Arabic Dialects Segmentation Guidelines


General Rules:
- It is mandatory to not correct any typos in the text under any circumstances
- Mark all the segments and separate them using the "+" mark

List of Prefixes:
- Determiner ل such as in: ال+رجل، ال+كتاب
- Prepositions such as in: ب+هل، مع+ه، ل+ه
- When the Determiner "ال" is preceded with the preposition "ل" the Alif "ل" is typically removed. In the segmentation will be restored such as: للرجل is segmented as: ل+ال+رجل
- Conjunctions such as: و+هو، ف+هم
- Future markers such as: ي+يكتب، س+يكتب

Progressive particles such as: ب+ي+كتاب، ك+ي+كتاب
Negation particles such as: ما+قال+ش، م+قال+ش
Interrogative particles such as: ما+هو

List of Suffixes:
- Feminine nouns marker such as: معلم+ة، مكتبة+ة
- Feminine plural nouns marker attached to nouns such as: معلمات+ة، مكتبات+ة
- Dual nouns markers such as: رجل+ان، رجل+ين، معلم+ون، معلم+ين
- Person pronouns verb affixes first, second and third such as: ﻗﺎل+ها، كتب+وا، كتب+ا، كتب+ت، كتب+ن
- Negation affixes such as: م+كاتب+ش، ﻣﺎ+ﮔﻻ+ت+ش
- Pronouns attached to nouns, verbs or particles: ﺷﻜﺮ+ه، كتاب+ه
- Dialectal pronoun و that is used as a pronoun in the case: ﻋﻨﺪ+و (والمقصود عنه)

Special Cases:
- Connected words needs to be separated such as: عبد+الله، كتاب+ل+ه، ما+اك+ي+دير+وا
- When a letter of the prefix or a suffix is omitted then the shared letter will be segmented with the affix against the core of the word such as: ب+ف+و+ل+ك
- If two words share some letters, the common letter should be segmented with the last word: ٌي+خر+بيت
- The hashtags should be one unit and should not separate its components: ﻏُلد+ه
- Emotion should be considered as one unit: :-)
- Mentions should be considered as one unit: @mohamed_ali
- Repeated characters should be preserved and segment the word according if there is no repetition: ﻤم+و+أخييييير+11111111١، دع+وووو+لل+يييييي
- In the case of spelling mistakes/typos, should deal with the word as if it is spelled correctly: ال+مهمة، مأسا+ه
- Where there is a non-Arabic alphabet, it should be replaced with its equivalent Arabic letter such as: گلام+ش، افوري+شwould be 
- Negation and interjection words should be treated as one unit such as: ڈش، بلاق، لی، ی، ایش، لیش، ی، معش
- Words that are possible to break should be separated such as: م+حد+ش، ما+حد+ش، م+فی+ش، م+فی+هو+ش
- Merge the letter "ن" in "کنا". کن+دا
- When there is an elongation -Short vowel becomes long, the long vowel should be attached to the original particle: معا+ه، لی+هم، بي+هم
- Diphthong or blending of two letters in the case of the first person pronoun with the preposition "ل", the letter should remain with the pronoun: ل+یا will become ل+یا

Release:

Data Set:
http://alt.qcri.org/resources/da_resources/

Last Update:
Mon Jun 12, 2017.