Morfix™ Insights
By melingo | Exeptional Language Solutions

Text Analysis for Semitic Languages

Melingo's *Morfix Insights* is a text-analysis system developed especially for the Semitic languages, addressing their rich morphology and high amount of ambiguity. This document describes the Hebrew version.

Morfix Insights is the state of the art solution for Hebrew analysis, used by leading enterprise, government and security organizations in Israel.
See our terms of use on https://www.melingo.com/tou-text-analysis-api/.

✔ Out-of-the-Box Entity Extraction

Morfix Insights can identify entities belonging to several **built-in categories** – right out of the box (generic entities).

✔ Morphological Analysis

All forms are identified in the text, including different inflections, conjugations, prefixes and suffixes, spelling alternatives.

✔ Disambiguation

Performs context analysis to choose the correct lemma in case of ambiguity.

✔ Intent Extraction

Extract the intent (meaning) of a sentence or a query trained by examples from your domain. Morfix Insights reaches high accuracy rates even with relatively small training sets, thanks to its unique pipeline that include normalization (unification) and disambiguation of the different inflections of Semitic words.

✔ User-Defined Entity Extraction

Define any custom entities to match the concepts in your organization's field. Morfix Insights will take care of identifying them in the text, overcoming inflections and ambiguities.

✔ Web API

Morfix Insights is available as web API-based product, accessible by any programming language.

1. Get Insights (Intents & Entities)

Create a POST request using the URL below with 2 parameters: "**Query**" & "**UserID**".
We will provide UserIDs upon request. We will also provide an authentication token specific to the UserID.

- **URL:** https://insights.morfix.com:8000/get-insights
- **Payload:** {"Query":"התורים לאורטופד בנתניה נורא ארוכים", "UserID": example@example.com}
- **Authorization token:** add new custom header for authentication.
  Header name: *Authorization*, Header value: *Token example123456789* **(note that the word "Token" is part of the token)**
- **Example output:**


```
{
  "input": "התורים לאורטופד בנתניה נורא ארוכים",
  "text_analysis": [
    {
      "intents": [
        {
          "intent": "זמינות רופאים",
          "confidence": 0.61
        }
      ],
      "entities": [
        {
          "entity": "תור",
          "value": "תור",
          "token": "התורים",
          "location": [
            0,
            5
          ],
          "confidence": 0.9
        },
        {
          "entity": "רופא מומחה",
          "value": "אורתופד",
          "token": "לאורטופד",
          "location": [
            7,
            14
          ],
          "confidence": 0.9
        }
      ],
      "generic_entities": [
        {
          "entity": "Location",
          "value": "נתניה",
          "token": "בנתניה",
          "location": [
            16,
            21
          ],
          "confidence": 1
        }
      ]
    }
  ]
}
```

2. Post Training Data

Create a POST request using the URL below with 2 parameters: "**Data**" & "**UserID**".

- **URL:** https://insights.morfix.com:8000/post-training-data
- **Payload:** {"Data": *see json below*, "UserID": example@example.com}
- **Authorization token:** add new custom header for authentication.
  Header name: *Authorization*, Header value: *Token example123456789* **(note that the word "Token" is part of the token)**
- **Example input:**

```
{"data":{
    "entities": [
        {
            "values": [
                {
                    "value": "רופא עור"
                },
                        {
                    "value": "רפואת עור"
                },
                        {
                    "value": "דרמטולוג"
                }
            ],
            "entity": "רפואת עור"
        }
    ],
    "intents": [
        {
            "intent": "זימון תור",
            "examples": [
                {
                    "text": "אני רוצה להזמין תור לרופא"
                },
                {
                    "text": "אני מבקש זימון תור"
                },
                {
                    "text": "אפשר לקבוע תור"
                }
            ]
        }
    ]
},
 "userID":"example@example.com"}
}
```

- **Example output:**

```
Data updated successfully
```

3. Get Training Data

Create a POST request using the URL below with a single parameter: "**UserID**".

