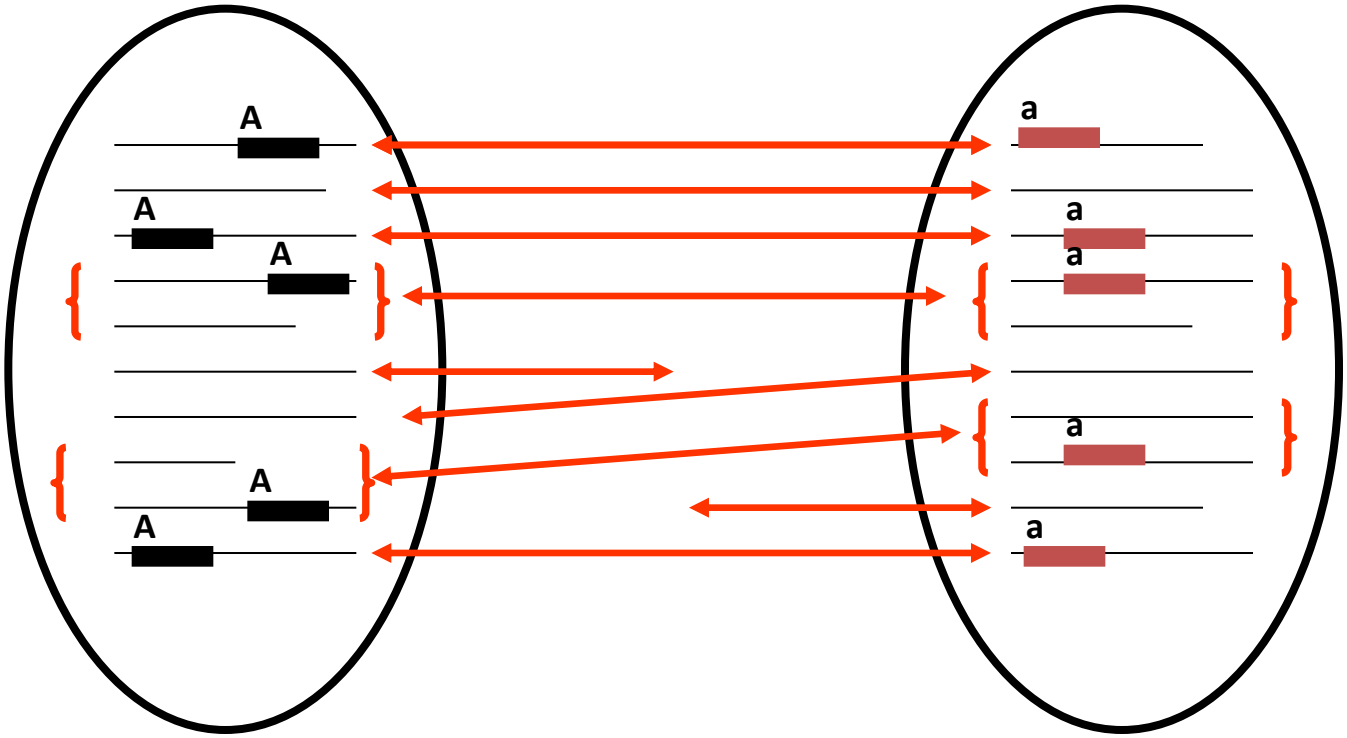
文のアライメント



日本語　　英語

単語のアライメント

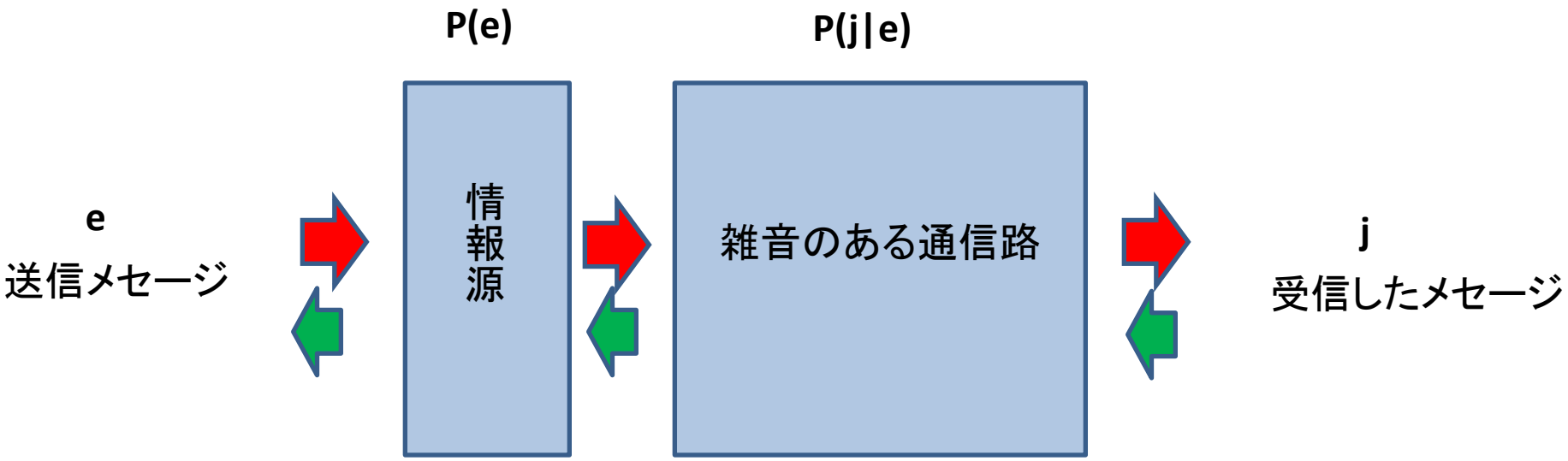
文のアライメント



単語のアライメント

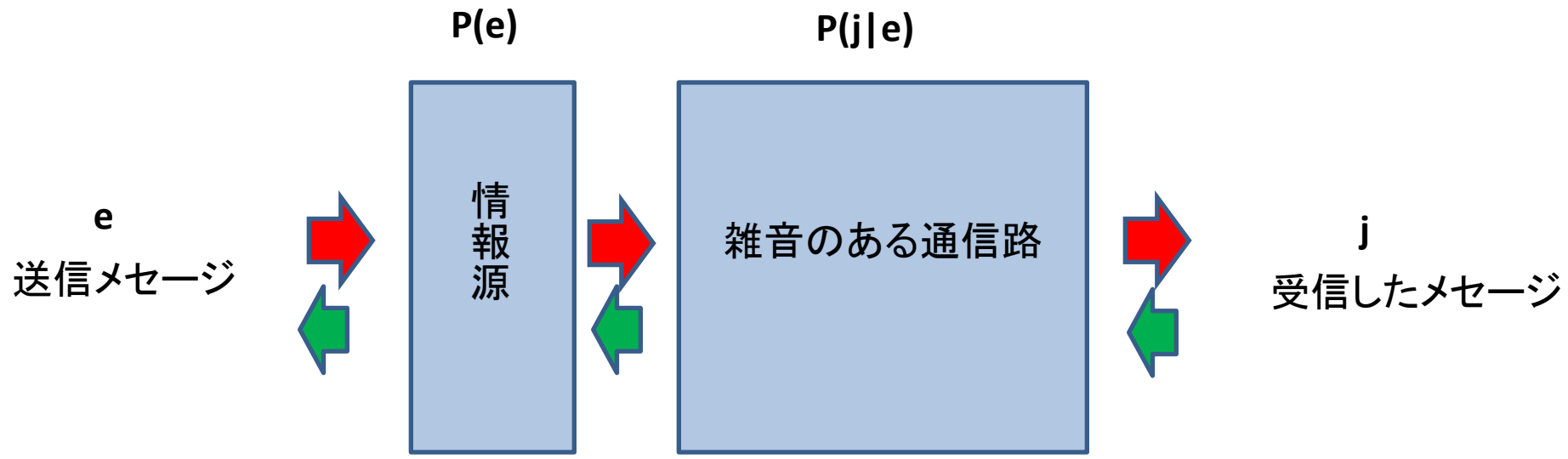
翻訳の統計モデル　：　$\underset{e \in E}{\mathrm{ARGMAX}} \{ P(e)P(j|e) \}$

雑音のある通信路
Noisy Channel Model



