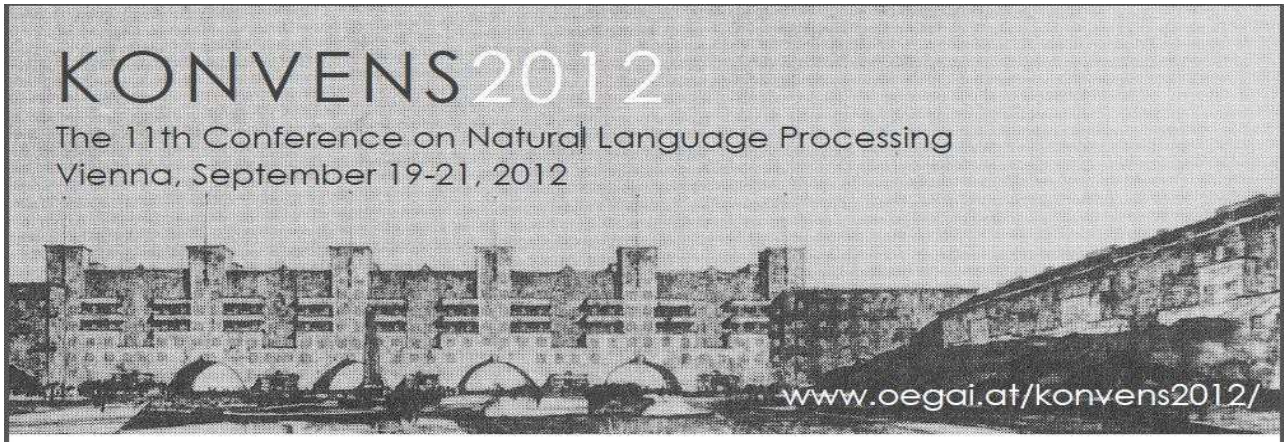
Formalising a dictionary of 17th century English with the NooJ software

Hélène Pignot



Formalising a dictionary of 17th century English with NooJ

First of all I'd like to thank my dear friend Thierry Declerck for his kind invitation. It is a pleasure and an honour to be here in the fair city of Vienna, attending this fascinating conference, gathering together researchers from all over the world, united by a common passion for the humanities and for new technologies, which can be used to benevolent ends, to disseminate and share knowledge, and not as an instrument of domination, as some hard-boiled and stuck-up pessimists try hard to have us believe!

Thierry asked me to present my work with the software NooJ, which was devised by a great French academic whom I admire for his unceasing, ground-breaking and, in a word, extraordinary work both as a linguist and as a computer scientist, Max Silberztein, Professor at the INALCO, the French Institute for the study of Oriental languages. He hates it when I eulogize him, so I take advantage of his absence to do so!

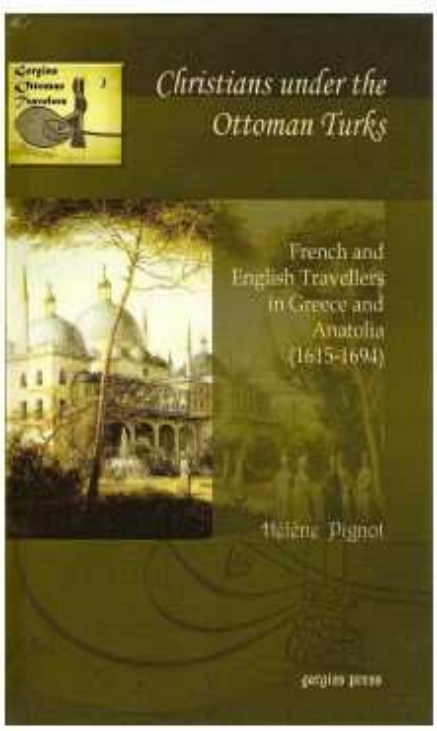
How did I discover NooJ? Life is often made up of boons and happy coincidences.

I graduated from the Sorbonne and the Ecole Normale supérieure, and did my PhD on Sarah Fielding, an English 18th century novelist, who was well-known in her day for being the beloved sister of Henry Fielding and the dear friend of Samuel Richardson, the author of *Clarissa*, admired and held in great esteem by the French philosopher Diderot.

I became a lecturer at the University of Paris I, where my students didn't have English as a major but history and philosophy. Therefore my research interests evolved towards historical and philosophical topics, and in an interdisciplinary domain in particular that is to say travel literature which is at a crossroads between several disciplines such as history, literature, geography, anthropology, history of ideas and even theology!

