Universal Dependencies プロジェクトと日本語チームの活動

はじめに

NINJAL 大学共同利用機関法人 人間文化研究機構 国立国語研究所
NATIONAL INSTITUTE FOR JAPANESE LANGUAGE AND LINGUISTICS

今から話す内容

「Universal Dependencies プロジェクト」

「言語における系統・変異・多様性とその数理」

Current UD Languages

- Afrikaans
- Ancient Greek
- Arabic
- Bambara
- Basque
- Belarusian
- Bulgarian
- Buryat
- Cantonese
- Catalan
- Chinese
- Coptic
- Croatian
- Czech
- Danish
- Dutch
- English
- Erzya
- Estonian
- Finnish
- French
- Galician
- German
- Gothic
- Greek
- Hebrew
- Hindi
- Hungarian
- Indonesian
- Irish
- Italian
- Japanese
- Kazakh
- Korean
- Kurmanji
- Latin
- Latvian
- Lithuanian
- Maltese
- Marathi
- North Sami
- Norwegian
- Old Church Slavonic
- Persian
- Polish
- Portuguese
- Romanian
- Russian
- Sanskrit
- Serbian
- Slovak
- Slovenian
- Slovenian
- Spanish
- Swedish
- Swedish Sign Language
- Tamil
- Telugu
- Thai
- Turkish
- Ukrainian
- Upper Sorbian
- Urdu
- Uyghur
- Vietnamese

Upcoming UD Languages

- Amharic
- Armenian
- Bengali
- Dargwa
- Faroese
- Georgian
- Kannada
- Kyrgyz
- Naija
- Old French
- Romansh
- Somali
- Sorani
- Yoruba

Universal Dependencies プロジェクト
http://universaldependencies.org/

- 自然言語処理の多言語化を目的としたプロジェクト
  - 多言語・言語横断的な依存構造アノテーション基準の策定
    - 言語横断的に一貫した文法的なアノテーション
  - 上記基準に基づく多言語の依存構造のタグ付きコーパスの開発
    - 特定の組織によらないオープンコミュニティ（200人以上）によるコーパス開発
    - 現在60言語100以上のTreebankが整備されている
      http://hdl.handle.net/11234/1-2515
    - 危機言語(Endangered)・ピジン／クレオール語・古語(Extinct)も対象
  - 上記コーパスに基づいて解析器を作る評価型WSの開催
    - CoNLL 2017/2018 Shared Task: Multilingual Parsing from Raw Text to UD
    - 参加者の出力結果も言語資源化
      http://hdl.handle.net/11234/1-2424

Universal Dependencies プロジェクト
http://universaldependencies.org/

- C. D. Manning があげる UD の6つの理念
  1. 個々の言語の言語学分析ができるものであるべき
  2. 言語ごとの比較をするのに適しているべき
  3. 人間が早く一貫性を保ってアノテーションできる構造であるべき
  4. コンピュータにとって高精度で解析できるものであるべき
  5. 言語の学習者・エンジニアも含めて、だれにとっても直感的な構造であるべき
  6. 関係抽出・機械翻訳など、後段の処理で使えるものであるべき

Universal Dependencies プロジェクト http://universaldependencies.org/

- 理念に沿うための基本的な方針
  - 依存構造(dependency structure)による表現
    - 簡潔で実用的な構造
    - 既存の多くのツリーバンクの活用
      - 言語固有の情報を失わないよう、言語ごとの拡張を許す
  - 語彙主義
    - 記法や音韻ではなく、文法上の単語(syntactic word)をアノテーションの基本単位とする
    - 単語に形態論情報をもたせる
    - 単語間に構文的な関係を持たせる
    - 存在しないものにはアノテーションしない
  - 再現性
    - 入力分から単語列の変換に透明性をもたせる

Universal Dependencies プロジェクト http://universaldependencies.org/

- 理念に沿うための基本的な方針
  - 依存構造(dependency structure)による表現
    - 簡潔で実用的な構造
    - 既存の多くのツリーバンクの活用
      - 言語固有の情報を失わないよう、言語ごとの拡張を許す
  - 語彙主義
    - 記法や音韻ではなく、文法上の単語(syntactic word)をアノテーションの基本単位とする
    - 単語に形態論情報をもたせる
    - 単語間に構文的な関係を持たせる
    - 存在しないものにはアノテーションしない
  - 再現性
    - 入力分から単語列の変換に透明性をもたせる

日本語で問題になる

Universal Dependencies Issue Tracker

各言語における問題が
他の言語にどのように波及するか
議論する場所

https://github.com/universaldependencies/docs/issues

Shared dependents and governors in coordination #524

Edit | New issue

Open | perrier54 opened this issue 9 days ago · 16 comments

perrier54 commented 9 days ago — Member

1. Basic dependencies.
   According to the annotation guide, shared dependents and governors are attached to the head of the first conjunct.
   This principle cannot apply when a dependent is not shared by the heads of the two conjuncts but by words more or less deeply embedded in the conjuncts. Consider the following example:
   Il pense obtenir et obtiendra sûrement son diplôme (word for word : *he thinks to get and will surely get his degree*)
   The word "*diplôme*" is an object shared by the verbs "*obtenir*" and "*obtiendra*". The problem is that "*obtenir*" is not the head of the first conjunct. I propose to attach the shared depend to the closest governor, "*obtiendra*" in this example. This proposal is consistent with the idea of surface dependencies.
   In the previous example, the closest governor is in the second conjunct, but if the shared dependent precedes the coordination, the closest governor is in the first conjunct, as in the following example:
   Le livre qu'il doit présenter mais connaît mal (the book he has to present but knows badly)

Assignees
No one—assign yourself

Labels
dependencies
enhancement
universal

Projects
None yet

Milestone
No milestone

Universal Dependencies アノテーションの作法

アノテーションの形式

CoNLL-U format: Tab separated

http://universaldependencies.org/format.html

sent_id = n01003013
text = Maybe the dress code was too stuffy.

