ACE (Automatic Content Extraction)
Arabic Annotation Guidelines for Values

Version 1.2.3 – 2005.05.31

Linguistic Data Consortium

http://www.ldc.upenn.edu/Projects/ACE/

1. Basic Concepts

1.1 What is a Value?

A Value is a string that further characterizes the properties of some Entity or Event.

We will only be interested in certain types of possible Values. Specifically, we will be annotating NUMERIC, CONTACT-INFO, TIMEX2, JOB-TITLE, CRIME and SENTENCE Values.

Additionally, we will only be interested in a subset of the possible subtypes for most of these types. For example, only MONEY and PERCENT will be annotated as NUMERIC Values. A complete list of subtypes for each of the Value types is provided in the section describing that type.

There will be no subtypes for the SENTENCE, CRIME and JOB-TITLE Value types. All examples of legal punishments, legal offenses and employment positions, respectively, will be taggable Values of these types whenever they are required as arguments of Events.

It is important to note that there are really two types of things that we are calling Values here:

1. Strings that provide potential characterizing information about entities. These strings will always be tagged when they occur in a document. They need not participate in any Relation or Event.
2. Strings that participate as arguments in Events. These strings are only tagged when they occur within the scope of a taggable Event.

	Entity Characterizing Values	***Event Argument Values***
TYPE ***SUBTYPE***	NUMERIC PERCENT MONEY CONTACT-INFO PHONE-NUMBER URL EMAIL TIMEX2	JOB-TITLE SENTENCE CRIME
TAGGABLE?	Always tagged when mentioned	Only tagged when used as an argument in an Event

1.2 Extent

The rules for identifying the extent of a Value mention will vary from type to type. The specific extent rules for a given type will be provided in the section describing that type. There are, however, some general properties of all Value extents that can be mentioned up front.

Many Values are mentioned by a noun phrase (NP). The Entity task guidelines (English-Entities-Guidelines_v5.4.doc) introduce a detailed account of the manner in which the full extent of an NP can be identified. That account is repeated here for convenience.

The extent of a mention consists of the entire NP. In case of structures where there is some irresolvable ambiguity as to the attachment of modifiers, the extent annotated should be maximally inclusive. In the case of a discontinuous constituent, the extent goes to the end of the constituent, even if that means including tokens that are not part of the constituent. Thus, in:

> *The terrorist was charged with conspiracy yesterday in the bombing of the USS Cole.*
>
> اتهم الارهابي بالتامر امس في تفجير الغواصة يو اس اس كول

the extent of the mention is the entire NP:

> *[conspiracy yesterday in the bombing of the USS Cole]*
>
> (بالتامر امس في تفجير الغواصة يو اس اس كول)

The extent includes all the modifiers of a NP, including prepositional phrases and relative clauses.

Generally speaking, tokens are broken at white space, and each item of punctuation is treated as a separate character. As a rule, we do not include punctuation such as commas, periods, and quotation marks in the extent of a mention unless words included within the extent continue on after the punctuation mark. Possessive endings ('s) are treated as separate tokens, and contractions are split (so that "*we're*" becomes the two tokens "*we*" and "*'re*"). Extents must begin at the beginning of a token and end at the end of a token.

These general rules will be most useful in identifying the extent of JOB-TITLE and CRIME Values. For the others, the specific rules may prove most useful, since many of these Values are expressed with highly formulaic constructions (e.g. a U.S. phone number will almost always be represented as *(XXX) XXX-XXXX* or *1-XXX-XXX-XXXX*)

When annotating Values, it will not be necessary to indicate a head. Identification of the extent will suffice.

2. Numeric Values

NUMERIC Values will be limited to the subtypes MONEY and PERCENT.

For NUMERIC Values there will be two important parts of the mention:

The first part is the *indicator,* which is used to express the subtype of the NUMERIC Value. For example, the symbol % would be an indicator that a NUMERIC Value is of the PERCENT subtype. This can be expressed either as a symbol or as a string of words.

The second part is the *number*. This can be either a numeral or a string of words expressing a number.

