LÍNGUA NATURAL 2023/2024

Mini-Project Nº 1 (MP1)

Should be done:	☐ individually	☒ **in group**
Submission:	☐ theoretical class	☒ **Fenix submission**
Submission deadline:	**till 23:59, October 3rd**	

OBJECTIVES

Learn to work with transducers, using them to solve a problem.

STATEMENT

Suppose you want to create a date reading module using only transducers to couple with a speech synthesis system. To do this, the module must convert a date in a specific format to the corresponding text, that is, to a corresponding sentence as it would appear if transcribed by a human.

a. Create a transducer that converts a date from mixed condensed format (**mmm/dd/yyyy**, e.g., *JAN/20/2018*) to numeric format (**mm/dd/yyyy**, e.g., *01/20/2018*). You should follow these steps and develop the following transducers:
   1. Create the transducer **mmm2mm.fst** that converts a month indicated in the format "mmm" into the corresponding numerical value (always two digits). Example: *SEP* is converted into *09*;
   2. Create the transducer **mix2numerical.fst** that converts a date from mixed condensed format to numeric format. The day and the year must remain unchanged. Example: *SEP/20/2018* is converted to *09/20/2018*.

b. Suppose now that the texts you are processing can be in both English and Portuguese:
   3. Create the transducer **pt2en.fst** that converts a date in a mixed condensed format from Portuguese to English. The day and the year must remain unchanged. Example: *SET/5/2018* is converted into *SEP/5/2018*;
   4. Create the transducer **en2pt.fst** that performs the inverse operation. The day and the year must remain unchanged. Example: *FEB/05/2081* is converted into *FEV/05/2081*;

c. Develop a transducer to convert a date from numeric format (**mm/dd/aaaa**) to text:
   5. Create the transducer **day.fst**, which converts any number from 1 to 31 into the corresponding ordinal text. A "-" must be used between words. Example: *22* is converted into *_twenty_-_second*.
   6. Create the transducer **month.fst**, which converts the month to the corresponding text. This transducer converts any number from 1 to 12 into the corresponding text. Example: *9* or *09* is converted into *_September*;
   7. Create the transducer **year.fst**, which converts the year to the corresponding text. This transducer converts any number from 2001 to 2099 into the corresponding text. The word *and* is (only) used after the word *thousand*. Example: *2025* is converted into *_two_thousand_and_twenty_five*;
   8. Create the transducer **datenum2text.fst**, which converts a date from numeric format to the corresponding text. A coma must be used between the day and the year. Example: *09/15/2055* is converted to *_September_fifteenth_,_two_thousand_and_fifty_five"*.

d. Based on the previous transducers, build a transducer to convert a date to the corresponding text:
   9. Create the transducer **mix2text.fst**, which converts a date from mixed condensed format (English or Portuguese) to text. Example: *MAY/12/2088* or *MAI/12/2088* is converted to *_May_twelfth_,_two_thousand_and_eighty_eight*;
   10. Create the transducer **date2text.fst**, which accepts a date in either numeric or mixed format (English or Portuguese) and returns the corresponding text. Example: *OCT/31/2025*, *OUT/31/2025* and

10/31/2025 are converted to *_October_thirty_-_first_,_two_thousand_and_twenty_five*;

e. Test the transducers **mix2numerical.fst**, **en2pt.fst**, **datenum2text.fst**, **mix2text.fst**, and **date2text.fst** with the dates on which the members of the working group turned 18.

Assume that:

- All the outputs in this statement that use text are generated using the "syms-out.txt" file (see the tiny example). The examples that contain digits use the "syms.txt" file;
- The days in the condensed form can have 1 or 2 digits (example: *1* or *01*) in the input but always have two digits on the output unless stated otherwise;
- The months in the condensed form can have 1 or 2 digits (example: *9* or *09*) in the input but always have two digits on the output;
- The months in the mixed condensed forms always have three capital letters. In Portuguese, the names are: *JAN*, *FEV*, *MAR*, *ABR*, *MAI*, *JUN*, *JUL*, *AGO*, *SET*, *OUT*, *NOV*, *DEZ*;
- The years in the condensed form always have four digits and belong to the interval [2001..2099];
- The transducers will only be tested with valid dates;
- You can use other transducers not mentioned in the statement;
- The "syms.txt" and "syms-out.txt" files contain the symbols to be manipulated by the transducers and cannot be changed;
- The "output tape" must always contain a single output (must be deterministic);
- The names of the transducers must be exactly: **mmm2mm**, **mix2numerical**, **pt2en**, **en2pt**, **day**, **month**, **year**, **datenum2text**, **mix2text**, and **date2text**.

