

# 統語解析データセット・モデルの公開

## UD Japanese の発展

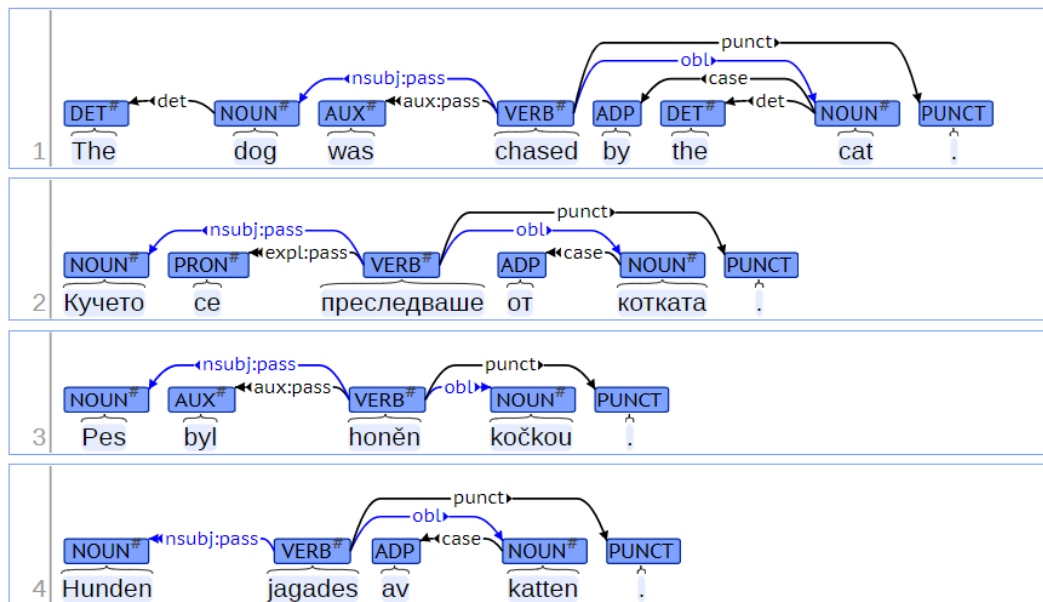
浅原正幸（国語研）

- Universal Dependencies とは「依存構造・係り受けアノテーション基準の国際標準化」  
分かち書き・品詞 (UPOS)・形態論情報・係り受け構造・係り受けラベル

UD (version 2) のアノテーション例

English, Bulgarian, Czech and Swedish の対訳

【<https://universaldependencies.org/introduction.html> より】



UD プロジェクト全体の目標

- 個々の言語の言語学的分析ができるものであるべき
- 言語ごとの比較をするのに適しているべき
- 人間が速く一貫性を保ってアノテーション出来る構造であるべき
- 言語の学習者やエンジニアを含めて、誰にとっても直感的な構造であるべき
- 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
- 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

UD を前に進めるために by C. D. Manning (2021/12/21 8:33)

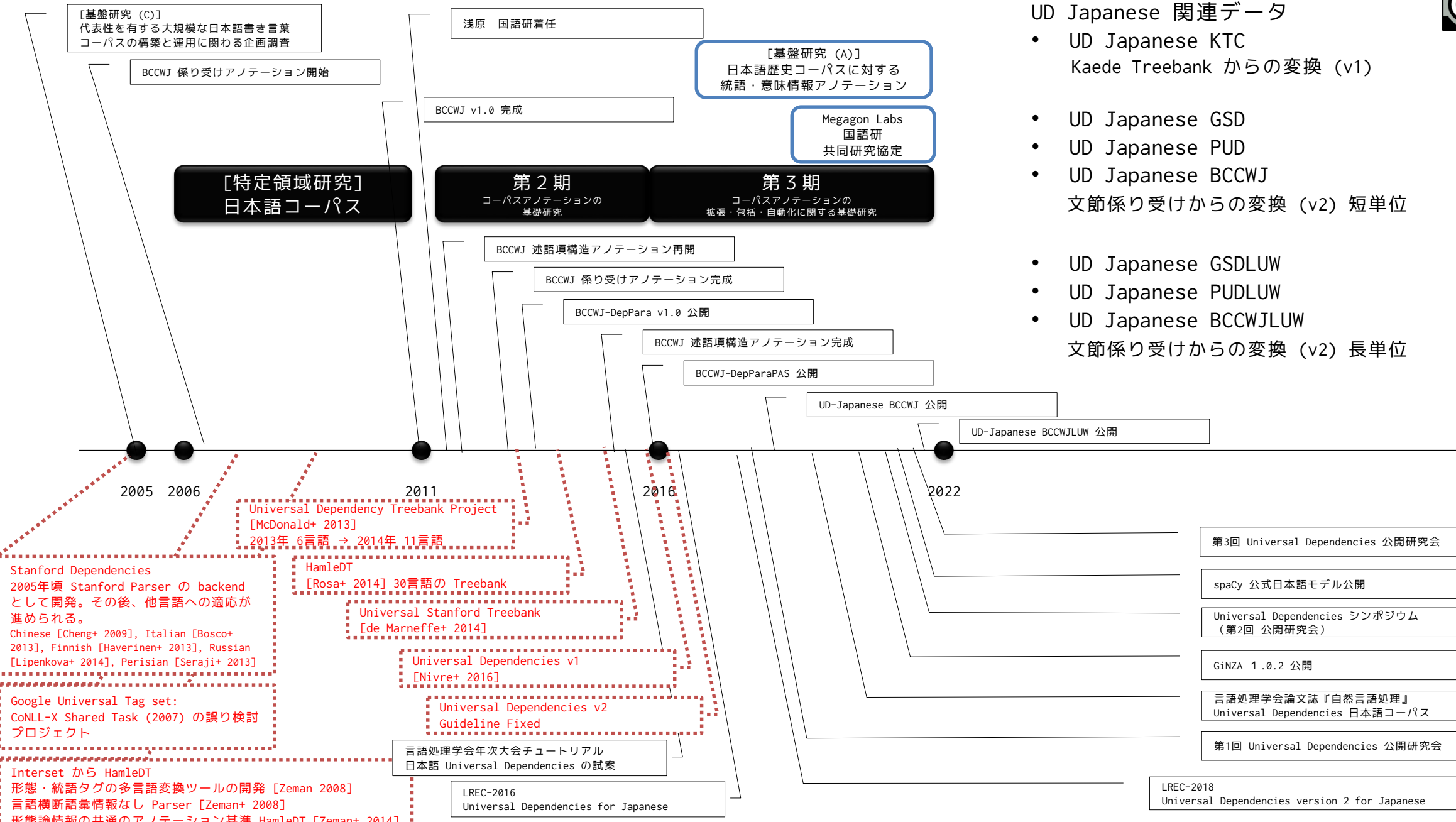
- 2023/05/01 まで、UD treebank に残すものはアクションが必要
- UD に残すために古い treebank を改善する
- Treebank の最小構成を 20文・100単語以上にする
- UD v2 のアノテーション基準の明確化・修正を順次実施

