Hebrew Coreference Resolution Guidelines (English Translation)

Hebrew Coreference Resolution Guidelines

Authors: Reut Tsarfati, Shaked Greenfeld
Emails: shakedgreenfeld@gmail.com | reut.tsarfaty@biu.ac.il

Task

The task of coreference resolution (איגוד אזכורים) is to link all specific mentions in the text that refer to the same entities and events.
A mention will be limited to noun phrases (Noun Phrases), construct structures (e.g., סגל הבית הלבן), and adverbs referring to time (e.g., אתמול) or place (e.g., כאן), when all mentions in the same cluster refer to the same entity.

Segmentation

Text segmentation means dividing the text into its smallest meaning-bearing units—words for sentence-level work, and further into morphemes when relevant. For example:

ח"כ אלי דיין (מערך)(הגיש הצעת חוק שלפיה יוטל היטל על מעסיקי עובדים זרים כדי למנוע את העדפתם על עובדים ישראליים.

Segmented into morphemes:

ח"כ, אלי, דיין, (מערך), ,הגיש, הצעת, חוק, ש_, לפי, היא, _יוטל, היטל, על, מעסיקי, עובדים, זרים, ישראליים. ,כדי, למנוע, _העדפה, של, _ , הם_, על, עובדים

Here, לפיה is split into לפי_ and _היא.

Mentions

Mentions are the units that make up a coreference cluster. A mention is an expression referring to some entity (named entity or non-named entity) that can be identified in everyday reality, whether it refers to a specific person, a group of people, an inanimate object, or abstract entities.

Examples of mentions (not all mentions in the following text are marked):

[ח"כ אלי דיין] submitted [הצעת חוק] according to which [מעסיקי עובדים זרים] היטל על]] will be imposed, in order to prevent their preference over Israeli workers. He is expected to promote [מס זה] in the upcoming session after returning from [החופש שלקח].

- ח"כ אלי דיין – specific person
- מעסיקי עובדים זרים – group of people
- הצעת חוק, מס זה, החופש שלקח – abstract entities

In the tagging task, mentions will be extracted automatically; however, since extraction is not perfect and some mentions will be missing or flawed, these missing ones must be marked according to the guidelines and the definition of "what is a mention." On the other hand, some extracted segments do not fall under the definition and must not be tagged at all—not even as a singleton mention group (a mention group of size one)—and will be marked as idiomatic. All valid mentions will be used in the second stage as candidates for coreference linking.

In this section, we define which expressions fall under the definition of a mention and which do not.

Noun Phrase

A noun phrase (NP) consists of a noun or a pronoun and the words attached to it in the syntactic parse tree. These phrases are presented to annotators as candidates for coreference clustering, and in general, the phrases are automatically extracted from the parse tree (if the document is annotated as part of the HTB) or by a model trained for this task.

Prepositions

Prepositions are not considered part of the noun phrase.
For example, in the sentence:
[ארמון] חי ב [ה מלך] — we do not include the preposition "ב" (in) as part of the mention *ארמון*, since it is not an inherent part of the phrase.

Alternatively, in the sentence:
אפשר לקנות קילו ענבים וחצי מלון, נורא זול שם, x[ב[מלון ברמת עמון.
The mention "מלון ברמת עמון" is one logical unit, and therefore we do include the second "ב" because it is part of the entire phrase. The first "ב," however, is not part of the mention and will not be tagged as part of the NP.

Appositive vs. Anaphoric Relation

An appositive is a type of modifier whose attachment to the head noun does not form a construct state; it is part of the sentence and serves as a synonym or an additional expression describing the same thing.
When tagging, we do not separate the annotation of the NP between one that is part of an anaphoric relation and one that is an appositive to the head—they are tagged identically. However, we annotate all components of the sentence: the head noun, the appositives, and the entire noun phrase.

Examples:

מאמצים דיפלומטיים לשחרור [ישראלי שכלוא בטייוואן]: "[הוא] חלש וחולה"

[[הדוקטור], [אבי כהן]], הצליח לרפא אותי.
[[ראש הממשלה], [בנימין נתניהו]], נאם אתמול בכיכר.

