Guidelines for Treebank Annotation of Speech Effects and Disfluency for the Penn Arabic Treebank
V1.0

(based on the English Switchboard Bracketing Guidelines by Ann Taylor 1996)

Mohamed Maamouri, Ann Bies, Fatma Gaddeche, Sondos Krouna, Dalila Tabessi Toub

December 16, 2009

1 Additional dash-tags and node-labels for speech

1.1 Dash-tags

The following are additional dash-tags related to speech:

1.1.1 –UNF (unfinished)

-UNF: stands for ‘unfinished’ and marks unfinished constituents. The –UNF tag applies for partial single words, phrases, clauses and for sentences, and is added to the lowest constituent possible that can be described as unfinished.

S-UNF:

```
(S (NP-TPC-1 أنا·>anA·I)
   (VP أقُولُ·>a+quwl+u·I+say+[ind.]
       (NP-SBJ-1 *T*)
       (NP-TMP الآنَ·Al+|n+a·the+time/moment+[def.acc.])
       (SBAR أنَّ·>an~a·that (INTJ أه·>ah·ah!/ouch!)
             (S-UNF (NP-SBJ المُفَوَّضِيَّةَ·Al+mufaw~aDiy~+ap+a·the+delegation
                            العُلْيا·Al+EuloyA·the+supreme/high/highest)
                    (INTJ أه·>ah·ah!)))))
```

أنا أقول الآن أنّ المفوضيّة العليا أه

>anA + >aquwl + Al|n + >an~a + Al+mufaw~aDiy~ap + Al+EuloyA + >ah
I + say + the+now + that + the + delegation + the+high + uh
I say now that the high delegation is uh

NP-UNF:

```
(NP (NP مُحادَثَتِ-·muHAdav+at+i-·discussion/talk/negotiation
        (NP -هِم·-him·their))
    (NP-ADV مَعَ·maE+a·with+[def.acc.]
            (EDITED (NP-UNF أل·>l·NO_GLOSS))
            (NP القِيادَةِ·Al+qiyAd+ap+i·the+leadership/command
              لإسْرائِيلِيَّةِ·Al+<isorA}iyliy~+ap+i·the+Israeli)))
```

محادثاتهم مع الـ القيادة الإسرائلية

muHadavatihim + maEa + Al- + Al+qiyadapi + Al+<isorA}iliy~api
negotiations+their + the- + the+leadership + the+Israeli
Their negotiations with the- the Isralei leadership

PP-UNF:

```
(S (VP أنْتَقِلُ·>a+notaqil+u·I+move/transfer+[ind.]
       (NP-SBJ *)
       (EDITED (PP-UNF ل·l·NO_GLOSS))
       (PP لِ-·li-·for/to
           (NP -الأُسْتاذِ·-Al+>usotA*+i·the+professor+[def.gen.]
               كَمال·kamAl·Kamal
               الساعدي·Al+sAEdy·the+NOT_IN_LEXICON))))
```

أنتقل لـ للأستاذ كمال السعيدي

>anotaqilu + li- + li+Al+>usotA*i + kamAl + Al+saEiydiy
I+move + to- + to+the+professor + Kamal + Alsaidi
I go to- to professor Kamal Alsaidi

Note that just like any other dash-tag, -UNF can never be added to a parent and its child constituent at the same time.

1.1.2 –ETC (et cetera and similar filler phrases)

-ETC: is used for fillers like in Arabic ألخ >alaxo (etc.), وكلش wa kul~i$ (and everything), وهذا wa ha*A (and such). The node label carrying the dash-tag -ETC needs to be attached at the same level as the phrase it modifies, and it can be attached even at VP level. The node marked –ETC can be coordinated with any type of constituent, and it does not affect the coordination level. For example, in the sentence below, the coordination level is still S, even though there is an NP-ETC conjunct at the end.

```
(S (S (VP <jAwA إجاوا
          (NP-SBJ * ) ) )
   wa وَ
   (S (VP fat~a$uwA فَتَّشُوا
          (NP-SBJ * )
          (NP-OBJ Albiyt البيت ) ) )
   wa وَ
   (S (VP >axadwA أخَدوا
          (NP-SBJ * )
          (NP-OBJ >aworAq أوْراق )
   wa وَ
   (NP-ETC ha*A هَذا ) )
```

إجاو وفتّشوا البيت وأخذوا أوراق وهذا

>iJawA + wu+fat~$wA + Al+biyt + wu+>a*uwA + >aworAq + wu+ha*A
Came they + and+searched + the+house + and+took they + papers + and+this
They came, searched the house, took some papers and so forth

1.2 Node Labels

1.2.1 INTJ (interjection and filled pauses)

INTJ: is used for common interjections (for list of interjections in Arabic go to 4.3.1.8 INTERJECTIONS in the POS guidelines) and <u>also</u> for filled pauses and hesitations such as اه uh ام umm (see section 2.1. below)

Filled pauses given the node label INTJ should be treated like punctuation and placed as high as possible in the tree.

```
(S (VP تَحَدَّثنا·taHad~av+nA·speak/discuss+we_[verb]
       (NP-SBJ *)
       (PP-CLR عَن·Ean·from,_off,_away_from,_about,_on,_over
              (NP مَسْألَةِ·maso>al+ap+i·issue/affair/matter/question
                  (NP الأبْنِيَةِ·Al+>aboniy+ap+i·the+structures/buildings
                      (INTJ أه·>ah·ah!)
                      المَدْرَسِيَّةِ·Al+madorasiy~+ap+i·the+scholastic/school)))))
```

تحدثنا عن مسألة الأبنية أه المدرسيّة

taHad~avnA + Eano + maso>alapi + Al+>aboniyapi + Al+madorasy~api
talked+we + about + matter + the+buildings + the+school
We talked about the issue of school buildings

Also annotated like interjections are the conjunctions **أو**>aw (or) and **بل**bal (but) when they are used for restarts (see section 2.3.4 below)

1.2.2 PRN (parenthetical)

PRN is used for filler sentences and clauses only, not for single words. The use of PRN is the same for broadcast news and speech annotation as it is for newswire and text annotation.

```
(SQ (EDITED (SQ (PRT هَلْ·halo·do?/is?)
                (VP-UNF (PRT سَ-·sa-·will)
                        -يُمْكِنُ·-yu+mokin+u·he/it+be_possible)))
    (PRT هَلْ·halo·do?/is?)
    (VP (PRT سَ-·sa-·will)
        -تَخْرُجُ·-ta+xoruj+u·it/they/she+go_out/exit/leave+[ind.]
        (NP-SBJ *)
        (PP-CLR بِ·bi·with/by
                (INTJ أه·>ah·ah!/ouch!)
                (PRN (S (VP يَعْنِي·ya+Eoniy+[null]·he/it+mean/concern+[ind.]
                        (NP-SBJ *))))
                (NP (NP مُفاجِئاتٌ·mufAji}+At+K·surprises+[fem.pl.]+[indef.gen.]
                        حَقِيقِيَّةٌ ·Haqiyqiy~+ap+K·real +[fem.sg.]+[indef.gen.])
                    أوْ·>awo·Or,_if_not,_unless,_except_if,_except_when
                    (NP قَراراتٌ·qarAr+At+K·decisions/resolutions
                        حاسِمَةٌ·HAsim+ap+K·decisive/critical)))))
```

هل سيمكن هل ستخرج بـ - أه يعني مفأجئات حقيقة أو قرارات حاسمة

Hal + sa+yumokinu + hal + sa+taxoruju + bi- + ah + yaEoniy + mufAja}AtK + Haqiyqy~apK + >awo + qarArAtK + HAsimapK
Do/does + will+be able + do/does + will+come out + with + means it + suprises + true + or + decisions + decisive
Will she come out with, uh, I mean, with real surprises or decisive decisions

