Arabic Dialects Segmentation Guidelines

http://alt.qcri.org/resources/da_resources/seg-guidelines.pdf

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 ال such as in: ال+رجل، ال+كتاب
- Prepositions such as in: ب+ه، مع+ه، ل+ه
- When the Determiner "ال" is preceded with the preposition "ل" the Alif "ا" 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: ه+يكتب، س+يكتب
Progressive particles such as: ب+ي+كتب، ك+ي+كتب
Negation particles such as: ما+قال+ش، م+قل+ش
Interrogative particles such as: ما+هو

List of Suffixes:

- Feminine nouns marker such as: معلم+ة، مكتب+ة
- Feminine plural nouns marker attached to nouns such as: معلم+ات، مكتب+ات
- Dual nouns markers such as: رجل+ان، رجل+ين، معلم+ون، معلم+ين
- Person pronouns verb affixes first, second and third such as: كتب+وا، كتب+ا، كتب+ت، كتب+ن، ي+كتب+ون، كتب+نا
- Negation affixes such as: م+كتب+ش، ما+قل+ت+ش
- Pronouns attached to nouns, verbs or particles: شكر+ه، كتاب+ه، علي+ك
- Dialectal pronoun و that is used as a pronoun in the case: (عند+و والمقصود عنده)

Special Cases:

- Connected words needs to be separated such as: عبد+الله، كتب+ل+ه، ما+ك+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+قو+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: #الحمد_لله
- 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: ووو+أخيييير+اااا، ادع+ووو+للل+ييييي
- In the case of spelling mistakes/typos, should deal with the word as if it is spelled correctly: ال+مياة، مأسا+ه

- Where there is a non-Arabic alphabet, it should be replaced with its equivalent Arabic letter such as: كلامه، اقورة would be كلام+ه، اقور+ة
- Negation and interjection words should be treated as one unit such as: مش، بلاش، ليه،إيه، إيش، ليش، ياه، معلش
-Words that are possible to break should be separated such as: م+حد+ش، ما+حد+ش، م+في+ش، م+في+هو+ش
- Merge the letter "ن" in "كنا"، كن+ا
- When there is an elongation -Short vowel becomes long, the long vowel should be attached to the original particle: معاه، لي+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: ليا will become ل+يا.

Release:

Younes Samih, Mohamed Eldesouki, Mohammed Attia, Kareem Darwish, Ahmed Abdelali, Hamdy Mubarak and Laura Kallmeyer. Learning from Relatives: Unified Dialectal Arabic Segmentation. CoNLL 2017. August 3-4, 2017. Vancouver, Canada.

Data Set:

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

Last Update:

Mon Jun 12, 2017.