Nested Phrases

Unlike *Base NP Chunking*, which is defined by splitting text into non-overlapping segments and marking only the non-recursive noun phrases:

[the nation] 's [manufacturing titans] typically jet off to [the sunny confines] of [resort towns] like [Boca Raton] and [Hot Springs].
אולם עולה [כאן] [השאלה], כיצד הסכימו [המנגנונים] ו[מושלי המחוזות] מלכתחילה הצגת מועמדות של [הם] של [ראשי כנופיות] ו[סוחרי סמים], כדי שייצגו את [מפלגת הרוב].

And unlike general noun phrases:

[the nation's manufacturing titans] typically jet off to [the sunny confines of resort towns like Boca Raton and Hot Springs].
אולם עולה [כאן] [השאלה], כיצד הסכימו [המנגנונים ומושלי המחוזות] מלכתחילה [הצגת מועמדות של הם של ראשי כנופיות וסוחרי סמים], כדי שייצגו את [מפלגת הרוב].

We allow annotation of **all phrases** that can refer to a specific entity in the world, and in particular, we allow **nested phrases**:

[[the nation] 's [manufacturing titans]] typically jet off to [[the sunny confines] of [resort towns] like [Boca Raton] and [Hot Springs]].
אולם עולה כאן השאלה, כיצד הסכימו [[המנגנונים] ו[מושלי המחוזות]] מלכתחילה [הצגת מועמדות של הם] של [[ראשי כנופיות] ו[סוחרי סמים]], כדי שייצגו את מפלגת הרוב.

Exceptions to this rule are *construct-state structures* (מבני סמיכות) and *demonstrative pronouns*.

Construct-State Structures

Although in Hebrew a construct state cannot be split while preserving the semantic meaning of the expression, we do not split a construct state into several separate mentions. However, we do allow it to be **nested** inside other mentions.
It should also be noted that a construct state can itself be part of another nested noun phrase.

Example:

ועורך [ה עיתון "אל - אחאבר"] תוקף על רקע זה את [ראשי [המפלגה]], אף שביטלה את מועמדותו

Nested Demonstrative Pronouns

A demonstrative pronoun is a pronoun used to indicate a specific thing to the addressee. Words such as *זה* ("this") or *אלה* ("these") are considered demonstratives.
Examples of expressions with demonstrative pronouns: *רקע זה* ("this background"),

מנהיגים אלה ("these leaders"), *ועדה זו* ("this committee"), *פעילות פוליטית כלשהי* ("some political activity").

Unlike construct-state structures, where we allow nesting, **for demonstrative pronouns (dem)** we do **not** allow nesting of the expression. Although the parts of speech of these expressions are in the form `NOUN PRON`, each word cannot stand alone without losing meaning; therefore, they cannot be split without semantic loss.

Example:
[הן] גם אינן אמורות לקחת [חלק] ב[פעילות פוליטית כלשהי].

Quantified Expressions

Quantified expressions refer to an entity *X* modified by a quantifier *Y*. These include:

- **Cardinals** – numbers: עשר (ten), מיליארד (billion), אלפים (thousands)
- **Partitives** – כמה (some), חצי (half), מעטים (few)
- **Measures** – ליטר (liter), קוב (cubic meter)
- **Collectives** – עדר (herd), צבא (army)

Examples:

- שלושה אנשים – cardinals
- הרבה חיות – partitives
- 6 מקטע של כביש – measures
- נחיל דגי דקר – collectives

These should be annotated both as a **single unit** and as a **nested expression**. That is, we mark the entire noun phrase, and also the internal phrase.

Examples:

- [אנשים] שלושה]] – cardinals
- [חיות] הרבה]] – partitives
- [כביש 6] של מקטע]] – measures
- [דגי דקר] נחיל]] – collectives

Adjectives

Adjectives are not annotated as separate mentions.
Example: *הילד היפה*– annotated as a single mention [הילד היפה], where the adjective *היפה* ("beautiful") is part of the mention *הילד* ("the boy").

Adverbs

Adverbs are generally not annotated as mentions (except for cases of time and place, detailed later).

If an adverb appears after the head of the mention, it will be included as part of the base mention.
Examples:

- "יותר" in *אנשים יפים יותר* ("more handsome people")
- "מאוד" in *הילד היפה מאוד* ("the very beautiful boy")
  These are included as part of the mention.