1.2.3 EDITED (restarts and repetition)

EDITED: The use of this tag shows repetition and restarts of constituents. Restarts of non-clausal elements like NPs, PPs, ADJPs, etc. need to be treated as <u>sisters</u> <u>of</u> the actual constituent. Restarts of Ss and SBARs need to be <u>inside</u> the S or the SBAR. (More on EDITED in section 2 and 2.4 below.)

```
(NP-ADV دَاخِلِ·dAxil+i·inside_of/interior/inside+[def.gen.]
        (EDITED (NP-UNF ال·Al·NO_GLOSS))
        (NP حَرَكَةِ·Harak+ap+i·movement/activity/organization
            (NP فَتْح·fatoH·Fatah_[PLO_branch])))
```

داخل الـ حركة فتح

dAxila + Al- + Harakapi + FatoH
inside + the- + movement + Fatah
Inside the- PLO

1.2.4 FRAG (fragment)

FRAG: Two or more constituents that need to be held together as a grammatical statement but that are not in a subject-predicate relationship are annotated as FRAG. The use of FRAG is the same for broadcast news annotation as it is for newswire annotation.

```
(FRAG (NP وائِل·wA}il·Wael/Wa'il الدحدوح·Al+dHdwH·the+NOT_IN_LEXICON)
      (NP الجَزِيرَة·Al+jaziyr+ap·the+Jazeera+[fem.sg.])
      (NP-LOC غَزَّة·gaz~+ap·Gaza+[fem.sg.]))
```

وائل الدحدوح الجزيرة غزة

wA}il + Al+daHduwH + Al+jaziyrap + gaz~ap
Wail + Aldahdouh + Aljazeera + Gaza
Wail Aldahdouh, Aljazeera, Gaza

```
(FRAG (NP-VOC إِيمان·<iymAn·Iman)
      (NP (NP شُكْراً·$ukor+AF·thankfulness/thanks+[acc.indef.]
              جَزِيلاً·jaziyl+AF·abundant+[acc.indef.])
          (PP لَ-·la-·to/for
              (NP -كَ·-ka·you_[masc.sg.]))))
```

إيمان شكراجزيلا لك

<imAn + $ukorAF + jaziylAF + la+ki
Iman + thank + abundant + to+you
Iman, thank you!

1.2.5 FRAG vs. -UNF

NOTE on FRAG vs. -UNF: A fragment is a proposition that is complete in meaning but missing syntactic constituents due to deliberate stylistic choices. In broadcast news we notice a high frequency of telegraphic style at the beginning and the end of the news programs, where anchors present the headlines, when they announce breaks, when they introduce reporters and outside reports, when they transfer from one news section to another, etc.

An unfinished item is a speaker mis-performance where a word, a phrase or a sentence is started correctly but ended abruptly before the global meaning of the proposition is conveyed properly. This tends to happen more frequently in broadcast conversation than in broadcast news.

Note that the distinction between FRAG and –UNF must be made in the context of the whole broadcast news file. It is possible that a "paragraph" could be properly annotated as FRAG in some contexts but –UNF in other contexts.

```
(S (EDITED (S-UNF (NP-TPC-1 نَحْنُ·naHonu·we)
                  (NP-SBJ دَوْر-·dawor·role/part
                          (NP (NP -نـا nA our )
                              (NP-1 *T*)))))
   (VP لَيْسَ·layos+a·not_be+he/it_[verb]
       (NP-SBJ دَوْرُ-·dawor+u-·role/part+[def.nom]
               (NP -نـا·-nA·our))
       (SBAR-PRD أَنْ·>ano·That,_to
                 (S (VP نُتابِعَ·nu+tAbiE+a·we+continue/follow/monitor+[sub.]
                        (NP-SBJ *)
                        (NP-OBJ الفَسادَ·Al+fasAd+a·the+corruption+[def.acc.])
                        (PP-MNR بِ-·bi-·in,_at, _with
                                (NP -شَكْل·-$akol+i·manner/form/configuration
                                    (NP التَفْصِيل·Al+tafoSiyl+i·the+detailing)
                                                                        )))))))
```

نحن دورنا ليس دورنا ان نتابع الفساد بشكل التفصيل

naHonu + daworu+nA + layosa + daworu+nA + >ano + nutAbiEa + Al+fasAda + bi+$akoli + Al+tafoSiyli

we + role+our + not + role+our + that + we follow + the+corruption + with+manner + the+detailing

Our role- our role is not trace corruption in its most detailed forms

2 Annotation of disfluency

Disfluency in speech is marked by hesitation sounds, partial words, repetitions and restarting of phrases or sentences. Each of these elements of disfluency can occur separately or together with other elements. The following is a brief description of most common disfluency features and the way they are treated in the Treebank by means of new and old node labels and function tags.

2.1 Hesitation sounds and filled pauses

Filled pauses are non-word sounds that speakers employ to indicate hesitation or to maintain control of a conversation while thinking of what to say next. At POS level they are annotated as INTERJ and at Treebank level they get the node label INTJ. Because of their free distribution, filler sounds should be treated like punctuation: they should be put as high as possible in the tree.

```
(S (INTJ أه·>ah·ah)
   (NP-TPC-3 حَوالَي·HawAlay·about,_approximately,_around,_roughly
             (NP ثَمانِيَة·vamAniy+ap·eight+[fem.sg.]
                 أَوْ·>awo·Or,_if_not,_unless,_except_if,_except_when
                 تِسْعَة·tisoE+ap·nine+[fem.sg.]))
   (VP ماتُوا·mAt+uwA·die/pass_away+they_[verb]
       (NP-SBJ-3 *T*)))
```

أه حوالي ثمانية أو تسعة ماتوا

>ah + HawAlayo + vamAniyapi + >awo + tisoEapi + mAtuwA
uh + around + eight + or + nine + died they
Uh, around eight or nine died

```
(PP (PP عَلَى·EalaY·on/above
        (NP لُبْنان·lubonAn·Lebanon))
    وَ-·wa-·and
    (PP -عَلَى·-EalaY·on/above
        (INTJ أه·>ah·ah)
        (NP سُوريا·suwriyA·Syria)))
```

على لبنان و على أه سوريا

EalaY + lubonAn + wa+EalaY + >ah + suriyA
On + Lebanon + and + on + uh + Syria
On Lebanon and on uh Syria

```
(FRAG (INTJ أُم·>um·um_[filled_pause])
      (NP-ADV طَيِّب·Tay~ib·good/pleasant)
      (PP عَلَى·EalaY·on/above
          (NP الصَعِيدِ·Al+SaEiyd+i·the+level/plane
              السِياسِيِّ·Al+siyAsiy~+i·the+political))
      (NP (NP {نْعِكاساتُ·{inoEikAs+At+u.repercussion+[fem.pl.]+[def.nom.]
              (NP (NP هٰذا·h`*A·this_[masc.sg.])
                  (NP الحادِثِ·Al+HAdiv+i·the+incident/event+[def.gen.])))
          (PP عَلَى·EalaY·on/above
              (NP الساحَةِ·Al+sAH+ap+i·the+scene/field/arena
                  السِياسِيَّةِ·Al+siyAsiy~+ap+i·the+political
                  اللُبْنانِيَّةِ·Al+lubonAniy~+ap+i·the+Lebanese))))
```

أم طيب على الصعيد السياسي انعكاسات هذا الحادث على الساحة السياسية اللبنانية

>um + Tay~ib + EalaY + Al+SaEiydi + AlsiyAsy~i + >inoEikAsAtu + ha*A + Al+HAdivi + EalaY + AlsAHapi + Al+siyAsiapi + AllubonAniy~api
umm + good + on + the+level + the+politic + the+ repercussions + this + the + incident + on + the+scene + the+political + the+ Libanese
Umm, good, at the political level, the repercussions of this incident on the political Lebanese scene

(Note that this last example was annotated as FRAG due to the context of the whole broadcast news file, even though in other contexts or in isolation it may appear unfinished.)

2.2 Noise

Transcription of noise and other sounds (other than filled pauses) like laughter, cough, music, etc. is considered as metadata and should not appear in our Treebank files. If present, they should be ignored (i.e., left as high as possible in the tree without any node label) and reported in the comment field and in the annotator’s email summary as a technical problem.

