Funded by the European Union

5. Texts and Corpora

This series of tutorials is based upon work from COST Action **Multi3Generation CA18231**, supported by COST (European Cooperation in Science and Technology).

COST (European Cooperation in Science and Technology) is a funding agency for research and innovation networks. Our Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career and innovation, cf. www.cost.eu

Texts and Corpora

- NooJ can open any type of text file in most formats, e.g., .txt, .doc, .docx, .html, .pdf, .rtf, .xml, etc., in all encodings: ASCII, ANSI, EBCDIC, UTF, etc.
- NooJ can process a single text, or a corpus. A corpus is a set of multiple texts, potentially thousands.

- All NooJ operations can be applied to a single text or a corpus
- All linguistic operations consist of applying some linguistic resource to a text (or a corpus), in order to add or remove annotations to the text (or corpus).
- Annotations are stored in the Text Annotation Structure (TAS). Annotated texts are stored in .not files; annotated corpora are stored in .noc files.

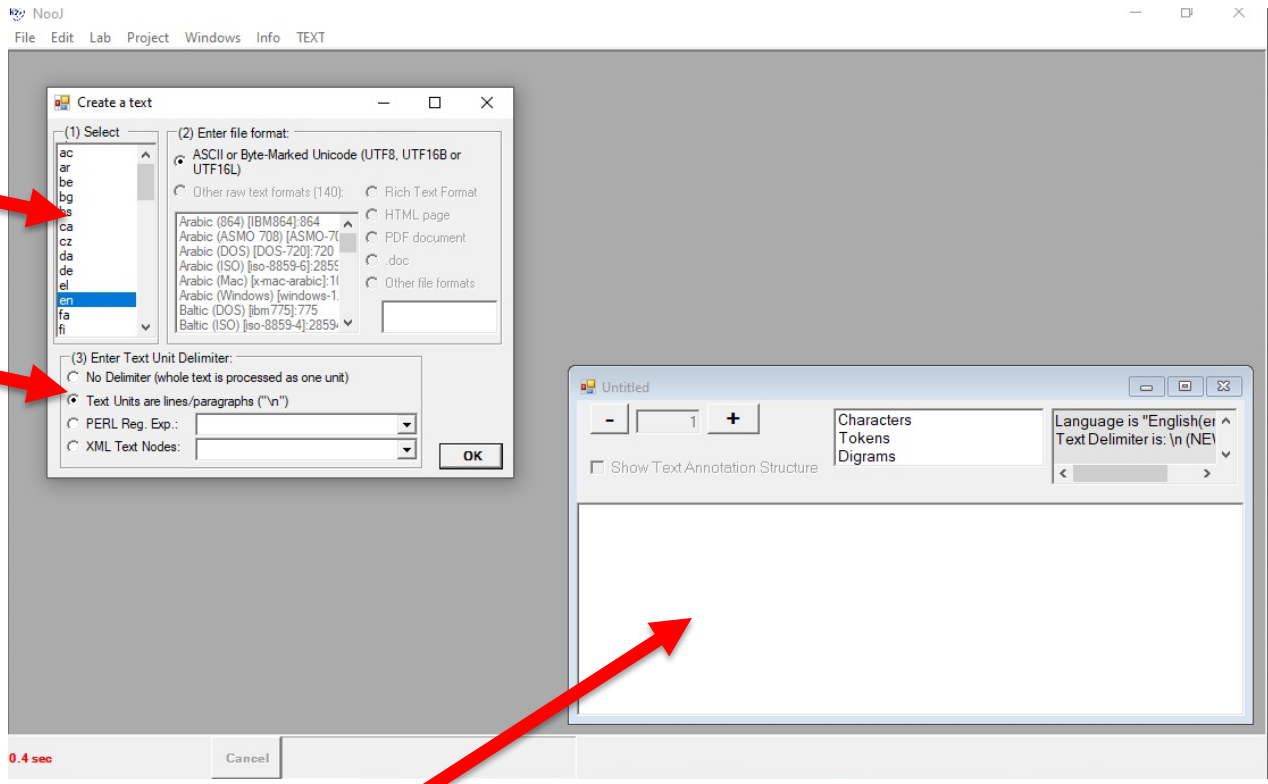
Import a text: two solutions

1. A quick and easy importation
   - File > New > Text to produce an empty text window
   - Select the text language, and the type of the text units
   - Copy & Paste the text content from any application (*e.g.*, MS Word) into NooJ empty text window

(a) Select the text language

(b) Nothing to select here: the text will be encoded as UTF8

(c) Select the type of text units



NooJ will process each text unit independently:

-- No delimiter: the whole file is processed as one single unit. Useful when texts are very small and queries need to find correlated pieces of information anywhere, e.g., in tweets and newsbits.

-- Text units are paragraphs delimited by the NEWLINE character ("\n"). This is the default setting.