With the exception of adverbs, those referring to **time** and **place** are also marked as mentions, since they can be part of a coreference chain.

For **time**, any expression referring to a defined time will be marked as a mention:
אתמול, *שלשום*, *אמש*, *מחר*, *אז*, *פעם*, [ה]יום, [ה]שבוע, [ה]חודש, [ה]שנה.

For **place**, we mark adverbs in a similar way:
פה, *כאן*, *שם*, *למעלה*, *למטה*, *קדימה*, *אחורה*, *בצד*, *מזרח*, *מערב*, *צפון*, *דרום*.
Also, any expression in a construct relation referring to place.

Any adverb of time or place not part of this closed list should be marked manually.

Example:

חברות המעסיקות עובדים זרים זוכות במכרזים, היות והן מציעות [שירותים זולים יותר]

Adjective vs. Mention

הוא רואה במס מעסיקים מס מעוות.
Here, *מס מעוות*("distorted tax") is an adjective describing *מס מעסיקים*("employers' tax"), so it is **not** marked as a mention.

Units of Measure

Units of measure that appear **on their own** are not marked as mentions. If something is described using a unit of size, time, or distance, we do not mark the unit itself as a mention.
Examples:

- העונש על רצח הוא כעשרים שנים – *עשרים שנים* not marked.
- הוא הולך כל יום 30 ק"מ – *30 ק"מ* not marked.
- מישהו גבוה יותר מעץ אלון – *עץ אלון* not marked.

Examples of non-mentions (idiomatic in this context):

- רנדי בארנס עלול להפסיד יותר מ*חצי מיליון דולר*
- שיאן העולם באכילה מהירה שוקל יותר מ400 קילוגרם *400 קילוגרם*
- המגלשה הזאת ענקית! היא יותר גבוהה מ*מגדל*

→ {*חצי מיליון דולר*, *400 קילוגרם*, *מנוף*} are **not** marked as mentions.

If the unit of measure is **part of the mention** or serves as a **synonym for a noun**, it **is** marked.

Example:
השיאן העולמי ב[400 מטר] הוא מייקל ג'ונסון
Here, *400 מטר* is marked as a mention because it functions as a synonym for " תחרות ריצת 400" ("400-meter race").

Construct-State Structures

In Hebrew, a construct state is a unique syntactic phenomenon connecting two nouns with a close relationship (e.g., "הבית של הנשיא", "בית הנשיא", "ראש ההר", "עורך העיתון"). In such a construct, we mark the **entire construct** as a single noun phrase.

Example:
[עורך ה עיתון "אל - אחאבר"] תוקף על רקע זה את [ראשי המפלגה], אף שביטלה את מועמודתם]

We do not split a construct state, so as not to create semantically meaningless expressions.
However, if the construct consists of more than two words, we also mark the nested phrase as a mention.

Possessive Expressions

For possessive expressions, similar to construct-state structures, we do not split them because splitting would cause loss of the original meaning. We also annotate the nested expression.

Example: *סיורים של צה"ל* ("tours of the IDF")

Marked as:

בגזרת עזה נערכים מדי יום [סיורים של [צה"ל]], הם נערכים בשעות הערב והבוקר

Other examples we mark:

- סיורים של צה"ל
- הכוונה המקורית של משה
- חלקים גדולים למדי של החברה
- רגעי החסד הגדולים שלו (הוא_ _של)
- התוכנית הראשונה שלהם (הם_ _של)
- הקומה השנייה של הבניין

Verbs

We allow NP clauses that contain VP clauses describing the noun to be part of the mention.

Example: *חברות המעסיקות עובדים זרים*
Marked as:

[[חברות המעסיקות [עובדים זרים]]]

Not as:

[חברות] המעסיקות [עובדים זרים]

On the other hand, we do not mark verbs in the present participle (בינוני) form as mentions.

Example:
[אנשים אלה] למעשה משמשים [עובדים שכירים זולים].
The verb *משמשים*("serve as") is **not** marked as a mention.