2.3 Partial words and unfinished phrases and sentences

Partial words and unfinished phrases and sentences should all carry the dash-tag –UNF.

```
(SQ (PRT هَلْ·halo·do?/is?)
    (VP يُمْكِنُ·yu+mokin+u·he/it+be_possible/make_possible_for+[ind.]
        (NP-ADV فِعْلاً·fiEol+AF·actually/in_effect+[acc.indef.])
        (PP فِي·fiy·in
            (NP (NP هٰذِهِ·h`*ihi·this/these)
                (NP المَرْحَلَةِ·Al+maroHal+ap+i·the+phase/stage/round)))
        (EDITED (NP-UNF ألتح·>ltH·NO_GLOSS))
        (NP-SBJ (NP الحَدِيثُ·Al+Hadiyv+u·the+discussion/talk/interview)
                (PP عَن·Ean·from,_off,_away_from,_about
                    (NP نْتِخاباتٍ·{inotixAb+At+K·elections/selections))))})
```

هل يمكن فعلا في هذه المرحلة ألتح- الحديث عن نتخابات

Hal + yumokinu + fiEolAF + fiy + ha*ihi + Al+maroHalapi + Al+taH- + Al+Hadiyvu + Ean + <intixAbAtK

Do/does + be possible + in effect + in + this + phase + the+ta- + the+talk + on + elections

Can we actually in this phase ta- talk about elections?

2.4 Repetition and Restarts

As a common feature of conversation, the term "restart" refers to the repetition of a word or string of words within an utterance by the same speaker. Restarts of syntactic constituents in Treebank are edited speech. They are given the node label EDITED and are treated as sisters of the restarted constituent if the constituent is smaller than a clause (NP, PP, ADJP, ADVP, etc.) and are embedded if the restarted constituent is a clause (SBAR or S). Note that EDITED cannot be the top node of the paragraph – any constituent type can have the EDITED node as a child rather than a sister if the constituent is the top node of the paragraph.

The fundamental theoretical linguistic and methodological principles for the treebanking of restarts are the same for the English Treebank and the Arabic Treebank. A note on these principles from the English Switchboard (speech) guidelines (Taylor, 1996):

> The linguistic theory behind restarts is that the speaker, on deciding that s/he wants to fix something already uttered, goes back one constituent and restarts from there. For this reason we try to make restarts sisters of the replaced constituent and have them contain a single constituent which is sometimes unfinished (and therefore labelled -UNF). There are two problems with implementing this approach, however: (1) exactly what counts as a "constituent" for restarting is not particularly clear; and (2) the way Treebank brackets things doesn't always provide bracketing at the appropriate level to do this accurately. We have therefore had to make some compromises, while to the extent possible keeping the system both internally consistent and linguistically plausible.
>
> Summary of restart policy
>
> - restarts of phrases (any label that doesn't start with S) are sisters
> - restarts of clauses (any label that starts with S) are embedded
> - if the restart starts at a point where in normal Treebank bracketing there is no label immediately to the left, there is no label inside the restart.
> - any label at top level is treated like a clause and the restart is embedded.

2.4.1 Restarts of constituents smaller than clauses

Restarts of non-clausal constituents are treated as sisters. If the restart starts at the beginning of a labeled constituent, the constituent child of the EDITED node has the same label (with –UNF if it is unfinished).

EDITED restart of an NP:

```
(NP مَبْعُوثُ·maboEuwv+u·envoy/representative+[def.nom.]
     (EDITED (NP-UNF وز·wz·NO_GLOSS))
     (NP (NP وِزارَةِ·wizAr+ap+i·ministry+[fem.sg.]+[def.gen.]
               (NP الخارِجِيَّةِ·Al+xArijiy~+ap+i·the+foreign/outside))
          (ADJP الأَمْرِيكِيَّةِ·Al+>amoriykiy~+ap+i·the+American)))
```

مبعوث وز وزارة الخارجية الامريكية

maboEuwvu + wiz- + wizArapi + Al+xArijiy~api + Al+>amoriyky~api
delegate + minis- + ministry + the foreign + affairs + the+American
The American depart- department of state’s delegate

```
(S (EDITED (NP-VOC-UNF أ·>·NO_GLOSS))
   (NP-VOC أُسْتاذ·>usotA*·professor حُسَيْن·Husayon·Hussein)
   (PP-TMP فِي·fy·in
           (NP الأَعْوام·Al+>aEowAm+i·the+years
               الماضِيَةِ·Al+mADiy+ap+i·the+past/bygone)
   (PP فِي·fy·in
       (NP كُلِّ·kul~+i·all,_entire,_every,_whole
           (NP بَرْلَمانٍ·barolamAn+K·parliament+[indef.gen.]
               مِصْرِيٍّ·miSoriy~+K·Egyptian+[indef.gen.]))))
   (VP كانَ·kAn+a·be/was/were+he/it_[verb]
       (NP-SBJ الإِخْوانُ·Al+<ixowAn+u·the+brothers+[def.nom.]))))
```

أ- أستاذ في الاعوام الماضية في كل برلمان مصري كان الاخوان ...

>u- + >usotA* + Husayon + fiy + Al+>aEowAmi + Al+mADiyapi + fiy + kul~i + barolamAnK + kAn + Al>ixwAnu…

Prof- + professor Husain + in + the+ years + the+previous + in + every + parliament + was + the + Brothers

Prof- professor Husain, in the past years and in every parliament, the Brothers were…

```
(S (PP فِي·fiy·in
       (NP (NP رَأْي-·ra>oy-·opinion/view/idea
               (NP -ي·-iy·my))
           (ADJP الشَّخْصِيِّ·Al+$axoSiy~+i·the+private/personal+[def.gen.])))
   (EDITED (NP-TPC (ADJP أَهَمُّ·>aham~+u·more/most_important+[def.nom.]
                          (SBAR-NOM (WHNP-3 ما·mA·what)
                                    (S (VP قالَ-·qAl+a-·said+he/it_[verb]
                                           (NP-OBJ (NP -هُ·-hu·it/him)
                                                   (NP-3 *T*))
                                           (NP-SBJ نَصْر·naSor·Nasr
                                                   الل`ّه·All~`h·Allah)
                                           (NP-TMP  اليوم Al+yawo+ma
                                                         the+today)))))))
   (SBAR-NOM-TPC-7 (WHNP-6 ما·mA·what)
                   (S (VP قالَ-·qAl+a-·said+he/it_[verb]
                          (NP-OBJ (NP -هُ·-hu·it/him)
                                  (NP-6 *T*))
                          (NP-SBJ السَيِّدُ·Al+say~id+u·the+Sir/Mr./Mister
                                  حَسَن·Hasan·Hassan
                                  نَصْر  naSor·Nasr
                                  اللّه·All~`h·Allah)
                          (NP-TMP اليَوْمَ·Al+yawom+a·the+today+[def.acc.]))))
   (VP يُعْتَبَرُ·yu+Eotabar+u·he/it+be_considered/be_regarded/be_believed+[ind.]
       (NP-SBJ-7 *T*)
       (S (NP-SBJ-7 *)
          (NP-PRD خِطاباً·xiTAb+AF·speech+[acc.indef.]
                  مَفْصِلِيّاً·mafoSiliy~+AF·pivotal/crucial/critical))))
```

في رأيي الشخصي أهم ما قاله نصر الله اليوم ماقاله السيد حسن نصر الله اليوم يعتبر خطابا مفصليا

fiy + ra>oy+iy + Al+$axoSiy + >aham~u + mA + qAla+hu + NaSr + All~ah + Alyaoma + mA + qAla+hu + Alsay~idu + Hasan + NaSr + All~ah + Al+yawoma + yuEotabaru + xiTAbAF + mufoSily~AF

in + opinon+my + the+personal + most important + what + said+it + Nasr+ Allah + the+today + what + said+it + mister + Hassan + Nasr + Allah + the+today + is considered + speech + crucial

In my personal opinion what Nasrallah said today, what mister Hasan Nasrallah said today is to be considered as a crucial speech