-- Text units are delimited by some character sequences that can be described by a PERL regular expression. This is useful when the text file consistes of little sections; each section starts with a title easily described by a PERL expression, *e.g.*, "===2023-11-14"

-- The text is encoded as an XML file. Text Units are stored in XML nodes, e.g., <abstract> ... </abstract> <S> ... </S>, <title> ... </title>. This functionality allows NooJ to process multilingual texts if each language is clearly annotated, e.g., <en> ... </en> <fr> ... </fr> <sp> ... </sp>

(d) Copy and Paste in there

(e) Do not forget to save the text File > Save. It will be saved as a text NooJ file (.not).

Import a text: two solutions

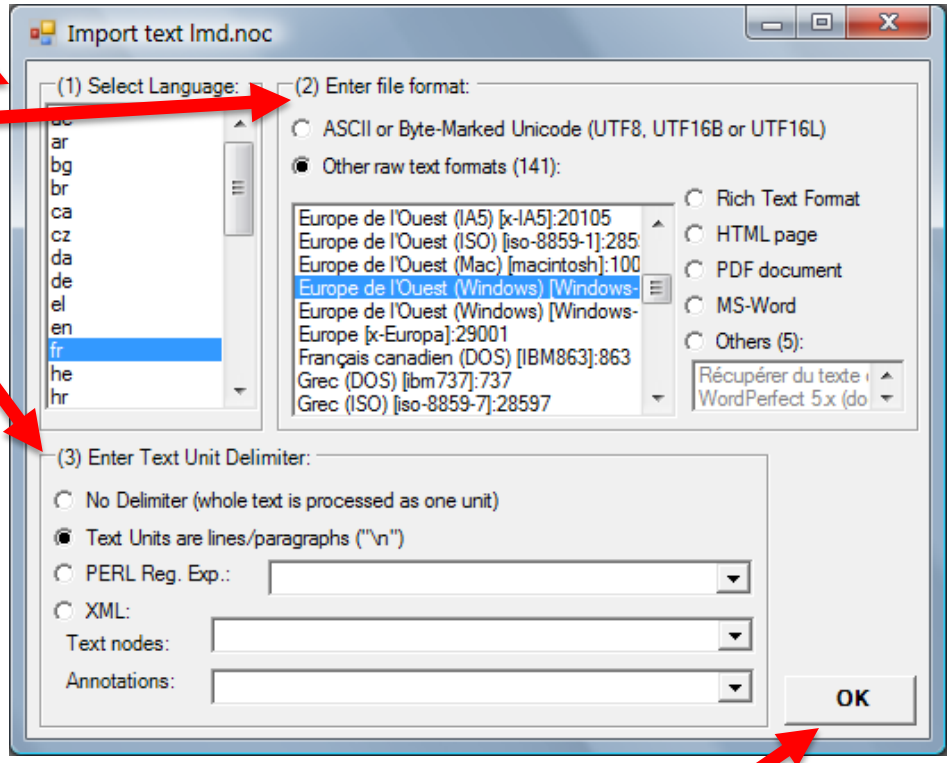
2. Import a file (more advanced option):
   - File > Open > Text
   - Select All types *.*
   - Select three parameters

Import a file: File > Open > Text > Import

(a) Select the language used in the text

(b) Select the file format

(c) Select how text units are separated.



NooJ will process each text unit independently:

-- No delimiter: the whole file is processed as one single unit. Useful when texts are very small and queries need to find correlated pieces of information anywhere, e.g., in tweets and newsbits.

-- Text units are paragraphs delimited by the NEWLINE character ("\n"). This is the default setting.

-- Text units are delimited by some character sequences that can be described by a PERL regular expression. This is useful when the text file consistes of little sections; each section starts with a title easily described by a PERL expression, *e.g.*, "===2023-11-14"

-- The text is encoded as an XML file. Text Units are stored in XML nodes, e.g., <abstract> ... </abstract> <S> ... </S>, <title> ... </title>. This functionality allows NooJ to process multilingual texts if each language is clearly annotated, e.g., <en> ... </en> <fr> ... </fr> <sp> ... </sp>

(d) Click OK

(e) Do not forget to save the text File > Save. It will be saved as a text NooJ file (.not).

Create a corpus: two steps

(1) Create an empty corpus

- File > New > Corpus
- Save (give it a file name: it will be stored as a .noc file.

- Select the three parameters: language, file format and text units separators

❑ These three parameters must be identical in all text files

(2) Add all the text files to the corpus

- Click “Add” or “Remove” to add/remove file to the corpus
- File > Save saves the corpus in a single .noc file.

❑ Make sure the text file names are ordered alphabetically, *e.g.*, ”Chapter 01”, “Chapter 02”, ... “Chapter 13” so that statistic analyses of the corpus will be ordered in a logical manner.