Action Nouns

An action noun is a noun derived from a verb but without tense or person. They usually indicate an action, process, state, or result (e.g., הלכתי – root 1 ,הלך st person past; action noun: הליכה).
We mark all action nouns as mentions.

Examples:

- [אם אמנם יאלץ להיעדר מפעילות שנתיים בגלל השעיה על [שימוש בסמים – *שימוש בסמים* is marked.
- אני אוהב לרקוד [ריקודי עם] עם אשתי – *ריקודי עם* is marked.

Punctuation

If the mention includes punctuation that requires symmetry (e.g., quotation marks, parentheses), we mark the mention in a way that preserves symmetry.

Examples:

- "עורך הדין" (correct) – not עורך הדין" (incorrect)
- "גיפריס" (correct) – not גיפריס" (incorrect)
- (הארץ) (correct) – not (הארץ (incorrect)

If the suggested mention is incorrect, mark it as **idiomatic**, and additionally create a new mention that includes both the opening and closing quotation marks or both parentheses.

Idiomatic Phrases That Are Not Mentions

If a noun is used as part of an idiomatic phrase, we do not mark it as a mention.
For example: *עם זאת*("however"), *על כך*("about that"), *על פי*("according to"), and their components (*פי*, *זאת*, *כך*) are not marked.

Example:

הילדים שוב הפריעו בכיתה ועולם כמנהגו נוהג, על פי המורה, החומר בכל אופן הועבר בהצלחה

Nouns Without Reference in the Text

An idiomatic expression may look like a candidate mention but is not a mention if it does not refer to the marked entity in the text.

Example:

שחקני [בית"ר [ירושלים]] חזרו ארצה למשחק הבית ב[עיר הבירה].

Here, *ירושלים*is idiomatic because the author refers to the team *בית"ר ירושלים*, not the city. Therefore, *עיר הבירה*and *ירושלים*are not in the same coreference chain, and *ירושלים*is marked as idiomatic.

Possessive-like Structures That Are Valid Mentions

On the other hand, expressions like *ממשלת ישראל* ("Government of Israel") represent a possessive-like structure equivalent to "הממשלה של ישראל" and are different from *בית"ר ירושלים* (which is not equivalent to "הבית"ר של ירושלים"). In this case, *ישראל* is a valid mention and not idiomatic.

Less Specific Expressions

If among the candidate mentions there are multiple segments with an inclusion relationship (one contains the other) and they refer to the same mention, we mark as the mention the **most specific expression**—usually the longest one—that refers to the entity. All other expressions are marked as **idiomatic**.

Example:

[[נשיא מועדון השחמט] שנבחר כבר 5 פעמים ברציפות] מתמודד על ראשות המועדון זה הפעם השישית]

Since "נשיא מועדון השחמט" is less specific than "נשיא מועדון השחמט שנבחר כבר 5 פעמים ברציפות", we mark:

- נשיא מועדון השחמט – idiomatic
- נשיא מועדון השחמט שנבחר כבר 5 פעמים ברציפות – valid mention

Example:

ברור ש_אנחנו עדים באחרונה ל[[[התפתחויות שלא הכרנו בה_עבר], בה_עולם המוסלמי הקיצוני שמשפיעות עלינו מאד]]

Here, the most specific expression is “ התפתחויות שלא הכרנו בה_עבר, בה_עולם המוסלמי הקיצוני שמשפיעות עלינו מאד”, so we mark:

- התפתחויות שלא הכרנו בה_עבר – idiomatic
- התפתחויות שלא הכרנו בה_עבר, בה_עולם המוסלמי הקיצוני – idiomatic
- התפתחויות שלא הכרנו בה_עבר, בה_עולם המוסלמי הקיצוני שמשפיעות עלינו מאד – valid mention

Separated by Comma

If the description is separated by a comma, we attach it to the beginning of the mention.

Example:

[יונתן, בן קיבוץ נחשולים], הלך אתמול לגן

- יונתן – idiomatic
- בן קיבוץ נחשולים – idiomatic
- יונתן, בן קיבוץ נחשולים – valid mention

Proper Nouns

Proper names are not split; we treat them as a single unit. This includes titles like “מר” (“Mr.”), “ד"ר” (“Dr.”), which are not separated from the noun they modify. The no-split rule also applies to names of places, songs, organizations, etc.