EDITED restart of a PP:

```
(S (NP-PRD -عِنْدَ-·-Einod+a-·with/at+[def.acc.]
           (NP -نا·-nA·our))
   (NP-SBJ (NPصُوَرٌ·Suwar+N·pictures/photographs/illustrations/photos)
           (EDITED (PP-UNF ل·l·for))
           (PP لِ-·li-·for/to
           (NP (NP -الناس·-Al+nAs·the+people) (NP دي·dy·this))
               (NP-ADV كُلِّ-·kul~+i-·all,_entire,_every,_whole+[def.gen.]
                       (NP -ها·-hA·its/their/her)))))
```

عندنا صور ل للناس كلها

Einoda+nA + SuwarN + li- + li+AlnAsi + kul~i+hA
at+us + pictures + for + for + the+people + all+their
We have pictures of of everybody

EDITED restart of an ADJP:

```
(PP-MNR بِ-·bi-·by/with
        (NP شَكْلِ-·-$akol+K·manner/form/configuration+[indef.gen.]
            (EDITED (ADJP-UNF جِذ·j*·NO_GLOSS))
            جِذْرِيٍّ·ji*oriy~+K·radical/root+[indef.gen.]))
```

بشكل جِذْ- جِذْرِيٍّ

bi+$akolK + ji*- + ji*riy~K
with+form + radic- + radical
In a radic- radical way

If the part inside would not normally have a node label, like one-token conjunctions, etc., the constituent child of EDITED does not have a node label.

```
(NP (NP الحَواجِزُ·Al+HawAjiz+u·the+obstacles/hurdles+[def.nom.])
    (EDITED وَ·wa·and)
    وَ-·wa-·and
    (NP المُعَوِّقاتُ·Al+muEaw~iq+At+u·the+obstacles/impediments))
```

الحواجز و- و المعوقات

Al+HawAjizu + wa- + wa+Al+maEuwqAtu
the+obstacles + and + and+ the+impediments
The obstacles and and the impediments

```
(S (S (NP-TPC-3 (NP ه`ذا·h`*A·this_[masc.sg.])
                (NP المِلَفُّ·Al+milaf~+u·the+file/dossier+[def.nom.]))
      (VP أَجِّلْ-·>aj~il+o-·postpone/delay+you
          (NP-SBJ *)
          (NP-OBJ (NP -هُ·-hu·it/him)
                  (NP-3 *T*))))
   (EDITED أَوْ·>awo·Or,_if_not,_unless,_except_if,_e)
   أَوْ·>awo·Or,_if_not,_unless,_except_if,_except_when
   (S (VP {تْرُكْ-·{utoruk+o-·leave/quit+you_[verb]
          (NP-SBJ *)
          (NP-OBJ -هُ·-hu·it/him))))
```

هذا الملف أجله أو أو ا تركه

ha*A + Al+mil~afu + >uj~ilo+hu + >aw + >aw + >utoruko+hu
this + the+folder + postpone+it + or + or + leave+it
This topic, postpone it, or or leave it

Single words that would not normally have a node label, such as a single adjective modifying a single noun, and that are simple restarts of complete words do not have a node label inside the

EDITED. If the adjective for example is itself an unfinished word, however, it needs to have the ADJP label inside the EDITED in order to have a node label to attach the –UNF dashtag to.

```
(NP-SBJ الأيّامُ·Al+>ay~Am+u·the+days+[def.nom.]
        (EDITED المُقْبِلَةُ·Al+muqobil+ap+u·the+next/coming/approaching(
        المُقْبِلَةُ·Al+muqobil+ap+u·the+next/coming/approaching)))
```

الايام المقبلة المقبلة

Al+>ay~amu + Al+muqobilapu + Al+muqobilapu
the+days + the future + the future
The future- future days

```
(PP-MNR بِ-·bi-·by/with
        (NP -شَكْلِ·-$akol+K·manner/form/configuration+[indef.gen.]
            (EDITED (ADJP-UNF جذ·j*·NO_GLOSS))
            جِذْرِيٍّ·ji*oriy~+K·radical/root+[indef.gen.]))
```

بشكل جِذْ- جِذْرِيٍّ

bi+$akolK + ji*- + ji*riy~K
with+form + radic- + radical
In a radic- radical way

EDITED restart of a VP:

```
(S wa وَ
   (PP فِي·fiy·in
       (NP الأيّامُ·Al+>ay~Am+u·the+days+[def.nom.]
           المُقْبِلَةُ·Al+muqobil+ap+u·the+next/coming/approaching))
   (EDITED (VP-UNF يَأْتِي·ya+>otiy+[null]·he/it+arrive/come/reach+[ind.]))
   (VP يَأْتِي·ya+>otiy+[null]·he/it+arrive/come/reach+[ind.]
       (NP-SBJ *)
       (PP-CLR بِ-·bi-·by/with
               (NP -شَيْءٍ·-$ayo'+K·something/thing+[indef.gen.]
                   جَدِيدٍ·jadiyd+K·new/modern+[indef.gen.]))
       (PP عَلَى·EalaY·on/above
           (NP السَاحَةِ·Al+sAH+ap+i·the+scene/field/arena
               الفِلَسْطِينِيَّةِ·Al+filasoTiyniy~+ap+i·the+Palestinian))))
```

وَ فِي الأيّامُ المُقْبِلَةُ يأتي يأتي بشئ جديد على الساحة الفلسطينية

wa + fiy + Al+>ay~Am+u + Al+muqobil+ap+u + ya>otiy + ya>otiy + bi+$ayo’K + jadiydK + EalaY + Al+sAHapi + Al+falasoTiyniy~api
and + in + days + coming + come + come + with + thing + new + on + the+scene + the+political
and in the coming days it brings brings something new to the Palestinian scene

Note: Non-clausal restarts at top-level are embedded just like clauses. EDITED cannot be the top node of the paragraph – any constituent type can have the EDITED node as a child rather than a sister if the constituent is the top node of the paragraph.

2.4.2 Restarts of clauses

All clause-level restarts are embedded.

EDITED restart of an S:

```
(S (EDITED (S-UNF (VP يَأْتِي·ya+>otiy+[null]·he/it+arrive/come/reach+[ind.])))
    (VP يَـأْتِـي·ya+>otiy+[null]·he/it+arrive/come/reach+[ind.]
        (NP-SBJ *)
        (PP-CLR بـ-·bi-·by/with
                (NP -شَيْءٍ·-$ayo'+K·something/thing+[indef.gen.]
                     جَدِيـدٍ·jadiyd+K·new/modern+[indef.gen.]))
        (PP عَلَى·EalaY·on/above
             (NP الـساحَةِ·Al+sAH+ap+i·the+scene/field/arena
                 الـفِلَسْطِينِيَّةِ·Al+filasoTiyniy~+ap+i·the+Palestinian))))
```

يأتي يأتي بشئ جديد على الساحة الفلسطينية

ya>otiy + ya>otiy + bi+$ayo’K + jadiydK + EalaY + Al+sAHapi + Al+falasoTiyniy~api
come + come + with + thing + new + on + the+scene + the+political
it brings brings something new to the Palestinian scene

EDITED restart of an SBAR

```
(S (VP (PRT لَمْ·lamo·did_not)
       يَكُنْ·ya+kun+o·he/it+be+[jus.]
       (NP-SBJ *)
       (SBAR-NOM-PRD (WHNP-1 0)
                      (S (VP متؤكدا·mt&kdA·NO_GLOSS
                             (NP-SBJ-1 *T*)
                             (UCP-MNR (NP تَماماً·tamAm+AF·completely)
                                       وَ-·wa-·and
                                      (PP -بِ-·-bi-·by/with
                                         (NP -{عْتِرافِ-·-{iEotirAf+i-
                                                      ·acknowledgment
                                              (NP -هِ·-hi·its/his))))
                             (SBAR
                                 (EDITED
                                      (SBAR
                                          (EDITED
                                              (SBAR-UNF أَنَّ·>an~a·that))
                                          أَنَّ-·>an~a-·that
                                          (S-UNF (NP-SBJ -ها·-hA·it))))
                                 أَنَّ-·>an~a-·that
                                 (S (NP-SBJ -ها·-hA·it)
                                    (ADJP-PRD نِهائِيَّةٌ·nihA}iy~+ap+N·
                                                      final)))))))
```