1	Maybe	maybe	ADV	RB	_	7	advmod	_	_
2	the	the	DET	DT	Definite=Def|PronType=Art	4	det	_	_
3	dress	dress	NOUN	NN	Number=Sing	4	compound	_	_
4	code	code	NOUN	NN	Number=Sing	7	nsubj	_	_
5	was	be	AUX	VBD	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	7	cop	_	_
6	too	too	ADV	RB	_	7	advmod	_	_
7	stuffy	stuffy	ADJ	JJ	Degree=Pos	0	root	_	SpaceAfter=No
8	.	.	PUNCT	.	_	7	punct	_	_

アノテーションの形式

CoNLL-U format: Tab separated

1. ID: 単語のインデックス: 1-origin
2. FORM: 表層形
3. LEMMA: 基本形
4. UPOSTAG: 品詞
   Universal part-of-speech tag.
5. XPOSTAG: 言語依存の品詞
6. FEATS: 形態論情報
   Universal feature inventory
7. HEAD: 係り先（Root は 0）
8. DEPREL: 係り受け関係ラベル
   Universal dependency relation
9. DEPS: 二次係り受け
10. MISC: その他

Google Universal POS [Petrov+ 2012]

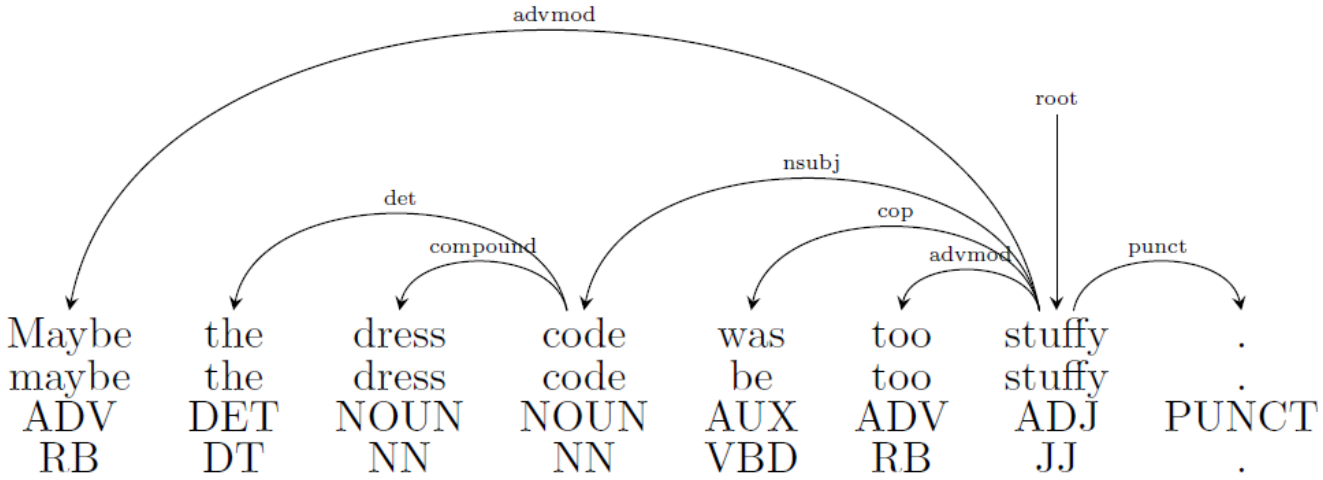
Interest interlingua [Zeman 2008]
HamleDT (Harmonized Multi-Language Dependency Treebank) [Zeman+ 2014]

Stanford Dependencies
[de Marneffe + 2006, 2008]
Stanford UD [de Marneffe+ 2014]

