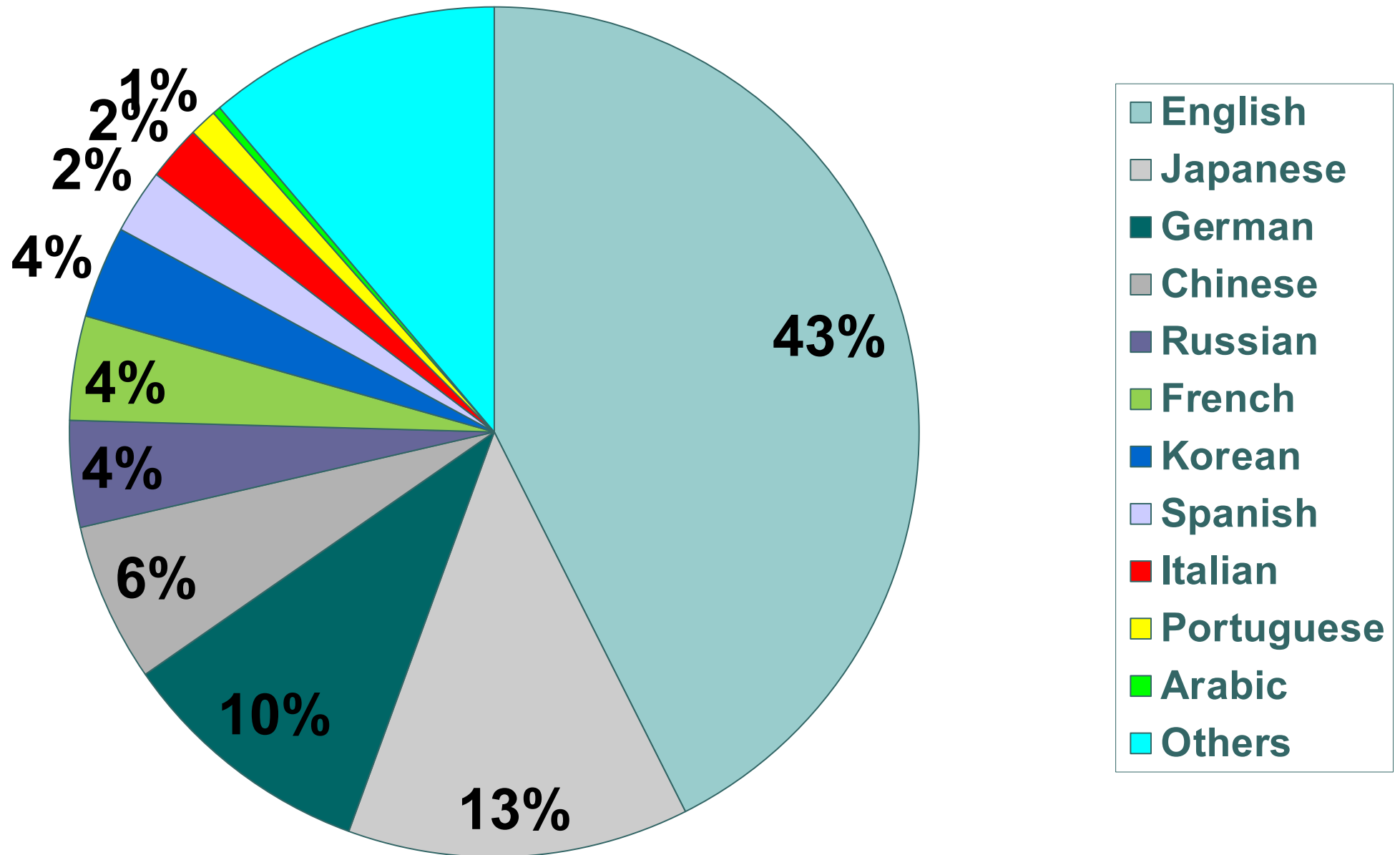




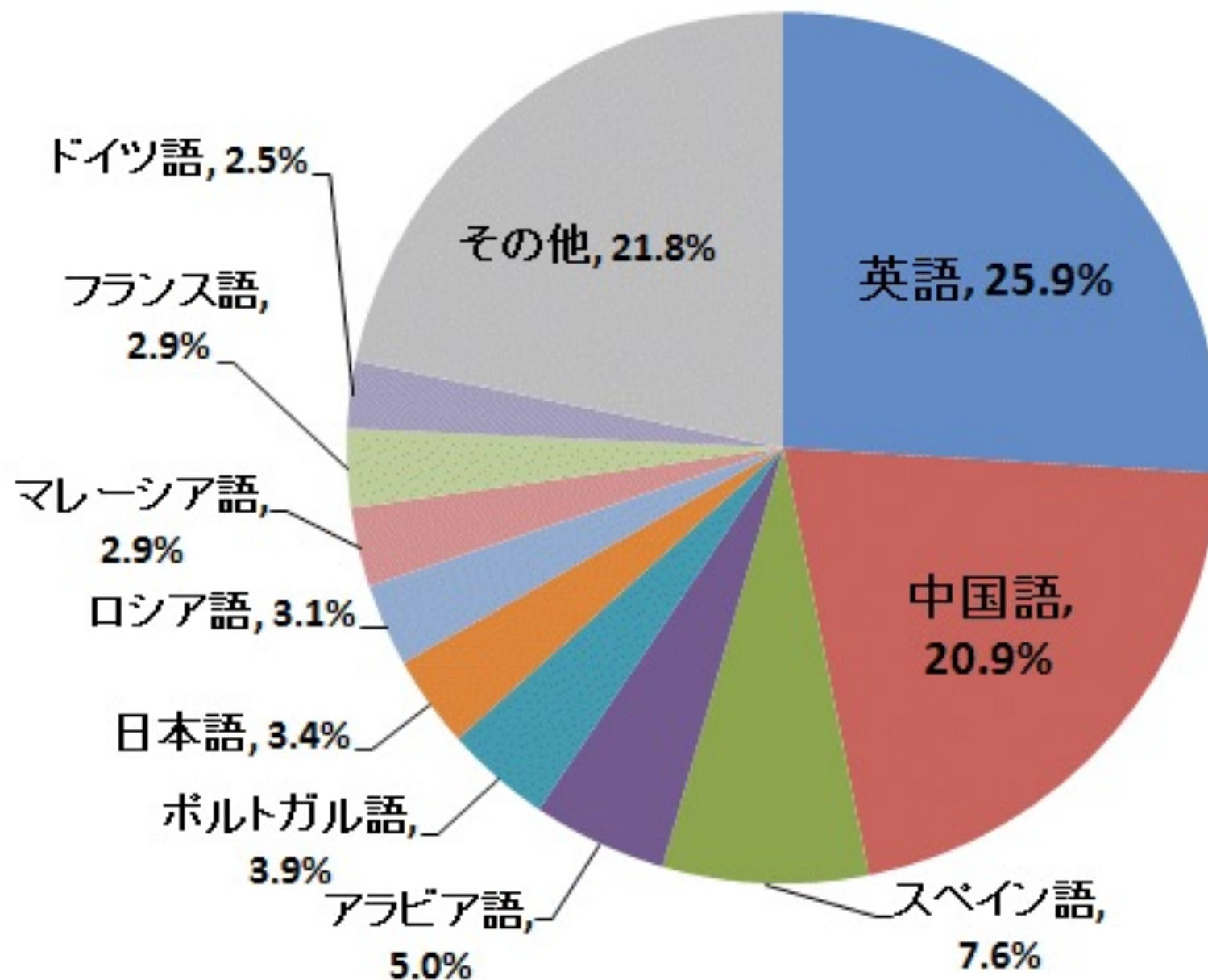