لم يكن متؤكدا تماما وباعترافه أن أنها أنها نهائية

lam + yakuno + muta>ak~iDAF + tamamAF + wa+bi+>iEorirAfi+hi + >an~a + >an~a + hA+ >an~a + hA + nihA}iy~apN

not + be he + sure + completely + and+with+acknowledgment+his + that + that + it + that + it + final

He is not completely sure, and in his own words, that that it that it is final

```
(S (NP-PRD -عِنْدَ-·-Einod+a-·with/at+[def.acc.]
           (NP -نا·-nA·our))
   (NP-SBJ (NP صُوَرٌ·Suwar+N·pictures/photographs/illustrations+[indef.nom.])
           (PP لِ-·li-·for/to
               (NP (NP (NP -النَاس·-Al+nAs·the+people)
                       (NP دي·dy·this))
                   (NP كُلِّ-·kul~+i-·all,_entire,_every,_whole+[def.gen.]
                       (NP -ها·-hA·its/their/her)))))
   (SBAR-ADV (EDITED (SBAR-ADV وَ-·wa-·while
                               (S-UNF (NP-SBJ -هِيَ·-hiya·it/they/she))))
             وَ-·wa-·while
             (S (NP-TPC-1 -هِيَ·-hiya·it/they/she)
                (VP (PRT ب-·b-·progressive)
                    -تِعمَل·-tiEmal·do
                    (NP-SBJ-1 *T*)
                    (NP-OBJ (NP الكَلام·Al+kalAm·the+speech/statement/talk)
                            (NP ده·dh·this))))))
```

عندنا صور للناس دي كلها وهي وهي بتعمل الكلام ده

Einoda+nA + SuwarN + li+AlnAsi + dih + kul~i+hA + wa+hya + wa+hya + bi+tiEomil + Al+kalAm + dah

at+us + pictures + for + the+people + this + all+their + while + they + while + they + in instance of + they + do + the+talk + this

We have pictures of all these people while they- while they are doing this stuff

```
(SBARQ (EDITED (SBARQ (WHNP ما·mA·what/which)
                      (S (NP-SBJ-UNF ه·h·NO_GLOSS))))
       (WHNP-1 ما·mA·what/which)
       (S (NP-SBJ (NP هِيَ·hiya·it/they/she)
                  (NP-1 *T*))
          (NP-PRD صَلاحِيّاتُ-·SalAHiy~+At+u-·privileges+[fem.pl.]+[def.nom.]
                  (NP -هُ·-hu·its/his))))
```

ما هـ ماهي صلاحياته

mA + hy + mA + hya + SalAHiyAtu+hu

what + sh- + what + she + privileges+his

What- what are his privieleges

Note that the EDITED node can never be a final node in the tree.

2.4.3 Note on restarts withأو aw (or) and بل bal (but) after repetition

We noticed in the ATB5 corpus that speakers sometimes correct themselves by repeating a structure *and* adding >awo or bal in between. >awo and bal in this context are actually not used in their coordination function – they are more like filled pauses, and so should be annotated as INTJ child(ren) of the EDITED node and sister(s) of the edited phrase TB level**.** **But** they will keep their POS value as CONJ.

```
(S (VP يَبْدُو·ya+boduw+[null]·he/it+appear/seem+[ind.]
   (SBAR-SBJ أنَّ·>an~a·that
             (S (S (NP-TPC-3 العَمَلِيَّةَ·Al+Eamaliy~+ap+a·the+operation/mission)
                   (VP (PRT سَ-·sa-·will)
                       تَسْتَمِرُّ-·-ta+sotamir~+u·it/they/she+continue/last
                       (NP-SBJ-3 *T*)
                       (PP-TMP لِ-·li-·for/to
                               (NP -ساعاتٍ·-sAE+At+K·hours))))
               وَ-·wa-·and
                (S (INTJ -أه·->ah·ah!/ouch!)
                   (EDITED (S (VP-UNF (PRT قَدْ·qado·may,_might,_perhaps
                                      تَكُونُ·ta+kuwn+u·it/they/she+be+[ind.]
                                      (NP-SBJ *)))
                           (INTJ أَوْ·>awo·Or,_if_not,_unless,_except_if))
                   (VP (PRT قَدْ·qado·may,_might,_perhaps,_maybe,)
                       يُرافِقُ-·yu+rAfiq+u-·he/it+accompany/escort+[ind.]
                       (NP-OBJ -ها·-hA·it/them/her)
                       (S-NOM-SBJ
                          (VP قِيامُ·qiyAm+u·undertaking/carrying_out
                              (NP-SBJ قُوّاتِ·quw~+At+i·armed_forces
                                      (NP ال{حْتِلالِ·Al+{iHotilAl+i·
                                                     the+occupation))
                              (PP-CLR بِ-·bi-·in,_at,_on,_by,_by_means_of
                                      (NP -تَصْعِيدٍ·-taSoEiyd+K·escalation
                                          أكْبَرَ·>akobar+a·
                                             larger/largest/greater)))))))))
```

يبدو ان العملية ستستمر لساعات واه قد تكون او قد يرافقها قيام قوات الاحتلال بتصعيد اكبر

yabodwA + >an~a + Al+Eamaly~apa + sa+tastamir~u + li+sAEAtK + wa + >ah + qad + takuwnu + >awo + qad + yurAfiqu+hA + qiyamu + quw~ati Al>iHtilAli + bi+taSoEidK + >akobara

it appear + that + the+operation + will+continue + for+hours + and + uh + might + she might + or + might + be accompanied + undertaking + the+forces + the+occupation + with+escalation + bigger

It appears that the operation will go on for hours and that uh it might be- or the occupation forces will escalate it further

```
(S (EDITED (NP-UNF أل·>l·NO_GLOSS))
   (NP-TPC-1 (NP الـقَـنـابِـلُ·Al+qanAbil+u·the+bombs/shells/grenades+[def.nom.]
                 الـعُـنْقُـودِيَّـةُ·Al+Eunoquwdiy~+ap+u·the+cluster_[bomb])
             (NP-ADV أساساً·>asAs+AF·primarily/basically+[acc.indef.]))
   (INTJ أه·>ah·ah!)
   (EDITED (VP أُعِدَّت·>uEid~+at·be_prepared/be_made_ready+it/they/she_[verb]
               (NP-SBJ-2 *)
               (NP-OBJ-2 *))
           (INTJ أَوْ·>awo·Or,_if_not,_unless,_except_if))
   (VP حُضِّرَت·HuD~ir+at·be_prepared+it/they/she_[verb]
       (NP-SBJ-1 *T*)
       (NP-OBJ-1 *)
       (PP-PRP لِ-·li-·for/to
               (NP (NP -ال{سْتِخْدام·-Al+{isotixodAm+i·the+usage/using)
                   (NP-ADV ضِدَّ·Did~+a·contrary/against/opposed/anti-/counter-
                           (NP الأَهْدافِ·Al+>ahodAf+i·the+targets+[def.gen.]
                               المَدَنِيَّةِ·Al+madaniy~+ap+i·the+civilian))))))
```

الـ القنابل العنقودية اساسا اه أعدّت أو حضرت للاستخدام ضد الاهداف المدنية

Al- + Al+qanAbilu + Al+Eunoquwdy~apu + >asAsAF + >ah + >uEid~at + >awo +
HuD`irat + li + Al}isotixdAmi + Did~a + Al>ahodAfi + Al+madany~api
the- + the+missiles + the_cluster + primarily + uh + be made + or + be prepared +
for+the+usage + against + the+targets + the+civil
The cluster missiles actually were made- or were put together to be used against civil targets

2.4.4 Internal structure of restarts (nesting of EDITED)

When a restarted constituent is repeated more than one time, all restarted elements are to be nested under one top node EDITED and organized hierarchically as follows:

Nesting of 1 EDITED:

```
(PP بِ-·bi-·by/with
    (NP (NP -النِسْبَةِ·-Al+nisob+ap+i·the+link/relation_[concerning/regarding])
        (EDITED (EDITED (PP-UNF لإنـتخ·l<ntx·NO_GLOSS -·-·nogloss))
                (PP-UNF لـل·ll·NO_GLOSS))
        (PP لِ-·li-·for/to
            (NP -القَوائِمِ·-Al+qawA}im+i·the+lists/indexes+[def.gen.]
                الأُخْرَى·Al+>uxoraY·the+other/another/additional))))
```

بالنسبة للإنتخ- لل للقوائم الأخرى

bi+Al+nisobapi + li+>inotx- + li + li+Al+qawA}imi + Al+>uxoraY
for+the+relation + for+the+elec- + for+the + for+the+lists + the+others
As for the elec-, for the for the other lists

```
(NP (NP الـقُـوّاتُ • Al+quw~+At+u·the+armed_forces+[fem.pl.]+[def.nom.])
    (EDITED (EDITED(ADJP-UNF الـع • AlE·NO_GLOSS -·-·nogloss))
            (ADJP-UNF ال • Al·NO_GLOSS))
    (INTJ أه • >ah·ah!/ouch!)
    (ADJP الْـمُـتَـعَـدِّدَةُ • Al+mutaEad~id+ap+u·the+multi-/poly-/manifold/numerous
          (NP الجِـنْـسِـيّـاتِ • Al+jinosiy~+At+i·the+nationality/citizenship)))
```

القوات العـ الـ أه المتعددة الجنسيات

Al+quwAtu + Al+E- + Al- + uh + Al+mutaEad~idapi + Al+jinosyAti
the+forces + the+A- + the- + uh + the+poly + the+nationalities
The A-, the-, uh, The multi-national forces

Nesting of 2 or more EDITED

```
(S (VP (PRT لَـمْ·lamo·did_not)
       يَـكُـنْ·ya+kun+o·he/it+be+[jus.]
       (NP-SBJ *)
       (SBAR-NOM-PRD (WHNP-1 0)
                     (S (VP مـتـؤكـدا·mt&kdA·NO_GLOSS
                           (NP-SBJ-1 *T*)
                           (UCP-MNR (NP تَـمـامـاً·tamAm+AF·completely)
                                    وَ-·wa-·and
                                   (PP -بِ-·-bi-·by/with
                                      (NP -{عْـتِـرافِ-·-{iEotirAf+i-
                                                  ·acknowledgment
                                          (NP -هِ·-hi·its/his))))
                           (SBAR
                               (EDITED
                                    (SBAR
                                        (EDITED
                                            (SBAR-UNF أَنَّ·>an~a·that))
                                        أَنَّ-·>an~a-·that
                                        (S-UNF (NP-SBJ -هـا·-hA·it))))
                               أَنَّ-·>an~a-·that
                               (S (NP-SBJ -هـا·-hA·it)
                                  (ADJP-PRD نِـهـائِـيَّـةٌ·nihA}iy~+ap+N·
                                                    final)))))))
```

لم يكن متؤكدا تماما وباعترافه أن أنها أنها نهائية

lam + yakuno + muta>ak~iDAF + tamamAF + wa+bi+>iEorirAfi+hi + >an~a + >an~a + hA+ >an~a + hA + nihA}iy~apN
not + be he + sure + completely + and+with+acknowledgment+his + that + that + it + that + it + final
He is not completely sure, and in his own words, that that it that it is final

```
(NP -إعادَةَ·-<iEAd+ap+a·return/repetition/re-[doing]+[fem.sg.]+[def.acc.]
    (INTJ أه·>ah·ah!)
    (EDITED (EDITED (EDITED (NP-UNF أل·>l·NO_GLOSS -·-·nogloss))
                    (NP-UNF أل·>l·NO_GLOSS -·-·nogloss))
            (NP-UNF أل·>l·NO_GLOSS))
    (NP البِناءِ·Al+binA'+i·the+structure/edifice/building+[def.gen.]))
```

إعادة أه الـ الـ الـ البناء

>iEAdapu + Al- + Al- + Al- + Al+binA'i
Redoing + the + the + the + the+building
the the the the rebuilding

```
(NP -الشِعار·-Al+$iEAr+i·the+slogan/motto+[def.gen.]
    (EDITED (EDITED (EDITED (ADJP-UNF أل·>l·NO_GLOSS -·-·nogloss))
                    (ADJP-UNF أل·>l·NO_GLOSS -·-·nogloss))
            (ADJP-UNF أل·>l·NO_GLOSS))
   السِياسِيِّ·Al+siyAsiy~+i·the+political+[def.gen.])
```

الشعار الـ الـ السياسي

Al+$iEari + Al- + Al- + Al- + Al+siyAsiy~i
the+slogan + the- + the- + the- + the+political
the the the the political slogan

```
(S (INTJ أه·>ah·ah!)
   (NP-TPC-1 (NP التِكْنُوقراطُ·Al+tikonuwqrAT+u·the+technocrat+[def.nom.])
             أوْ·>awo·Or,_if_not,_unless,_except_if
             (NP غَيْرُ·gayor+u·other+[def.nom.]
                 (NP التِكْنُوقراطِ·Al+tikonuwqrAT+i·the+technocrat+[def.gen.])))
   (VP (EDITED (EDITED (PRT سي·sy·NO_GLOSS -·-·nogloss))
               (PRT س·s·NO_GLOSS))
       (PRT سَ-·sa-·will)
       -يَفْشَلُونَ·-ya+fo$al+uwna·they_[people]+fail/be_unsuccessful+[masc.pl.]
       (NP-SBJ-1 *T*)))
```

أه التكنوقراط أو غير التكنوقراط سي- سـ- سيفشلون

>ah + Al+tikonuwqrAt + >awo + gayori + Al+tikonuwqrAt + siy + sa- + sa+yafo$aluwna
uh + the+technocrats + or + other + the+technocrats + will- + will- + will+fail
Uh, the technocrats and the non-technocrats will- will- will fail

```
(S (INTJ أه • <h·ah)
   (EDITED (EDITED (EDITED وع • wE·NO_GLOSS -·-·nogloss)
                   وَ-·wa-·and)
           (VP-UNF -استخدم • -Astxdm·NO_GLOSS)
           (INTJ وَ-·wa-·and))
   (VP -يَسْتَخْدِمه • -yasotaxodimh·NO_GLOSS
       (NP-SBJ *)
       (NP-OBJ (NP المُسْتَقِلِّينَ • Al+musotaqil~+iyna·the+independent/autonomous)
               (NP-ADV خارجَ • xArij+a·outside/outer_part+[def.acc.]
                       (NP الإخْوان • Al+<ixowAn+i·the+brothers+[def.gen.]))))))
```

أه وعـ واستخدم ويستخدمه المستقلون خارج الإخوان

>ah + waE + wa+Aistxdm + wa+yasotaxodimu+hu + Al+musotaqil~uwna + xArija + Al+>ixowAni

uh + NO_GLOSS + and+NO_GLOSS + and+use+it + the+independents + outside + the +Brothers

Uh, and the independent candidates other than the Brothers, use- use- used it

3 Special syntactic constructions in ATB broadcast corpora

This section presents annotation guidelines of usages and structures that are:

- specific to spoken language like greeting, thanking, etc., or
- specific to certain broadcast stylistic choices (vs. newswire style)
- specific to certain new MSA usages, or
- simply miscellaneous features that we came across during annotation of ATB5.