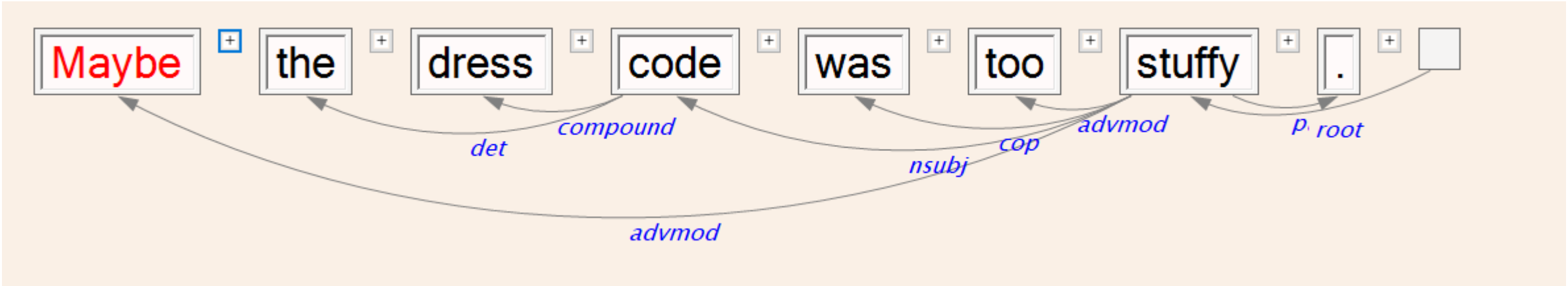
可視化

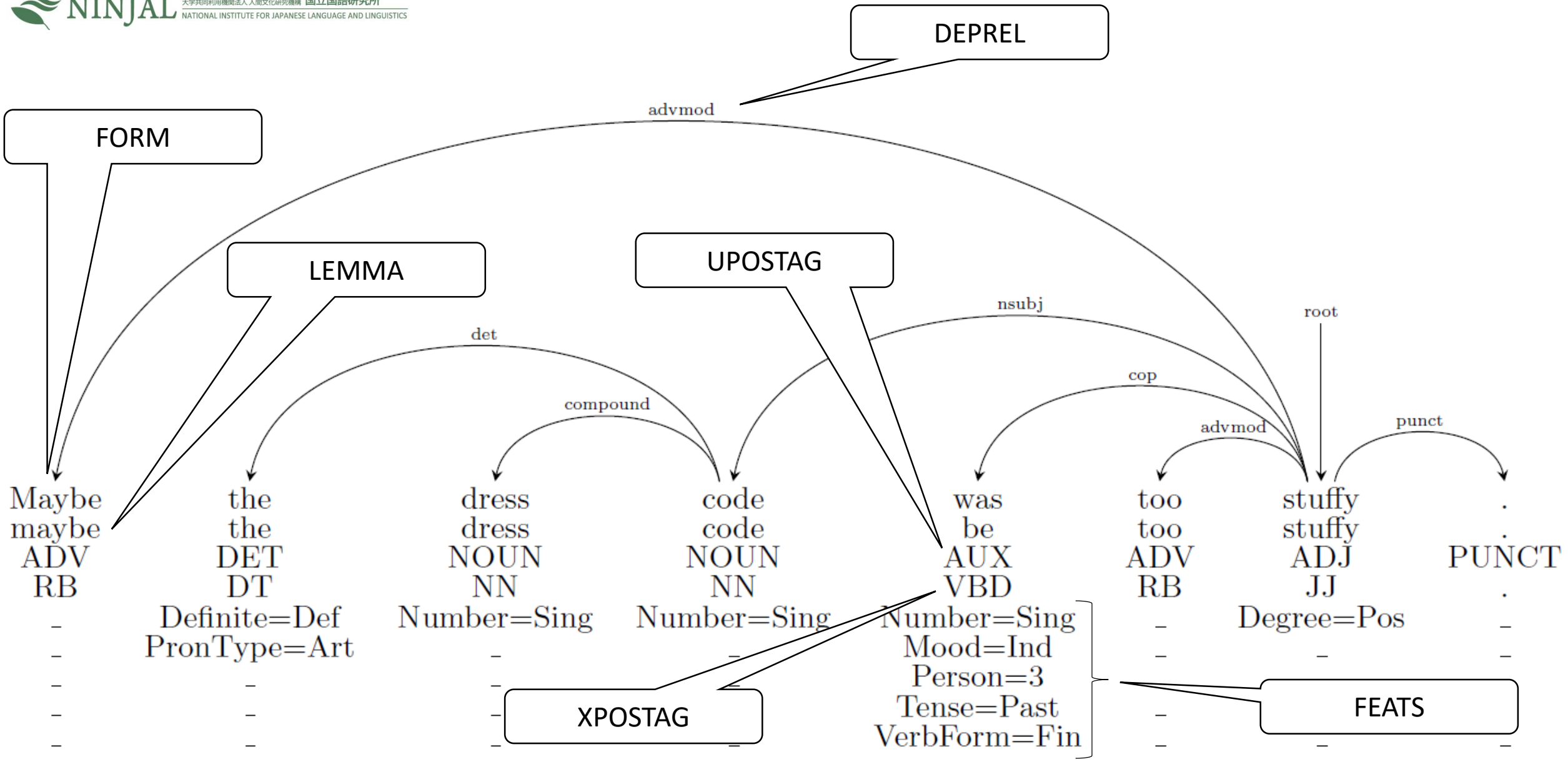
http://universaldependencies.org/tools.html

LaTeX: tikz-dependency



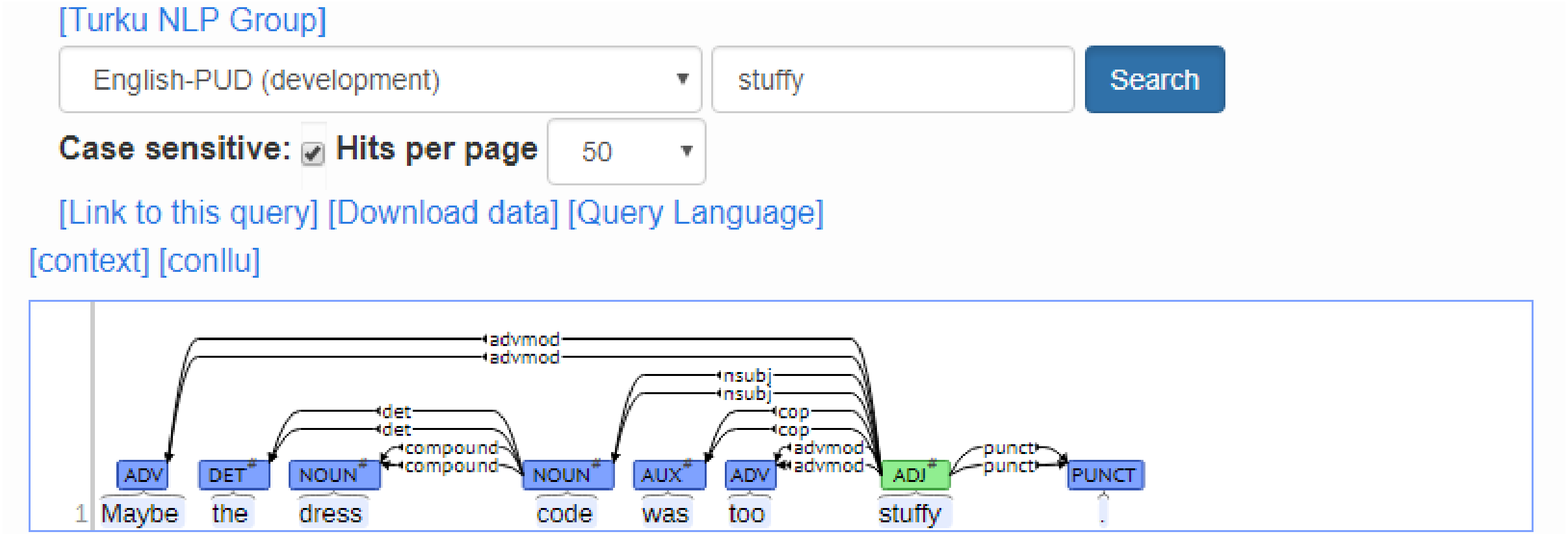
Viewer/Annotation Tool: ChaKi.NET [Asahara 2016]