# 情報アクセス論 第11回

## 「多言語情報アクセス」

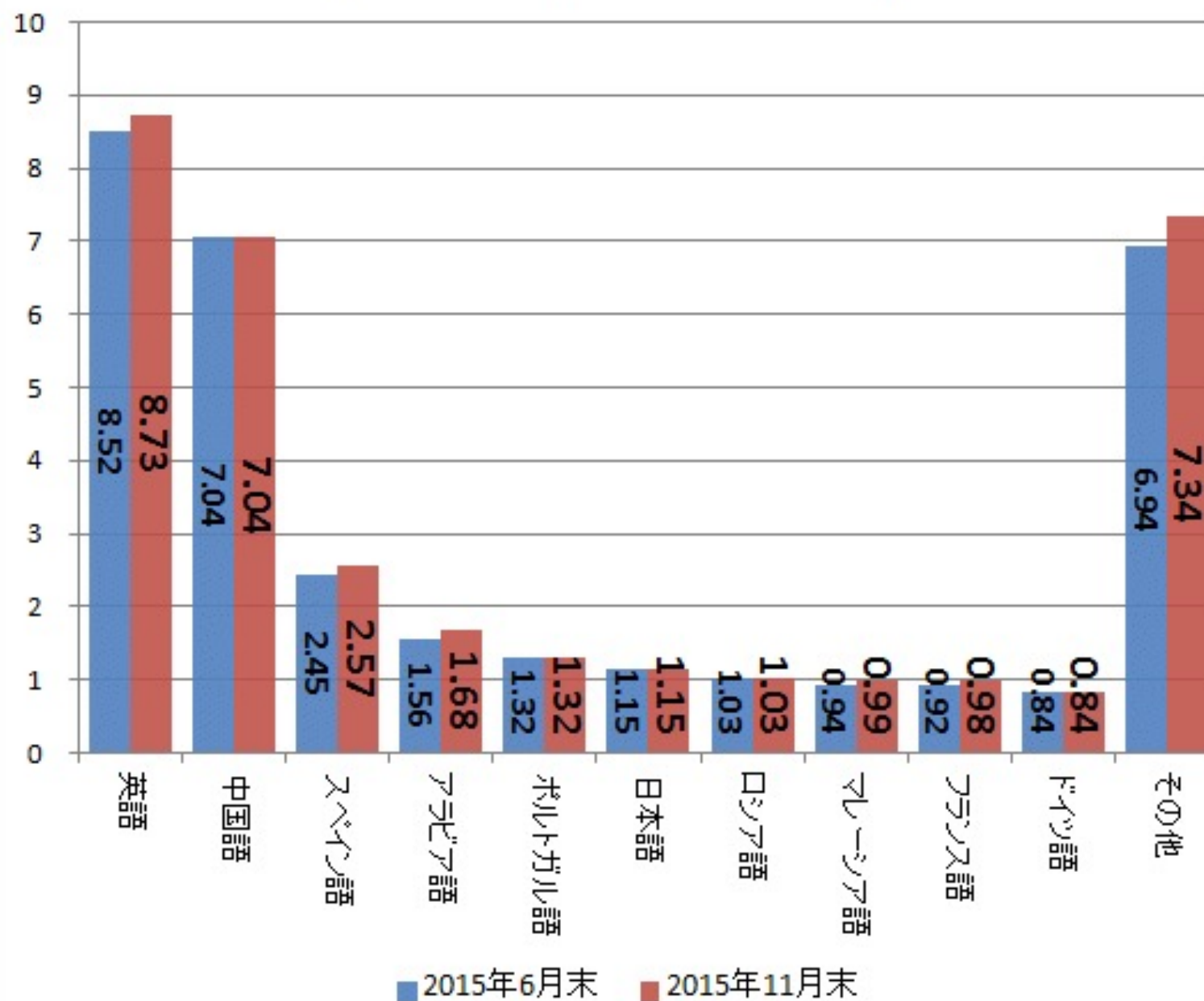
# Webページの言語別割合(2006)



