分析用データベースの設計と活用

国立国語研究所 プロジェクト非常勤研究員
井上 雄太（いのうえ ゆうた）

NINJAL
National Institute for Japanese Language and Linguistics

もくじ

1. 分析の対象

2. データベースの全体像

3. 解析結果の活用

１．分析の対象

今回使用したデータ

- クラウドワークス社のデータベース

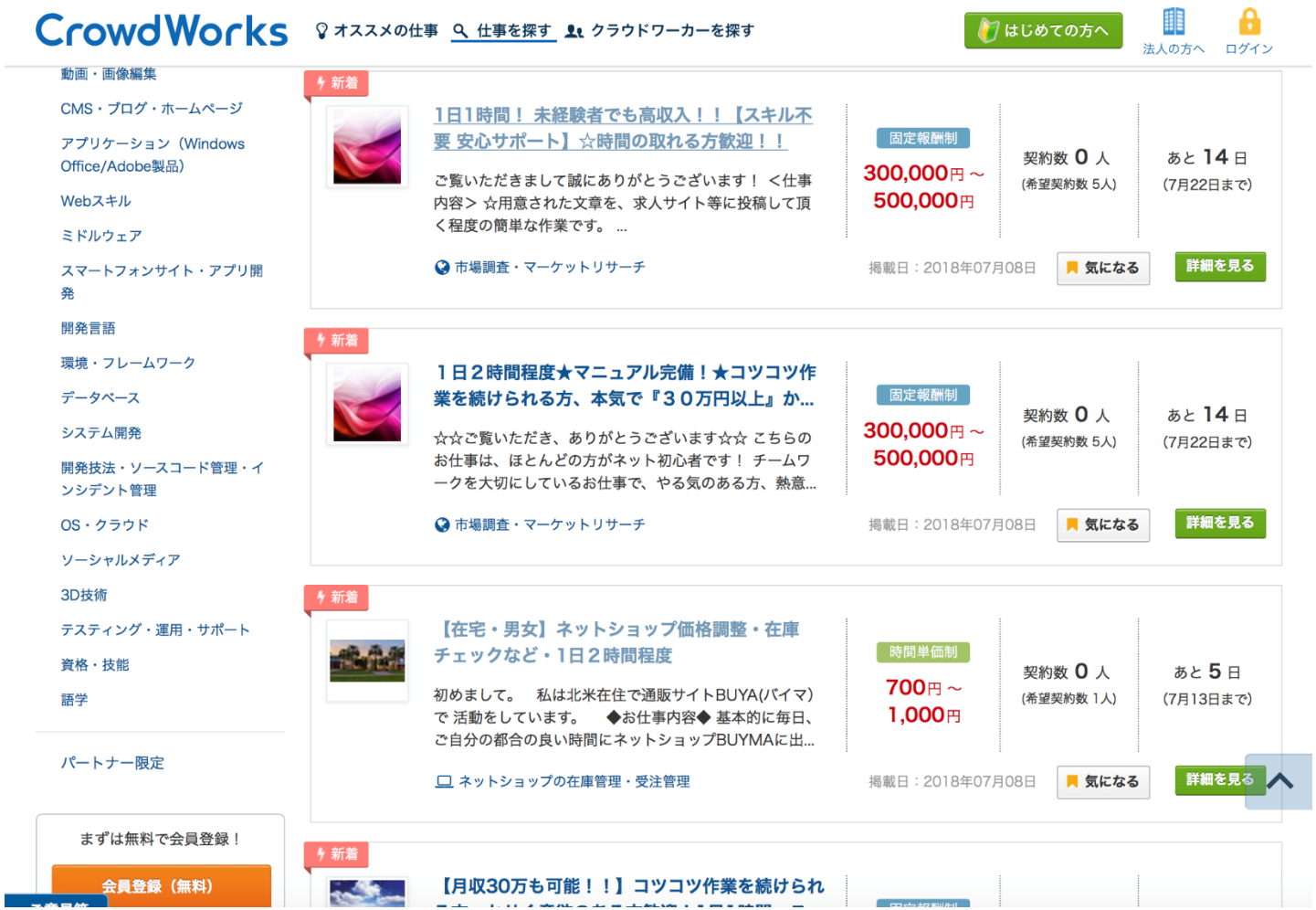
115,262件（2017年8, 11月, 2018年2, 5月）

本文（発注文）と件名（依頼タイトル）

1．分析の対象

クラウドワークス 「仕事を探す」 2017/07/ 08 取得
(https://crowdworks.jp/public/jobs/u/professionals?ref=root_pages-index-header)

1．分析の対象



クラウドワークス 「仕事を探す」 2017/07/ 08 取得
(https://crowdworks.jp/public/jobs/u/professionals?ref=root_pages-index-header)

1．分析の対象

クラウドワークス 「仕事の詳細」 2017/07/ 08 取得
(https://crowdworks.jp/public/jobs/2271029)

1．分析の対象

クラウドワークス 「仕事の詳細」 2017/07/ 08 取得
(https://crowdworks.jp/public/jobs/2271029)