(a) Select the text language

(b) Nothing to select here: the text will be encoded as UTF8

(c) Select the type of text units

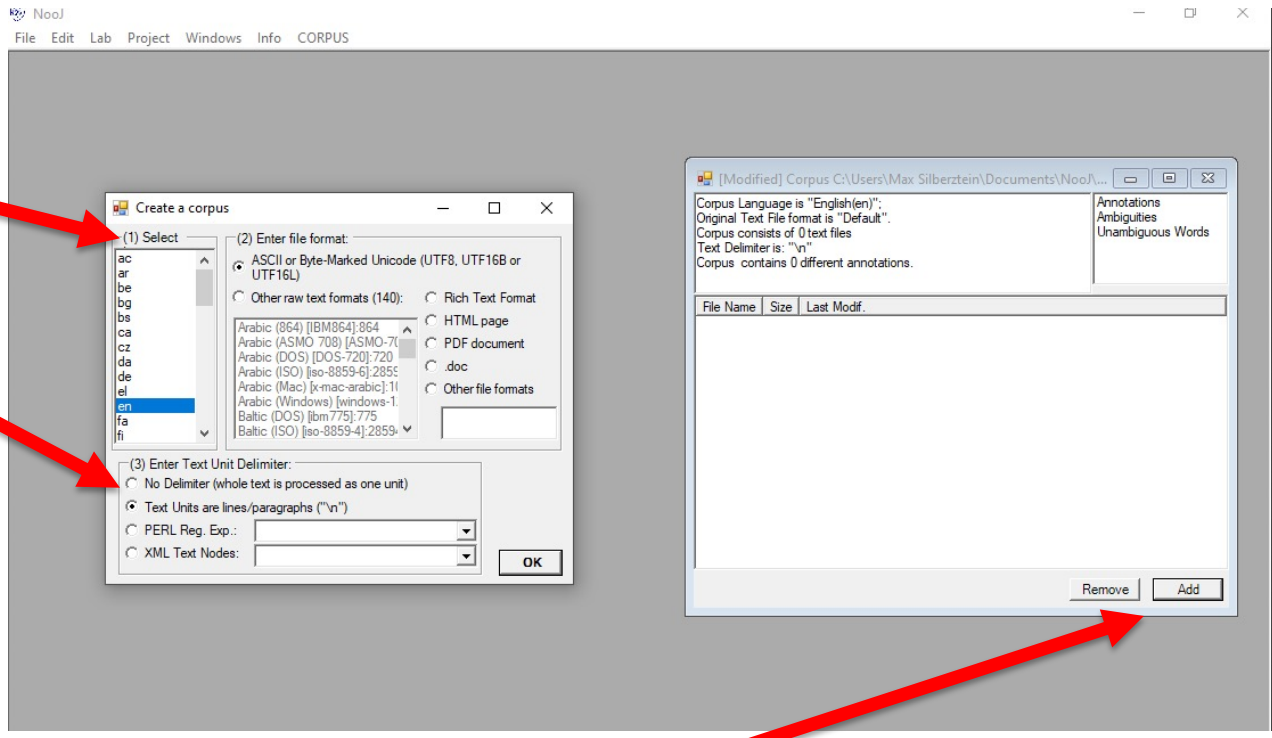
NooJ will process each text unit independently:

-- No delimiter: the whole file is processed as one single unit. Useful when texts are very small and queries need to find correlated pieces of information anywhere, e.g., in tweets and newsbits.

-- Text units are paragraphs delimited by the NEWLINE character (“\n”). This is the default setting.

NooJ creates an empty corpus

(d) Click Add to add any number of text files



Open & Save NooJ texts or corpora

- To open a NooJ text (.not file): File > Open > Text
- To open a NooJ corpus (.noc file): File > Open > Corpus

- To save the current text as a NooJ text (.not file): File > Save
- To save the current corpus as a NooJ corpus (.noc file): File > Save

- Default save folder is Documents\NooJ\<language>\Projects, *e.g.*,

Documents\NooJ\en\Projects for English texts,
Documents\NooJ\pt\Projects for Portuguese texts,
etc.

NooJ files and directories

- All your personal files should be stored in Documents\NooJ
- Never modify the content of the application folder (NooJApp)
- Inside the Documents\NooJ folder:
  - there are language folders, *e.g.*, **en**, **fr**, **it**, **pt**
  - By default, the pre-installed language folders are **en** and **fr**. Download other language resources from https://nooj.univ-fcomte.fr/resources.html
- File extensions:
  - NooJ annotated texts are stored in .not files
  - annotated corpora are stored in .noc files
  - dictionaries are stored in .dic files
  - agglutinative morphological grammars in .nom files
  - inflectional and derivational grammars in .nof files
  - syntactic grammars in .nog files
- Associate these file extensions with the **Nooj.exe** application, so that double-clicking them will launch NooJ.



CONGRATULATIONS


You know how to manage your texts and corpora to process them with NooJ