"Lectomorphic" translation aid

The Project.

This is a project to do the tedious work of dictionary look-up for an early stage student of Polish who wants to engage with mature texts from the beginning. The student should first have obtained a good grasp of formal grammar, for instance from Oscar Swan's Grammar of Contemporary Polish. Learning the grammatical structure of a language is not arduous compared with learning the details. Software can look up every word in the text and this involves two dictionaries – firstly the very extensive dictionary of inflections from sgjp.pl and then the English word equivalents from the Lektorek Polish-English online dictionary (lektorek.org/polish). With this capability in hand it can be applied to exhaustively annotate every word in a Polish text with its various interpretations (and there are frequently multiple interpretations). This does not amount to translation. It is merely the mechanical steps leading up to the point where higher intelligence is applied in piecing together the meaning of the text.

The annotated text is presented in an html document in a way that makes it comprehensible for a user to interact with, both exploring the competing alternative word interpretations from the dictionaries and editing them to reduce to the chosen meaning. (The software relies on javascript processing by the user's browser. It does not make any use of html hyperlinks.)

Screen shots, all in a companion document, show what has been done so far. The project is not ready to be released to unsuspecting users but it could be released as a prototype to those who do not expect perfection but are prepared to contribute towards maturing the project.

Browser support. This work has been tested with Firefox, chromium-browser and Microsoft Edge. Whenever a browser is tried for the first time minor problems have to be fixed. Correct behaviour can be determined by reproducing the screenshots. An unsolved issue is that the software needs to know the fontwidth in pixels of whichever 12pt monospace font the browser will use. I do not know how to determine this except empirically, case by case. Whatever the value, it should be edited into the line

```
     const fw= 7.23;
```

in the main html file (currently line 57).

How to use -Introduction

In this example, the text is a historical document from the Gutenberg Project, O Sprawie Glodowej W Galicyi 1866 by Karol Langie (approx 80 pages). The title page is shown in screen shot *title.png*.

The four line display. In *Gdy.png,* the user has selected a sentence, Gdy bieda ..., and the sentence is displayed in a box in four lines at the top of the page. The first line is the selected text itself. The lines below are a word by word analysis and translation. The second line has the analysis of each word as found in the Grammatical Dictionary of Polish (internet address sgjp.pl). The

results from this dictionary are used directly except that a concise set of symbols has been substituted for brevity. For example the word kark is shown as sNA.mq_ which means, letter by letter, substantive (ie noun) Nominative or Accusative singular masculine impersonal inanimate. The next line down shows the lemma to which the word belongs. In about 50% of cases the word appears in its lemma form in any case. To cut down on clutter ~ signifies the word is in the lemma form. We see that lezie, the word coming after kark, analysed as v=.3(is an inflection of leźć. v=.3(stands for verb present-tense third-person and (means it is an imperfective verb.

The last line shows the English equivalents for the Polish word as listed in lektorek.org/polish from the University of Pittsburgh. Lektorek, though it is published on the internet is not a computerised reference from the beginning and a considerable amount of effort is required to get its contents on board. That work is not complete. Missing words are indicated with ?. In the example there are no words missing the English from Lektorek but sie is missing the analysis from SGJP. (This omission may be corrected by now but the examples still serve a purpose.)

Asterisks *, also plus signs +, have special significance to be explained presently.

The screenshot *czekaja.png* focuses on the word czekają. It has the same analysis as lezie except plural (:) rather than singular (.). The lemma is czekać. The complete translation given by Lektorek is as follows: "wait. +G await. ~ na+A wait for". The screenshot shows *wait in the compact four-line presentation and below we can see that the user has discovered that it expands to first 'wait +await' and then that +await expands to 'wait +G'. In this way the detail of the definition is mostly initially hidden from the user though available when needs be. [1] [2]

There will be a little digression before explaining the user interface.