Examples:
[פרופסור [דני אורבך]] התארח הערב בתוכנית פגוש את העיתונות.
[[בית המלון ב[קליפורניה]] – from the song sung by *להקת הנשרים*
[בנק [ירושלים]]

Marked as valid mentions:

- פרופסור דני אורבך
- בית המלון בקליפורניה
- בנק ירושלים

Marked as idiomatic:

- דני אורבך
- קליפורניה
- ירושלים

Errors in the Text (Gibberish)

If a sequence of characters is not a Hebrew word and is not understandable, followed by a valid word, we mark **only** the valid word as the mention.

Example:

קפידבינהמלך קפינדרה הסכים לבטל את החוקה

We mark only: "המלך קפינדרה" as the mention (not "קפידבינהמלך קפינדרה").

Results and Additional Details

If game results or other details appear that add to the base mention and are themselves a shortened form of a valid mention, we mark them as well.

Example:

קייזרסלאוטרן (61 נק) מחכה לדיסלדורף (מקום 21)

We mark:

- [קייזרסלאוטרן ([61 נק])]
- [דיסלדורף ([מקום 21])]

Marking Mentions

Although most potential mentions are pre-marked automatically, there are cases where some mentions are not offered as candidates in the text. In such cases, the mentions must be marked manually.

When marking mentions, pay attention to the following guidelines:

- Once a mention is marked, it cannot be deleted. If you later decide the mention is incorrect, mark it as **idiomatic**.
- Do not mark a mention that has already been marked.
- A mention that should be marked is any sequence of words that meets the definition of a valid mention and was not automatically extracted as one.
- A mention must always be within the same sentence. Do not mark a sequence of words that crosses sentence boundaries.

Pronouns and Demonstratives

Inside quotations

Pronouns and demonstratives within quotations are linked to the entity they refer to, even if the pronoun is part of the quoted text.

Example:

כפי שאמר [חוסה מרטי] לאחר שהותו בארה"ב: "חייתי בתוך המפלצת ו[אני] יודע זאת אפילו מעבר לקרביו".

Cluster: {חוסה מרטי, אני}

Mentions resulting from tokenization

When a pronoun appears as a result of tokenization—meaning the pronoun did not appear in the original text and was added during tokenization—we mark **only** the pronoun itself as the mention, linking it to its head noun, rather than marking the entire expression.

Example:
Original sentence:

הם עודדו חברות לתעל את הפילנטרופיה שלהן

After tokenization:

הם עודדו [חברות] לתעל את הפילנטרופיה של [הן]

Cluster: {חברות, הן}

Here, "שלהן" is not marked—only "הן" is linked to "חברות".

Expletive (non-referential) pronouns

Pronouns that do not carry referential meaning or contribute directly to it are not marked. They can be identified using the **replacement test**:

- If the pronoun cannot be replaced by a noun phrase (NP chunk), it is expletive.

Example:

[זה]_1 נראה שחם היום. [זה]_2 מקשה על הנשימה.

- "זה_1" cannot be replaced by another NP → idiomatic.
- "זה_2" can be replaced by "המסיכה שלי" or "חום" → can be linked.

Copula vs. Pronoun

In nominal sentences (no verb), the hidden pronoun "הוא/היא/הם/הן" acts as a **copula** linking the subject and predicate. The demonstrative "זה" can also function as a copula.

Examples:

- הוראה זה מקצוע מאתגר
- כדורגל הוא הספורט הכי פופולרי בעולם

Copular elements are **not** considered mentions and are **not** part of coreference chains.

Example:

[11 היום בן [2_הוא] .[[יוסי] האח של [1_הוא דני].

Cluster: {2_הוא ,יוסי האח של ,דני}

Copular constructions

A copula is an "empty" verb linking the subject to a non-verbal predicate. These sentences are composed of the subject, a property of the subject, and the copula connecting them.
In such cases, **the property is not marked as a mention of the subject**.

Example:

[רקדנית בלהקת המחול של מוסקבה] הייתה [דנית].

Here, "דנית" (subject) is linked via "הייתה" (copula) to "רקדנית בלהקת המחול של מוסקבה" (property). Since this is an adjectival property, "רקדנית בלהקת המחול של מוסקבה" is **not** marked as a mention.

