Coreference Guidelines

coreference	إشارة متبادلة	הוראה משותפת
antecedent	اسم سابق يعود عليه ضمير لاحق	קודמן
pronoun, anaphora	ضمير، ضمير انعكاسي	כינוי חבור, אנאפורה
cataphora/cataphoric pronoun	ضمير يسبق الاسم الذي يعود عليه	אזכור לְפָנִים, קטאפורה
apposition	بَدَل	תמורה, ביטוי תמורה
ellipsis, zero-anaphora	حذف نحوي، ضمير محذوف/مستتر	הֶשְׁמֵט
index	مؤشّر	צַיָן, אִינְדֵקְס
mention	ذِكر	אזכור
nominalization	مصدر الفعل	העצמה
NP - noun phrase	مركّب اسميّ	צירוף שמני
deictic element	"عنصر إشاريّ"، يعتمد على السياق	ישות דאיקטית
quantifier, quantification	مُكمّم، تكميم	כמת

The aim of this project is to collect data on the way context is constructed and used in natural language, particularly with regard to anaphoric reference and deictic reference.

Basic concepts in syntax:

1. **Index:** a number that aims to clarify the relations between nouns and pronouns. Two morphemes that share the same index refer to the same entity.
   - أحبت [الطالبة]$_1$ [أخت[ها]$_1$]$_2$ المجتهدة، [سامية]$_2$.
   - يعتقد [زيد]$_1$ أن [سميرة]$_2$ تحب أن تساعد [___]$_2$ في عمل[ـه]$_1$.
   - جمعية الصليب الأحمر قدمت هدية لـ[لطبيبة]$_1$ التي قال الصحفيون أنـ[ـها]$_1$ ساعدت الجرحى.
   - [התלמידה]$_1$ אוהבת את [אחות]$_2$[ה]$_1$ החרוצה, [יסמין]$_2$
   - [גדעון]$_1$ חושב ש[סמירה]$_2$ תשמח [___]$_2$ לעזור ל[ו]$_1$ בעבודת[ו]$_1$.
   - [שרון]$_1$ הבטיח ל[רקפת]$_2$ ש[הוא]$_1$ בא ל[מסיבה]$_3$ של[ה]$_2$ ש[___]$_3$ מתוכננת למחר
   -
2. **Antecedent:** generally, an antecedent is a noun phrase (NP) headed by a common noun, proper noun, or a pronoun and it refers to a specific entity in context. Subsequent NPs such as the pronoun "هو" and the clitic pronouns "ـه" as in the first example refer to the definite NP "كيس البطاطا".
   - يجب أن تحمل [كيس البطاطا]$_1$ وأن تضع[ـه]$_1$ في الثلاجة. احذر فـ[هو]$_1$ ثقيل.
   - ستشهد [فكتوريا تشن]$_1$، المديرة المالية لMegabucks Banking، ارتفاعًا ملحوظًا في راتب[ـها]$_1$ بعد توليـ[ـها]$_1$ المنصب. (Jurafsky and Martin, 2023)
   - אתה צריך לקחת את [שק תפוחי האדמה הזה]$_1$ ולהכניס אות[ו]$_1$ למזווה. תיזהר, [זה]$_1$ אחד כבד.
   - [ויקטוריה צ'ן]$_1$, סמנכ"לית הכספים של [מגאבקס בנקינג]$_2$, תראה עלייה משמעותית בשכר של[ה]$_1$ לאחר ההשתלטות על [החברה]$_2$.

