ellexis.

An integrated computational environment for the study and processing of language

User Guide

Corpus Page & Tools Page Documentation

Version 2.0 — July 2026

A. Fiotaki & V. Tsakalis

https://ellexis.gr

Table of Contents

1 Introduction

2 Getting Started

3 Corpus Page

3.1 Search Tab

3.2 Statistics Tab

4 Tools Page

4.1 Upload Tab

4.2 Edit Tab

4.3 Search Tab

4.4 Statistics Tab

5 File Formats

6 Keyboard Shortcuts & Tips

7 Troubleshooting

8 Citation

1. Introduction

ellexis is a web-based computational environment for corpus linguistics, designed for the study and processing of the Greek language. The platform provides two main pages for working with linguistic data:

Corpus Page (ellexis.gr/corpus) — Browse, search, and analyze the pre-loaded ellexis research corpus, focusing on ότι/να subordinate clauses in Modern Greek.

Tools Page (ellexis.gr/tools) — Upload your own annotated corpus files, edit annotations in a spreadsheet-style editor, perform multi-field searches and token sequence pattern matching, and run comprehensive statistical analysis including frequency distributions, custom frequency queries, pivot tables, collocations, and hapax/dis legomena analysis. Supports CoNLL-U, CSV, TXT, and XLSX formats.

All corpus data uses the CoNLL-U annotation format with 10 columns: ID, Form (word), Lemma, UPOS (Universal Part-of-Speech), XPOS (language-specific POS), Features, Head, Deprel (dependency relation), Deps, and Misc.

2. Getting Started

Access ellexis

1. Open your web browser (Chrome, Firefox, Edge, or Safari)
2. Navigate to **https://ellexis.gr**
3. Use the navigation bar to switch between pages: **Corpus**, **Tools**, **About**, **Contact**

System Requirements

ellexis runs entirely in your web browser. No installation is required. For optimal performance with large corpus files (50,000+ tokens), use a modern browser with at least 4 GB of RAM.

> ***Tip:*** *For large files, ellexis shows a loading indicator while processing. Do not close the browser tab during processing.*

3. Corpus Page

The Corpus page provides access to the pre-loaded ellexis research corpus. It has two tabs: **Search** and **Statistics**.

3.1 Search Tab

Search for words, lemmas, or annotations

1. Navigate to the **Search** tab
2. Enter a search term in one or more filter fields: **Word**, **Lemma**, **UPOS**, **Deprel**, **XPOS**, or **Features**
3. Click **Search**
4. Results appear as sentence cards with matching tokens highlighted

View token annotations

1. Click on any sentence card to expand it
2. The annotation table shows all tokens with ID, Form, Lemma, UPOS, XPOS, Features, Head, and Deprel
3. Matching tokens are highlighted in the table

Show/Hide all annotations

1. Click **Show All Annotations** to expand every sentence
2. Click **Hide All Annotations** to collapse all

Download search results

1. Click the **Download** button after searching
2. Choose a format: **CSV**, **TXT**, **XLSX**, or **DOCX**

3.2 Statistics Tab

The Statistics tab displays pre-computed metrics about the ellexis corpus including frequency distributions, POS distribution, collocations, and sentence length analysis.

4. Tools Page

The Tools page lets you work with your own annotated corpus. It has four tabs: **Upload**, **Edit**, **Search**, and **Statistics**.

4.1 Upload Tab

Upload a corpus file

1. Navigate to the **Upload** tab
2. Drag and drop a file onto the upload area, or click **Choose File**
3. Accepted formats: **.conllu**, **.conll**, **.txt**, **.csv**, **.xlsx**
4. All formats must follow the 10-column CoNLL-U structure
5. After upload, a corpus preview appears with pagination

Note: A loading indicator appears while processing large files.