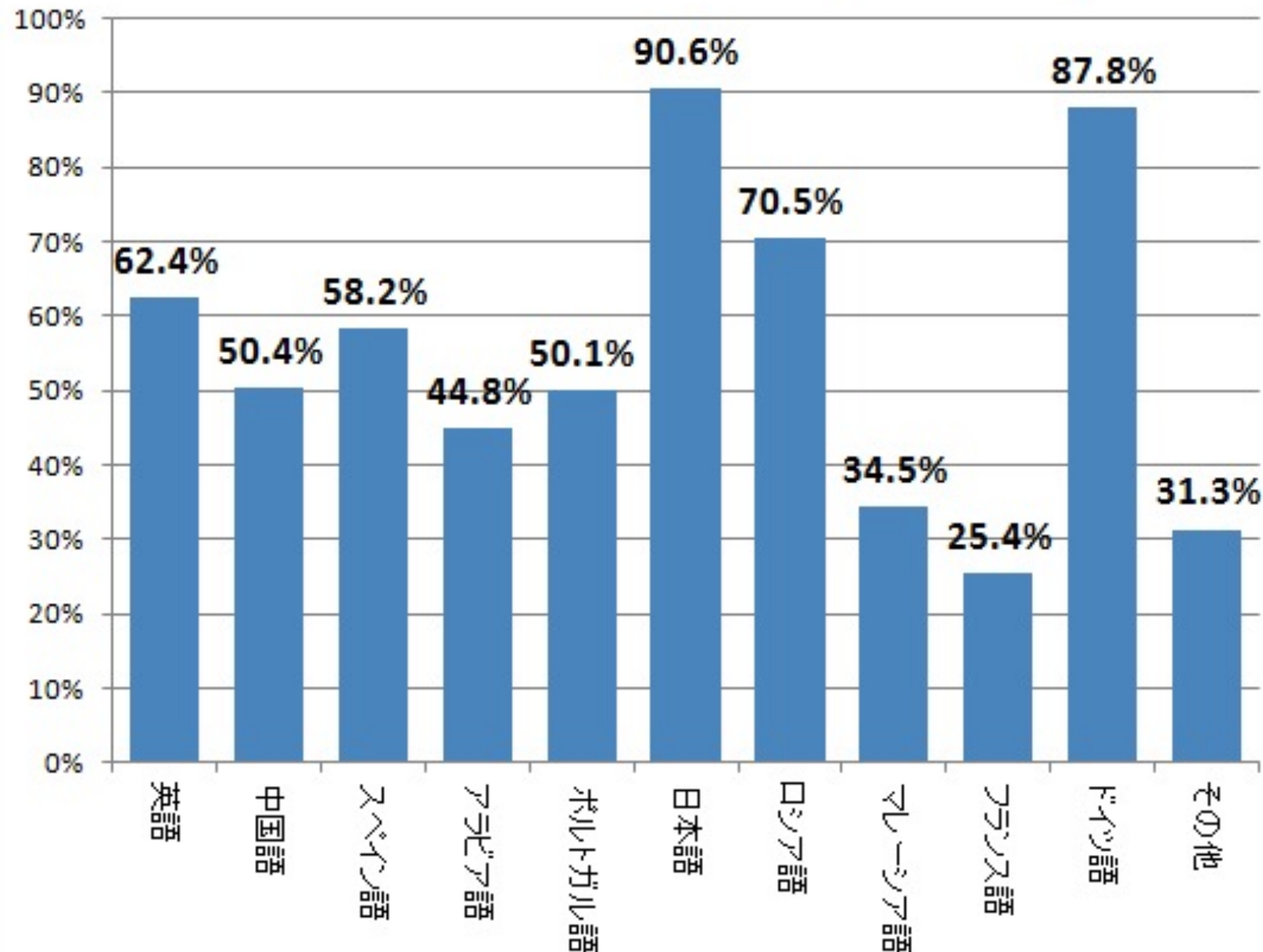
+その他(全体に占める比率)(2015年11月末時点)