For *NUMERIC* Values, the extent will be the smallest string of words that includes both the *number* and the *indicator* and also any additional quantifiers that might be present such as '*nearly*', '*almost*' and '*over*'. The number and indicator may occur in either order and need not be contiguous. The usual rules about whitespace and punctuation apply (see Section 1.2, above). The usual rules about annotating the full extent of the NP **do not**. We will only annotate the extent relevant to the NUMERIC Value itself. For instance:

[25%] of all respondents
(%25) من جميع المستجيبين
[almost $400 Million] in stock
(حوالي 400 مليون دولار) احتياطي

2.1 Percent

A PERCENT Value is mentioned whenever numeric information is presented as a fraction of one-hundred (100).

[Twenty-five Percent]
(خمسة وعشرون بالمائه)
[over 25 Per Cent]
(اكثر من خمسة وعشرون بالمائه)
[25%]
(%25)

2.2 Money

A MONEY Value is mentioned whenever capital is described in terms of the currency of some country or region (e.g. *US Dollars* or *Euros*).

> *[nearly $400 Million]*
> (400 مليون دولار تقريبا)
> *[50 dollars]*
> (50 دولارا)
> *[20 Euros]*
> (20 يورو)

3. Contact-Info

CONTACT-INFO Values will be limited to examples of the PHONE-NUMBER, EMAIL and URL subtypes.

The extent for mentions of CONTACT-INFO Values will be described independently for each of the subtypes.

3.1 Phone-Number

A PHONE-NUMBER Value is mentioned whenever there is a string of numerals that can be used to make contact via phone.

The extent of a PHONE-NUMBER mention is the smallest sequence of tokens such that all of the numerals in the string are included. This will frequently include internal punctuation such as '-', '.', '(' and ')' and prefixes such as '+'.

> *[215.555.1111]*
>
> *[(215) 555-1111]*
>
> *[+44 23 345-1234]*

3.2 Email

An EMAIL Value is mentioned whenever there is a string of tokens that can be used to make contact via electronic mail.

The extent of an EMAIL mention is the smallest sequence of tokens such that all of the tokens in the string are included. EMAIL Value mentions should be contiguous, but will be punctuated by the symbols '@' and '.'.

> *chwalker@ldc.upenn.edu*

president@whitehouse.gov

vice.president@whitehouse.gov

3.3 URL

A URL Value is mentioned whenever the virtual location of a webpage is provided. It is not important that the webpage be the front page (or index page) of some site. A URL mention can point to any page directly accessible using a web-browser and the URL mentioned.

The extent of an URL mention is the smallest sequence of tokens such that all of the tokens in the string are included. URL Value mentions should be contiguous but will contain tokens of various types (e.g. symbols, numbers and letters).

http://www.ldc.upenn.edu/Projects/ACE/

www.ldc.upenn.edu/Projects /ACE/

google.com

4. Time

The extent of a TIME mention will be identified as in the TIMEX2 annotation guidelines. The general rules for identifying NP extents provided in Section 1.2 above can serve as a guide in making decisions about TIME mention extents.

For a more detailed discussion, please consult the TIMEX2 guidelines (Ferro, L., Gerber, L., Mani, I., Sundheim, B. and Wilson G. (2005). "TIDES 2005 Standard for the Annotation of Temporal Expressions" April 2005).

The sponsor arrived at [ten minutes to 3].

وصل الممول الساعة (الثالثة الا عشر دقائق)

I returned to work at [twelve o’clock January 3, 1984].

عدت الي العمل في تمام الساعة (الثانية عشر يوم ثلاثه يناير 1984)

On the nineteenth I am in class until [eleven in the morning].

في التاسع عشر ساكون في النادي حتي (الحادية عشر صباحا)

At [11:59 p.m.], Mayor Rudolph W. Giuliani sat on a stage at 45th Street and Broadway and pushed a button.

[April 11, 1996 11:13 GMT]
(11 ابريل 1996 الساعة 11:13)

The two collaborated closely during the [1994] crisis over Haiti.
تعاون الاثنين خلال اضطرابات عام (1994) في هاييتي

After an emergency meeting in [November], relations began to improve.

بعد اجتماع طارئ في (نوفمبر) بدأت العلاقات في التحسن

I was sick [yesterday].

كنت مريضا (بالامس)

The bombing took place on [the second of December].

بدأ القذف (الثاني من ديسمبر)

Dancing deteriorated in [the 1960s] into group chaos.