翻訳の統計モデル　：　ARGMAX { P(e)P(j|e) }

雑音のある通信路
Noisy Channel Model



Spring has come

P(e)
英語のテキストで
Spring has come
という文がでる確率

≒

P(spring|*)P(has|*, spring)P(come|spring, has)

P(j|e)
Spring has comeが
「春が来た」となる確率

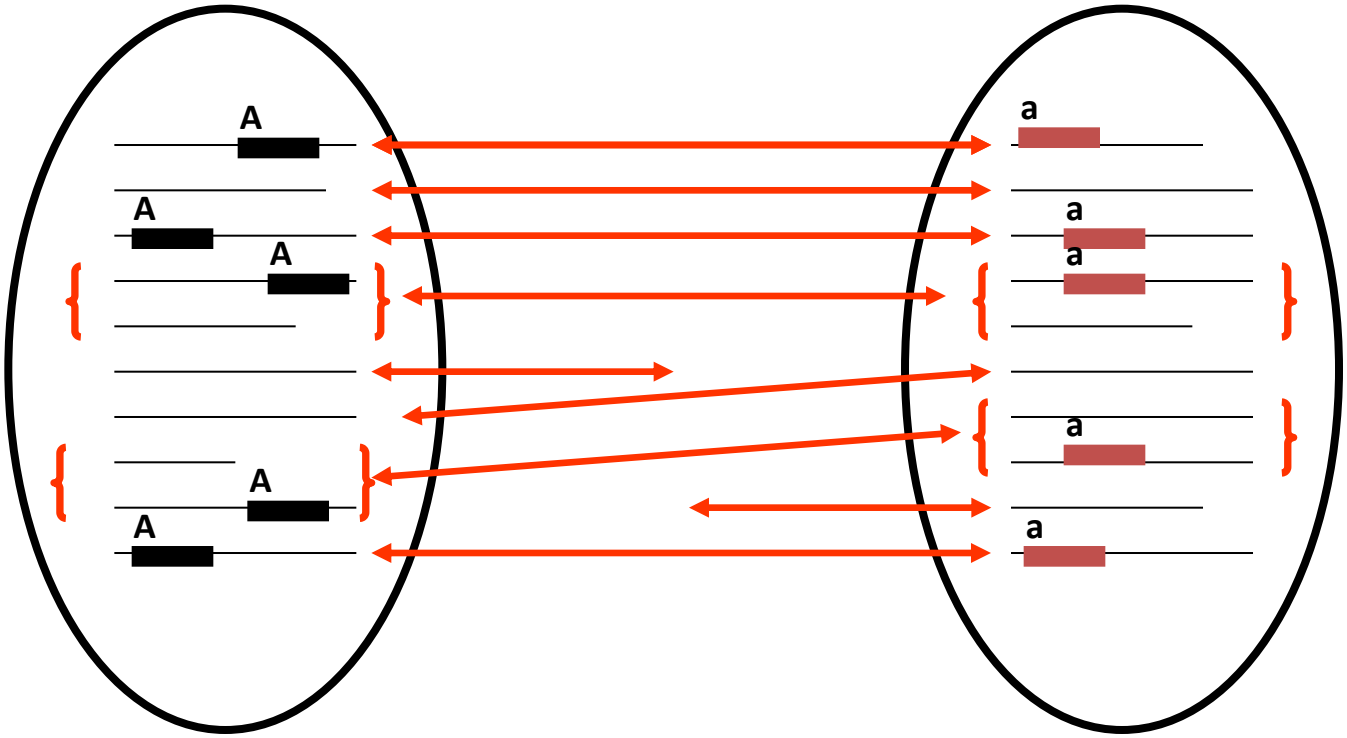
≒

P（春|spring)P(くる|Come)P(た|has)