インターネット上の主要言語トップ10+その他(億人)  
(2015年6月末/2015年11月末時点)



インターネット上の主要言語トップ10+その他における、  
各言語利用者毎のインターネット普及率(2015年11月末)





# 多言語情報アクセスとは？

- 多言語の情報が混在する情報源に対する情報アクセス技術
  - 欧州連合(EU)の公用語は23言語
  - インドの公用語は19言語
- Web上では言語・国家などによる区分が存在せず、世界中の様々な言語の情報が混在
  - Wikipediaは281の言語版が存在(2011/6/1現在)
- 利用者の母国語だけでなく、他の言語に対するアクセスを実現する技術



# 多言語情報アクセスの課題1: 多言語情報検索

- 利用者が使う言語によって、検索対象が制限される
  - 日本語だとWeb全体の一部しか検索できない
- 検索要求によっては、他の言語も探したい
  - ある国のニュースは、その国の言語によるニュースサイトのほうが情報が豊富
- 言語横断情報検索(**Cross-Language IR**)
  - ある言語で書かれた文書群を別の言語による問合せで検索

# 言語横断検索へのアプローチ

## ○ 検索対象の文書を翻訳

- Webのように、大規模で多言語かつ更新が頻繁な文書群に対しては非現実的

## ○ 利用者の問合せを翻訳

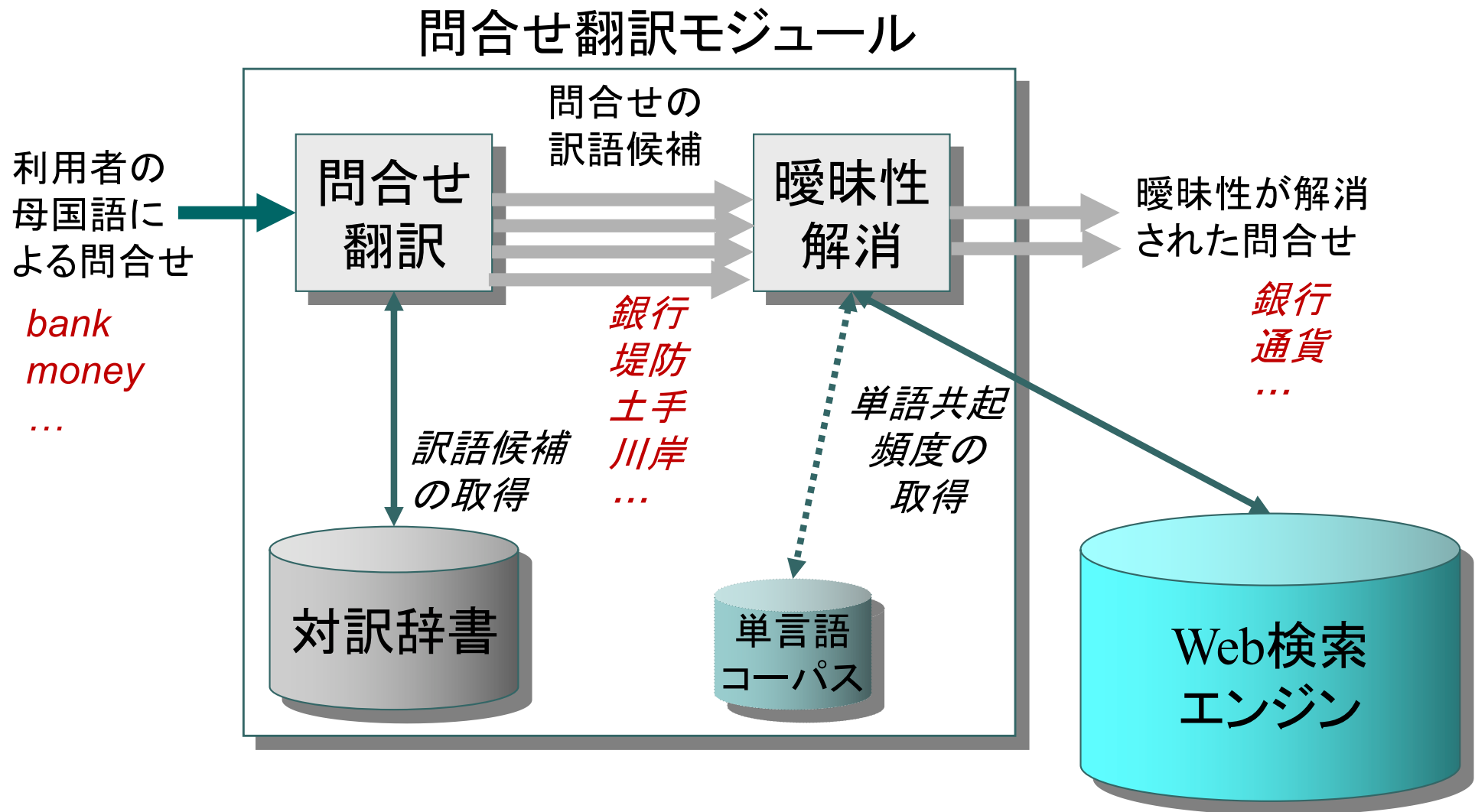
- 翻訳された問合せは、既存の検索エンジンにそのまま適用可能
- 辞書で翻訳しただけでは、訳語の曖昧性が生じる
  - bank: 銀行, 堤防, 土手, 川岸...
  - crane: 鶴(ツル), 起重機(クレーン)