DEPREL
FORM
LEMMA
UPOSTAG
XPOSTAG
FEATS
advmod
nsubj
root
det
cop
compound
advmod
punct
Maybe maybe ADV RB _ _ _ _ _
the the DET DT Definite=Def PronType=Art _ _ _
dress dress NOUN NN Number=Sing _ _ _ _
code code NOUN NN Number=Sing _ _ _ _
was be AUX VBD Number=Sing Mood=Ind Person=3 Tense=Past VerbForm=Fin _ _
too too ADV RB _ _ _
stuffy stuffy ADJ JJ Degree=Pos _ _
. . PUNCT . _ _ _

とりあえず Tree をみたい方

- http://bionlp-www.utu.fi/dep_search/



とりあえず Parsing がしたい方

- CoNLL-2018 Shared Task (1月26日 Registration open!)
  http://universaldependencies.org/conll18/

  訓練データ/開発データ/評価データ
  ＋生データ
  ＋並行コーパスデータ

  ＋オープントラックは以下も利用してよい
  - Word Atlas of Language Structures (WALS, http://wals.info/)
  - Wikipedia dumps (https://dumps.wikimedia.org/backup-index-bydb.html)
    - Word vectors for 90 languages trained on Wikipedia have been released by Facebook
  - WMT 2016 parallel and monolingual data (http://www.statmt.org/wmt16/translation-task.html)
  - Morphological transducers in Apertium and in Giellatekno
  - Unimorph (https://unimorph.github.io/)

アノテーションの作成

基本的には既存の TreeBank からの Conversion が多い

- 句構造木からの組み換え
  - 主辞規則
- 依存構造木からの組み換え
  - 単位の変換
  - 依存構造木の向きの変換
    - 自立語主辞 (UD) vs 付属語主辞 (HamleDT)

- 最初からアノテーション
  - 危機言語・古語などに対しては、1000文単位で付与することが多い

アノテーションする情報

1. ID: 単語のインデックス: 1-origin
2. **FORM: 表層形**
3. **LEMMA: 基本形**
4. **UPOSTAG: 品詞**
   Universal part-of-speech tag.
5. XPOSTAG: 言語依存の品詞
6. **FEATS: 形態論情報**
   Universal feature inventory
7. **HEAD: 係り先（Root は 0）**
8. **DEPREL: 係り受け関係ラベル**
   Universal dependency relation
9. DEPS: 二次係り受け
10. MISC: その他

FORM 表層形

- 単位認定が重要
  - Syntactic word を言語横断的に規定する必要がある
  - かならずしも分かち書きに左右されない
  - 複単語表現の取り扱い

Manning の6つの理念　再掲

1. 個々の言語の言語学分析ができるものであるべき
2. 言語ごとの比較をするのに適しているべき
3. 人間が早く一貫性を保ってアノテーションできる構造であるべき
4. コンピュータにとって高精度で解析できるものであるべき
5. 言語の学習者・エンジニアも含めて、だれにとっても直感的な構造であるべき
6. 関係抽出・機械翻訳など、後段の処理で使えるものであるべき

LEMMA 基本形

- 何に対する基本形なのか

語彙素（LEXEME)の定義

品詞が変わることを許す
統語属性が変わることを許す
発音が変わることを許す
実体が変わることを許す

PTB

- “was”
  - “is” or “be”
- “could”
  - “could” or “can”

BCCWJ

- “松本” “松元”
  - “マツモト”
- “日本”
  - “ニホン” or “ニッポン”
- “日本橋”
  - “ニホンバシ” or “ニッポンバシ”

UPOSTAG

- Universal POS (Google Universal POS から改変)
  17種類

NOUN	名詞	PRON	代名詞	PUNCT	句読点
PROPN	固有名詞	NUM	数詞	SYM	記号
VERB	動詞	DET	冠詞	X	その他
ADJ	形容詞	AUX	助動詞		
ADV	副詞	ADP	接置詞		
INTJ	間投詞	PART	接辞		
		CCONJ	等位接続詞		
		SCONJ	従属接続詞		

FEATS: UD Features

Lexical features

PronType	代名詞型
NumType	数値型
Poss	所有（性・数）
Reflex	再帰
Foreign	外国語
Abbr	省略形

Inflectional features

Nominal (approximate)

Gender	性
Animacy	有生性
Number	数
Case	格
Definite	定性
Degree	級

Verbal (approximate)

VerbForm	形
Mood	法
Tense	時制
Aspect	相
Voice	態
Evident	証拠性
Polarity	極性
Person	人称
Polite	敬体

FEATS: UD Features (言語依存属性)