3. **Anaphora:** this category encompasses all types of pronouns; nominative subject pronouns (هو), accusative/genitive clitics (ـه، ني، ـها), reflexives (نفسي، نفسه، أنفسهم), or separate accusative pronouns (إياه، إياي، إياكم) that follow the NP with which they corefer.
   Anaphoras are called "context-dependent entities"; understanding them and analyzing their relation to the other entities mentioned before usually depends on the context and on the previous NP to which they refer, directly or indirectly. It is also called a "deictic element" because it refers to/points at a single, context-specific NP.
   - بعد أن توفي الشاعر [محمود درويش]$_1$، بدأت بقراءة قصائد[ه]$_1$.
   - [الكثيرون]$_1$ اعتبرو[ه]$_2$ [منقذًا]$_2$ لـ[هم]$_1$.
   - رأت أم [الرجل]$_1$ [ابنتـ[ـه]$_1$]$_2$ و[هي]$_2$ تسقط أرضًا.
   - قالت [شركة إلكو]$_1$ أنـ[ـها]$_1$ تتوقع أن تُقدّر أرباحـ[ـها]$_1$ بـ1.65 مليون دولار.
   - לאחר שהמשורר [מחמוד דרוויש]$_1$ נפטר, התחלתי לקרוא את שירי[ו]$_1$.
   - אמ[ו]$_1$ של [אמנון]$_1$ ראתה את [בת]$_2$[ו]$_1$ בדרכ[ה]$_2$ אל בית הספר.
   - מ[החברה]$_1$ נמסר כי רווחי[ה]$_1$ יעמדו על 1.65 מיליון דולר.
   - [רבים]$_1$ רואים ב[ו]$_2$ [המושיע]$_2$ של[הם]$_1$.

4. **Cataphora:** a special case of pronouns that precede the noun with which they corefer, usually because of a syntactic movement—the movement of part of a sentence to the beginning.
   - بعد تلقيـ[ـها]$_1$ بلاغًا عن بائعي مخدرات، حضرت [الشرطة]$_1$ إلى المكان المشبوه.

5. **Ellipsis, zero-anaphora:** this type of pronoun is usually omitted due to the conjugation of the verb as in the first and second examples; the conjugation of the jussive verb "تكون" and the imperative verb "ساعِد" obviates the mention of the pronouns "هي" and "أنتَ". In generative linguistics, the deleted pronoun is called PRO.

   - رشَّح أوباما كلينتون لتكون ___ وزيرة خارجيته يوم الإثنين. اختارها ___ لأنها ذات خبرة في مجال الشؤون الخارجية. (Al-Oraini et. al, 2020)
   - ساعِد ___ الناس! (فعل أمر، الضمير "أنت" محذوف)
   - تمنى المعلم أن يكتب ___ النص.
   - ذهب [بوش]$_1$ إلى موسكو لمقابلة [بوتين]$_2$ وتناقش [___]$_1$ معـ[ـه]$_2$ في عدة مواضيع مهمة وأشار [___]$_1$ إلى أهمية العلاقات الأمريكية-الروسية. (Al-Oraini et. al, 2022)
   - היום הצעירים חכמים יותר - ___ כבר לא הולכים להוראה.
   - ארגון הצלב האדום העניק פרס ל[רופאה] שהעיתונאים אמרו ש[___]עזרה לפצועים.
   - [גרשון]$_1$ חושב ש[סמירה]$_2$ תשמח [___]$_2$ לעזור ל[ו]$_1$ בעבודת[ו]$_1$.
   - כש[___]ראה את החשבון, [פנחס]$_1$ העמיד פנים שהוא שכח את הארנק בבית.

Concepts and decisions in coreference:

1. Annotation-related decisions:

1. All pronouns must be annotated as mentions, regardless of whether they co-refer with any other mention or not. This applies to pronouns found within fixed expressions (رغمًا عن ذلك 'in spite of that', بعد ذلك 'after that'), expletive pronouns (قال أنه 'he said that'), etc.

2. Proper nouns within other proper nouns should have their own mention:
   'The Atlantic Ocean Institute for Neurology'

3. Adjectives should be included within the mention's span.
   a. توصّل البحث الجديد إلى أن مادة الأميلويد هي إحدى المواد التي تزيد احتمالية فقدان الذاكرة.

   'The new research concluded that amyloids increase the chance of developing dementia.'