春が来た

翻訳の統計モデル　：　ARGMAX｛P(e)P(j|e)｝

文のアライメント

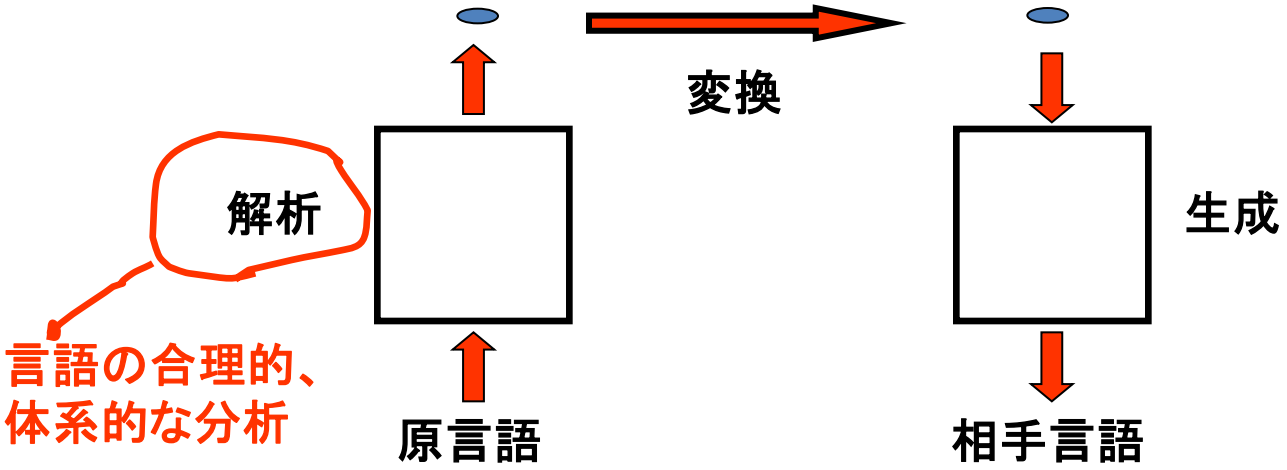


単語のアライメント

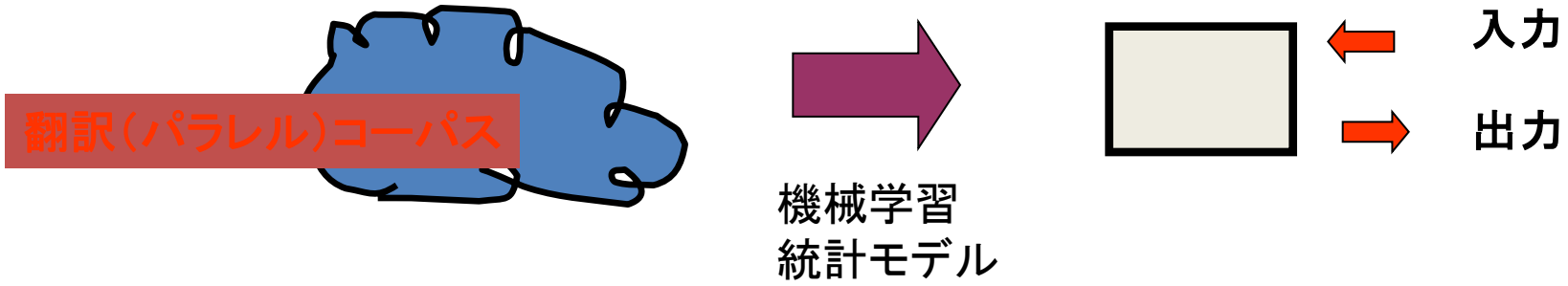
翻訳の統計モデル　：　$\underset{e \in E}{\mathrm{ARGMAX}} \{ P(e)P(j|e) \}$

