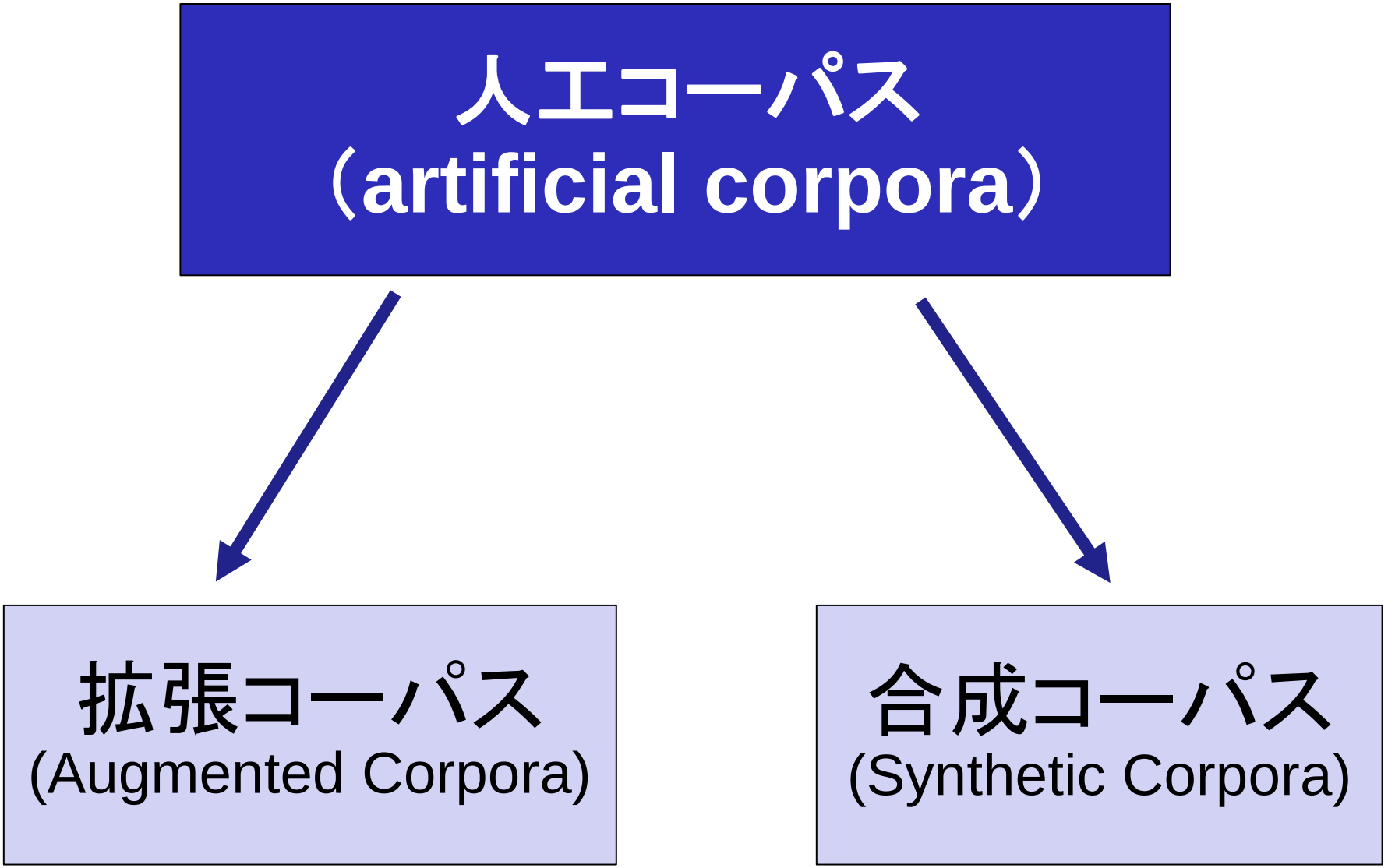
辞書データベースに基く
大規模合成コーパスの生成

春遍雀來（Jack Halpern, CEO
The CJK Dictionary Institute）
日中韓辭典研究所

AAMT 2023, Tokyo
AP虎ノ門, 2023年11月29日

日中韓辭典研究所

- 1993年設立、日本が拠点
- 日中韓諸語とアラビア語の計算言語学を専門とする
- 固有名詞と専門用語の大規模辞書データベースを開発・拡張


人工コーパス
（artificial corpora）
拡張コーパス
(Augmented Corpora)
合成コーパス
(Synthetic Corpora)

PASCとは?

- 日中韓諸語とアラビア語の包括的な語彙データベースに基づく
- PASCは、固有名詞等低リソース分野で直面するNLPアプリ開発の課題を解決する
- NMT、ASR、TTSを含む言語モデルの品質向上に貢献する

特徴

- **文アラインメント**：文、句、語彙、文法及び構文レベルでの完全なアラインメント
- **翻訳の精度**：原文に忠実、正確かつ自然な訳文
- **正確な音素表記**：非ラテン言語に対しほぼ100%正確な音素表記を付与
- **多言語対応**：単言語、二言語及び多言語に対応
- **完全な注釈**：原文と訳文に、部分的または完全な注釈を付与
- **形式の統一性**：厳密に言語規則に基づいて、完全な統一性を実現