Lists. With languages it is surely the case that exceptions are the rule. When we ask what is the analysis of a word or the meaning of a word we have to be prepared for in some cases no answer but very frequently a multiplicity of possible answers instead of a single answer. The asterisk * before wait indicates that there is more than one unique translation for czekać. "wait" is the first of a list. The user can click on *wait to see the list. In fact in Lektorek there are three identifiable levels of list. Lektorek presents lists of synonyms separated by commas. If Lektorek determines that there are sufficiently distinct applications of the word it may separate lists of synonyms with "." (full stop or period). If there are distinct meanings these are separated with numerals 1, 2, 3... It takes as many clicks as there are levels to display the word translations.

How to display tree-structured data. As described above the data that needs to be displayed is “tree-strucured” (though the tree is inverted, branching out downwards). For each word in the Polish text that word is the base of a tree. Depending on that is the analysis of it found in SGJP and there are typically multiple analyses – multiple branches. Each analysis comes with one only lemma so there is no branching at this point though the lemma is treated as depending on the analysis. Each lemma then branches to the alternative definitions as founf in Lektorek. The definitions themselves are tree-structured. However a different principle has been adopted for displaying the definition tree, the definition being expanding below the four line display.

---

1 In this case it is perhaps regrettable that 'wait for na+A' does not also appear. It was excluded by policy for the reason that was presented as an idiom.

2 When a word is expanded the expansion usually needs to be displaced so as not to cover up anything important. The displacement is either up-down or left-right. If it were possible to show a line connecting the word and its expanded list it would be clearer. In the example the expansion of *wait was moved down then the expansion of +await was moved sideways.

Editing the data. The display is interactive in either case, there being two aspects to user intreaction - choosing what to display and choosing between alternative translations. The user may edit the data to reflect a final choice between the alternatives that were offered (the asterisk disappears when a list is reduced to one choice). The user can make a more tentative choice without editing data. This typically is a matter of selecting one item with a single click. The selected item is then treated as the default choice for display.

Following the examples, it can be seen that the four-line display shows only one of the alternate analyses, lemmas and definitions at any time. Which one is displayed is controlled by user selection. The definition by contrast is expanded as the user decides. The whole definition tree may be displayed at one time.

Working from either the four line display or the definition display * indicates there is a list hidden behind the entry as displayed. Clicking on the entry expands the list. At that point the list is colour-coded yellow and the user should drag it somewhere sensible with the mouse. Dragging done, the list is displayed white. Items in the list may also be expandable if marked with * or may be expandable for another reason if marked with +. There is an example in +await. In this example, clicking + reveals some additional information: Genitive case follows (+G) when the verb czekac means "await". In the screenshot await +G is yellow as the user has not finally positioned it.

Two words after czekają is the word pora -see *pora.png*. The analysis is marked with an asterisk meaning that sG.mq_ (of lemma por) is merely the first of a list of possibilities. In fact SGPL lists 6 alternatives as follows, and I have aligned [3] the Lektorek translations:

sG.mq_	por	leek (Lektorek por 2, SGJP por:Sm2:Sm3)
sGA.mq^	por	-SGJP treats this word as alternatively animate and inanimate.
sG.mq_	por	pore (Lektorek por 1, SGJP por:Sm3)
sN.f	pora	time
pred		-appears to be an idiomatic use of pora, “it is time”
v=.3(	porać	-not listed in Lektorek

(Lektorek entries may have subsequently changed; the illustrative value remains.)

The translation of this sentence requires some intelligence on the part of user. Studying the full Lektorek entry for pora would normally help since it gives idioms and examples that are likely to be relevant. (The translation aid will not reproduce idioms or examples from Lektorek. For an enhancement, we could assist with a link to the complete entry.) When necessary these days the user can usually get accurate help from Google Translate. In this case Google gives for the surrounding clause "it is not time for us to go into broad arguments”. This fits, interpreting pora to be the nominative feminine alternative and this aligns also with the pred interpretation. Thus: [It is] not [the] time for us into broad go-into arguments. The adjective szeroki is correctly given as as wide or broad. sie is a particle meaning self that often functions as part of the verb. The adjective broad being separated from the noun argument is telling us something about the flexibility of word order in Polish.

