Языки для сравнения (на основе кириллицы):

- русский;
- белорусский;
- болгарский;
- сербский;
- татарский;
- украинский;
- казахский

Тексты для примеров:

<u>Сербский</u>

“Члан 3.

Свако има право на живот, слободу и безбедност личности.

Члан 4.

Нико се не сме држати у ропству или потчињености: ропство и трговина робљем забрањени су у свим облицима.

Члан 5.

Нико не сме бити подвргнут мучењу или свирепом, нечовечном или понижавајућем поступку или казни.”

<u>Казахский</u>

“3 бап

Әр адам өмір сүруге, бостандықта болуға және оның жеке басына қол сұғылмауына құқылы.

4 бап

Ешкім де құлдықта немесе кіріптарлықта ұсталуы тиіс емес. Құлдық пен құл саудасына, қандай түрде болса да, тыйым салынады.

5 бап

Ешкім де азапталуға немесе қадір-қасиетін қорлайтындай адамшылыққа жатпайтын қатыгездік жолмен жәбірленуге, немесе жазалануға тиіс емес.”

<u>Украинский</u>

“Стаття 3.

Кожна людина має право на життя, на свободу і на особисту недоторканність.

Стаття 4.

Ніхто не повинен бути в рабстві або у підневільному стані; рабство і работоргівля забороняються в усіх їх видах.

Стаття 5.

Ніхто не повинен зазнавати тортур, або жорстокого, нелюдського, або такого, що принижує його гідність, поводження і покарання."

1 этап: разбиваем все тексты на токены при помощи TreebankWordTokenizer. Получаем следующее:

Сербский

['члан', '3.', 'свако', 'има', 'право', 'на', 'живот', ',', 'слободу', 'и', 'безбедност', 'личности.', 'члан', '4.', 'нико', 'се', 'не', 'сме', 'држати', 'у', 'ропству', 'или', 'потчињености', ':', 'ропство', 'и', 'трговина', 'робљем', 'забрањени', 'су', 'у', 'свим', 'облицима.', 'члан', '5.', 'нико', 'не', 'сме', 'бити', 'подвргнут', 'мучењу', 'или', 'свирепом', ',', 'нечовечном', 'или', 'понижавајућем', 'поступку', 'или', 'казни', '.']

Казахский

['3', 'бап', 'әр', 'адам', 'өмір', 'сүруге', ',', 'бостандықта', 'болуға', 'және', 'оның', 'жеке', 'басына', 'қол', 'сұғылмауына', 'құқылы.', '4', 'бап', 'ешкім', 'де', 'құлдықта', 'немесе', 'кіріптарлықта', 'ұсталуы', 'тиіс', 'емес.', 'құлдық', 'пен', 'құл', 'саудасына', ',', 'қандай', 'түрде', 'болса', 'да', ',', 'тыйым', 'салынады.', '5', 'бап', 'ешкім', 'де', 'азапталуға', 'немесе', 'қадір-қасиетін', 'қорлайтындай', 'адамшылыққа', 'жатпайтын', 'қатыгездік', 'жолмен', 'жәбірленуге', ',', 'немесе', 'жазалануға', 'тиіс', 'емесэ']

Украинский

['стаття', '3.', 'кожна', 'людина', 'має', 'право', 'на', 'життя', ',', 'на', 'свободу', 'і', 'на', 'особисту', 'недоторканність.', 'стаття', '4.', 'ніхто', 'не', 'повинен', 'бути', 'в', 'рабстві', 'або', 'у', 'підневільному', 'стані', ';', 'рабство', 'і', 'работоргівля', 'забороняються', 'в', 'усіх', 'їх', 'видах.', 'стаття', '5.', 'ніхто', 'не', 'повинен', 'зазнавати', 'тортур', ',', 'або', 'жорстокого', ',', 'нелюдського', ',', 'або', 'такого', ',', 'що', 'принижує', 'його', 'гідність', ',', 'поводження', 'і', 'покарання', '.', '"']

2 этап: отбираем из токенов непосредственно слова, чистим от знаков препинания при помощи регулярных выражений (при этом будем использовать выражение \w, а не ['А-я'], так как помимо кириллицы в выбранных языках используются уникальные для конкретного языка символы, а также элементы латиницы:

Сербский

члан 3 свако има право на живот слободу и безбедност личности члан 4 нико (...)