# 訳語曖昧性の解消手法

- 検索対象言語コーパス中における単語の**共起傾向**を用いる
  - **コーパス**: 大量のテキストを集めた言語データ
  - **共起傾向**: 単語間の関連の強さの統計量
- 既存のコーパスは分野が限られている
  - 新聞記事, 文学, 特許, 論文, 国会議事録, etc.
- **Web検索エンジン**をコーパスとして利用
  - 多様な分野にわたる膨大な量の言語資源
  - 訳語候補の組をWeb検索エンジンで検索した検索文書数を共起頻度とみなす

# 問合せ翻訳の流れ



# 共起傾向による曖昧性解消の 手順

問合せ	辞書による訳語候補リスト
bank	<span>銀行</span> , 貯金箱, 岸, 浅瀬, 土手, 堤防. . .
	AND AND AND 9.10
money	<span>富</span> , 財産, <span>資産</span> , <span>通貨</span> . . .
	AND 7.61 AND AND AND
trade	<span>商売</span> , 同業者, <span>貿易</span> , 交換, 道. . .

問合せ：（銀行 AND 通貨 AND 貿易）OR（銀行 AND 資産 AND 商売）

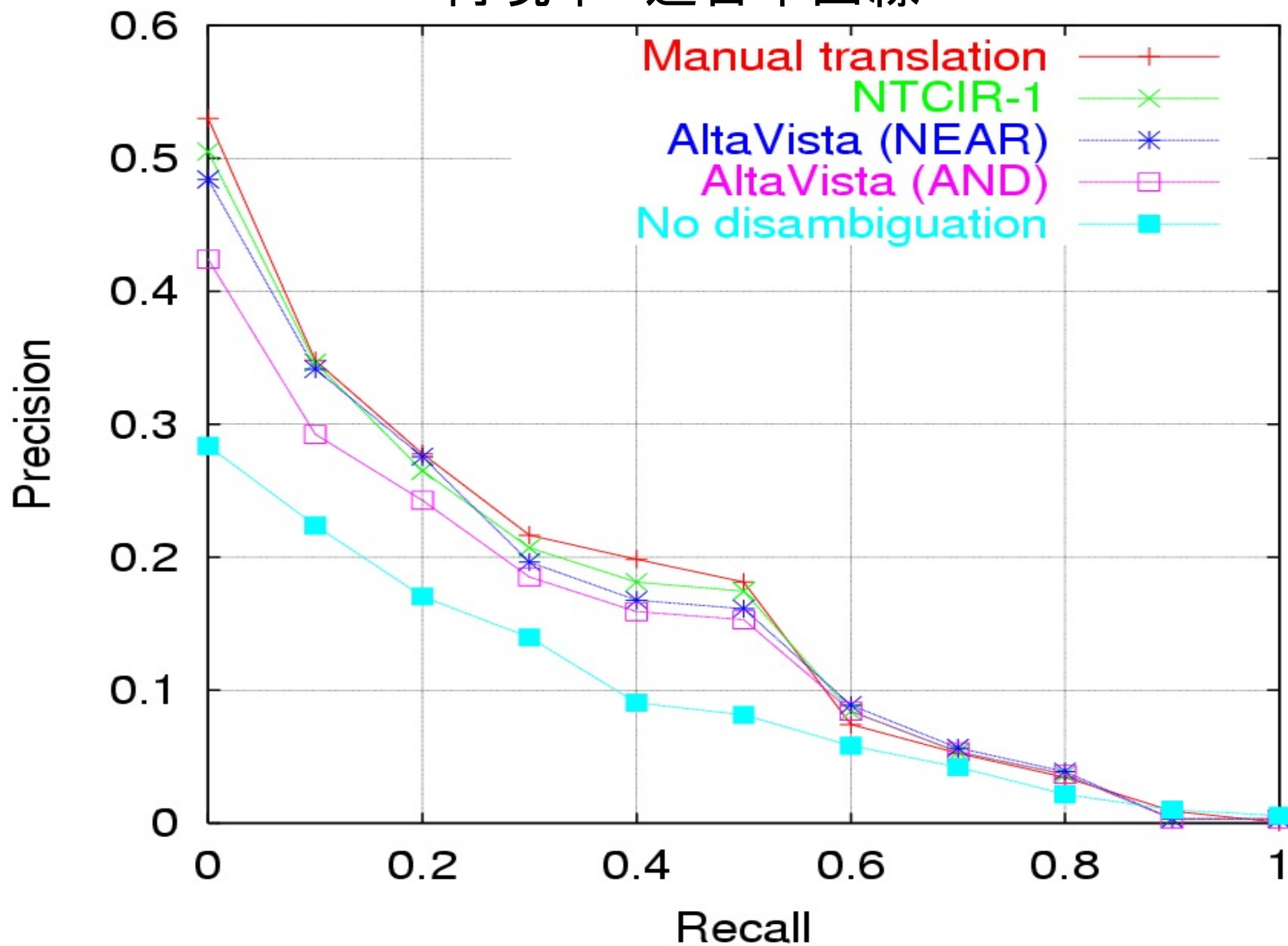
## ●曖昧性解消の例

●「神経 再生」という問合せに対して、各単語の英訳の組み合わせの共起傾向（相互情報量）を計算

●上位7件がすべて正解，それ以下はすべて不正解

順位	訳語候補の組		共起傾向
	神経	再生	
1	nerve	regeneration	2.20
2	nerve	regrowth	1.82
3	“nervous system”	regeneration	1.12
4	nerves	regeneration	0.54
5	nerves	regrowth	0.43
6	“nervous system”	regrowth	-0.17
7	sensation	regrowth	-1.52
8	sensation	reincarnation	-2.33
9	“nervous system”	resuscitation	-2.65
10	sensitivity	playback	-2.95
11	sensitivity	regeneration	-3.07
12	nerve	resuscitation	-3.07
13	sensation	regeneration	-3.09
14	worry	read	-3.27
15	sensation	rebirth	-3.91

再現率-適合率曲線

