日本語の語彙平易化の評価のためのデータセットの構築　http://www.jnlp.org/resources/

① 語彙的換言データセット

IPA辞書 ∩ JUMAN辞書 の内容語(名詞,動詞,形容詞,副詞)
平易な語を削除(学習基本語彙：小学生のための理解語彙)
換言がない語を削除　(内容語換言辞書に含まれる語のみ)
低頻度語を削除 (新聞記事15年分での出現頻度が10以上)

各対象語について10文脈ずつ語彙的換言を収集
クラウドソーシングにより各５人の作業者を募集
文脈中での対象語の換言を思いつく限り列挙
辞書の参照は可、無記入も可、他人に聞くのは不可
37人の作業者から平均5.38語の語彙的換言を収集

クラウドソーシングにより新たに５人の作業者を募集
３人以上が「適切な換言である」と回答した換言を採用
83人の作業者によって平均4.50語の換言が認められた
作業者間の一致度は66.4%（十分に高い一致率）

② 語彙平易化データセット

【背景】語彙平易化は、文中の難解な語をより平易な同義語に置換する技術。子どもや言語学習者をはじめとする幅広い読者の文章読解を支援する。

【貢献】語彙的換言や語彙平易化のアルゴリズムの精度や再現率を計算し、F値を求める自動評価の枠組みを提供 → 研究のサイクルが素早く回るように！

文脈中で対象語とその換言を平易な順に並び替え
クラウドソーシングで各５人の作業者を募集
83人の作業者間の一致度は33.2%

難易度順位の平均をとってデータを統合
平均値が同じ単語同士には同じ順位を割り当てる
平均4.94段階の難易度が割り当てられた

データセットの文脈依存性	英語版１	日本語版
①：対象語が同じ文脈の組	9,045(100%)	10,485(100%)
②：①のうち換言リストが等しい組	302(3.3%)	1,593(15.2%)
③：②のうち難易度順位が異なる組	163(54.0%)	948(59.5%)
④：③のうち最も平易な語が違う組	57(35.0%)	463(48.8%)

規模	総文数	名詞(%)	動詞(%)	形容詞(%)	副詞(%)
英語版１	300	80(26.7)	80(26.7)	90(30.0)	50(16.7)
英語版２	430	100(23.3)	60(14.0)	160(37.2)	110(25.6)
日本語版	2,330	630(27.0)	720(30.9)	500(21.5)	480(20.6)

語彙的換言と語彙平易化の評価のためのデータセット

・820，悪気，親は悪気で言ったわけではなく、子供をあやすということを本当に知らなかった様子。
・820，悪気，意地悪 １；悪い考え １；悪意 ４；
・820，悪気，{意地悪，悪意} {悪気} {悪い考え}

語彙的換言の評価のためのデータセット

・ID，対象語，換言1 投票頻度1；換言2 投票頻度2；

語彙平易化の評価のためのデータセット

・ID，対象語，{最も平易な語}…{　}…{最も難解な語}