FEATURE SET A

(INFLECTIONAL FORMS LIST)

includes all the standard inflectional forms for nouns, verbs, adjectives, prepositions, conjunctions, etc. Each form will be annotated with the lemma (root form), POS, and morphological attributes (perfective, tense, person, number, gender, case, degree, negative).

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 (perfective, tense, person, number, gender, case, degree, negative).

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 (perfective, tense, person, number, gender, case, degree, negative 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: 45,000 lemmas
- Total number of forms: 1,500,000 million forms
  - Verbs: 300,000 forms (20%)
  - Nouns: 300,000 forms (20%)
  - Adjectives: 900,000 forms (59%)
  - Other: 20,000 forms (1%)