3.1 MSA greetings, thanking phrases, vocatives and interjections

The following are common and high frequency expressions in spoken MSA with TB solutions. The list is not exhaustive and is not meant to be so. If annotators come across a structure that cannot fit in one of these patterns, **they should bring it to discussion before deciding on their annotation.**

Arabic/Buckwalter	Gloss	TB
Greetings: All greetings are annotated separately from any further text, whether delimited by final punctuation or not. Example: أهلا أنا رشيد/ AhlAF >anA ra$iyd/ Hi I am Rasheed is annotated as follows: (NP AhlAF اهلا) (S (NP-SBJ >anA أنا) (NP-PRD ra$iyd رَشيد))		
أهلا/>aHolAF	Welcome	(NP AholAF اهلا)

مرحبا/ maroHabAF	Welcome	(NP maroHabAF مَرْحَباً)
أهلا وسهلا/>aHolAF wa saholAF	welcome and welcome	(NP >aholAF أهْلاً wa وَ saholAF سَهْلاً)
أهلا بكم/>AholAF bikum	welcome+with (to)+you	(NP (NP >AholAF أاهْلاً) (PP bi بِ (NP kum كُم)))
صباح الخير/SabAhu Alxayori (idhafa)	morning+ goodness	(NP SabAHu صَباحُ (NP Alxayori الخَيْرِ))
مساء الخير/masA'u Alxayori (idhafa)	evening+ goodness	(NP masA'u مَساءُ (NP Alxayori الخَيْرِ))
أمسية طيبة/>umsyap Tay~ibap	evening+nice	(NP >umsyap أمسيَة Tay~ibap طَيِّبَة)
مع السلامة/maEa AlsalAmap	with+safety	(NP-ADV maEa مَعَ (NP AlsalAmap السَلامَة))
إلى اللقاء/<ilaY+AlliqA'	to/until+meeting	(PP <ilaY إلَى (NP AlliqA'i اللِقاءِ))
Thanking:		
شكرا/$ukrAF	Thank	(NP $ukrAF شُكراً)
شكرا لك/$ukrAF la+ka	thank to you	(NP (NP $ukrAF شُكراً) (PP la لَ (NP ka كَ)))
شكرا جزيلا لك/$ukrAF jaziylAF la+ka	thank abundant to you	(NP (NP $ukrAF شُكراً jaziylAF جَزِيلاً) (PP la لَ (NP ka كَ)))
Interjections: Unless clearly separated by a final punctuation, interjections are annotated inside the next sentence unit. For a complete list of interjections refer to POS guidelines section **4.3.1.8**.		
نعم/naEam	Yes	(INTJ naEam نَعَم)
لا/lA	No	(INTJ lA ••)
Vocatives: Unless clearly separated by a final punctuation, vocatives are annotated inside the next sentence unit. They have to carry the dash-tag –VOC even if they come up as a separate unit.		
مشاهدينا/mu$Ahidiy+nA	Viewers+our	(NP-VOC mu$Ahidiy مُشاهِدِي (NP nA نا))

مشاهدينا الكرام / mu$Ahidiy+nA AlkirAm	Viewers+our the distinguished/ Noble	(NP-VOC (NP mu$Ahidiy مُشاهِدِي (NP nA نا)) (ADJP AlkirAm الكِرام))
أعزائي المشاهدين />aEiz~A}+iy Almu$Ahidiyn	dear/precious+ my+the viewers	(NP-VOC (NP mu$Ahidiy مُشاهِدِي (NP nA نا)) (ADJP AlkirAm الكِرام))

3.2 MSA Filler phrases and clauses

Filler phrases and clauses are to be annotated according to the nature of their internal structure. Depending on their position in a sentence, they are annotated as PRN when they appear in positions that disrupt the basic syntax of certain structures (e.g., between NP complements).

The following are examples of typical filler phrases and clauses in MSA:

3.2.1 Phrases

الله All~Ah God! → `(NP-VOC Allh الله )`

والله wa+All~Ah by God! →

```
(PP wa وَ
     (NP Allhi اللهِ ) )
```

فعلا FiEolAF indeed → `(NP-ADV fiEolAF فِعْلاً )`

طيب Tay~ib good/well → `(NP-ADV Tay~ib طَيِّب )`

طبTabo (non-MSA) good/ well → `(NP-ADV Tab طب )`

الحقيقة AlHaqiyqap the truth (truth be said) → `(NP-ADV AlHaqiyqap الحَقِيقَة )`

3.2.2 Ss and SBARs

يعني yaEoniy (he/it) means

```
(PRN (S (VP yaEoniy يَعْنِي
            (NP-SBJ * ) ) ) )
```

إن شاء الله <in $A'a Al~Ah if God wills

```
(SBAR-ADV <in إن
          (S (VP $A'a شاءَ
                 (NP-SBJ Al~Ahu الّاهُ ) ) ) )
```

لا سمح الله lA samaHa Al~Ah God forbid

```
(S-ADV (VP (PRT lA لا )
           samaHa سَمَحَ
           (NP-SBJ Al~Ahu اللّاهُ ) ) )
```

الحمد لله AlHamodu lil~Ah thank to God

```
(S-ADV (NP-SBJ AlHamodu الحَمْدُ )
       (PP-PRD li لِ
               (NP llhi لله ) ) )
```

4 Dialectal items in broadcast news

Due to the diglossic nature of the Arabic language, the use of different vernaculars even in contexts where speakers are supposed to use only MSA leads to the presence of lexical items and syntactic structures in our corpora that are not shared with MSA. Our global decision on how to deal with dialectal items in our overall MSA data is not to ignore these sections in dialect and to treat dialectal items as described in the coming sections.

4.1 Partial use of dialect

4.1.1 Single lexical items

At POS level, if speakers use single words that are not necessarily specific to dialects and have an equivalent unvoweled form in MSA and in SAMA (Maamouri, et al. 2009), we decided to give those words an MSA POS value, even if its voweled form is slightly diverging from the phonetic utterance of the word by the speaker.

If the case ending is dropped by the speaker, POS annotators give the most suitable case ending according to syntactic context.

In the below utterance, some use of pure Egyptian words like ده dah (this) and the progressive form lower بتعمل bi+tiEimilo (is doing) tell us that the speaker is mixing Egyptian dialect and MSA, but as you see from the POS analysis, most of tokens are given MSA POS values:

```
(S (S (EDITED (S (EDITED (EDITED (NP-UNF أل·>l·NO_GLOSS -·-·nogloss))
                         (NP-UNF أل·>l·NO_GLOSS))
                 (NP-TPC-2 (NP السُيُوف·Al+suyuwf·the+swords)
                           وَ·wa·and
                          (NP (NP كُلُّ·kul~+u·all,_every,_whole+[def.nom.])
                              (NP ده·dh·this)))
                 (VP كانَ·kAn+a·be/was/were+he/it_[verb]
                     (VP-UNF يَسْتَخْدِمُ-·ya+sotaxodim+u-·he/it+use/utilize/employ
                             (NP-OBJ (NP -هُ·-hu·it/him)
                                     (NP-2 *T*))))))
      (VP يَسْتَخْدِمُ-·ya+sotaxodim+u-·he/it+use/utilize/employ/operate+[ind.]
          (NP-OBJ -هُ·-hu·it/him)
          (EDITED (NP-SBJ (NP مُرَشَّحِي·mura$~aH+iy·candidate/nominee
                              (NP الإخْوان·Al+<ixowAn+i·the+brothers))
                          (ADVP أيْضاً·>ayoDAF·also,_too,_as_well_[as])))
          (NP-SBJ مُرَشَّحِي·mura$~aH+iy·candidate/nominee+[masc.pl.gen.]
                  (NP الإخْوان·Al+<ixowAn+i·the+brothers+[def.gen.]))))
   وَ-·wa-·and
   (S (NP-PRD -عِنْدَ-·-Einod+a-·with/at+[def.acc.]
              (NP -نا·-nA·our))
      (NP-SBJ صُوَرٌ·Suwar+N·pictures/photographs/photos+[indef.nom.])
      (EDITED (PP-UNF ل·l·NO_GLOSS))
      (PP لِ-·li-·for/to
          (NP (NP (NP -الناس·-Al+nAs·the+people)
                  (NP دي·dy·this))
              (NP كُلِّ-·kul~+i-·all,_entire,_every,_whole+[def.gen.]
                  (NP -ها·-hA·its/their/her))))
      (SBAR-ADV (EDITED (SBAR-ADV وَ-·wa-·while
                                 (S-UNF (NP-SBJ -هِيَ·-hiya·it/they/she))))
               وَ-·wa-·while
               (S (NP-TPC-1 -هِيَ·-hiya·it/they/she)
                  (VP (PRT ب-·b-·progressive)
                      -تِعمَل·-tiEmal·do
                      (NP-SBJ-1 *T*)
                      (EDITED (NP-OBJ-UNF (NP الكَلام·Al+kalAm·
                                                  the+speech/statement)
                                          (NP ده·dh·this)))
                      (NP-OBJ (NP الكَلام·Al+kalAm·the+speech/statement)
                              (NP ده·dh·this))))))))
```

4.1.2 Substitution of single MSA words by dialectal words within an overall MSA structure

If speakers decide to use a word in its dialectal form instead of its MSA most appropriate counterpart like إحنا <iHonA instead of نحن naHonu (we), that single word is given a minimal POS value as a dialectal word and is treated at Treebank level like its counterpart in MSA.

```
(NP  ده dah/ DEM_PRON)
```

If the use of dialect consists only in using a slightly phonemic variation of an MSA word, or in substituting single lexical items by their counterparts in their own dialects, this remains of very little effect on our Treebank annotation.