Minority Languageの機械翻訳への関心

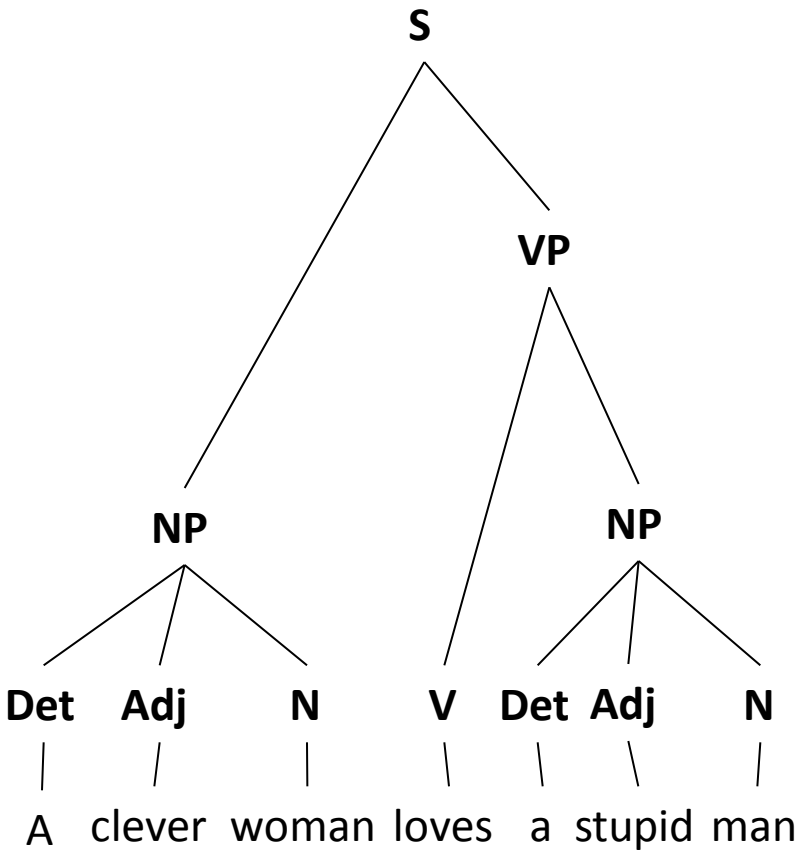
（1） 合理主義の機械翻訳： 規則に基づく翻訳

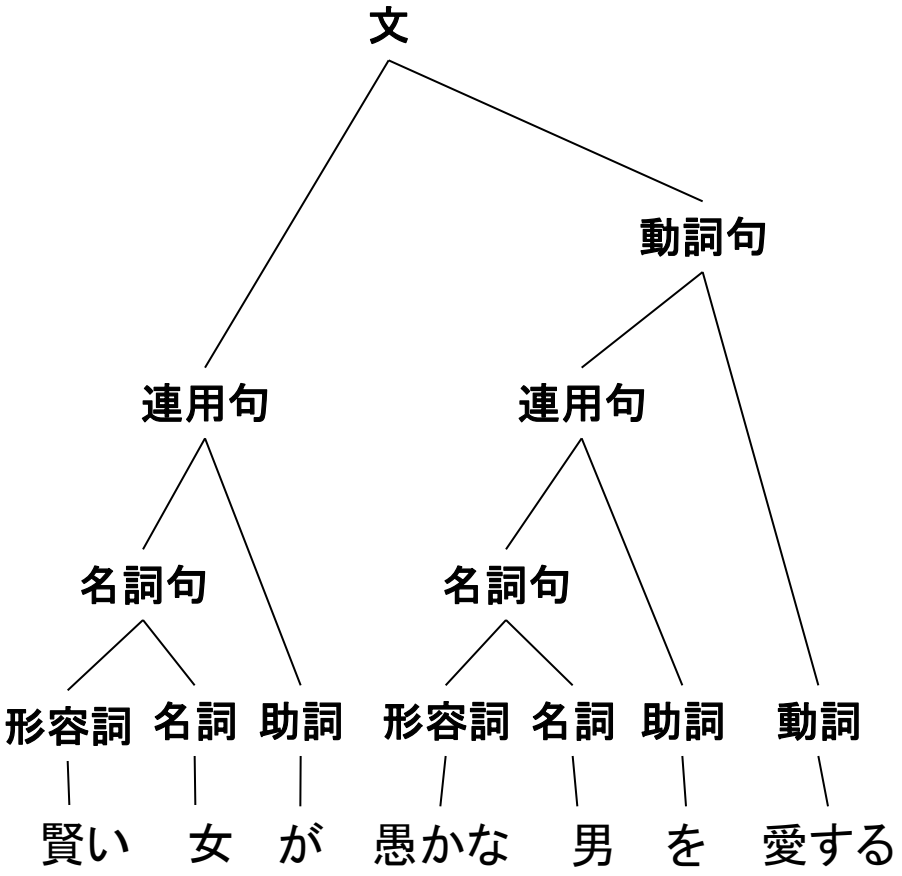


（2） 経験主義の機械翻訳： データに基づく翻訳



構成的な翻訳


S
VP
NP
NP
Det
Adj
N
V
Det
Adj
N
A
clever
woman
loves
a
stupid
man


文
動詞句
連用句
連用句
名詞句
名詞句
形容詞
名詞
助詞
形容詞
名詞
助詞
動詞
賢い
女
が
愚かな
男
を
愛する

言語の規則性

- 統語的な規則性　シンタックス、Syntax
  言語の形に関する規則性
  - 形態的な規則性
  - 統語的な規則性

無限

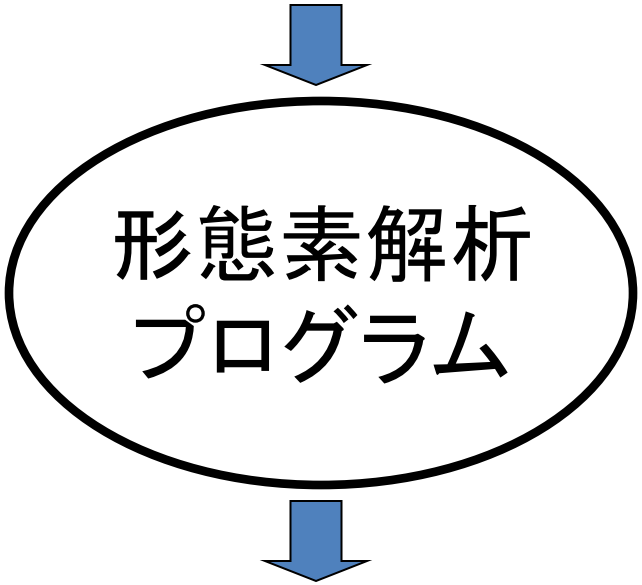
- 意味的な規則性
  言語と表現するものとの間の規則性
- 語用論的な規則性　セマンティックス、Semantics
  言語、表現するもの、環境（使い手、聞き手）との間の規則性　プラグマティックス、Pragmatics

形態的な規則性(Morphology)

- 屈折（Inflection）、接続関係
- 遊ぶ―遊ば ない (未然）　Asob　a
  遊び ます（連用）　Asob　i
  遊ぶ とき（連体）　Asob　u
  遊ぶ　　(終止）　Asob　u
  遊べ ば　（仮定）Asob　e
  遊べ　　(命令）　Asob　e
  語幹　語尾