انحدر الرقص في (التسعينات)

NATO is debating how the Atlantic security partnership should define its strategic interests for [the next century].

الناتو تعتزم مناقشة خططها في (القرن القادم)

Angkor Wat, the fabled [11th century] temple, is Cambodia's main tourist attraction.

(القرن العشرين) اعظم عصور النهضة

We are entering what is popularly regarded as [the last year of [this millennium]].
نحن الان في مقتبل (العام الاخير من (هذا القرن))

Note: We will also use the TIMEX2 standard as the basis for TIME normalization, but we will not perform this process at the same time as Value annotation.

5. Crime

A CRIME Value will be mentioned whenever the offense associated with some JUSTICE Event is explicitly expressed. **Note that there must be a taggable instance of the corresponding JUSTICE Event for there to be a taggable CRIME Value.**

Since most CRIME Value mentions will be expressed in the form of an NP, the extent of CRIME Value mentions will be defined generally, as in Section 1.2 above. In the examples that follow, square brackets are used to indicate the CRIME Value mention and **bold** font is used to indicate the trigger of the corresponding JUSTICE Event.

*The terrorist was **charged** with [conspiracy in the bombing of the USS Cole].*

اتهم الارهابي (بالتامر في تفجير الغواصه يو اس اس كول)

*46-year-old Abu Talib was **sentenced** to life imprisonment in 1990 in Sweden for [terrorist acts in Amsterdam, Copenhagen and Stockholm between 1985 and 1986].*

حكم علي ابو طالب 46 عام بالسجن مدي الحياة ....

6. Sentence

A SENTENCE Value will be mentioned whenever a punishment for some JUSTICE Event is explicitly expressed. **Note that there must be a taggable instance of the corresponding JUSTICE Event for there to be a taggable SENTENCE Value.**

Since most SENTENCE Value mentions will be expressed in the form of an NP, the extent of SENTENCE Value mentions will be defined generally, as in Section 1.2 above. In the examples that follow, square brackets are used to indicate the SENTENCE Value mention and **bold** font is used to indicate the trigger of the corresponding JUSTICE Event.

*46-year-old Abu Talib was **sentenced** to [life imprisonment] in 1990 in Sweden for terrorist acts in Amsterdam, Copenhagen and Stockholm between 1985 and 1986.*

حكم علي ابو طالب 46 عام **بعقوبة** السجن مدي الحياة ....

*Solomon could be **sentenced** to [up to 211 years in prison].*

يمكن ان يواجه سليمان **عقوبة** (تصل الي 211 عام سجن)

*Hutomo ``Tommy'' Mandala Putra, 37, was **sentenced** to [18 months in prison] on Sept. 22 by the Supreme Court, which overturned an earlier acquittal by a lower court.*

حكم علي حمدي 37 عام **بعقوبه** (18 عام في السجن)...........

7. Job-Title

A JOB-TITLE Value will be mentioned whenever the office associated with some PERSONNEL Event is explicitly expressed. (For a complete discussion of Events and Event scopes, please see the Events Guidelines).

Please note that JOB-TITLE Value mentions will often be co-extensive with PERSON Entity mentions. When this happens, both the Value and the Entity will be annotated. For a complete discussion of the annotation of entities, please see the Entity Guidelines.

Since most JOB-TITLE Value mentions will be expressed in the form of an NP, the extent of JOB-TITLE Value mentions will be defined generally, as in Section 1.2 above. In the examples that follow, square brackets are used to indicate the JOB-TITLE Value mention and **bold** font is used to indicate the trigger of the corresponding PERSONNEL Event.

> *The company **hired** Steve Jobs as [the new CEO].*
> عينت الشركة ستيف (كرئيس مجلس ادارة)
> *In 1997, the company **hired** John D. Idol to take over as [chief executive].*
> في عام 1997 **عينت** الشركة جون ليشغل منصب (المدير التنفيذي)
> *The question of which party controls the Texas Senate is especially important this year because the Senate will redraw congressional and legislative districts and could elect the next lieutenant governor if [Gov.] George W. Bush is elected president and is **succeeded** by Lt. Gov. Rick Perry.*
> جورج بوش انتخب كرئيس و **نجح** بمجهود الحزب