UD Japanese の立場

- 文節係り受けからの自動変換によるデータ整備
- 分かち書き単位について、短単位・長単位の2つを提案

### UD Japanese 関連データ

- UD Japanese KTC  
Kaede Treebank からの変換 (v1)
- UD Japanese GSD
- UD Japanese PUD
- UD Japanese BCCWJ  
文節係り受けからの変換 (v2) 短単位
- UD Japanese GSDLUW
- UD Japanese PUDLUW
- UD Japanese BCCWJLUW  
文節係り受けからの変換 (v2) 長単位



**[特定領域研究] 日本語コーパス**

**Stanford Dependencies**  
2005年頃 Stanford Parser の backend として開発。その後、他言語への適応が進められる。  
Chinese [Cheng+ 2009], Italian [Bosco+ 2013], Finnish [Haverinen+ 2013], Russian [Lipenkova+ 2014], Perisian [Seraji+ 2013]

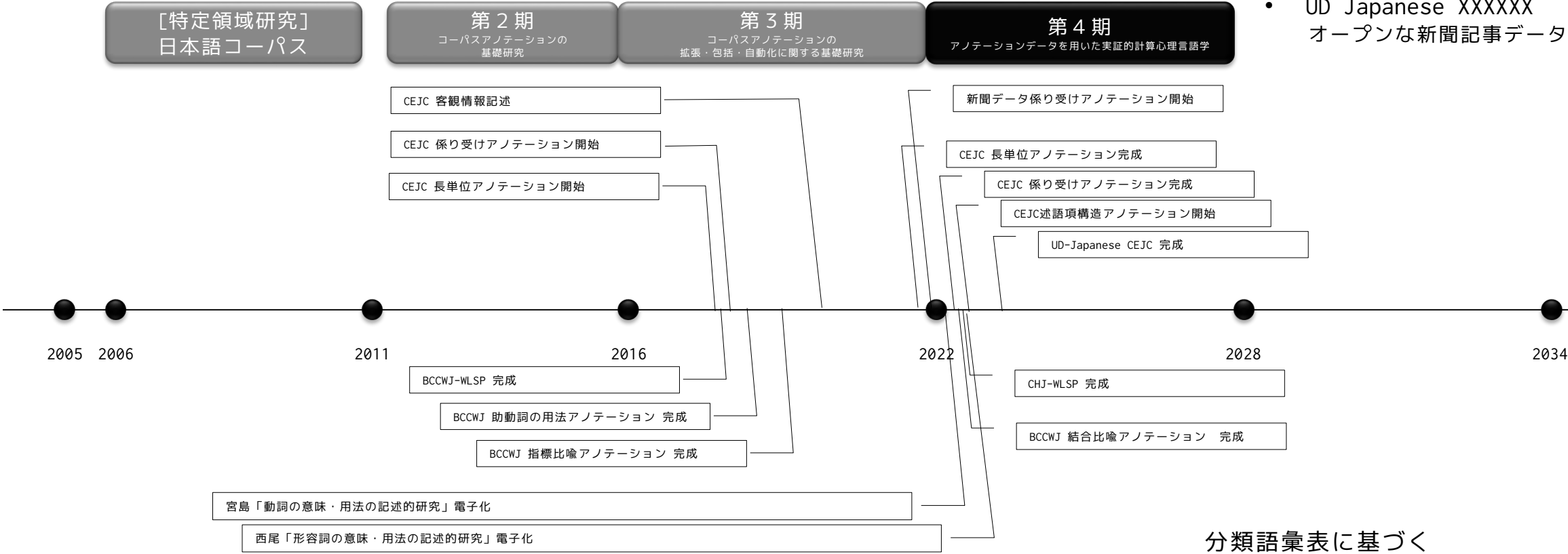
**Google Universal Tag set:**  
CoNLL-X Shared Task (2007) の誤り検討プロジェクト

**Intersect から HamleDT**  
形態・統語タグの多言語変換ツールの開発 [Zeman 2008]  
言語横断語彙情報なし Parser [Zeman+ 2008]  
形態論情報の共通のアノテーション基準 HamleDT [Zeman+ 2014]

2021/03/18

UD Japanese 関連データ

- UD Japanese CEJC 話し言葉データの構築
- UD Japanese XXXXXX オープンな新聞記事データ



分類語彙表に基づく  
結合価データの拡充