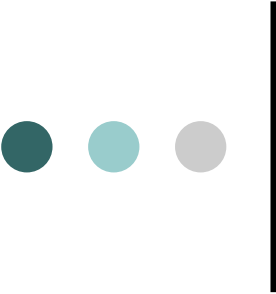




# 言語横断情報検索の実例

## ○ Googleの「翻訳して検索」機能


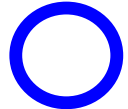


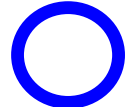
- 利用者が入力した問合せを翻訳して検索
- 52言語への翻訳に対応
- 検索結果のスニペットやページ自体も翻訳して表示
- 後述の機械翻訳機能を用いて問合せを翻訳しているため、曖昧性解消がうまくいかない場合がある
  - 「river bank」→「川銀行」



## 多言語情報アクセスの課題2: 機械翻訳

- 言語横断情報検索が実現できても、検索結果の文書を読めなければ意味がない
- **機械翻訳**: コンピュータで、ある言語の文を他の言語の文に翻訳する技術
- 自然言語処理技術の集大成
  - 形態素解析, 構文解析, 意味解析, 文生成, 辞書, 統計的言語処理, etc.
- 古くから研究されているが、難しい課題
  - 人間にとっても高度に知的な処理

## 機械翻訳の例

- 「あなたのオフィスに明日行きます」
  - 「I will go to your office tomorrow.」 
  - 「I will **come** to your office tomorrow.」 
- 「Time flies like an arrow.」(光陰矢のごとし)
  - 「**時バエ** (time flies) は矢を好む」 
  - 「時は矢のように飛ぶ」 
  - 「月日が経つのは矢のように速い」 



# 言語構造の違い(1)

## ○ 語順が異なる

- 日本語はSOV型, 英語はSVO型
- 「彼はスーツを着ている」→「He wears a suit.」

## ○ 必ずしも一対一に対応しない

- 「play」→「(ピアノを)弾く」「(フルートを)吹く」  
「(野球を)する」
- 「服を着る」「眼鏡をかける」「帽子をかぶる」「靴を履く」→「put on one's glasses [hat, coat, shoes, ring, eye shadow]」
- 「スープを飲む」→「eat soup」

## 言語構造の違い(2)

- 対応する語の品詞が異なる
  - 「3冊の本」(数詞＋の＋名詞)→「three books」  
(形容詞＋名詞)
- 1語が2語に対応する
  - 「湯」→「hot water」, 「牛・肉」→「beef」
- 1語が節に対応する
  - 「efficient」→「効率が良い」