The sentence selected here illustrates the motivation behind the translation aid

---

3 We see in this example a case where there are two lemmas with the same form 'por'. They happen to be the same part of speech though it need not be so. As yet there is no alignement between the SGJP and Lektorek dictionaries of distinct lemmas with the same form. As a result we have to offer both Lektorek definitions of por for each occurrence of por turned up from SGJP - even when the parts of speech do not match. So for example na appears twice as a lemma, as preposition with accusative and with locative and the corresponding translations should be 'onto' and 'on'. At present this correspondence has been lost.

- it does the hard work of dictionary look-up for the user. It leaves the intelligent and more interesting work to the user. It complements and does not compete with comprehensive translation services. Thus Google translate can deliver the meaning if the user has difficulties with the meaning but Google itself does not illuminate the structure of the sentence word by word.

There can be multiple 4-line displays. It is best to display a long sentence in pieces with several displays. *Bieda.png* is an example.

User interface reference

4 line display

To initiate a display, merely select judiciously the words of interest with the mouse.

The cell at the left of the display is the “key” (see Pora.png illustration). Click on X to close, Dr to drag, ? for help information.
Display modes, by background colour
  Yellow -display may be dragged vertically with the mouse.
  White -display is in normal operating condition

Symbols. The following symbols are used in the second line.

Part of speech
  s noun s (substantive), to leave n for neuter
  v verb
  vs verbal noun
  a adjective
  @ adverb
  & conjunction
  e preposition pr_e_, since p stands for personal
  , particle
  o pronoun pr_o_
  % participle
  %^ active participle
  %_ passive participle

Gender
  m male
  f female
  n neuter
  p personal
  q non-personal
  ^ animate
  _ inanimate

Case
  N nominaive
  G genitive
  D dative
  A accusative
  I instrumental
  L locative
  V vocative

Number, etc
  . singular
  : plural
  # superlative

List column

Lists are generated for items in the four line display or in the definition display by clicking on the item. The definition display itself consists of a collection of lists.

Entry prefixed with *. To display a list, click on any entry in the four line display that is prefixed with * (in the second or fourth lines) or on . Normally you should move the list to a convenient place on the screen.
To close a list click on the same entry (an unobscured part of it).

Display modes, by background colour
  Yellow -display may be draggged vertically or horizontally with the mouse.
  White -normal mode where user can select items on the list with the mouse.
  Pink   -editing mode. User may modify the list.

  double-click to select an item once and for all. There is no longer a list with alternatives.

4 line display again
When the user selects an item out of the list then lower items in the display will change appropriately. This effect is noticeable with respect to second line items, since each analysis may have a distinct lemma and distinct meanings.

A click causes an item to expanded if it can be expanded or selected if it can be selected.

Entry prefixed with +. This expands to an expanded entry rather than to multiple entries.

Editing of single entries. User can edit a single entry for whatever reason by clicking on it. (An editable entry, pink, appears. The editable entry disappears after the user goes on to something else after making a change. The editable entry can be closed also by clicking on the original word but that is likely to be obscured.)

Maximal use of the Translation Aid

The translation aid has been put together with the idea that a student of Polish though barely a beginner in terms of his or her Polish vocabulary might wish to make his own translation of a significant text. This the reason for offering comprehensive editing facilities, which allow the user to record a choice for every interpretative decision. A choice may be made and recorded for every case where there is a list of alternatives (every *). Having reduced all lists to a single entry then even these choices can and should be edited in place as a single entry. The most normal reason to edit a single entry is that it still contains alternatives as for instance, something like sNAV:n would be reduced to sN:n if the user analyses it as nominative. Another reason is to substitute a translation other than what Lektorek offers or to correct an occasional typographical error in the original text.

When a 4 line display is closed, for now the changes are lost. It is desirable that the changes be retained. After multiple sessions a user may have gone through the entire text. In that case the English words alone could be emitted into a file and that could be edited into translation in correct English.
Bieda2.png is an example of a (more or less) fully processed sentence.