2. データベースの全体像

1. 元データ：クラウドワークス社から提供された約11万5,000件の発注文書

2. ランダム1000：ランダムに抽出した1,000件の発注文書（70万3,428文字）

2. データベースの全体像

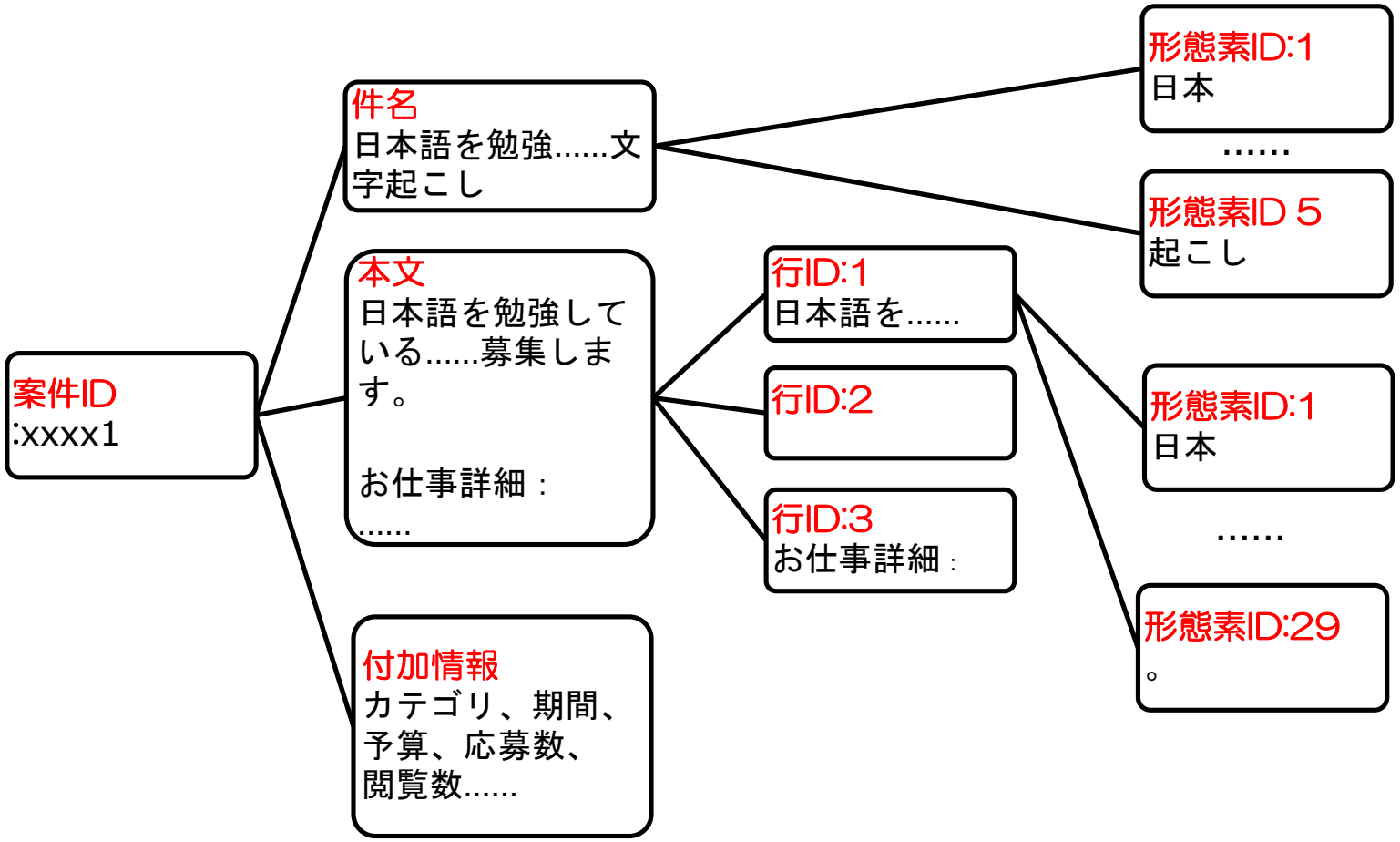
3. 実例50：さらに丁寧な質的分析を行うために取り出した50件の発注文書

4. やりとり文書：実際に発注を行ったテキストチャット

2. データベースの全体像

1. 形態素解析：件名
2. 形態素解析：本文
3. 付加情報：支払い形式、サイト上での案件のカテゴリ、期間、予算、応募数、閲覧数等

2. 形態素解析データベース



IDの階層構造

2. 形態素解析データベース

job_offer_ID	行ID	形態素ID	書字形	語彙素	語彙素読み	品詞大分類	品詞中分類
2122039	1	1	弊社	弊社	ヘイシャ	名詞	普通名詞

品詞小分類	品詞細分類	語種	語彙素ID	一致列
一般		漢	33870	弊社_ヘイシャ_名詞_普通名詞_一般_NA

2. 形態素解析の方法

形態素解析: Python 3.6.4を使用

形態素解析の辞書: unidic-cwj-2.2.0

2. 形態素解析の結果

解析結果

- 件名の全形態素: 2,778,529
- 仕事の詳細(本文)の全形態素: 14,419,821

2. 形態素解析の結果

詳しい内容は

石黒圭編（2020）

『ビジネス文書の

応用言語学的研究』

第２章に掲載されています

3．解析結果の活用

デザインの仕事の件名

ライティングの仕事の件名

ご清聴ありがとうございました。

井上 雄太（いのうえ ゆうた）
国立国語研究所 プロジェクト非常勤研究員