When the copula indicates a unique identifier

If the copula introduces a unique, context-specific identification of the subject, we do link them.

Example:

[בנימין נתניהו] הוא [ראש ממשלת ישראל].

Here, "ראש ממשלת ישראל" refers exclusively to Benjamin Netanyahu in this context—it is not just a property. Therefore, they **are** clustered together.

Coreference Linking

Once the potential mentions have been defined, this section describes how we group them into a **coreference chain** (or cluster). We will define which mentions are grouped together, detail the different types of chains, and illustrate with representative examples to ensure optimal annotation.

Identity Link

Name mentions, pronouns, compound nouns, adverbs, and demonstrative pronouns referring to the **same entity, event, or concept** are linked together in an identity relationship. There is no restriction on the type of connection as long as they refer to the same referent.

Example:

היתרון העצום של [ויליאמס_1] בסקרים פחת והלך, ו[הרפובליקאים] מרטו בייאוש את שערותיהם. [הם] התחננו לפני [ויליאמס_2] לשתוק.

Clusters:

- Cluster 1: {ויליאמס_1, ויליאמס_2}
- Cluster 2: {הרפובליקאים, הם}

Appositive Link (תמורות)

An appositive link is a sequence of noun phrase mentions, usually separated by a comma, dash, or parentheses, describing the **same entity**.

In this project:

1. We first mark the **longest continuous expression** as a single mention.
2. We do not distinguish between different subtypes of appositive relationships.
3. We also mark each part of the appositive separately and add it to the same cluster.

Example:

איש בטקסס לא פיקפק ש[[קלייטון ויליאמס], [חוואי ואיש נפט]], יביס אותה בקלות. [הוא] הופיע בתשדירי הבחירות שלו רכוב על סוס, עם מגבעת רחבת תיתורת, ופרט על נימי המאציזמו הטקסני.

Cluster: { "הוא", "חוואי ואיש נפט", "קלייטון ויליאמס", חוואי ואיש נפט, "קלייטון ויליאמס" }

Compound Nouns

If an error occurred in the first part of the task and, for a certain expression, we have additional **less specific** sub-mentions, we cluster the **most specific** expression with the other mentions pointing to the same entity.

The **less specific** expression is placed in its **own singleton cluster** (and not with the correct mentions).

Example:

תופעה זו התבררה אתמול ב[[וועדת העבודה והרווחה של הכנסת], שדנה בנושא העסקת עובדים זרים]. יו"ר [הוועדה], ח"כ אורה נמיר (מערך), טענה כי "מביאים עובדים זרים לישראל על תקן של מתנדבים מתאילנד, רק כדי לא לשלם להם שכר מינימום".

Clusters:

- Cluster 1: {"וועדת העבודה והרווחה של הכנסת, שדנה בנושא העסקת עובדים זרים", "הוועדה"}
- Singleton cluster: {"וועדת העבודה והרווחה של הכנסת"}

Here, "וועדת העבודה והרווחה של הכנסת" is **less specific** than "וועדת העבודה והרווחה של הכנסת, שדנה בנושא העסקת עובדים זרים". We choose to cluster the more specific mention with the correct chain and place the less specific one in its own singleton cluster, marked as idiomatic (no other mention points to it).

Generic Mentions

Definition
Generic mentions are expressions that do not refer to any specific entity. They may be abstract or generalized, and are linked only via **anaphoric** (or **cataphoric**) connection to a pronoun or a definite expression—but not to other generic mentions without such a connection.

Plural generic expressions

Plural forms such as “בכירים” (senior officials), “ילדים” (children), or “אריות” (lions) are considered generic mentions and are linked only to pronouns with an anaphoric connection. Since generic mentions are not linked to one another, each new generic mention starts a new coreference cluster.