Other Examples

Here are some examples (using the "syms-out.txt" file):

- *9/09/2001* = *_September_ninth_,_two_thousand_and_one*
- *01/3/2011* = *_January_third_,_two_thousand_and_one*
- *02/24/2022* = *_February_twenty_-_fourth_,_two_thousand_and_twenty_two*
- *10/01/2099* = *_October_first_,_two_thousand_and_ninety_nine*
- *12/22/2043* = *_December_twenty_-_second_,_two_thousand_and_forty_three*
- *OCT/30/2025* = *_October_thirtieth_,_two_thousand_and_twenty_five*
- *DEZ/13/2069* = *_December_thirteenth_,_two_thousand_and_sixty_nine*
- *FEV/25/2071* = *_February_twenty_-_fifth_,_two_thousand_and_seventy_one*
- *MAR/21/2060* = *_March_twenty_-_first_,_two_thousand_and_sixty*

Software

To test the proposed solution use, in a Linux environment, the tools:

- "OpenFST" from Google (http://www.openfst.org/twiki/bin/view/FST/FstDownload).
- "Graphviz" (http://www.graphviz.org/);

Submission

Submit in Fenix, project *MP1*, a zip file with, and only with:

- A shell script [the name has to be "run.sh"] with **all** the commands used to generate all transducers, either in binary and in graphical format (PDF, PS, or PNG) from the ".txt" files;

- A folder "sources" containing all the text files used to define the transducers (extension ".txt");
- A folder "tests" with all the source test files (name has to start with "t-" and use extension ".txt"). A minimum of two tests must be included in this folder to be used with the date2text.fst transducer;
- A folder "compiled" containing all the compiled versions of all the transducers used, including the tests (extension ".fst");

- A folder "images" containing the graphical versions of all the transducers, including the tests (extension ".pdf", ".ps" or ".png");
- A folder "scripts" with the files needed to run your shell script. This folder is not mandatory;
- A short report with the following requirements:
    - The filename has to be "report.txt" or "report.pdf";
    - Must not exceed one page;
    - Must identify the group members with an estimate of each element's contribution to the work. For example, Peter: 60%, John: 40%, along with a short justification;
    - Must contain a brief description of your options;

You can make several submissions: a new one replaces the previous one.

Attention:

- Developed transducers must have exactly the same names as above;
- The four folders "sources", "tests", "compiled", and "images" should not contain sub-folders.

Evaluation criteria

The following criteria will be taken into account in the assessment (maximum = 20 points):

1. Use of unnecessary writing of transducers (up to -5 points). For example, you should avoid to write a transducer that can be generated using FST operations on existing transducers;
2. Correct operation of each transducer (1,5 points each);
3. Run.sh operating correctly (3 points);
4. Submission of the graphic versions of all transducers, as well as the examples, in their different forms, that is, before and after being fed as input to the **date2text** transducer (1 point);
5. Quality of the report [in Portuguese or English] including spelling and syntactic correction (1 point);

Non-compliance with any rule implies a minimum discount of 4 points (in 20 points).

During the evaluation of the correct operation of any transducer, it does not take into account the origin of the errors (e.g., when testing a transducer B, every time the expected output is not obtained, an error is taken into account, even when the origin of the error is the malfunction of another transducer used to generate B). So, malfunctions on the first transducers may impact the evaluation of the **date2text** transducer.

"*Academic Integrity*" in Língua Natural

In this course, each student is expected to subscribe to the highest standards of academic honesty. This means that every idea that is not the student's must be explicitly accredited to the respective author. Failure to do so constitutes plagiarism.

Plagiarism includes using ideas, code, or sets of solutions from other students or individuals, or any sources other than the course texts, without crediting those sources. Students are encouraged to discuss the problems with other students and should mention this discussion when they submit their results. This mention will NOT influence the grade. Students should not, under any circumstances, show to their classmates, even temporarily, their solutions to the quizzes or projects subject to evaluation. They should not even throw away drafts of the solutions without destroying them first, nor leave the developed code on shared-use computers.

Academic dishonesty also includes copying in exams. In this discipline, these should be taken without consulting any text or other classmates. Receiving or giving help during these exams is an act of academic dishonesty. Situations that could give rise to suspicions of dishonesty (opening backpacks to get paper, looking around instead of concentrating on the exam paper, etc.) should be avoided.

In this course, academic dishonesty is considered fraud, with all the legal consequences. Any fraud will have the immediate consequence of failing all students involved (including those who enabled it to occur). Any suspicion of academic dishonesty will be reported to the higher bodies of the school for disciplinary action. This may result in failure of the subject, failure of the year, temporary or permanent suspension from IST, or even from the University of Lisbon.

Copyright in Moodle and Fenix