4.2 Entire sentences in dialect

It happens in interviews and reports that speakers chose to speak in their vernacular using lexical as well as syntactic dialect-specific items. That leads in our Treebank data to the presence of entire sentences in dialect. In this case, annotators should use POS values to guide them in their syntactic decisions. For an overview about POS value of most frequent dialect words found in ATB5 refer to section ????? in the POS guidelines.

We had to agree, however, on policies regarding certain specific syntactic features.

4.2.1 Some annotation policies for certain dialectal structures

4.2.1.1 The use of bid~ in Levantine

The Levantine word: بد *bid~* (wish) like in بدو يصير محام bid~w ySiyr muHAmiy (his wish is to become a lawyer), bid~hA fustAn jdyd, بدها فستان جديد (her wish is a new dress, she wants to have a new dress) is treated in Treebank annotation as a verb, and as such it takes either a subject and an S complement, or a subject and an object, even though it does not morphologically inflect like a verb. See the Arabic Treebank Annotation Guidelines for more on pseudo-verbs in the Arabic Treebank.

With clausal complements:

a. If the clausal complement has a pro-drop subject:

```
(S (VP bid~
       (NP-SBJ-1 w)
       (S (VP ySiyr
              (NP-SBJ-1 *)
              (NP-PRD muhAmiy)))))
```

بدّو يصير محامي

bid~+w + ySiyr + muHamiy
wish+his + become + lawyer
He wants to become a lawyer

b. or an overt subject (if it ever happens)

```
(S (VP bid~
       (NP-SBJ w)
       (S (VP ySiyr
              (NP-SBJ <ibon
                      (NP w))
              (NP-PRD muhAmiy)))))
```

بدّو يصير إبنو محامي

bid~+w + ySiyr + <ibon+w + muHamiy
wish+his + become + son+his + lawyer
He wants his son to become a lawyer

c. If the clausal complement has topicalized subject:

```
(S (VP bid~
       (NP-SBJ w)
       (S (NP-TPC-1 <ibon
                    (NP w))
          (VP ySiyr
              (NP-SBJ-1 *T*)
              (NP-PRD muhAmiy)))))
```

بدّو إبنو يصير محامي

bid~+w + <ibon+w + ySiyr + muHamiy
wish+his + son+his + become + lawyer
He wants his son to become a lawyer

If it takes an object:

```
  (S (VP bad~
         (NP-SBJ hA)
         (NP-OBJ fusotAn
                 jdyd)))
```

بدّها فستان جديد

bad~+hA + fusotAn + jdiyd
wish+her + dress + new
She wants a new dress

4.2.1.2 The use of ‘fyh’

Analysis of ***fyh/fy***: fy used as dialectal with the meaning of ‘there is’ as in: فيه صدامات يومية fih SidAmAt yawomy~ap (in it daily clashes/ there are daily clashes) should be transcribed as ‘fiy+h and TB annotated as following:

```
(S (PP-PRD fy-
          (NP-h))
   (NP-SBJ SidamAt
          yaomy~ap))
```

فيه صدامات يوميّة

fiy+h + SidAmAt + yawomy~ap
in+it + clashes + daily
There are daily clashes

fyh transcribed as fy only and the –h clitic is missing is annotated as follows:

```
(S (PP-PRD-UNF  fy)
   (NP -SBJ SidamAt
           yaomy~ap))
```

في صدامات يوميّة

Fiy + SidAmAt + yawomy~ap
in (it) + clashes + daily
There are daily clashes

4.2.1.3 Annotation of progressive particle in Levantine, Iraqi and Egyptian dialects

```
(S (INTJ أه·>ah·ah!)
   (NP-ADV الحَقِيقَةُ·Al+Haqiyq+ap+u·the+truth/reality+[fem.sg.]+[def.nom.])
   (NP-TPC-1 أنْتَ·>anota·you_[masc.sg.])
   (VP (PRT ب-·b-·is)
       -تِسْأَلْ-·-tiso>alo-·asking
       (NP-SBJ-1 *T*)
       (NP-DTV -نِي·-niy·me)
       (NP-OBJ سُؤالَيْن·su&Al+ayoni·question/inquiry+two_[acc.])))
```

أه الحقيقة إنت بتسألني سؤالين

Al+Haqiyqap + >anota + bi+tiso>alo+niy + su&Aliyn
the+truth + you + in instance of + ask+me + question two
Truth is, you are asking me two questions (not one)

4.2.1.4 Annotation of active participle in Levantine, Iraqi and Egyptian dialects

The active participle in dialectal forms of Levantine, Iraqi and Egyptian Arabic is treated in exactly the same way as active participles in MSA, and according to the ATB MSA guidelines[1].

```
(S (VP يُمْكِنْ • yu+mokin+o·he/it+be_possible/make_possible_for+[jus.]
       (SBAR-SBJ *0*
                 (S (VP يَكُونُ • ya+kuwn+u·he/it+be+[ind.]
                        (VP تُوُفِّيَ • tuwuf~iy+a·die/pass_away/expire+he/it_[verb]
                            (NP-SBJ-1 *)
                            (NP-OBJ-1 *)
                            (NP-TMP اليَوْم • Al+yawom·the+today)))))))
(S (NP-SBJ *)
   (PRT مش • m$·not)
   (ADJP-PRD عارِفِينَ • EArif+iyna·knowing/having_knowledge_of))
```

يمكن يكون توفيّ اليوم مش عارفين

yumokinu + yakuwn + tuwuf~iya + Al+yawom + mu$ + EArfiyn
be possible + he be + pass away + the+today + not + knowing us
Maybe he could have died today. We don't know.

5 References

Arabic Treebank Morphological and Syntactic Annotation Guidelines. 2008. Mohamed Maamouri, Ann Bies, Sondos Krouna, Fatma Gaddeche and Basma Bouziri. http://projects.ldc.upenn.edu/ArabicTreebank/. Linguistic Data Consortium, University of Pennsylvania.

Ann Bies, Mark Ferguson, Karen Katz and Robert MacIntyre (Eds.). 1995. *Bracketing Guidelines for Treebank II Style*. Penn Treebank Project, University of Pennsylvania, CIS Technical Report MS-CIS-95-06.

Mohamed Maamouri, Ann Bies, Seth Kulick and Fatma Gaddeche. 2009. *Arabic Treebank part 5 - v1.0* (ATB5), LDC Catalog Number: LDC2009E72. Linguistic Data Consortium, University of Pennsylvania.

Mohamed Maamouri, David Graff, Basma Bouziri, Sondos Krouna and Seth Kulick. 2009. *LDC Standard Arabic Morphological Analyzer (SAMA) v. 3.0*. LDC Catalog No.: LDC2009E44. Special GALE release to be followed by a full LDC publication.

Ann Taylor. 1996. *Bracketing Switchboard: An addendum to the TREEBANK II Bracketing Guidelines*. Penn Treebank Project, University of Pennsylvania.

[1] In previous work on the JHU Levantine corpus, we gave a different treatment to the active participle in dialect. It is clear that the more we work on Arabic dialects, the more we will know to what extent different treatments for dialect forms will be necessary.