4. Time expressions' spans should be maximal:
   [Tuesday], …, [Tuesday 26th of April].

5. When proper names are preceded by titles, these titles should **not** be included in the span.
   a. التقى ~~النائب الأسبق لرئيس الوزراء~~ زياد عمرو بوزير الصحة.

   '~~Former Deputy Prime Minister~~ Ziyad Amro met with the Minister of Health.

6. Relative pronouns may be included in the mention's span when they disambiguate the head in the mention.
   a. النساء اللواتي يعشن في مناطق الصراع هن أكثر عرضة للعنف المنزلي.

   'Women who live in conflict-ridden areas are more vulnerable to domestic violence.'

   A special case of this is when there is a pronoun within the relative clause that co-refers with the *head*, and another one with the *disambiguated noun phrase*:

   b. **[[النساء]**$_1$ اللواتي لا يتحكمن في مصدر رزقهن]$_2$ هن أكثر عرضة للعنف لأن[هن]$_2$ يعتمدن ماديًا على ذكور العائلة.

   '[**[Women]**$_1$ who do not control **[their]**$_1$ own income]$_2$ are more prone to violence because [they]$_2$ are financially dependent on their male relatives.

7. Nouns with pronominal suffixes are necessarily specific and referring in context and should be annotated as mentions. However, abstract/idiomatic nouns with pronominal suffixes should not:

a. تتميز **بلغاريا** بقلاعها الجذابة. صمدت هذه القلاع لمئات السنين.

'**Bulgaria** is known for **its** fascinating castles. These castles are at least a hundred years old.

b. لقد تجاوزت الأمور حدودها.

'עצם קיומו של חוק כזה הוא הבעיה.'

2. Mentions

2.1. NP mentions:

Head-sharing NPs: in many cases, an NP (especially if it is long) is briefly referred to using its 'head', as in the first example. The entire NP must be tagged with its head as coreferring.
The first "long" NP (broadly speaking, the antecedent) between the head-sharing NPs should have its head annotated.

- هناك إشاعات مؤكدة عن قمة إسرائيلية-فلسطينية مشتركة في مصر خلال الأيام المقبلة. هذه القمة ستناقش عملية السلام. (BBN Technologies, 2008)

- רציחתו הדרמטית של שליטה העריץ של הרפובליקה הרומית גאיוס יוליוס קיסר הפתיעה רבים. קיסר היה בן 55 במותו.

2.2. Possessives: does this apply to Hebrew/Arabic at all?

OntoNotes:

1.1.2 Possessives

Possessive nouns should be co-referenced to other mentions. Possessive proper nouns *(Fred's)* are extracted from the treebanked data; however, possessive pronouns *(his)* must be manually extracted by the annotator and added to the list of mentions:

(1) [Fred's]$_x$ wife is Wilma, and [his]$_x$ daughter is Pebbles.

2.3. NP premodifiers:

OntoNotes: A premodifier (PreMod) is a word that precedes and modifies a noun. Proper noun PreMods can be co-referenced to existing noun phrases and/or other proper PreMods, and should be manually extracted by the annotator and added to the list of mentions.

Non-proper and adjectival premodifiers are not eligible for co-reference. Articles in proper noun premodifiers should not be included in the span.

Note that **only the premodifying noun itself is included in the PreMod span, since any preceding articles (*the, a, an*) belong to the full noun phrase.

(19) But [the Army Corps of Engineers]$_x$ expects the river level to continue falling this month. "The flow of the Missouri River is slowed," an [Army Corps]$_x$ spokesman said.

- IDENT chain: [the Army Corps of Engineers], [Army Corps](proper PreMod, manually extracted)

Acronymic premodifiers should be co-referenced unless they refer to nationality (see example (29) below). In the examples (24) and (25), "FBI" and "U.N." are eligible for co-reference.

(24) the [FBI] spokesman

(25) the [U.N.] Secretary General

Nationality acronyms and other adjectival forms of GPEs, however, are **not** eligible for co-reference as premodifiers. (Although nationality acronyms can always occur as proper noun phrases, as in (26) below.) Thus, only example (27) below contains a linkable PreMod.