Example:

[ילדים] אומרים ש[הם] אוהבים את החופש הגדול.
Cluster: {ילדים (generic mention), הם}

Generic mentions linked via definite expressions

Example:

מכל החיות אני הכי אוהב נחשים. חיות אלה מרגשות אותי מאד. הם מאד יפות.
Cluster: {נחשים (generic mention), הם (anaphoric pronoun), חיות אלה (definite expression)}

Multiple generic mention clusters

Example:

[הורים]1 צריכים להיות מעורבים בחינוך ילדיהם בבית, לא בבית הספר. [הם]1 צריכים לדאוג שהילדים של [הם]1 לא יכלו זמנם במשחקי מחשב; [הם]1 צריכים לוודא שהילדים מבלים מספיק זמן בהכנת שיעורי בית; [הם]1 צריכים להגיע לאספות הסמסטריאליות בבית הספר. [הורים]2 נוטים להאשים את בתי הספר במגבלות החינוכיות של הילדים של [הם]2. אם [הורים]3 אינם מרוצים מבית ספר, צריכה להיות ל[הם]3 אפשרות לעבור לבית ספר אחר.

Clusters:

- Cluster 1: {הורים, הם, הם, הם, הם}
- Cluster 2: {הורים, הם}
- Cluster 3: {הורים, הם}

Generic noun phrases not linked to synonyms

Noun phrases that are generic and non-specific, such as "תנור לבנים" (brick oven) or "מחשב מקבוק פרו" (MacBook Pro), are not linked to one another even if they refer to similar objects.

Example:

[תנור לבנים], המכונה לעיתים [טאבון] הוא מתקן עשוי אבנים או לבנים שמסיקים בו אש ומיועד בעיקר לחימום, לאפייה ולבישול.

Clusters:

- Cluster 1: {תנור לבנים}
- Cluster 2: {טאבון}

Generic pronouns

Generic pronouns refer to a group or abstract entity rather than a specific one. They are not part of any other coreference cluster and form a **singleton** cluster.

Example:

(תהיה [אתה] השינוי ש[אתה] רוצה לראות בעולם. (מהטמה גנדי

Modifying descriptors (תיאורים מאייכים)

- A modifying descriptor is a linguistic element indicating a specific quality of an action or state, often expressed via an adverb (e.g., "היטב" – well, "מאד" – very, "בשקט" – quietly).
- Modifiers do not stand alone as coreference mentions and are not linked to other mentions unless they are nouns.

- Like compound nouns, non-noun modifiers are included as part of the noun phrase mention they modify.

Example:

[ילד קטן מאד] הלך לגן, [הוא] התיישב במעגל ליד הגננת.
Cluster: {ילד קטן מאד, הוא}

When modifiers are adverbs
Adverbs that act as modifying descriptors are not linked to other mentions even if the meaning is similar.

Example:

הילדים הלכו הבייתה ב[שמחה]. [השמחה] הייתה רבה כשהופתעו לגלות שגם מחר אין בית ספר.
Here, the first "שמחה" (as an adverbial phrase) is not linked to the second "השמחה" (noun).

When semantically related but distinct
Even if two mentions seem related in meaning, if one is a modifying descriptor and the other is a generic noun, they are not clustered together.

Example:

ביום ראשון בערב שני תושבי צפון הארץ, אב ובנו, פשטו על [מקשת אבטיחים] השייכת לחקלאי מטירת צבי והעמיסו לטנדר שלהם אבטיחים בשווי כולל של עשרות אלפי שקלים. מתנדבי ארגון "השומר החדש" ולוחמי מג"ב הבחינו ברכב החשוד עמוס ב[אבטיחים] נוסע ליד אחד היישובים בצפון הארץ.
Here, "מקשת אבטיחים" and "אבטיחים" are not part of the same cluster.

Exceptions for temporal modifiers
Time-related modifying descriptors (e.g., "יולי" – July, "השנה" – this year, "מחר" – tomorrow) **can** be clustered when contextually linked.

Example:

ב[דו"ח [יולי][1] מצויין כי רווחי החברה עלו ב־20% ב[חודש זה][1].
Cluster: {יולי, חודש זה}

Entity-type modifiers
If a modifier refers to an entity, it is marked separately and can be linked to other mentions in the text.

Example:

[מזכיר [הקיבוץ החדש] נשא את נאומו לפני הקהל שהורכב ברובו מיוצאי [הקיבוץ].
Cluster: {הקיבוץ החדש, הקיבוץ}

Names

- Names are not split and are treated as a single unit. This includes titles such as "Mr.", "Dr.", etc., which are not separated from the noun they modify.
- This rule also applies to place names or song titles, which will not be linked to other mentions.