Download format templates

1. Find the **Example format templates** section below the upload area
2. Click any template: **CoNLL-U**, **TXT**, **CSV**, or **XLSX**
3. Open the file to see the required format with example data

Replace or remove a file

1. Click **Replace file** to upload a different corpus
2. Click **Remove** to clear the current corpus

4.2 Edit Tab

Edit a token value

1. Click on any cell in the spreadsheet-style table to edit it
2. Double-click a cell to open a popup editor (useful for long Features strings)
3. Press **Apply Changes** to save your edits

Tip: *The Features column is wider than others to accommodate long feature strings.*

Full Screen mode

1. Click **Full Screen** in the toolbar to expand the editor
2. Click **Exit Full Screen** to return to normal view

Find & Replace

1. Enter a search term in the **Find** input
2. Select a field: **All Fields**, **Form**, **Lemma**, **UPOS**, etc.
3. Choose a match mode: **Contains**, **Exact Match**, **Starts With**, or **Ends With**
4. Optionally enable **Enable Regex** for regular expression matching
5. Click **Find** — matching cells are highlighted
6. Enter replacement text and use **Replace Next**, **Replace Page**, **Preview All**, or **Replace All**

Example: *Find UPOS = 'COMP', Replace with 'SCONJ' to rename a POS tag across the entire corpus.*

Note: ID fields are protected. Head fields must remain numeric.

Conditional Replace

1. Click **Conditional Replace** to expand the rule panel
2. Add IF conditions (AND logic): e.g., IF Lemma = 'X' AND Features contains 'Y'
3. Set THEN action: e.g., set UPOS to 'VERB'
4. Click **Preview** to review changes, then **Apply**

Example: *IF Lemma Exact 'λέω' AND Features Contains 'Tense=Past' THEN set Deprel to 'root'.*

Undo, Import & Save

After any operation, click the green **Undo** button to revert. Use **+ Import CoNLL** to paste and validate new sentences. Use **Save & Download** to export in CoNLL-U, CSV, TXT, or XLSX.

4.3 Search Tab

Basic search

1. Enter a search term in **Find**
2. Select a field and match mode
3. Optionally enable **Enable Regex**
4. Click **Search**

Example: *Search for Lemma Exact 'φαίνομαι' to find all forms of the verb.*

Advanced Filters (multi-field AND search)

1. Click **Advanced Filters** to expand the filter panel
2. Click **+ Add filter** to add conditions (each with field, match mode, value)
3. All filters use AND logic — a token must match ALL conditions
4. Click **Search**

Example: *Main query: Lemma Contains 'μαθ' + Filter: UPOS Exact 'VERB' + Filter: Features Contains 'Tense=Past'. Finds all past-tense verbs with lemma containing 'μαθ'.*

Token Sequence Search

1. Click **Token Sequence Search** to expand
2. Define **Word 1**: field, match mode, value
3. Set **Distance** (1 = adjacent, 2 = one token apart)
4. Define **Word 2**
5. Optionally add **Word 3**
6. Click **Search Pattern**

Example: *Word 1: Lemma Exact 'ότι', Distance 1–3, Word 2: UPOS Exact 'VERB'. Finds 'ότι' followed by a verb within 3 token positions.*

Example: *Word 1: UPOS Exact 'DET', Distance 1, Word 2: UPOS Exact 'ADJ', Distance 1, Word 3: UPOS Exact 'NOUN'. Finds adjacent DET+ADJ+NOUN sequences.*

4.4 Statistics Tab

The Statistics tab auto-computes comprehensive metrics when first visited. All statistics reflect the current state of your corpus, including any edits.

Scorecards

Nine summary metrics are displayed in two rows:

Metric	Description
Sentences	Total number of sentences in the corpus
Word Types	Total number of tokens (all words including repetitions)
Lemmas	Number of unique base forms (dictionary entries)
Unique Tokens	Number of distinct word forms
Avg. Words/Sent	Average number of tokens per sentence
Lexical Density	Ratio of content words (NOUN, VERB, ADJ, ADV) to total tokens
Hapax Ratio	Percentage of unique tokens appearing exactly once (with count)
Dis Legomena	Percentage of unique tokens appearing exactly twice (with count)
TTR	Type-Token Ratio: unique tokens / total tokens

Word Types Frequency

Ranked list of all word forms in the corpus, sorted by frequency (descending). Shows rank, word form, count, and percentage of total tokens.

> ***Example:*** *Row 1: φαίνεται — 1,049 — 1.2%. The most frequent word form appears 1,049 times.*

Lemma Frequency

Ranked list of all lemmas (base forms), sorted by frequency. Useful for identifying the most common vocabulary regardless of inflection.

> ***Example:*** *Lemma 'φαίνομαι' covers forms like φαίνεται, φαίνεσαι, φαινόταν.*

POS Distribution

Frequency of each Universal Part-of-Speech (UPOS) tag in the corpus. Shows how the corpus is distributed across grammatical categories.