(26) relations between [the U.S.] and Japan – proper noun phrase

(27) the [United States] policy – proper noun PreMod

(28) the American policy – nationality adjective (no coref.)

(29) the U.S. policy – nationality acronym (no coref.)

2.4. Nominalization/verbs: sometimes a verb mentioned earlier is referred to by using its nominalized form, as in the first example, or a synonym of it, as in the second example. The verb and the gerund/nominalized form (or its synonym) must be tagged as coreferring entities.

Only tag if:

- The verb is nominalized later on in the document ("The prices rose. Very few people dealt well with the rise.")
- There is a pronoun that refers back to a VP ("Kim arrived. That annoyed me.")
- In both of these cases, only link the verb itself not the full VP with the NP.

We have to decide whether we allow for nouns that have different "roots" than the verb or stick to derivational morphology. Let's annotate near-synonyms - to be able to QA we can declare a new coref-type.

- أصر العمّال على مطالبهم. هذا الإصرار أدى إلى تحصيلهم لحقوقهم.

- ارتفعت مبيعات السيارات اليابانية ب18% من السنة الماضية. النمو القوي سيلحقه ارتفاع سنوي بنسبة 21%. (BBN Technologies, 2008)

- בית המשפט הרשיע את יהודית גונן בתקיפה הגורמת לחבלה של ממש. בין הראיות המרכזיות שהובילו להרשעה היתה עדותו של בעלה לשעבר.
- מחירי המכוניות היפניות נסקו ב-18% בשנה שעברה. העלייה החדה האטה את המכירה השנתית בשיעור של 21%.

Note: the nominalization (or its synonym) must be **definite**. In the following example, "تقليص" cannot be considered as co-referring with "تقلل" because it is **indefinite**:

- قالت الأرجنتين أنها ستطلب من البنوك الدائنة أن تقلل ديونها الخارجية التي تبلغ 64 مليار دولار. تطمح الأرجنتين للوصول إلى تقليص بنسبة 50% من قيمة ديونها الخارجية. (BBN Technologies, 2008)

- ארגנטינה הודיעה כי תבקש מהבנקים הנושים להפחית את החוב הזר שלה בסך 64 מיליארד דולר. ארגנטינה שואפת להגיע להפחתה של 50% בערך החוב הזר שלה.

3. Singletons

Singletons are defined as mentions that

(1) are never coreferent, or
(2) are referring expressions but do not have an antecedent, or
(3) are potentially coreferent but simply occur once in the document.

- We went to the Victoria and Albert Museum. We enjoyed learning about the artifacts a lot.

4. **Apposition** بَدَل: An NP that describes the proper noun and can replace it in the context later on. Appositive coreference requires two NPs which are adjacent and can be placed in either order, which fulfill the same grammatical function simultaneously.

- ستشهد فكتوريا تشن (main NP) ارتفاعًا ملحوظًا في راتبها بعد تولّي المديرة المالية لMegabucks Banking (بدل apposition) منصبها. (Jurafsky and Martin, 2023)
- ولد ابن خلدون في تونس ثم هاجر العالِم إلى مصر.
- لندن مدينة مميزة. تتميز العاصمة البريطانية بكثرة سحابها. (Poesio et. al., 2021)
- למרות שיהודה עמיחי נולד בגרמניה, את מרבית חייו העביר המשורר בירושלים.

- קהיר נחשבת כיעד תיירותי מועדף על האירופאים. ניתן למצוא בבירה המצרית מגוון מוזיאונים ואטרקציות מסקרנות.

5. Synonyms: We have a case of two synonymous nouns, where the second one has a demonstrative and they together corefer to the first noun:

אטימותו של השר הובילה לצמצום תקציבים של משרד הרווחה. נראה שהנוקשות הזו מייצגת היטב את האג׳נדה החברתית-כלכלית של הממשלה הנוכחית.