- **URL:** https://insights.morfix.com:8000/get-training-data
- **Payload:** {"UserID": example@example.com}
- **Authorization token:** add new custom header for authentication.
  Header name: *Authorization*, Header value: *Token example123456789* **(note that the word "Token" is part of the token)**
- **Example output:**


```
{
	"intents":
	[{
		"intent": "זימון תור",
		"examples":
			[{
				"text": "אני רוצה להזמין תור לרופא"
			},
			{
				"text": "אני מבקש זימון תור"
			},
			{
				"text": "אפשר לקבוע תור"
			}
		]
	}],

	"entities":
	[{
		"entity": "רפואת עור",
		"values":
			[{
				"value": "רופא עור"
			},
			{
				"value": "רפואת עור"
			},
			{
				"value": "דרמטולוג"
			}
		]
	}]
```

4. Get Insights (Morphology)

The 'morphology' key is enabled for workspaces that require this information.
Morphology analysis can be either included on its own or combined with intents & entities analysis.
In the following examples, intents & entities are disabled and morphology is enabled.

- **URL:** https://insights.morfix.com:8000/get-insights
- **Payload:** {"Query":"התורים לאורטופד", "UserID": example@example.com}
- **Authorization token:** add new custom header for authentication.
  Header name: *Authorization*, Header value: *Token example123456789* **(note that the word "Token" is part of the token)**
- **Example output:**


```
{
  "input": "התורים לאורטופד",
  "text_analysis": [
    {
      "intents": [],
      "entities": [],
      "generic_entities": []
    }
  ],
  "morphology": [
    {
      "token": "התורים",
      "token_len": 6,
      "lemma": "תור",
      "lemma_voc": "תּוֹר",
      "lemma_id": 7629,
      "lemma_POS": 1,
      "lemma_root": "תור",
      "family_id": 7629,
      "score": 98,
      "person": 0,
      "gender": 1,
      "number": 2,
      "tense": 0,
      "prefix_len": 1,
      "construct": 0,
      "conjugation": 0,
      "phrase_id": 0,
      "phrase_lemma": "",
      "phrase_num_of_tokens": 0,
      "token_num_in_phrase": 0,
      "alternatives": []
    },
    {
      "token": "לאורטופד",
      "token_len": 8,
      "lemma": "אורתופד",
      "lemma_voc": "אוֹרתוֹפֶד",
      "lemma_id": 14683,
      "lemma_POS": 1,
      "lemma_root": "",
      "family_id": 14681,
      "score": 92,
      "person": 0,
      "gender": 1,
      "number": 1,
      "tense": 0,
      "prefix_len": 1,
      "construct": 0,
      "conjugation": 0,
      "phrase_id": 0,
      "phrase_lemma": "",
      "phrase_num_of_tokens": 0,
      "token_num_in_phrase": 0,
      "alternatives": []
    }
  ]
}
```

- In case of alternative analyses to a word, they will be presented in descending order according to the score.


```
{
  "input": "הפקיד נחמד",
  "text_analysis": [
    {
      "intents": [],
      "entities": [],
      "generic_entities": []
    }
  ],
  "morphology": [
    {
      "token": "הפקיד",
      "token_len": 5,
      "lemma": "פקיד",
      "lemma_voc": "פָּקִיד",
      "lemma_id": 19152,
      "lemma_POS": 1,
      "lemma_root": "פקד",
      "family_id": 19152,
      "score": 99,
      "person": 0,
      "gender": 1,
      "number": 1,
      "tense": 0,
      "prefix_len": 1,
      "construct": 0,
      "conjugation": 0,
      "phrase_id": 0,
      "phrase_lemma": "",
      "phrase_num_of_tokens": 0,
      "token_num_in_phrase": 0,
      "alternatives": [
        {
          "lemma": "הפקיד",
          "lemma_voc": "הַפָּקִיד",
          "lemma_id": 19123,
          "lemma_POS": 8,
          "lemma_root": "פקד",
          "family_id": 19123,
          "score": 89,
          "person": 3,
          "gender": 1,
          "number": 1,
          "tense": 1,
          "prefix_len": 0,
          "construct": 0,
          "conjugation": 6,
          "phrase_id": 0,
          "phrase_lemma": "",
          "phrase_num_of_tokens": 0,
          "token_num_in_phrase": 0
        }
      ]
    }
  ]
}
```

5. Add Entities To Training

Create a POST request using the URL below with 2 parameters: "**Data**" & "**UserID**".