Казахский

3 бап әр адам өмір сүруге бостандықта болуға және оның жеке басына қол сұғылмауына құқылы 4 бап ешкім де құлдықта немесе кіріптарлықта ұсталуы тиіс емес құлдық пен (...)

Украинский

['стаття', '3', 'кожна', 'людина', 'має', 'право', 'на', 'життя', 'на', 'свободу', 'і', 'на', 'особисту', 'недоторканність', 'стаття', '4', 'ніхто', 'не', 'повинен', 'бути', 'в', 'рабстві', 'або', 'у', 'підневільному', 'стані', 'рабство', 'і', 'работоргівля', 'забороняються', 'в', 'усіх', 'їх', 'видах', 'стаття', '5', 'ніхто', 'не', 'повинен', 'зазнавати', 'тортур', 'або', 'жорстокого', 'нелюдського', 'або', 'такого', 'що', 'принижує', 'його', 'гідність', 'поводження', 'і', 'покарання']

3 этап: при помощи регулярного выражения \D убираем цифры:

Сербский

['члан', 'свако', 'има', 'право', 'на', 'живот', 'слободу', 'и', 'безбедност', 'личности', 'члан', 'нико', 'се', 'не', 'сме', 'држати', 'у', 'ропству', 'или', 'потчињености', 'ропство', 'и', 'трговина', 'робљем', 'забрањени', 'су', 'у', 'свим', 'облицима', 'члан', 'нико', 'не', 'сме', 'бити', 'подвргнут', 'мучењу', 'или', 'свирепом', 'нечовечном', 'или', 'понижавајућем', 'поступку', 'или', 'казни', 'члан', 'свако', 'има', 'право', 'на', 'живот', 'слободу', 'и', 'безбедност', 'личности', 'члан', 'нико', 'се', 'не', 'сме', 'држати', 'у', 'ропству', 'или', 'потчињености', 'ропство', 'и', 'трговина', 'робљем', 'забрањени', 'су', 'у', 'свим', 'облицима', 'члан', 'нико', 'не', 'сме', 'бити', 'подвргнут', 'мучењу', 'или', 'свирепом', 'нечовечном', 'или', 'понижавајућем', 'поступку', 'или', 'казни']

Казахский

['бап', 'әр', 'адам', 'өмір', 'сүруге', 'бостандықта', 'болуға', 'және', 'оның', 'жеке', 'басына', 'қол', 'сұғылмауына', 'құқылы', 'бап', 'ешкім', 'де', 'құлдықта', 'немесе', 'кіріптарлықта', 'ұсталуы', 'тиіс', 'емес', 'құлдық', 'пен', 'құл', 'саудасына', 'қандай', 'түрде', 'болса', 'да', 'тыйым', 'салынады', 'бап', 'ешкім', 'де', 'азапталуға', 'немесе', 'қадір', 'қасиетін', 'қорлайтындай', 'адамшылыққа', 'жатпайтын', 'қатыгездік', 'жолмен', 'жәбірленуге', 'немесе', 'жазалануға', 'тиіс', 'емесэ']

Украинский

['стаття', 'кожна', 'людина', 'має', 'право', 'на', 'життя', 'на', 'свободу', 'і', 'на', 'особисту', 'недоторканність', 'стаття', 'ніхто', 'не', 'повинен', 'бути', 'в', 'рабстві', 'або', 'у', 'підневільному', 'стані', 'рабство', 'і', 'работоргівля', 'забороняються', 'в', 'усіх', 'їх', 'видах', 'стаття', 'ніхто', 'не', 'повинен', 'зазнавати', 'тортур', 'або', 'жорстокого', 'нелюдського', 'або', 'такого', 'що', 'принижує', 'його', 'гідність', 'поводження', 'і', 'покарання']

4 этап: сохраняем обработанные тексты без учета начального 31 слова при помощи среза [31:], так как это метаинформация о статье на английском языке:

Universal Declaration of Human Rights - Kazakh