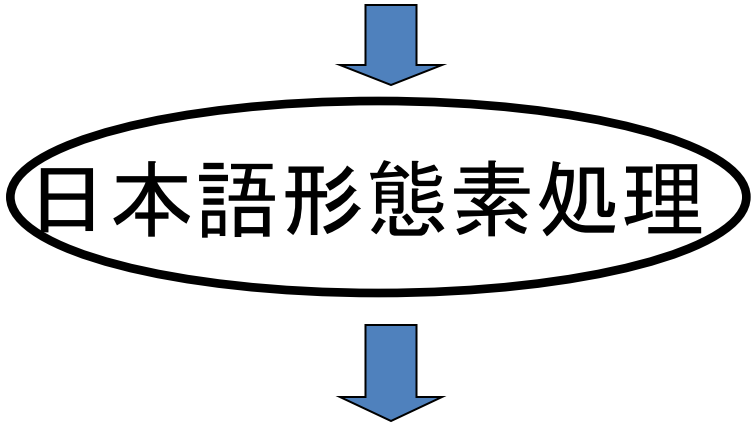
日本語の形態素処理

私は計算言語学の国際学会に
代理出席しなければなりません。



私・は・計算・言語・学・の・国際・学会・に・
代理・出席し・なけれ・ば・なり・ませ・ん。

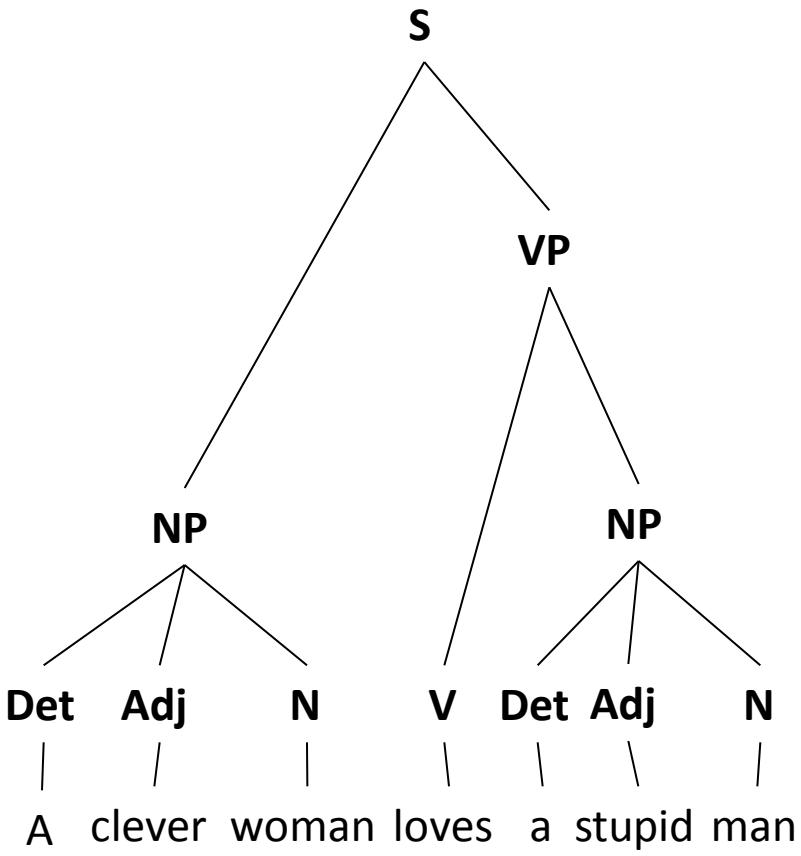
ここではきものを脱いでください。

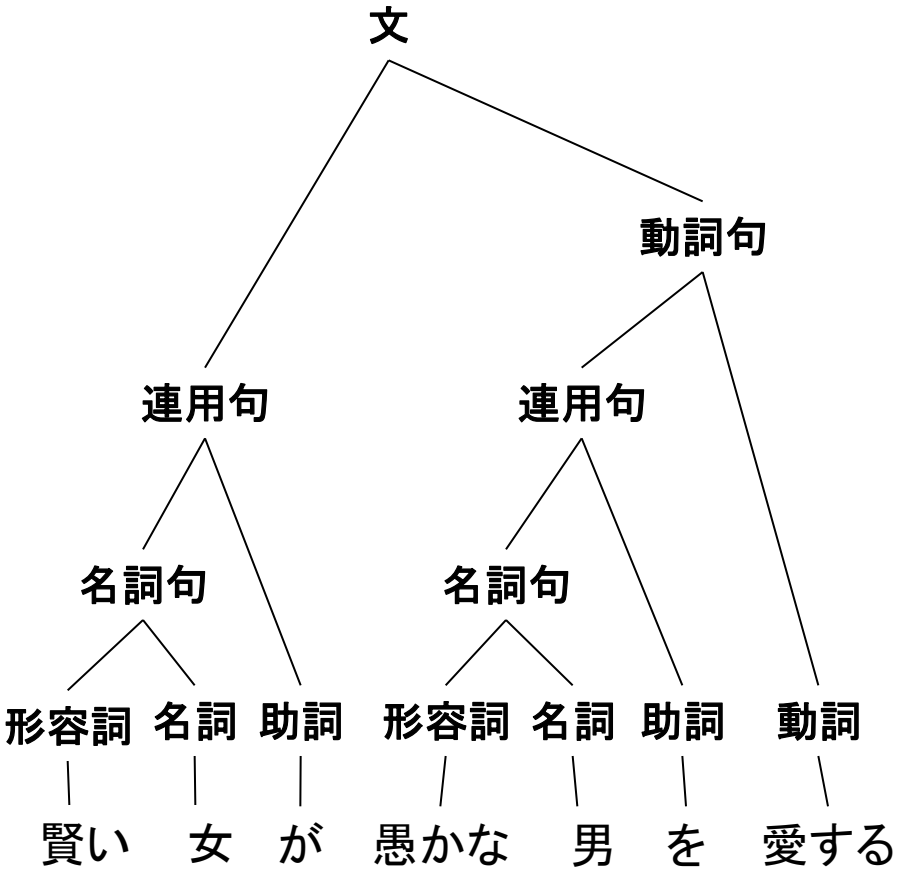


ここ･で･は･きもの･を。。。。。。

ここ･で･はきもの･を。。。。。。

構成的な翻訳

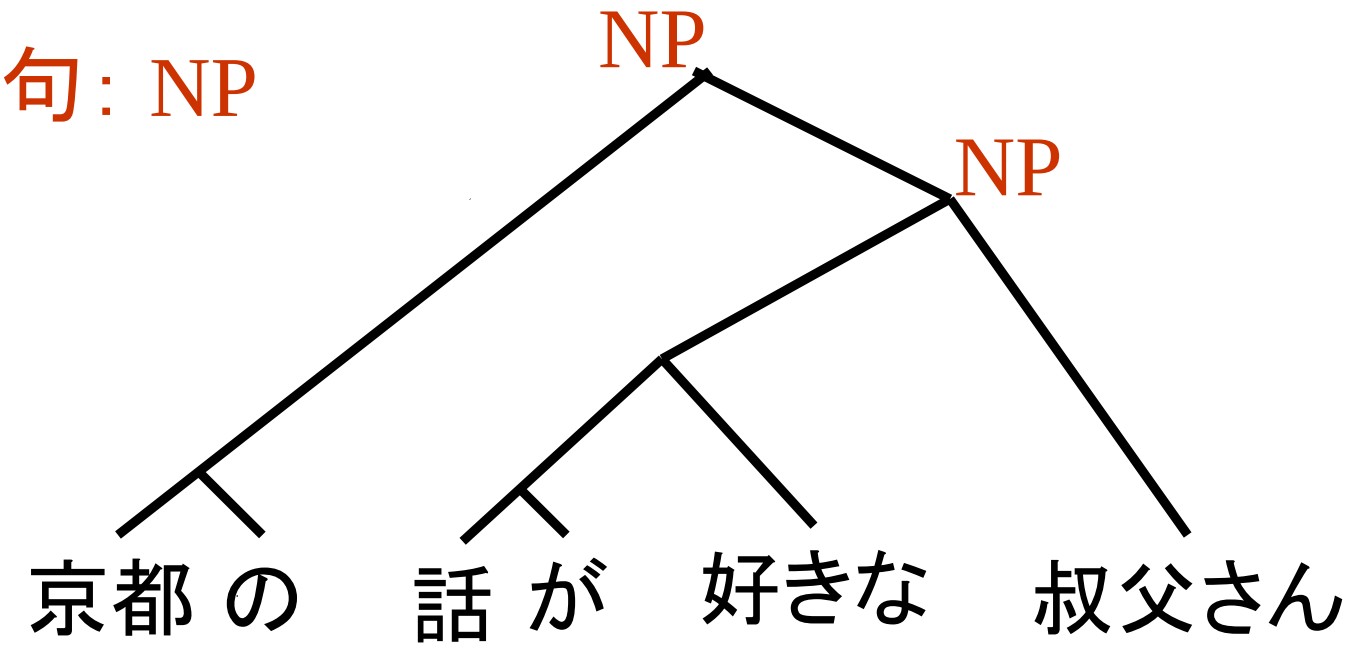

S
VP
NP
NP
Det
Adj
N
V
Det
Adj
N
A
clever
woman