- AbsErgDatNumber 数の一致
- AbsErgDatPerson 人称の一致
- AbsErgDatPolite 敬体の一致
- AdpType 接置詞の一致(前・中・後・母音化)
- AdvType 副詞型（様態・場所・時間...）
- Clusivity 包括・除外
- ConjType 接続型（比較・数学演算子）
- Echo 畳語
- ErgDatGender 性の一致
- Hyph ハイフン付き
- NameType 固有名型（地名・人名・組織名）
- NounType 助数詞など
- NumForm 数形式(word, digit, roman)
- NumType 数型(基数・序数・範囲・集合..)
- NumValue 数値(1, 2, 3 or 4)(5未満)
- PartType 接辞型(法, 強調, 不定, 分離動詞)
- PossGender 所有者の性
- PossNumber 所有者の数
- PossPerson 所有者の認証
- PossedNumber 所有されるものの数
- Prefix 接頭表現
- PrepCase 格が前置詞によって与えられるか
- PunctSide 句読点位置 (initial, final)
- PunctType 句読点型（句点・読点・コロン..）
- Style 使用域(詩・口語・文語...)
- Subcat 自他
- Typo 誤記
- VerbType 動詞型（助動・繋辞・法・軽）

HEAD, DEPREL = UD Relations (37 labels)

	Nominals	Clauses	Modifier words	Function words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case

Coordination	MWE	Loose	Special	Other
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

日本語対応における論点

- 境界認定
  - 何を単語（Syntactic word）とするのか
  - 節と句の区別
- ラベル認定
  - 品詞認定（語彙主義 vs 用法主義）
  - 表層的関係と意味的関係
- 構造認定
  - 格交替（受身・使役・与格交替・場所格交替...）
  - 並列構造の扱い
  - dislocated（主語優勢言語と主題優勢言語）
  - 省略要素に対する係り受け

Universal Dependencies 日本語の言語資源

Universal Dependencies 日本語チーム

- 植松すみれ　（NII）
- 大村舞　（国語研）
- 金山博　（日本IBM）
- 田中貴秋　（NTT CS研）
- 松本裕治　（NAIST）
- 宮尾祐介　（NII）
- 村脇有吾　（京都大学）
- 森信介　（京都大学）
- 浅原正幸　（国語研）

基本的に
手弁当

（2018年2月現在）

Universal Dependencies for Japanese

- UD Japanese KTC [Tanaka+ 2016]
  - 句構造木 (Kaede treebank [Tanaka+ 2013])から
- UD Japanese BCCWJ [大村+ 2017]
  - 文節係り受け (BCCWJ-DepPara [Asahara+ 2016]) から
- UD Japanese Modern [Omura+ 2017]
  - 文節係り受け (BCCWJ-DepPara 互換 [浅原・高橋 2016]) から
- UD Japanese (無印)
  - 文節係り受けを再タグ付け中
- UD Japanese PUD (並行コーパス)
  - 文節係り受けを再タグ付け中

日本語における Syntactic Word

現状は
国語研短単位(SUW)

- 何を単位にするか
  - 国語研短単位 (SUW)
  - 国語研長単位 (LUW)
  - 国語研文節
  - 京大コーパス JUMAN 形態素単位
  - 京大コーパス 文節単位

Manning の6つの理念　再掲
1. 個々の言語の言語学分析ができるものであるべき
2. 言語ごとの比較をするのに適しているべき
3. 人間が早く一貫性を保ってアノテーションできる構造であるべき
4. コンピュータにとって高精度で解析できるものであるべき
5. 言語の学習者・エンジニアも含めて、だれにとっても直感的な構造であるべき
6. 関係抽出・機械翻訳など、後段の処理で使えるものであるべき

- Spicks & Specks (Quasi-Blog by Greg 18th Sept. 2016)

http://www.cjvlang.com/Spicks/udjapanese.html

  - “Thoughts on the Universal Dependencies proposal for Japanese”
    The problem of the word as a linguistic unit

日本語における品詞

現状は
国語研短単位(SUW)
に対する語彙主義

- 語彙主義
  - 可能性に基づく品詞体系
    - UniDic POS for 国語研短単位
      例）「名詞-普通名詞-サ変形状詞可能」
    - IPADIC

Manning の6つの理念　再掲
1. 個々の言語の言語学分析ができるものであるべき
2. 言語ごとの比較をするのに適しているべき
3. 人間が早く一貫性を保ってアノテーションできる構造であるべき
4. コンピュータにとって高精度で解析できるものであるべき
5. 言語の学習者・エンジニアも含めて、だれにとっても直感的な構造であるべき
6. 関係抽出・機械翻訳など、後段の処理で使えるものであるべき

- 用法主義
  - 文脈に基づく品詞体系
    - UniDic POS for 国語研長単位
      BCCWJ の長単位データには「用法」の情報がついている
      例）「名詞-普通名詞-サ変形状詞可能」＋「する」→「動詞」

用法主義的な情報は
係り受けラベルで対処

日本語における格関係

- 表層格
  - 京都大学テキストコーパス
  - NAIST テキストコーパス (BCCWJ-PAS)