I have always been passionate about Greece, and reading 17th century literature, I realized that there wasn't any recent book giving access to travel accounts, which are primary sources for the history of Greece in the 17th century. I was also intrigued by the perception of the Greeks' religious otherness by French and English travellers. So I read dozens of travel accounts and gathered what I thought the pearls of this literature, sometimes wrongly considered as minor by traditional literary criticism. Two books were

published, one in French and one in English.



As I collected and translated these texts into French or English for publication, I realized access to this type of literature in the original text might be difficult for non-native speakers. Many of my footnotes attempted to shed light on the historical context of this travel literature, but were also philological notes, notes explaining the lexicon of these beautifully written texts.

Indeed for French students, but even for native speakers (who can read Shakespeare without notes?), 17th century literature may be difficult to read in the original. Many masterpieces of English Literature and philosophy were written in what in France we call le Grand Siècle (the Metaphysical Poets, Milton, Thomas Hobbes, John Locke and what have you), but they are only read by students having English as their major. How could I make this literature more accessible? There is no grammar or lexicon of 17th century English available in France, not to mention of course any free resource that could be downloaded from the Internet, except for lexicons of the King James Bible for instance. So the only thing a French student can do whenever is confronted with archaic words or idioms is to consult the OED, and not the two volume edition, but the twenty volume one. Nevertheless, the more I read 17th century texts downloaded from EEBO, the more I realized that some spelling variants were absent, and even some words did not feature in the OED.

The 20 volume OED can be consulted at the French National Library, and online too, but only if you subscribe to it of course.

I wrongly thought, as quite a few people do, that the first dictionary of the English language is Samuel Johnson's, which he compiled in the 18th century. Discovering the work of early lexicographers like the Puritan Robert Cawdrey, the polyglot Randle Cotgrave, the scholars Elisha Coles and John Ray, I realized that there were dictionaries in the 17th century which were naturally in the public domain, and could be used to create an

electronic dictionary of 17th century English.

Why electronic? An electronic dictionary has many advantages, it is endlessly revisable, it may be updated and completed with new entries all the time, it can be easily downloaded and shared.

What software could I use to create this wonderful resource, a task which is labour intensive and would probably take years to carry out?

I spoke of happy coincidences or boons, or --the English have a very nice word for that-- serendipity.

I was also working in a department called AES (devoted to the study of economics and labour issues) where a colleague of mine had been teaching computer science for quite some time just as me, but I didn't know her, she had simply escaped my notice! This colleague decided to organize a NooJ workshop at Paris I, and another colleague who taught economics put us in touch: she knew I was working on 17th century literature and interested in historical linguistics. This is how I discovered NooJ.

I was a bit scared in the beginning, because computer science was not my field, I was more literature inclined and a bit impressed by all these codes I had to master to be able to use this software efficiently. So I had to overcome my fear, and finally I read the NooJ manual and realized everything was crystal clear, and it was just a question of concentration and hard work!

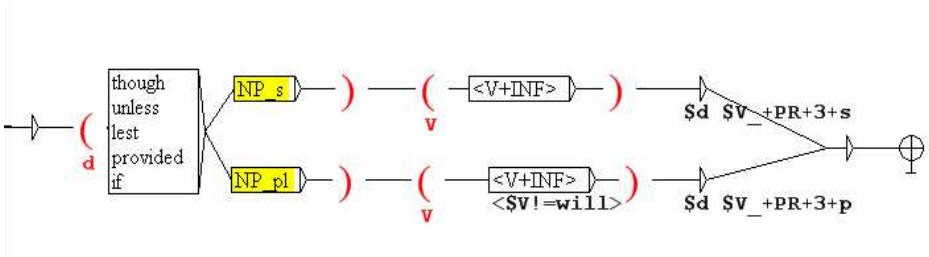
Now that I have explained how I discovered NooJ in 2006, I'd like to show you where my research stands now, what I have so far been able to achieve with this wonderful software.

Mainstream taggers have a hard time with historical corpora: the spelling was very different in the 17th century, and therefore many words cannot be identified and tagged. Our colleagues at the university of Lancaster have created a software called VARD which now exists in a new version called VARD2. This software can detect and display variants and suggest a transcription in modernized spelling. It is a wonderful tool, but it does not deal with semantics, only with the identification of spelling variants.