# 機械翻訳方式の分類

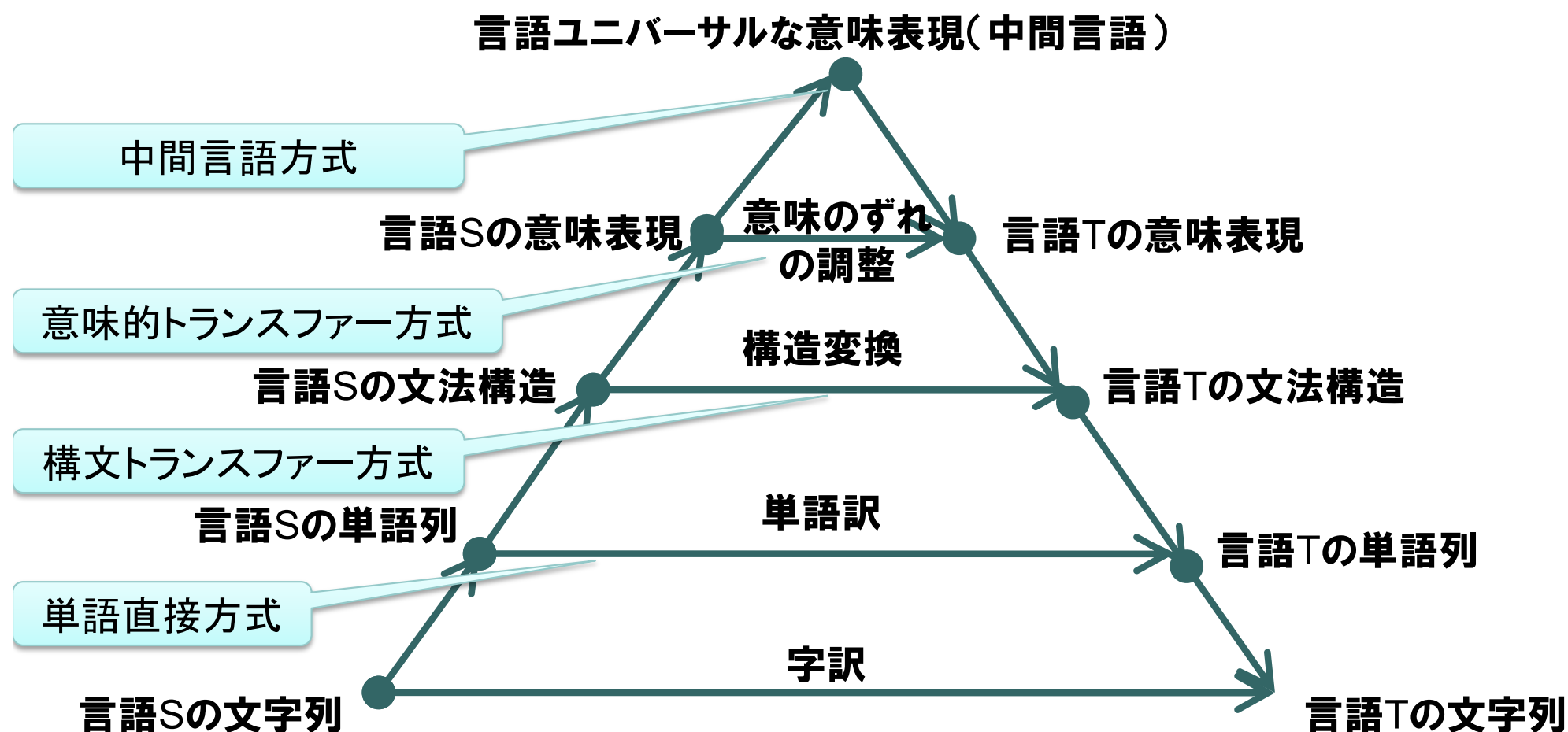
## ○ 用語の定義

- **原言語** (source language) : 翻訳元の言語
- **目的言語** (target language) : 翻訳先の言語

## ○ 主な機械翻訳方式

- 単語直接方式
- **トランスファー方式**
- 中間言語方式
- 実例型機械翻訳
- 統計的機械翻訳

# 解析／生成のトライアングル

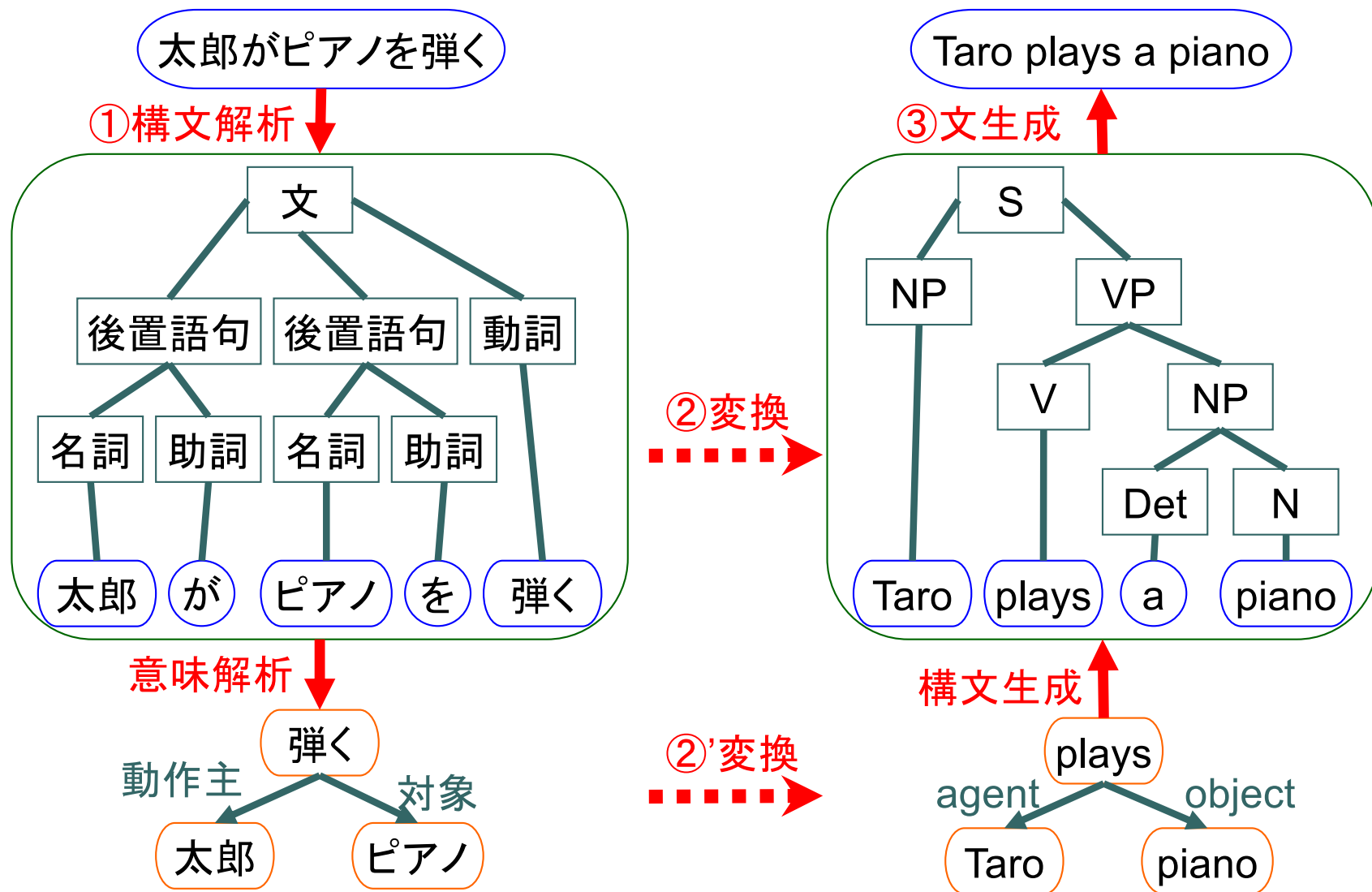




# トランスファー方式

- 機械翻訳の伝統的な方法
- **解析, 変換, 生成**の3過程からなる
  1. 原言語の文を解析して文構造を求める
  2. 原言語の文構造を目的言語の文構造に変換
  3. 目的言語の文構造から訳を生成
- 構文レベルで変換する場合, **構文トランスファー方式**と呼ばれる
  - 多くのシステムでは, 意味解析も含む

# トランスファー方式の概念図



# ● ● ● | トランスファー方式の処理過程 (1)

## ①解析

- まず形態素解析が行われ、次に構文解析

## ②変換

- 文の変換
  - 文が主節や従属節からなる場合、部分構造を変換し、それらの結果と部分構造の関係により文全体の変換が行われる
  - “He likes mathematics but she doesn’t like it.”
  - 「彼は数学が好きだが、彼女はそれを好きではない」

# トランスファー方式の処理過程 (2)

## ○ 節(格構造)の変換

- **格フレーム**(動詞がどのような名詞に修飾されるかを示す知識)を用いる