  ガ・ヲ・ニは nsubj/csubj, dobj, iobj なのか？
  格交替はどう扱うのか？

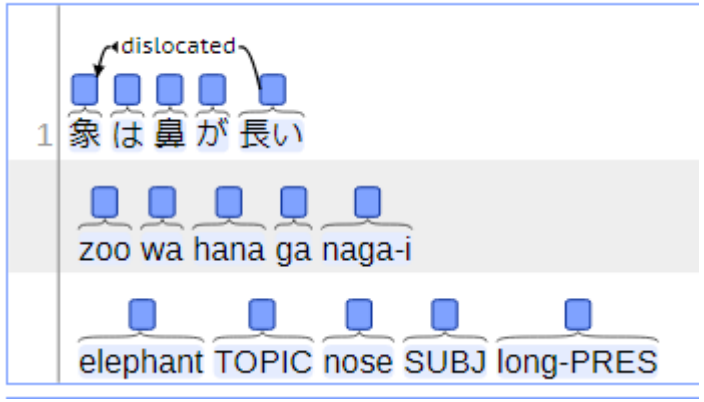
- 深層格
  - 動詞項構造シソーラス
  - 日本語フレームネット

日本語における格関係

- 「は」と「が」

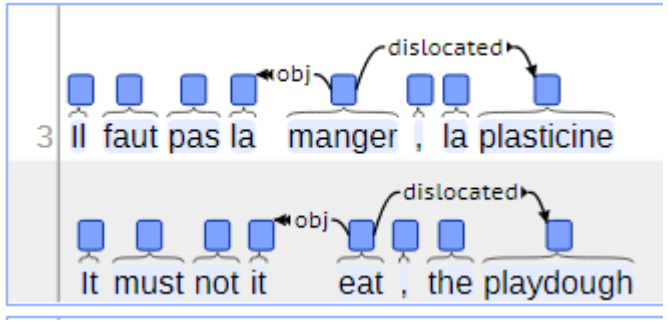
dislocated and nsubj/csubj

主語優勢言語と主題優勢言語



dislocated（転移）

「分裂文」(cleft)との混同



日本語における節

- 句と節（nsubj/名詞主語 or csubj/節主語）

		声-が	思い出さ-れる
	野太い	声-が	思い出さ-れる
	野太かっ-た	声-が	思い出さ-れる
部長-の	野太い	声-が	思い出さ-れる
部長-の	野太かっ-た	声-が	思い出さ-れる

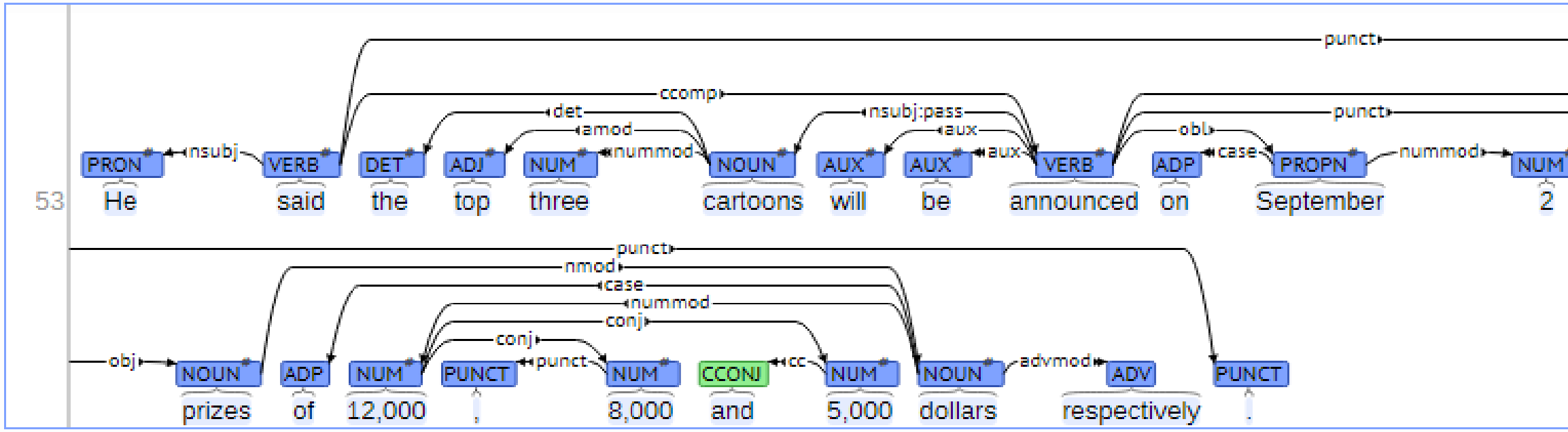
どこかで nsubj と csubj の線引きを行う必要がある

日本語における並列構造

- 基本的に並列構造と係り受けは親和性が低い
  - 入れ子
  - 統語範疇を共有しない並列
  - 係り元を共有する並列
  - 非構成素並列
  - 共有要素省略
  - 係り元の共有

UD における並列

- 左側構成素が主辞
- 接続助詞・並列内部の読点は直後の要素が主辞



Universal Dependencies
その他のアジア言語

- UD for Ainu [Senuma+ 2018]
- UD Korean
- UD Chinese
- UD Cantonese
- UD Vietnamese
- UD Thai
- UD Buryat
- UD Kazakh
- UD Hindi
- UD Urdu
- UD Sanskrit
- UD Tamil

系列や木の評価

「Universal Dependencies プロジェクト」

「言語における系統・変異・多様性とその数理」

系列のはる距離（空間）

“文書間類似度について”[浅原・加藤 2016]

https://doi.org/10.5715/jnlp.23.463

距離・類似度・系列カーネル・順序尺度・相関係数を整理

距離空間の再考

- テキストから得られる特徴量がはる距離空間
  - 文字列・単語列　　系列に対する距離空間
  - 係り受け木・句構造木　　木に対する距離空間
  - 一般のアノテーション　　有向非循環グラフに対する距離空間

系列に対する距離空間と関連尺度

- テキストから得られる特徴量がはる空間
  - 類似度（スコア）　　[0,1]↑
  - 距離　　[0,∞)↓
  - カーネル　　[0,∞)↑
  - 相関係数　　[-1,1]↑
  - 言語処理で用いられる評価尺度
    - 機械翻訳・文書要約などで用いられる評価尺度