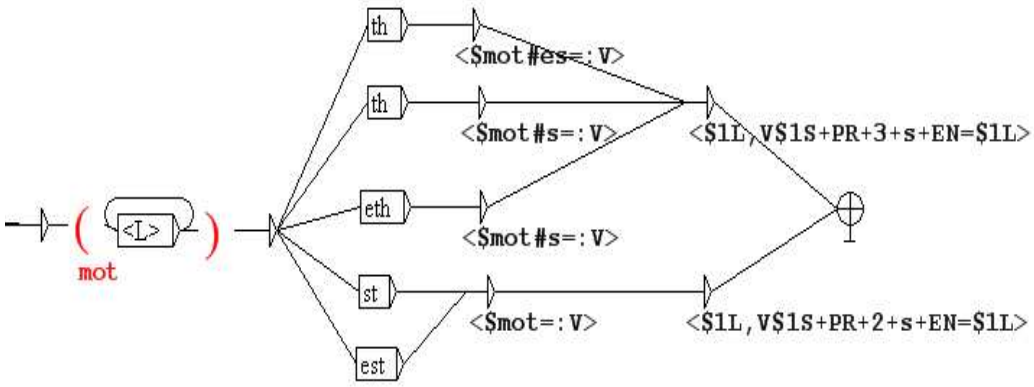
I. Recognition and transcription of forms that are specific to 17th century English.

Part of the work I did with my colleague was to create graphs to enable the machine to recognize the words with spelling variants and suggest a transcription. To treat forms that are specific to 17th century English, the NooJ software uses both morphological and syntactic graphs and dictionaries. I will show you only four of these graphs, because this is all we have time for today.

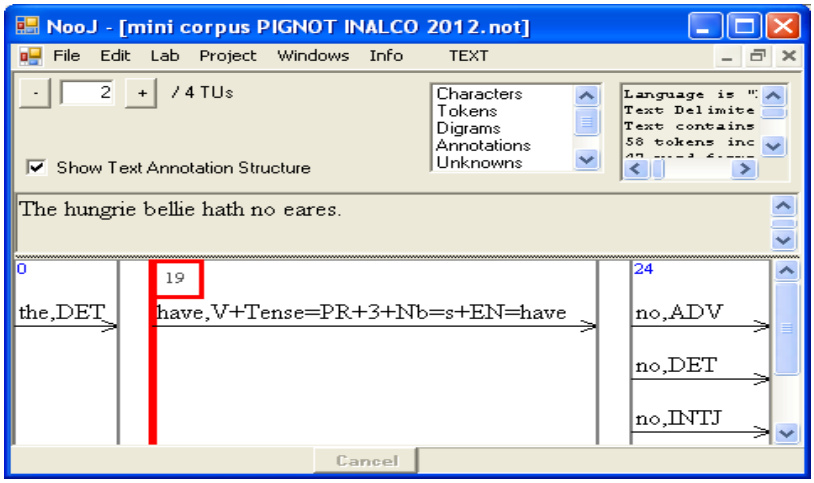
1. This syntactic graph spots and annotates the use of the subjunctive in a sentence:



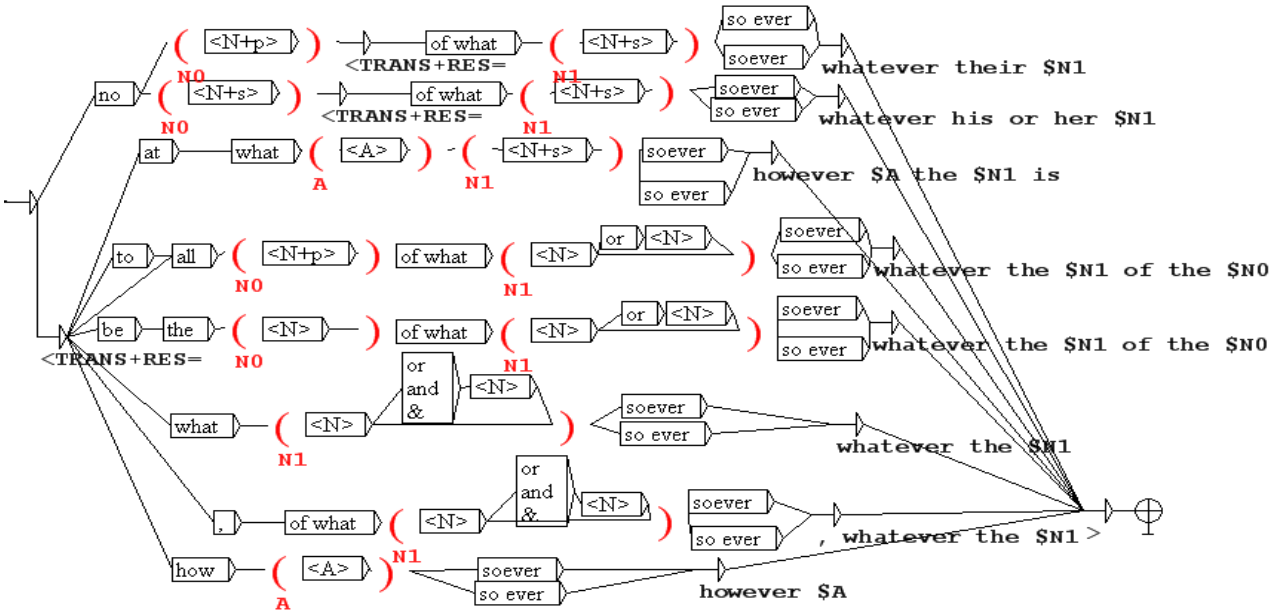
2. This morphological graph recognizes and annotates the archaic forms of the second person and third person present, forms such as “the Holy Spirit *proceedeth* from the Father”, “thou *satisfiest* the desire of every living thing”:



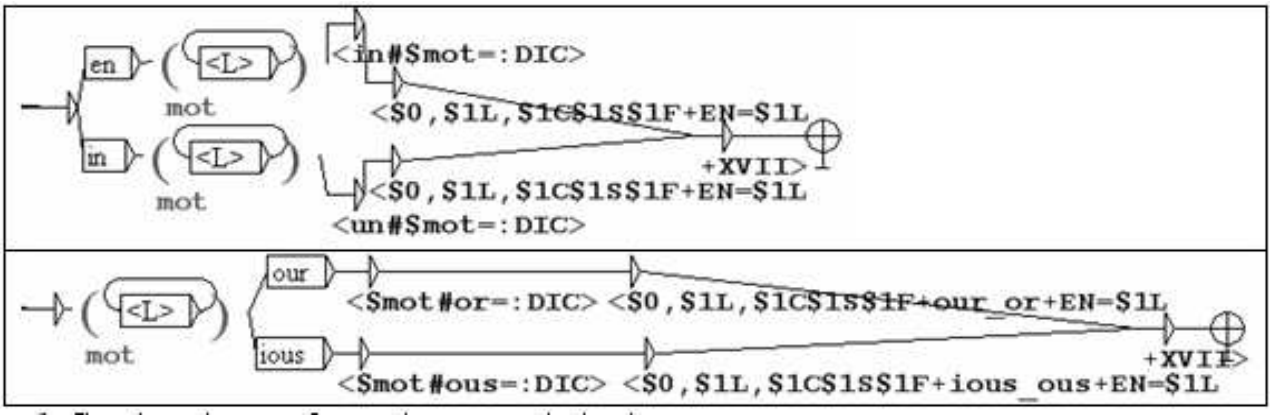
The first path of this graph tells the machine to look for a sequence in which a word is ending with the suffix th, take the variable mot (word), add the letters es thanks to the concatenation operator # and check in the NooJ dictionary if this words exists as a verb. If it does, it then produces the following annotation:this is a verb with syntactic constraints signalled by the letter S (for instance is the verb transitive or not?) whose inflectional code is PR+3+s or the third person singular present. These annotations may be visualized in the TAS (*Text Annotation Structure*) by ticking the box on the left.



3. The syntactic graph below recognizes forms like “any Christian of what communion soever” and transcribes them into contemporary English as “any Christian, whatever his communion”.



4. This morphological graph locates variations in prefixes, such as en for in, and in suffixes such as our for or:



However, we came up against difficulties as many words in English are ambiguous for the machine as they can be both nouns or verbs grammatically. A word like bookes can be both the archaic plural of the noun book, spelled with a mute e (all words could be spelled with a mute e in the 17th century) or it may be the third person of the present of the verb in the singular, so the transcription will be correct (books instead of bookes, that's easy, but the analysis and tagging of the form may be wrong). To fix this, we realized it would take dozens of disambiguation grammars, which is the work of a lifetime, and frankly not a very useful task! What would still be very useful and rewarding both for students and for the researcher was to carry on with the task of recording the singularities of the 17th century lexicon with NooJ, and this is what I elected to do.