- **URL:** https://insights.morfix.com:8000/add-entity-to-training
- **Payload:** {"Data": *see json below*, "UserID": example@example.com}
- **Authorization token:** add new custom header for authentication.
  Header name: *Authorization*, Header value: *Token example123456789* **(note that the word "Token" is part of the token)**
- **Example input:**

```
{"data": {
            "values": [
                {
                    "value": "רופא ילדים"
                },
                {
                    "value": "רופא משפחה"
                }
            ],
            "entity": "רופאים"
        },
        {
            "values": [
                {
                    "value": "טופס התחייבות"
                },
                {
                    "value": "17 טופס"
                }
            ],
            "entity": "טפסים"
        }
"userID":"example@example.com"}
```

6. Add Entity Values To Training

Create a POST request using the URL below with 2 parameters: "**Data**" & "**UserID**".
This function inserts values only to **existing** entities.

- **URL:** https://insights.morfix.com:8000/add-entity-values-to-training
- **Payload:** {"Data": *see json below*, "UserID": example@example.com}
- **Authorization token:** add new custom header for authentication.
  Header name: *Authorization*, Header value: *Token example123456789* **(note that the word "Token" is part of the token)**
- **Example input:**

```
{"data": {
            "values": [
                {
                    "value": "רופא עור"
                },
                {
                    "value": "רופא נשים"
                }
            ],
            "entity": "רופאים"
        }
"userID":"example@example.com"}
```

7. Add Value Synonyms To Training

Create a POST request using the URL below with 4 parameters: "**UserID", "entity", "value" & "synonyms**".
This function inserts synonyms only to **existing** entity values.

- **URL:** https://insights.morfix.com:8000/add-value-synonyms-to-training
- **Payload:** {"UserID": example@example.com, "entity": "רופאים", "value": "רופא נשים", "synonyms": ["גניקולוג"]}
- **Authorization token:** add new custom header for authentication.
  Header name: *Authorization*, Header value: *Token example123456789* **(note that the word "Token" is part of the token)**

JSON output description

The JSON output for a given text input includes a list of analyses per each token in the input text.

Keys relating to the token

- **token** – The exact text of the **token** as received: the string that the current reply relates to (including punctuation marks).
- **offset** – The offset (in characters) from the start of the input to the beginning of the token in the input text. This can be useful for display or indexing purposes.
- **token_len** – The length of the token in characters. This can be useful for display purposes.
- **token_num** – The index of the token in the list of tokens returned per text input, starting from 0.