“文書間類似度について”[浅原・加藤 2016]
https://doi.org/10.5715/jnlp.23.463

系列に対する距離空間と関連尺度
n-gram 系/p-mer 系

	スコア [0,1]↑	距離 [0,∞)↓	カーネル [0,∞)↑	相関係数[-1,1]↑	指標
部分文字列系	K all_str		(加重)全部分文字列		IMPACT
(連続, n-gram)	K n-gram		n-スペクトラム		ROUGE-N
					BLEU
					LRscore
最長一致部分文字列長	LCStr				
部分列系	K all_seq		(加重) 全部分列		
(非連続, p-mer)	K p-mer		p-mer 部分列		ROUGE-S(U)
	K w_p-mer		加重 p-mer 部分列		ESK
(加重)最長一致部分列長	w_LCS				ROUGE-W
最長一致部分列長	LCS	d Ulam			ROUGE-L

“文書間類似度について”[浅原・加藤 2016]
https://doi.org/10.5715/jnlp.23.463

系列に対する距離空間と関連尺度
順序尺度系

	スコア [0,1]↑	距離 [0,∞)↓	カーネル [0,∞)↑	相関係数[-1,1]↑	指標
順序（ベクトル）系	||Rank||Θ	d||Rank||Θ			
	Footrule	d Footrule (Θ=1)		Spearman’s ρ	
	Spearman	d Spearman (Θ=2)^2			
	Hamming	d Hamming			
順序（編集）系	Kendall	d Kendall		Kendall’s τ	RIBES, LRScore
		d Cayley			
(=最長一致部分列長)	LCS	d Ulam			ROUGE-L

“文書間類似度について”[浅原・加藤 2016]
https://doi.org/10.5715/jnlp.23.463

系列に対する距離空間と関連尺度
順序尺度系(編集系)

隣接要素交換
昔、係り受けの non-projective 対応で
この swap を使ったものがあったような

Kendall

$$d_{\mathrm{Kendall}}((1,4,3,2),(1,2,3,4)) = 3$$

$$\begin{pmatrix} 1 & 4 & \underline{3} & \underline{2} \\ 1 & 4 & \underline{2} & \underline{3} \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & \underline{4} & \underline{2} & 3 \\ 1 & \underline{2} & \underline{4} & 3 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 2 & \underline{4} & \underline{3} \\ 1 & 2 & \underline{3} & \underline{4} \end{pmatrix}$$

$$d_{\mathrm{Kendall}}((2,3,1,4),(1,2,3,4)) = 2$$

$$\begin{pmatrix} 2 & \underline{3} & \underline{1} & 4 \\ 2 & \underline{1} & \underline{3} & 4 \end{pmatrix} \Rightarrow \begin{pmatrix} \underline{2} & \underline{1} & 3 & 4 \\ \underline{1} & \underline{2} & 3 & 4 \end{pmatrix}$$

“文書間類似度について”[浅原・加藤 2016]
https://doi.org/10.5715/jnlp.23.463

系列に対する距離空間と関連尺度
順序尺度系(編集系)

Caylay

$$d_{\mathrm{Caylay}}((1,4,3,2),(1,2,3,4)) = 1$$

$$\begin{pmatrix} 1 & \underline{4} & 3 & \underline{2} \\ 1 & \underline{2} & 3 & \underline{4} \end{pmatrix}$$

$$d_{\mathrm{Caylay}}((2,3,1,4),(1,2,3,4)) = 2$$

$$\begin{pmatrix} \underline{2} & 3 & \underline{1} & 4 \\ \underline{1} & 3 & \underline{2} & 4 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & \underline{3} & \underline{2} & 4 \\ 1 & \underline{2} & \underline{3} & 4 \end{pmatrix}$$

任意要素交換

ケーリーグラフ
（頂点推移グラフ）の
最短経路

“文書間類似度について”[浅原・加藤 2016]
https://doi.org/10.5715/jnlp.23.463

系列に対する距離空間と関連尺度
順序尺度系(編集系)

いわゆる、本棚入れ替え

最長一致部分列長と等価

Ulam

$$d_{\mathrm{Ulam}}((1,4,3,2),(1,2,3,4)) = 2$$

$$\begin{pmatrix} 1 & \boxed{4} & \boxed{3} & \underline{2} \\ 1 & \underline{2} & \boxed{4} & \boxed{3} \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 2 & \boxed{4} & \underline{3} \\ 1 & 2 & \underline{3} & \boxed{4} \end{pmatrix}$$

$$d_{\mathrm{Ulam}}((2,3,1,4),(1,2,3,4)) = 1$$

$$\begin{pmatrix} \boxed{2} & \boxed{3} & \underline{1} & 4 \\ \underline{1} & \boxed{2} & \boxed{3} & 4 \end{pmatrix}$$

誤り訂正符号にも（ビット飛びに頑健）
Farnoud, F., Skacheck, V. and Milenkovic, O., Errorcorrection in ash memories via codes in the Ulam metric, IEEE Transaction on Information Theory, 59 [5] (2013), 3003-3020.