II. Formalising the dictionary.

What is very convenient about an electronic dictionary is that you may describe not only the meaning of the word, but also it morphology and the rules that govern it, and the machine may recognize the word and annotate it whatever its grammatical form.

For instance a word like “toung” a spelling variant of “tongue” which cannot be recognized even by our graphs because the word is too distorted will be described in this way in a NooJ dictionary:

toung, N+FLX=APPLE+EN=“tongue”+Dic_EN_17th

divell,N+FLX=APPLE+EN=“devil”+Dic_EN_17th

The code N stands for noun, so the description reads: this is a noun, whose flexion follows the paradigm APPLE (therefore it takes an s in the plural) and whose transcription in English is “tongue”. The NooJ uses what is called nof files; in a nof file the user includes all the flexions for nouns, for verbs, for adjectives, so that the machine can recognize all the forms that it comes across.

Hence the machine knows that better is the superlative of the adjective good, that spake is the archaic preterite of speak, and that the word dictionarie is a spelling variant and the singular of the word dictionary.

When the word is archaic, we suggest an equivalent whenever possible, for example:

camelopardall,N+FLX=APPLE+EN=“giraffe”+Dic_EN_17th

improperation,N+FLX=APPLE+EN=“reproach”+Dic_EN_17th

sithence that,CONJS+EN=“since”+Dic_EN_17th

the weaker vessel,N+Struct=AN+s+BibleQuote+NOTE=“the weaker sex King James Bible 1 Peter 3:7”+Dic_EN_17th

When the word has disappeared from CE, we provide a definition. Hyphenated words that have fallen into obsolescence need to be indexed. For instance here is the layout of the entry “by-design”:

by-design,N+FLX=APPLE+NOTE=“incidental design or purpose”+ Dic_EN_17th

When the notion is complex, and needs an in-depth explanation, we provide a link with Wikipedia (when the article is reliable) thanks to the command +http=

For historical events, we propose this formalisation, using the functionalities +HistEvent and +NOTE= which provides a short definition whenever possible. With NooJ thanks to the command find, and by typing +HistEvent we can extract all the definitions containing an historical event in the dictionary.

Annus Mirabilis,N+A+s+HistEvent+NOTE="1666 or the year of wonders memorable for the Great Fire of London"+Dic_EN_17th

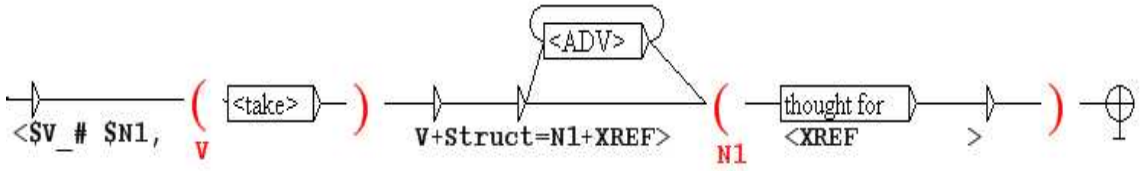
Phrasal verbs are indexed this way, thanks to our colleague Peter Machonis who invented it: here is the example of the phrasal verb to sin out … God's mercy.

sin,V+PV+FXC+Part="out"+N0Hum+N1example="God's mercy"+FLX=BEG+Dic_EN_17th

The most difficult aspect of our work is to formalise frozen expressions, as we need to build grammars for the machine to be able to recognize them.

Here is a grammar that recognizes the verb construction "to take thought for" something, which means to worry about it:

The first part of the graph enables the machine to recognize all the conjugated forms of the verb "take" followed by any chain of adverbs or not and by the noun phrase "thought for".



We would like to make this dictionary as simple as possible, by providing synonyms for each meaning; when the word is polysemous, it is up to the user to determine what is its correct meaning in context. When we look up words in the matchless dictionary that is the OED, looking for the meaning of a word, we often need to sift through several layers of definitions before finding the ones that are really relevant to our century, this is what makes it difficult to use sometimes.

In the near and distant future, we are hoping to digitize 17th century dictionaries presenting entries in the NooJ format and create a website, where other colleagues will be able to contribute by posting words and new entries.

We would also like to make this resource multilingual, as the equivalent in 17th century German and French for instance, could be indicated by using the +DE= or +FR= functionality. Again a lot of hard work lies ahead, and this is really exciting!

Thank you so much for your attention!

Hélène Pignot