loves
a
stupid
man

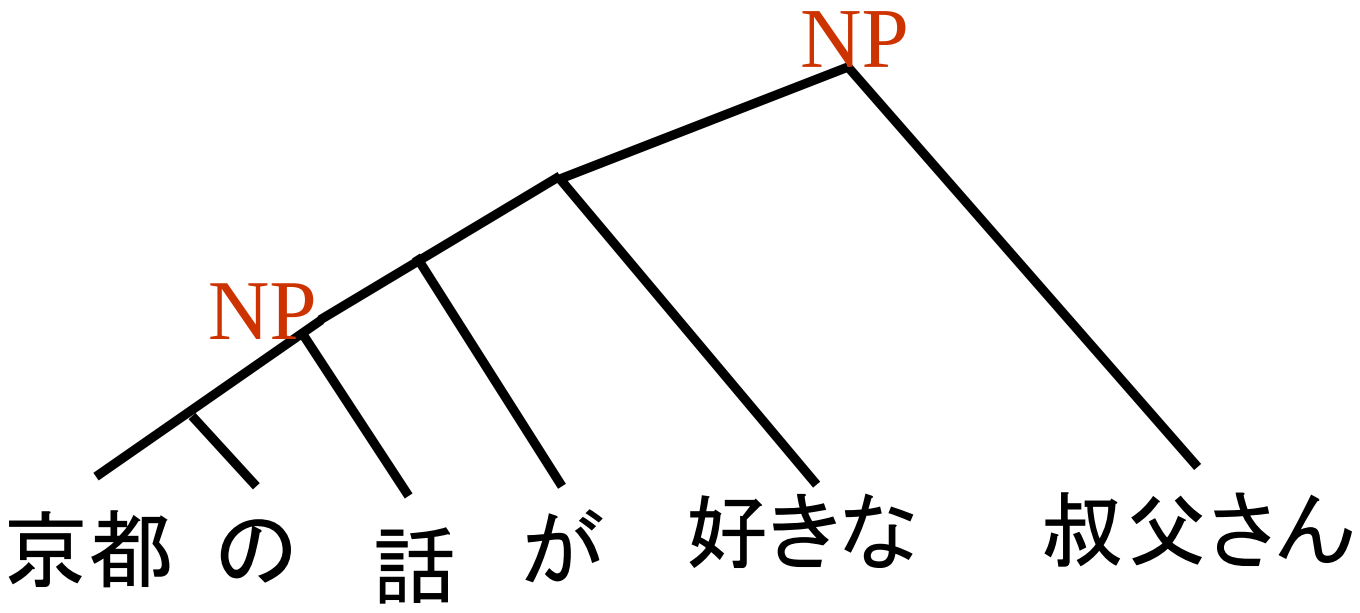

文
動詞句
連用句
連用句
名詞句
名詞句
形容詞
名詞
助詞
形容詞
名詞
助詞
動詞
賢い
女
が
愚かな
男
を
愛する

言語の持つ構造

- 京都の話が好きな叔父さん
- （京都の（（話が 好きな）叔父さん））
- （（（京都の 話）が 好きな）叔父さん）

名詞句: NP

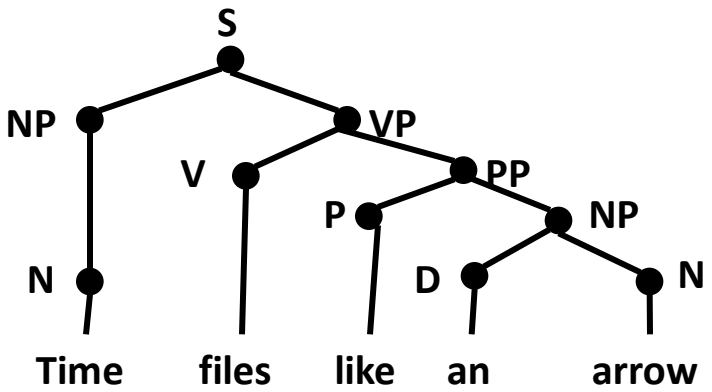




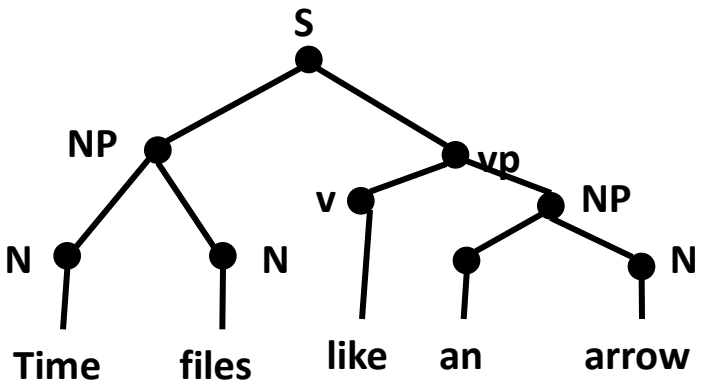
HMM（Hidden　Morkov　Model）による形態素解析（POS Tagger)

Time flies like an arrow.

