# Kazakh (KK) Language Data

*by Bitext*

bitext we help AI understand humans

## About Bitext

Bitext has broken through the barriers that block multi-language text analysis. The company's Deep Linguistics Analysis Platform supports 77 languages at a lexical level and +20 at a syntactic leveland makes the company's technology available for a wide range of applications in Artificial Intelligence, text analytics and the new wave of products designed for voice interfaces such as chatbots and assistants.

## FEATURE SET A

(INFLECTIONAL FORMS LIST)

includes all the standard inflectional forms for nouns, verbs, adjectives, postpositions, conjunctions, etc. Each form will be annotated with the lemma (root form), POS, and morphological attributes (tense-mood, affirmative, person, number, case, degree, possessive-person, possessive-number, cooperative, modality, copulative).

## FEATURE SET B

(DERIVATIONAL FORMS LIST):

includes all the standard derivational forms including comparatives and superlatives. Each form will be annotated with the lemma (root form), POS, and morphological attributes (tense-mood, affirmative, person, number, case, degree, possessive-person, possessive-number, cooperative, modality, copulative).

## FEATURE SET C

(EXTENDED FORMS LIST)

includes the result of extending the inflectional and derivational forms lists as a result of considering additional morphological phenomena such as clitic pronouns and modality affixes. Each form will be annotated with the lemma (root form), POS, and morphological attributes (tense-mood, affirmative, person, number, case, degree, possessive-person, possessive-number, cooperative, modality, copulative).

# Kazakh (KK) Language Data

*by Bitext*

bitext we help AI understand humans

info@bitext.com

www.bitext.com

## FEATURE SET D

(NAMED ENTITIES FORMS LIST)

includes the data regarding named entities comprising person names, places, companies and organizations. Each form will be annotated with the lemma (root form), POS, and morphological attributes (tense-mood, affirmative, person, number, case, degree, possessive-person, possessive-number, cooperative, modality, copulative and entity-type).

## FEATURE SET E

(FREQUENCY INDICATION)

includes the data regarding the relative frequency of appearance for the words in the above lists in the given language. The relative frequency could be in the range of 0-255, or as requested.

## FEATURE SET F

(OFFENSIVE LANGUAGE FLAG)

includes information per word indicating if the word might be considered offensive in certain contexts.

## VOLUME OF LANGUAGE DATA

- Total number of lemmas: 10,000 lemmas
- Total number of forms: 2 million forms
  - Verbs: 1,000,000 forms (50%)
  - Nouns: 800,000 forms (40%)
  - Adjectives: 150,000 forms (7%)
  - Other: 50,000 forms (3%)