西洋人名サンプル

ENGLISH	JAPANESE
My full name is [Michael Owen].	私の姓名は[オーウェン・マイケル]です。
[Michael] is my given name and [Owen] is my surname.	[マイケル]は私の名前で、[オーウェン]は私の苗字です。
I'm called [Michael Owen].	[オーウェン・マイケル]と言います。
Both [Michael] and [Owen] are personal names.	[オーウェン]と[マイケル]は両方とも人名です。
[Michael Owen] is my full name.	[オーウェン・マイケル]とは私のフルネームです。
[Michael Owen] is what's written on my ID.	旅券に記載されている姓名は[オーウェン・マイケル]です。
I've never heard of anyone called [Michael Owen].	[オーウェン・マイケル]と言う人のことを聞いたことがない。
I go by the name [Michael Owen].	[オーウェン・マイケル]と言う名前で呼ばれています。
Do you know of anyone who goes by the name of [Michael Owen]?	[オーウェン・マイケル]という人を知っていますか。

日本人名サンプル

JAPANESE	ENGLISH
私の姓名は[森隆大]です。	My full name is [Takahiro Mori].
[隆大]は私の名前で、[森]は私の苗字です。	[Takahiro] is my given name and [Mori] is my surname.
[森隆大]と言います。	I'm called [Takahiro Mori].
[森]と[隆大]は両方とも人名です。	Both [Takahiro] and [Mori] are personal names.
[森隆大]とは私のフルネームです。	[Takahiro Mori] is my full name.
旅券に記載されている姓名は[森隆大]です。	[Takahiro Mori] is what's written on my ID.
[森隆大]と言う人のことを聞いたことがない。	I've never heard of anyone called [Takahiro Mori].
[森隆大]と言う名前で呼ばれています。	I go by the name [Takahiro Mori].
[森隆大]という人を知っていますか。	Do you know of anyone who goes by the name of [Takahiro Mori]?

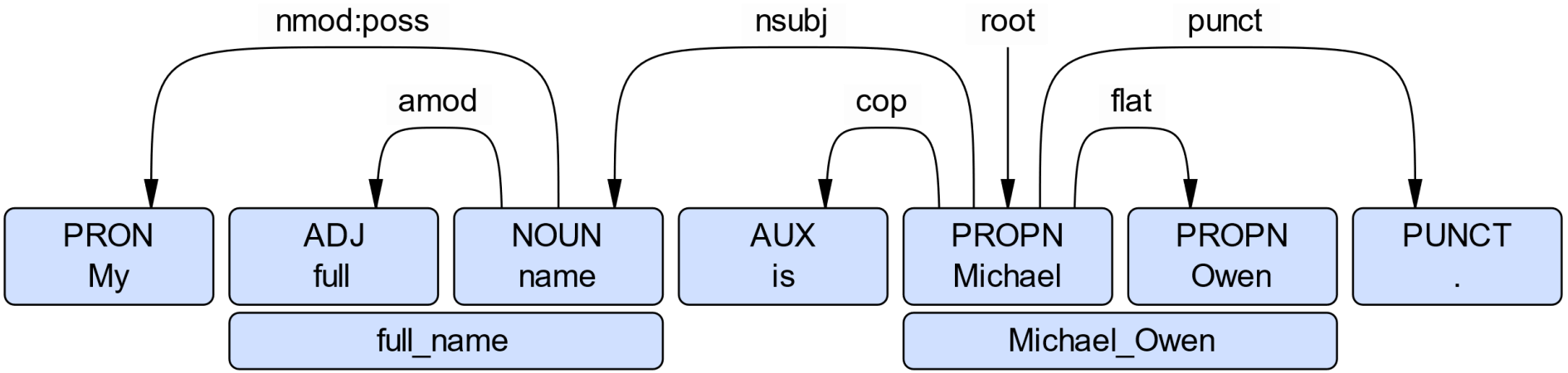
中国人名サンプル

CHINESE	ENGLISH
我的姓名是[张小东]。	My full name is [Xiaodong Zhang].
[小东]是我的名字，[张]是我的姓。	[Xiaodong] is my given name and [Zhang] is my surname.
我叫[张小东]。	I'm called [Xiaodong Zhang].
[小东]和[张]都是人名。	Both [Xiaodong] and [Zhang] are personal names.
[张小东]是我的姓名。	[Xiaodong Zhang] is my full name.
我的身份证上的姓名是[张小东]。	[Xiaodong Zhang] is what's written on my ID.
我从未听过叫[张小东]的人。	I've never heard of anyone called [Xiaodong Zhang].
我叫[张小东]。	I go by the name [Xiaodong Zhang].
你知道叫[张小东]的人吗？	Do you know of anyone who goes by the name of [Xiaodong Zhang]?

CoNLL-U解釈の例

ID	FORM	UPOSTAG	MISC
# sent_id = en-ja-0001-01 # text = My full name is Michael Owen. # text_ja = 私の姓名はオーウェン・マイケルです。			
1	My	PRON	Gloss=私の
2-3	full_name	_	Gloss=姓名
2	full	ADJ	_
3	name	NOUN	_
4	is	AUX	Gloss=は
5-6	Michael_Owen	_	Gloss=オーウェン・マイケル |NamedEntity=B-PER
5	Michael	PROPN	Gloss=マイケル|NamedEntity=B-PER:FN
6	Owen	PROPN	SpaceAfter=No|Gloss=オーウェン |NamedEntity=I-PER:LN
7	.	PUNCT	SpaceAfter=No

セマンティックツリー


nmod:poss
amod
nsubj
cop
root
punct
flat
PRON
My
ADJ
full
NOUN
name
AUX
is
PROPN
Michael
PROPN
Owen
PUNCT
.
full_name
Michael_Owen

Thank You

有難う御座います