最長一致部分列長（Longest Common Sequence）と親和性があり
（最長一致部分文字列(Longest Common String)ではないことに注意）

“文書間類似度について”[浅原・加藤 2016]
https://doi.org/10.5715/jnlp.23.463

木のはる距離

“木の半正定値カーネル” [申 2009]
Hausseler の畳み込みカーネルの分類・拡張 [Haussler 1999]

- 解析木カーネル [Collins 2001] 飽和・連続
- 弾性木カーネル [Kashima 2002] 非飽和・非連続
- 依存性カーネル [Zelenko 2003] 木上の経路
- 3-mer 連続部分木カーネル [Hizukuri 2005] サイズ3の連続部分木
- 畳み込み木カーネル [Suzuki 2005] 根以外の頂点削除
- 編集コスト分布カーネル [Kuboyama 2006] 編集コストのトレース
- q-gram 木カーネル [Kuboyama 2007] 長さ q の木構造の経路
- 文脈依存解析木カーネル [Zhou 2007] 木＋ルート経路の合同

ふるきよき黒田標準形

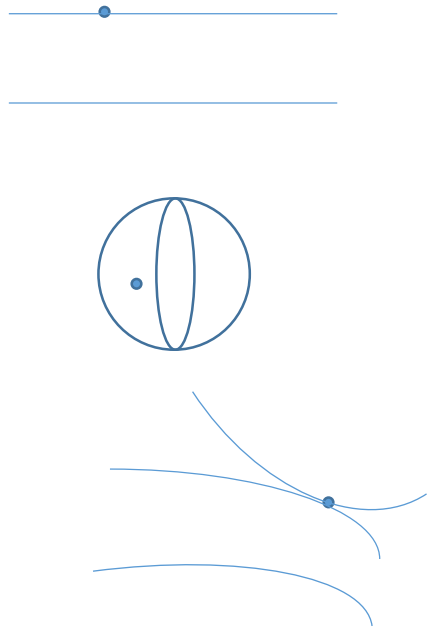
埋め込み

- 単語埋め込み（word2vec, glove）[Mikolov+ 2013]
  - 単語の分散表現を数百次元で
- ポワンカレ埋め込み [Nickel+ 2017]
  - 双曲空間への埋め込み
  - 階層構造の取得
    - たぶん係り受け木よりも、句構造木（非終端記号）のほうが親和性がある

双曲幾何

「1本の直線」 l とその直線上にない「1つの点」A とが与えられた際に、A を通る l と交わらない直線は何本引けるか?

- ユークリッド空間 (Euclidean Space)
  - 1本 (平行線)
- 球面空間 (Spherical Space)
  - 0本
- 双曲空間 (Hyperbolic Space)
  - ∞本



カギ針編みと双曲幾何

珊瑚の形（the frilly crenulated forms）は双曲幾何（Hyporbolic geometry）の具体例

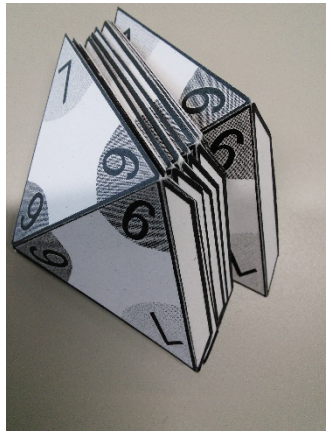
フォンノイマン型コンピュータでモデル化できず 3Dプリンタでも作れないが、「カギ針編み」では作れる

- Daina Taimina, (2009) ‘Crocheting Adventures with Hyperbolic Planes’ CRC Press.
- Sarah-Marie Belcastro and Carolyn Yackel (Eds), (2007) ‘Making Mathmatics with Needlework: Ten papers and Ten Projects’ CRC Press.

双曲面の可視化

ハイプレイン

- 多面体による模式化
  - 曲率を頂点の角度として表現



- 阿原 一志，(2008) ‘ハイプレイン-のりとはさみでつくる双曲平面’日本評論社.

ファーガソン毛布

- 双曲空間上の直角正五角形
  - 曲率を広範囲に分布させる

http://helasculpt.com/

双曲空間に対する計量とテキスト評価のための計量

カギ針編みの目数

双曲空間

- 距離
  - 有向非巡回グラフに相当

→DAG カーネルのバリエーション

- 等長変換と曲率
  - 曲線に対する曲率と捩率
    ＝平面とのちがいの評価

テキストに対する計量

- 距離＝内容評価（論理）
  - テキストの1次元尺度への写像
    - 正規化類似度スコア [0,1]↑
    - 距離[0,∞)↓

- 曲率＝表現評価
  - 等距離にある無数の線のちがいの評価
  - 同一の意味を表現する多様な表現の評価

カギ針編みの目数の増分

おわりに

おわりに

- Universal Dependencies の紹介
  - 危機言語も含めた言語横断的なアノテーション

イベント（国内）：

- 2018/06/16（土） Universal Dependencies 公開研究会（at 京都）
  節関連：　有田節子先生（立命館大学）
  並列構造関連：　中俣尚己先生（京都教育大）
  UD メンバーから数名

イベント（国外）：

- 2018/Oct-Nov　CoNLL-2018 Shared Task (at Brussels)
- 2nd Workshop on Universal Dependencies

機構間連携プロジェクト　採択3件中の以下の2件

- 「言語における系統・構造・変異とその数理」　（持橋P）
  今日のシンポジウム

- 「テキストを刺激とした視線計測データ収集とその利用
  ---人の文処理機構の解明と工学応用---」　（相澤P）

このスライドにない「おはなし」のつづきは以下で

2018/03/03（土）シンポジウム
「日本語学習者はどのように文章を理解しているのか
---目の動きから見えてくるもの---」
http://www.ninjal.ac.jp/event/specialists/project-meeting/m-2017/20180303-sympo/

どうもありがとうございました

2018/03/03（土）　シンポジウム（at 国語研）
「日本語学習者はどのように文章を理解しているのか
---目の動きから見えてくるもの---」
http://www.ninjal.ac.jp/event/specialists/project-meeting/m-2017/20180303-sympo/

2018/06/16（土）　Universal Dependencies 公開研究会（at 京都）
節関連：　有田節子先生（立命館大学）
並列構造関連：　中俣尚己先生（京都教育大）
UD メンバーから数名