- 動詞と名詞の訳し分け
  - “take a walk”「散歩する」
  - “take a cold”「風邪を引く」
- 表層格の対応
  - “take a picture”「写真を撮る」
  - “take a bus”「バスに乗る」



# トランスファー方式の処理過程 (3)

## ○ 名詞句の変換

- 名詞句の構成要素である修飾部と主要部の構造を変換
- 修飾部が関係節(埋め込み文)
  - “picture that Mary painted”「メアリが描いた絵」
- 修飾部が前置詞句(後置詞句)
  - “girl with blond hair”「金髪の少女」

## ③生成

- 意味構造／構文構造を単語列, 文字列に変換
  - 語順, 態, 活用形, 人称, 数的一致, 冠詞など

# 「機械翻訳」のまとめ

- 単なる自然言語処理ではなく、自然言語**理解**
- 高度な処理であり、かつ膨大な知識(変換規則, 辞書)が必要
- 用途を限定すれば、ある程度実用的な精度が得られる
  - 公文書の翻訳(EC諸国)
  - 英語の天気予報をフランス語に(カナダ)
  - 科学技術論文の日英翻訳(日本)
  - 企業の決算速報の英訳(日本)
- 最近では無料で利用できる検索サービスがある
  - Yahoo!, Google, excite, OCNなどが提供



## まとめ

- 多言語情報アクセス技術は，母国語以外の情報へのアクセスを実現し，問題解決の可能性を向上させる重要な技術
- 言語横断情報検索は，ある言語による問合せで別の言語で書かれた情報を検索する技術
- 機械翻訳技術は，自然言語処理技術の集大成であり，まだ改良の余地は大きい