Example:

[הוא] רואיין בנוגע לספרו החדש בנוגע [פרופסור דני אורבך. [התארח הערב בתוכנית *פגוש את העיתונות*] ליפן.
Cluster: {פרופסור דני אורבך, הוא}

Example:

בשיר [בית המלון בקליפורניה] שרה להקת הנשרים על מלון ב[קליפורניה]. זהו [השיר המפורסם ביותר של הלהקה].
Cluster: {בית המלון בקליפורניה, השיר המפורסם ביותר של הלהקה}
Note: "קליפורניה" inside the song name is **not** linked to the standalone mention of "קליפורניה."

Example:

[אני כבר שנים לקוח ב[בנק ירושלים] מאז שגרתי ב[בירת ישראל].
Note: "ירושלים" inside "בנק ירושלים" is **not** linked to "בירת ישראל."

Quantified Expressions

Quantified expressions, as defined in the mention definition section, are not linked in the same cluster with their base noun.

Examples:

- [שלושה [אנשים]]
- [הרבה [חיות]]
- [מקטע של [כביש 6]]
- [נחיל [דגי דקר]]

The longer mention (with the quantifier) and the shorter mention (without it) are **not** clustered together. References to the entity should use the quantified form. This applies also to quantifiers like "כל" (all).

Example:

[כל [הילדים בגן]] הלכו לישון. [הם] קמו לקראת ארוחת הצהריים.
Cluster: {כל הילדים בגן, הם}

Example:

[שלושת [הילדים בגן]] הלכו לישון. [הם] קמו לקראת ארוחת הצהריים.
Cluster: {שלושת הילדים בגן, הם}

Lists

When a list appears in the text and is referred to later, the entire list is marked as one mention, and the cluster contains:

1. The full list mention
2. The reference to it

Example:

את הקורס *מבוא למדעי המחשב* באוניברסיטה העברית מלמדים השנה [פרופ' אביב זהר, פרופ' ג'פרי רוזנשיין וד"ר אריה שלזינגר]. [הם] מעבירים את הקורס יחדיו כבר שנים רבות.

Cluster: { "פרופ' אביב זהר, פרופ' ג'פרי רוזנשיין וד"ר אריה שלזינגר", "הם" }

Special Cases

Organizations and Sub-Organizations

If an organization and a sub-unit within it appear, they are not clustered together.
Example:

[2[הם] ישבתו מחר אחרי שהמשא ומתן בין 2[200 מעובדיו] שכל הודיע אתמול 1[משרד התחבורה]
לא צלח 1[משרד].
Cluster 1: {משרד התחבורה, משרד}
Cluster 2: {מעובדיו, הם 200}

Gender or Number Mismatch

If two mentions differ in gender or number but clearly refer to the same entity, they are still clustered together.
Example:

הזמנתי [כוס קפה] בארומה, אך [הוא] עדיין לא מוכן.
Cluster: {כוס קפה, הוא}

GPEs (Countries) and Governments

Countries and their governments are **not** linked. Countries and their citizens are treated separately.

Example:

עם קום המדינה קיוותה [ממשלת [ישראל]] שעצרת האו"ם תקבל את תוצאות מלחמת העצמאות ואת חלוקת העיר בין [ישראל] לירדן…
Cluster: {ישראל, ישראל}
Note: “ממשלת ישראל” is not clustered with “ישראל.”

Example:

לאחר מלחמת ששת הימים בה ניצחה [ישראל], [אזרחי [המדינה]] חגגו ברחובות.
Cluster: {ישראל, המדינה}

Negation

A noun phrase with a negation quantifier is not clustered with other mentions, even if they appear to refer to the same object. This is because the negated NP denotes an empty set that cannot be referred to in the real world.

Example:

אף ילד] מבית מהכיתה לא הגיע היום לבית ספר. [כולם] היו חולים].
Cluster: {כולם}

Example:

אין לי [כדורגל]. ממש הייתי רוצה לקנות [אחד].
Cluster: {אחד}