6. **Temporal/numeric phrases:** coreference between two NPs may include a number of years, amount of money, or a number of people. The entire temporal/monetary/number phrases should be tagged along with the succeeding entity to which it refers.
   - قضى مروان عشرة سنوات في السجن. خلال ذاك الوقت أنهى دراسته الجامعة.
   - عرض الشركة لمبلغ 150 دولار لم يكن متوقعًا. صُدم الجميع بالسعر.
   - نحو 650 جنديًا أمريكيًا سينضمون إلى القوات الفلبينية. حُذّر الجنود من المخاطر.
     (BBN Technologies, 2008)
   - 70 שנה מלכה המלכה אליזבת השניה בממלכה המאוחדת, במהלכן ביקרה ברחבי העולם.
   - אני מרוויח 40 ש״ח לשעה. עם הסכום הזה אין לי סיכוי לגמור את החודש.
   - מתוך 800 הסטודנטים שניגשו למבחן, רק כמחציתם עברו אותו בהצלחה.

7. **Quantification:** coreference must be tagged between the NP that includes the quantifier (כל, רוב, אף, מספר, שום, כמה…) (كل، أغلب، أحد، جميع، بعض...) in its entirety with the pronouns.
   Each part of the coordination should be tagged individually.
   - [أغلب الطلاب]$_1$ دخلوا الصف. [هم]$_1$ سعيدون. (Poesio et. al., 2021)
   - جلس [كل طالب]$_1$ في درج[ـه]$_1$. (Poesio et. al., 2021)
   - [כל התלמידות]$_1$ יצאו לחצר. [הן]$_1$ שמחו שהמבחן נגמר.
   - [רוב המורים]$_1$ שותים קפה במקום עבודת[ם]$_1$.
   - ראיתי את **יפעת** ועמנואל. שתיהן היו שמחות. **יפעת** הייתה רעבה.

Examples for the absence of coreference between an NP and/or an NP and a pronoun (non-referring NPs):

1. **Appositive NP**:
   Mark the more specific term (usually, the proper noun) as the head.
   - الناطقة باسم وزارة الخارجية السويسرية (appos)، دانييال ستوفل (head NP) ستَحضر العشاء.
     (BBN Technologies, 2008)
   - كارلوس (head NP)، ابن عمي (appos)، هو شاب لطيف. (Poesio et. al., 2021)
   - פאינה, חמותי, היא אישה מקסימה וזמרת בחסד.

- מזכיר ועדת הבחירות המרכזית, ינון גמל השתתף בפתיחת התערוכה.

2. **Predicative NP**: in the following examples, the NPs in red are predicative or explanatory rather than phrases that introduce a new entity in discourse - they should not be tagged.
   - إليزابيث الثانية هي ملكة مملكة بريطانيا العظمى. (Poesio et. al., 2021)
   - ادعت مصادر أن منطقة معركة برونانبوره هي بالفعل برومبوروه. (Poesio et. al., 2021)
   - اكتشفنا أن القاتل كان مستر راي. (Poesio et. al., 2021)
   - כל המנכ"לים ברוב החברות הם אנשים מבוגרים עם המון ניסיון.
   - קיוויתי שהמנחה יהיה דודו טופז.

3. **Generic NP:** in some cases the antecedent is generic or indefinite, meaning it does not have the property of referring to another entity. In the first example the pronoun "ـها" refers to mangos as a fruit rather than to a specified set of mangoes, and in the second example the pronoun "هم" does not refer to a specific set of woodwind players.
   - أحب [المانجا]$_1$. إنـ[ـها]$_1$ فاكهة لذيذة. (Jurafsky and Martin, 2023)
   - على [عازفي آلات النفخ الخشبية]$_1$ أن يكونوا مبدعين إذا أرادوا النجاح لأن جمهور[هم]$_1$ محدود. (Poesio et. al., 2021)
   - אני אוהב [חתולים]$_1$. [הם]$_1$ נעימים ומגרגרים.
   - [אנשים]$_1$ נוטים לחשוב שהדעה של[הם]$_1$ מעניינת אותי.