N　　N　　V　　D　　N
V　　V　　P

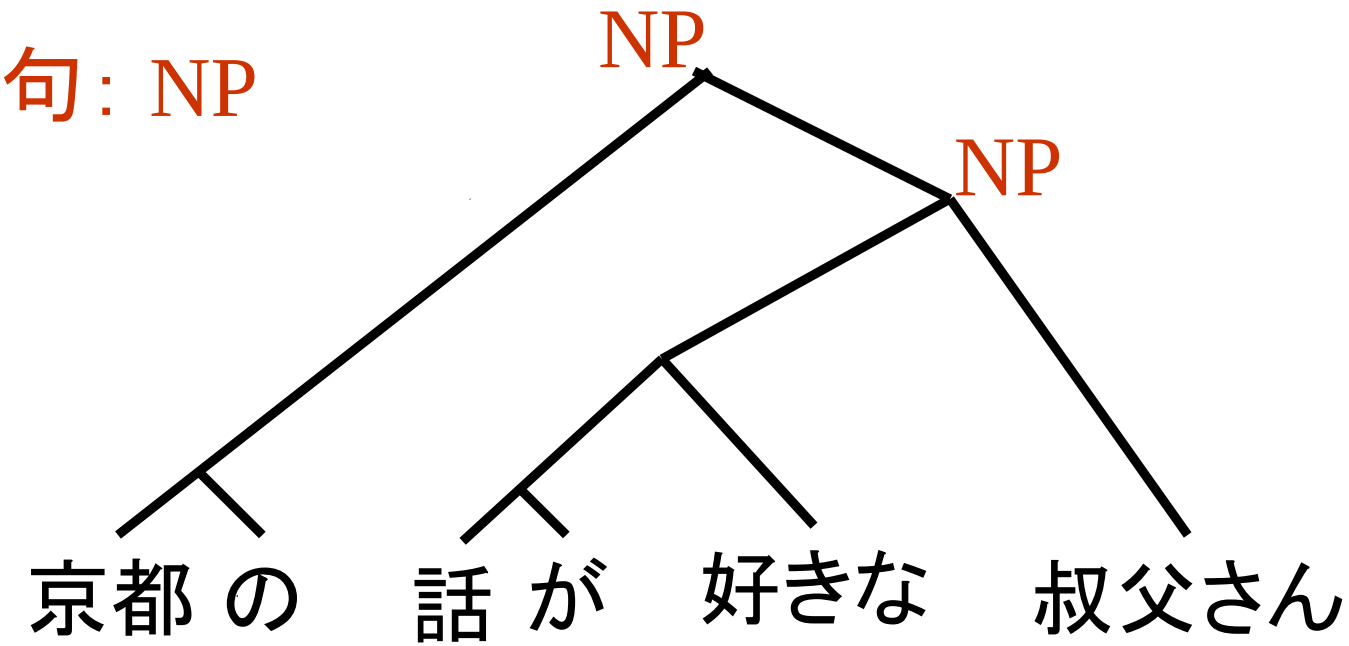


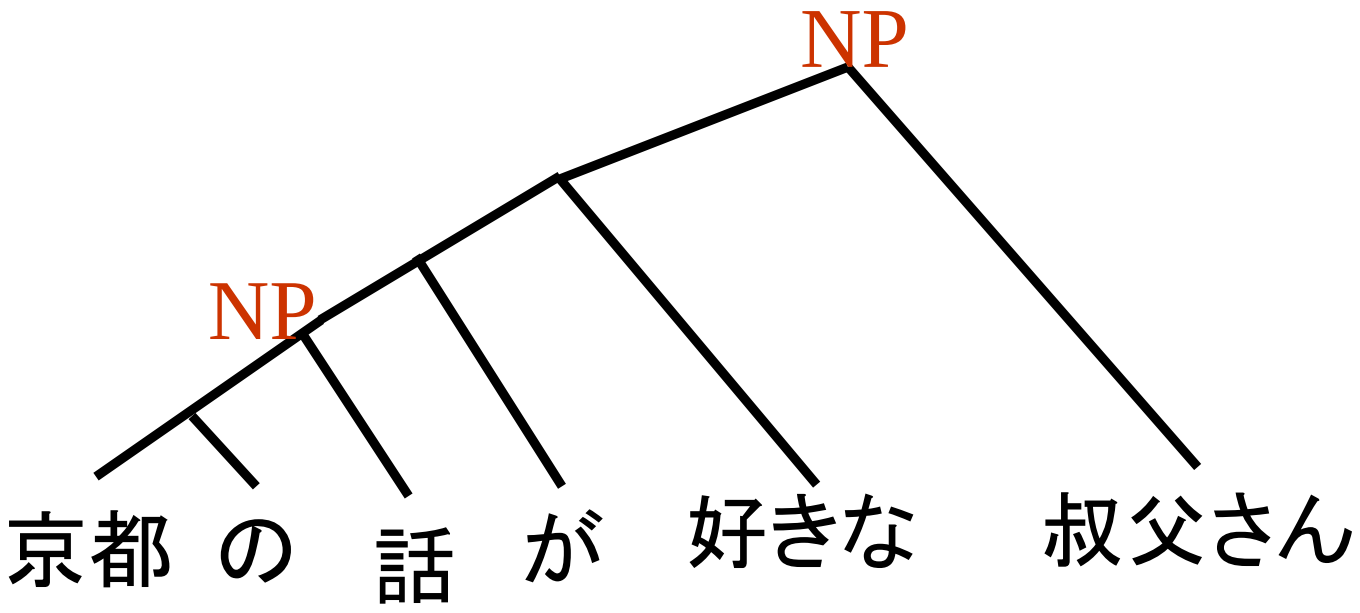
N　　V　　P　D　　N



N　　N　　V　D　　N

名詞句: NP





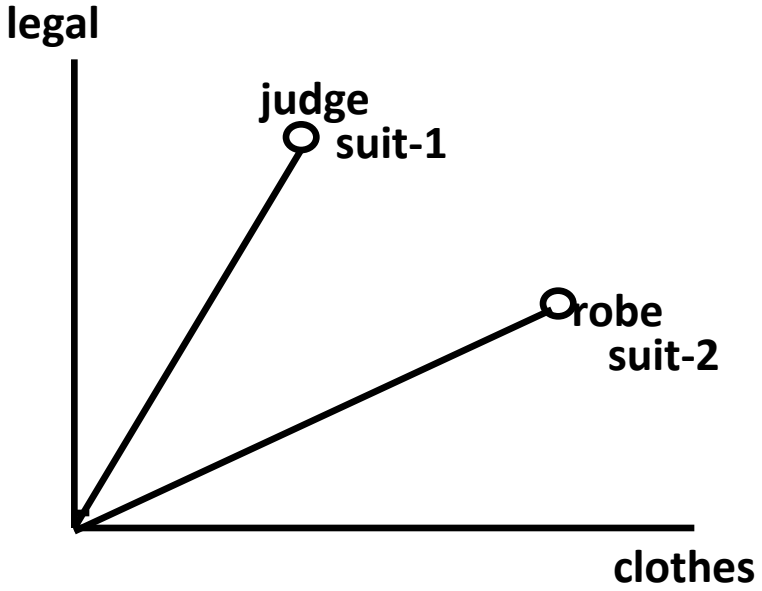
語彙の曖昧さと統計モデル

- Spring: 春、バネ、泉、。。。
- Bank: 銀行、川岸、。。。。

単語の意味空間

- 単語の意味は、共起する単語で定義される

単語 / 次元	suit-1	suit-2
	judge	robe
legal	300	133
clothes	75	200



言語と(意味・文脈・記憶・構造・解釈)

- 幼児、絵本
- Bilingualの人
- 公正＝to be fair
- Freedom Fighters , Terrorists
- 遊び、甘えの構造

スパゲッティ問題

- スパゲティ、スパゲッティ, スパゲッティー, スパゲッテイ, スパゲティー, スパゲテイ,スパゲッティ, ….

Microsoft Academic Search Beta

Explore over 5,381,371 papers, 155,073 were added last week.

tsujii

Advanced Search

Papers | Authors | Conferences | Journals

Jun'ichi Tsujii

Publications: 84 | Citations: 359 | G-Index: 17 | H-Index: 10

Research Interest: Natural Language & Speech, Bioinformatics and Computational Biology, Machine Learning and Pattern Recognition

University of Manchester, Manchester, United Kingdom

Bing

Freedom Fighters とTerrorists

d to check author trend. Please enable it in your browser or click here to install.

Papers | Citations

Year 2006

K. Masuda, T. Ninomiya, Y. Miyao, T. Ohta, **J. Tsujii** : Nested region algebra extended with variables , 2006 (Citations: 1)

Year 2005

Akane Yakushiji, Yusuke Miyao, Yuka Tateisi, **Jun'ichi Tsujii** : Biomedical information ex-traction with predicate-argument structure patterns , 2005 (Citations: 13)

Takashi Ninomiya, Yoshimasa Tsuruoka, Yusuke Miyao, **Jun'ichi Tsujii** : Efficacy of beam threshold-ing , 2005 (Citations: 1)

Year 2004

Microsoft Corporationのガイドラインに従って画面写真を使用しています‡