> ***Example:*** *VERB: 16,748 (19.9%), NOUN: 14,743 (17.6%), DET: 11,520 (13.7%).*

Dependency Relations

Frequency of each dependency relation label. Shows the syntactic structure patterns in the corpus.

> ***Example:*** *det: 12,230 (14.6%), aux: 7,862 (9.4%), root: 6,802 (8.1%).*

Feature Analysis

Morphological feature distributions. Use the **Feature** dropdown to select a feature key (e.g., Aspect, Case, Gender, Mood, Number, Person, Tense, Voice). The table shows the frequency of each value for the selected feature.

> ***Example:*** *Select 'TENSE': Pres: 10,207 (61%), Past: 6,454 (39%).*

Table Controls (all tables)

1. Click any table header to **collapse/expand** it
2. Use **Show** dropdown: 10, 20, 30, 50, 100, 150, 200, or All rows
3. Click **Count** or **%** to toggle bar chart visualizations
4. Export table data: **CSV**, **TXT**, **XLSX**
5. Export charts: **PNG**, **JPEG**, **SVG**

Note: Charts always render from the full dataset regardless of the table row limit.

Frequency Explorer (Custom Frequency Query)

Build custom multi-filter queries to explore token subsets. All filters use AND logic. Results show total count, percentage, and a detailed breakdown by a chosen dimension.

1. Select a field (UPOS, Lemma, Features, etc.), match mode, and value
2. Click **+ Add** to add the filter (appears as a tag pill)
3. Add more filters as needed
4. Click **Compute**
5. Click **Show Details** to expand the breakdown
6. Use **Group by** to change the breakdown dimension (Word Type, Lemma, UPOS, etc.)

Frequency Explorer examples:

Query	Filters	Group by	What it shows
Past imperfective verbs	UPOS = VERB Features contains Tense=Past Features contains Aspect=Imp	Lemma	Which verbs appear in past imperfective
Root verbs only	UPOS = VERB Deprel = root	Lemma	Most common main predicates
Accusative nouns	UPOS = NOUN Features contains Case=Acc	Word Type	Which noun forms appear in accusative
Specific lemma features	UPOS = NOUN Lemma = 'παιδί' Features contains Case=Acc	Word Type	Accusative forms of 'παιδί'

Pivot Table (Cross-Tabulation)

Create cross-tabulation tables with up to 4 dimensions. Each dimension can be assigned as a Row or Column axis. Supports value filtering via search box, dropdown, or feature sub-key selection.

1. Each dimension shows a **ROW** (blue) or **COL** (teal) badge
2. Click **Switch to Row/Column** to change axis
3. Select a field (UPOS, XPOS, Deprel, Lemma, Word Type, Features)
4. For **Features**: select a Feature key (e.g., Tense), then add specific values in Key=Value format
5. For other fields: type in the search box to filter, select from dropdown, click **+ Add**
6. Click **+ Add Dimension** for up to 4 total
7. Click **Generate**
8. Use **Stacked Count** or **Stacked %** for visualizations

Pivot Table examples:

Rows	Columns	What it shows
UPOS	Deprel	Which POS tags serve which syntactic roles
UPOS (VERB, NOUN only)	Tense (all values)	Tense distribution across verbs and nouns
Deprel	UPOS	POS composition within each dependency relation
Lemma (search: μαθ)	Tense (all values)	Tense usage for verbs containing μαθ

Collocations

Two tables showing 2-word and 3-word collocations. Built from actual adjacent tokens (only punctuation excluded). Only collocations with count > 2 are shown.

> ***Example:*** *2-word: 'φαίνεται να' (1,011), 'μάθει να' (347). 3-word: 'δεν φαίνεται να' (211).*

Hapax & Dis Legomena

Two searchable tables listing words that appear exactly once (hapax legomena) or exactly twice (dis legomena). Sorted alphabetically. Punctuation is excluded based on UPOS tags.

1. Use the **search/filter input** at the top of each table to find specific words
2. The filter updates instantly as you type

> ***Example:*** *Type 'απο' in the hapax search to find all hapax legomena starting with 'απο'.*

Sentence Length

Two tables: **Length Stats** (Min, Average, Max sentence length) and **Length Frequency** (how many sentences have each length, sorted by frequency).