4. **Expletive pronouns**: pronouns that do not refer to any entities but rather play a syntactic role:
   - إنه زيدٌ قادم. It is Zeid who is coming
   - من المهم أن يدرك المجتمع أنهِ من الصعب على الفئات الضعيفة أن تطالب بحقوقها.
   - זה חשוב לאכול בריא.

A worked example:

[فيكتوريا تشن]$_1$، المديرة المالية لـ[ميغابكس بانكينغ]$_2$، ستشهد ارتفاعًا ملحوظًا في [راتبـ[ـها]$_1$]$_3$ بعد أن أصبحت [الثلاثينية]$_1$ رئيسة [الشركة]$_2$. من المعروف أنـ[ـها]$_1$ جاءت إلى [ميغابكس]$_2$ من الشركة المنافسة [لوتسابكس]$_4$.

A summary of the coreferring entities in the example:

1. فيكتوريا تشن، ـها، الثلاثينية، ـها
2. ميغابكس بانكينغ، الشركة، ميغابكس
3. راتبها
4. لوتسابكس

[שמרית לביא]$_1$, קופאית ב-"[שופרסל]$_2$" עתידה לקבל העלאה משמעותית בשכר[ה]$_1$ לאחר שתהפוך למנכ"לית [החברה]$_2$ בספטמבר הקרוב. ידוע כי [היא]$_1$ עברה אל "[שופרסל]$_2$" מהחברה המתחרה "[רמי לוי]$_3$".

1. שמרית לביא, -ה, היא
2. שופרסל, החברה, שופרסל
3. רמי לוי

(Jurafsky and Martin, 2023)

Decisions/technical issues:

1. Demonstrative pronouns and determiners should be part of the span, so when the antecedent mention is "הרשות לנייר ערך", followed later by "הרשות", the second mention should be **הרשות** not רשות.

2. When a mention is long (i.e., includes necessary adjuncts), one might use the head. Ex: תשקיף שאישרה הוועדה may have its label as תשקיף.

3. Annotators mustn't create chains of full mentions, meaning there is no need to annotate a certain mention if there isn't a pronoun/term-denoting expression that refers to it.
   Ex: if נייר ערך is mentioned so many times along the text, there is no need to chain them all together.

4. A full date "April 19, 1989" is a full mention - no nesting is required.

5. When an entity is written in English right after its full name in Hebrew/Arabic then it is not necessary to annotate the English name.
   Example: كنيسة ساغرادا فاميليا ~~Basilica of the Holy Family~~

6. <u>To be discussed</u>: when a coreference chain contains multiple grammatical features:
   [The Palestinian people] has … [They] have… [We] must …
   We will take [you.SG] on a trip to the Maldives. [You.PL] might want to try…

References:

1. Al-Oraini, Abdulrahman., Yu, Juntao., Peosio, Massimo. 2020. Neural Coreference Resolution for Arabic. Spain: Association for Computational Linguistics. Online version: https://aclanthology.org/2020.crac-1.11/
2. Al-Oraini, Abdulrahman., Pradhan, Sameer., Poesio, Massimo. 2022. Joint Coreference Resolution for Zeros and non-Zeros in Arabic. Spain: Association for Computational Linguistics. Online version: https://aclanthology.org/2022.wanlp-1.2.pdf
3. Arabic Co-reference Guidelines For OntoNotes. 2008. BBN Technologies. Online version: https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/arabic-coreference-guidelines.pdf
4. Jurafsky, Dan & Martin, James. 2023. Coreference Resolution. In *Speech and Language Processing*. Online version: https://web.stanford.edu/~jurafsky/slp3/
5. Poesio, Massimo., Camilleri, Maris., Carretero-Garcia, Paloma., Artstein, Ron. 2021. ARRAU 3 Annotation Manual Version 1.0.