Keys relating to morphological analysis

- **lemma_id** – A unique identifier for the lemma (the base form) of the token. The id serves to identify the lemma in case of ambiguous lemmas (e.g. the lemma טיפוס, noun, relates to three different ids, with different meanings: climbing, a disease, and "a character").
- **family_id** – The id of the head of the semantic family of the current lemma (if there is one). For example, for the noun טיפוס in the sense of climbing, the family id is the id of the lemma טיפס, verb.
- **lemma** – The basic form of the word. This is the form shown as the entry in most dictionaries, and it serves as the common denominator for a set of forms that are inflectionally related. Lemmas are shown in their Ktiv Male form (e.g. for טפול, והטפול – the lemma is טיפול). Also see lemma_voc
- **lemma_voc** – the vocalized form of the lemma (lemma with Hebrew Nikud diacritics).
- **lemma_POS** – A code value representing the Part of Speech of the lemma. Major POSs are 1: noun, 5: adjective, 8: verb. For the complete list see Part of Speech table.
- **lemma_root**– The root of the lemma (usually, a three letter string:טיפול טפל)
- **score** – The relative score of the morphological analysis.
- **person** – The grammatical person of the result. Relevant for verbs. Possible values are 1 -first person, 2 - second person, 3 - third person. 0 – not relevant.
- **gender** – The grammatical gender of the result; possible values are 1 - masculine, 2 - feminine, 3 – masculine and feminine, 0 –not relevant.
- **number** – The grammatical number of the result; possible values are 1 - singular, 2 - plural, 3 - dual, and 0 – not relevant.
- **tense** – The tense of the result. Relevant for verbs. Possible values are 1 - past, 2- present, 3-future, 4-imperative, 5- infinitive, 6-beynoni pa'ul, 7-makor nismach, or 0-not relevant.
- **prefix_len** – the length of the prefix if it exits (e.g. לכרטיס – prefix_len = 1)
- **construct** – The result's role within a construct state (סמיכות). Possible values are: 1- נסמך, 0- נפרד. Example: *מספר חשבון*– the word *מספר* is in construct state (which means it forms a syntactic unit with the next word).
- **conjugation** – The conjugation (Binyan) of the result. Relevant for verbs. Possible values are 1 - kal, 2 – nif'al, 3 – pi'el, 4 – pu'al, 5 – hif'il, 6 – huf'al, 7 – hitpa'el
- **alternatives** – Alternative morphological analyses for the input token, if there are any alternatives to the chosen analysis. Each alternative analysis contains all the morphological keys as the main analysis (but not the token keys)

Keys relating to phrases

Examples are בית ספר, לקח בחשבון.

- **phrase_lemma** – The canonical form of the collocation, equivalent to the lemma for single word tokens (e.g. בית ספר for בתי הספר).
- **phrase_id** – a unique identifier for the collocation.
- **phrase_num_of_tokens** - The total number of words in the collocation (e.g. in בית ספר there are two words).
- **token_num_in_phrase** – the position of the current word within the collocation (e.g. for הספר in בתי הספר - **token_num_in_phrase** = 2).

Keys relating to entities

If the current token is identified as an entity or a part of an entity, the extraction information appears as follows:

- **type** – The type of the entity: a string label (examples: "Location", "Organization"). See Entity Type Options table above for the list of default types extracted.
- **entity_counter**- The index of the entity in the total entities returned per input text.
- **entity_phrase** – The string in the input text that is associated with the current entity. Can include the current token only or the current token with neighboring tokens.

Value Tables

Part of Speech	
0	NO POS
1	NOUN
2	PROPER_NOUN
3	NUMERIC
4	NUMERIC_NOUN
5	ADJECTIVE
6	PROPER_ADJECTIVE
7	ADVERB
8	VERB
9	PRONOUN
10	INTERROGATIVE_PRONOUN
11	PREPOSITION
12	CONJUNCTION
13	INTERJECTION
14	INTERROGATIVE
15	PREFIX
16	ACRONYM
17	POSSIBLE_PROPER_NAME
18	SUFFIX
19	PHRASE
20	CHEMICAL_SIGN
21	DETERMINER

Gender	
0	NO GENDER
1	MASCULINE
2	FEMININE
3	BOTH GENDERS

Generic Entities
PERSON / OTHER
ORAGANIZATION
LANGUAGE
MEDIA
TIME EXPRESSION
LOCATION
NATIONALITY / AFFILIATION
CURRENCY
WEAPON
MILITARY
ILLEGAL DRUG
MISCELLANEOUS
SEA / AIRCRAFT
MEDICAL TERM
CATASTROPHE
DANGEROUS SUBSTANCES
TRAVEL
VEHICLE
NET ADDRESS
MEASUREMENTS
PHONE NUMBERS
CITY
STREET
NEIGHBORHOOD
HOUSE NUMBER
CREDIT CARD NUMBER
ID NUMBER
FLIGHT NUMBER
UNIVERSITY
COMPANY