Export Entire Statistics Tab

1. Click **Download PDF** for a branded report with all tables (top 10 rows each), all feature distributions, and hapax/dis legomena
2. Click **Download PPTX** for a PowerPoint with title slide, scorecards, all frequency tables with charts, and a closing slide

5. File Formats

All file formats must follow the 10-column CoNLL-U structure. Download a template from the Upload tab to see the exact format.

Supported Formats

Format	Extension	Separator	Headers	Notes
CoNLL-U	.conllu .conll	Tab	# sent_id = ... # text = ...	Standard format. Blank line between sentences.
Plain Text	.txt	Tab	Same as CoNLL-U	Identical structure, different extension.
CSV	.csv	Comma	Column header row: sent_id, id, form...	UTF-8 with BOM. Excel-compatible.
Excel	.xlsx	Spreadsheet cells	Column header row: sent_id, id, form...	Parsed via server. Same columns as CSV.

CoNLL-U Column Reference

#	Column	Description	Example Values
1	ID	Token index within sentence	1, 2, 3, 4...
2	Form	Surface word form	φαίνεται, το, να
3	Lemma	Dictionary/base form	φαίνομαι, ο, να
4	UPOS	Universal POS tag	VERB, NOUN, DET, ADJ, ADV, PRON, ADP, CONJ
5	XPOS	Language-specific POS	VbMn, NoCm, AtDf
6	Features	Morphological features (pipe-separated Key=Value)	Aspect=Imp|Mood=Ind| Number=Sing|Tense=Pres
7	Head	ID of syntactic head (0 = sentence root)	0, 3, 5, 2
8	Deprel	Dependency relation to head	root, nsubj, obj, det, amod, advmod, mark
9	Deps	Enhanced dependencies	_ (typically empty)
10	Misc	Miscellaneous annotation	_ (typically empty)

6. Keyboard Shortcuts & Tips

• Press **Enter** in any search field to execute the search immediately.

• Use **Ctrl+Shift+R** (Windows) or **Cmd+Shift+R** (Mac) to force-reload the page.

• In the Edit tab, **double-click** any cell to open the popup editor for long values.

• The **scroll-to-top** button appears in the bottom-right when you scroll down.

• All tables can be **collapsed** by clicking their header bar.

• Charts always render from the **full dataset**, regardless of visible table rows.

• Export buttons export **all data**, not just visible rows.

• Use **Enable Regex** for advanced pattern matching (see Section 4.2).

• The **Token Sequence Search** distance of 1 means adjacent tokens (no gap).

• The **Frequency Explorer** Group By dropdown re-computes instantly when changed.

• In the **Pivot Table**, use the search box to find specific lemmas or word types from large vocabularies.

• The **PDF export** includes all Feature distributions (one table per feature, top 10 values each).

7. Troubleshooting

Page shows old content

Press Ctrl+Shift+R (Windows) or Cmd+Shift+R (Mac) to force-reload. Ensure the script URL includes the latest version query string (e.g., ?v=48).

Upload fails or 'No valid sentences found'

Ensure your file follows the 10-column CoNLL-U structure. Download a template from the Upload tab. Each sentence must have a # sent_id line in CoNLL-U format.

Browser becomes slow with large corpus

For corpora over 100,000 tokens, some operations may take a few seconds. The loading indicator prevents browser freezing. Use specific search filters to narrow results.

CSV shows garbled Greek characters

CSV files include a UTF-8 BOM for Excel compatibility. If still garbled, open in Excel via File > Open and select UTF-8 encoding.

Statistics tab takes time to load

Statistics are computed client-side. First load of a large corpus may take 2–3 seconds. Subsequent visits are faster because data is cached until the corpus changes.

PDF or PPTX export fails

These features load libraries from CDN on demand. Ensure you have an active internet connection. Corporate firewalls may block CDN access — try a different network.

8. Citation

If you use ellexis in your research, please cite it as follows:

APA 7th Edition

Fiotaki, A., & Tsakalis, V. (2026). *ellexis: An integrated computational environment for the study and processing of language* (Version 2) [Web application]. https://ellexis.gr

BibTeX

```
@misc{fiotaki2026ellexis,
  author = {Fiotaki, A. and Tsakalis, V.},
  title = {ellexis: An integrated computational environment for the study and
processing of language},
  year = {2026},
  note = {Version 2, Web application},
  howpublished = {\url{https://ellexis.gr}}
}
```