Unitex : a free software for NLP

Sébastien Paumier

Penguin from http://tux.crystalxp.net/

A brief overview…

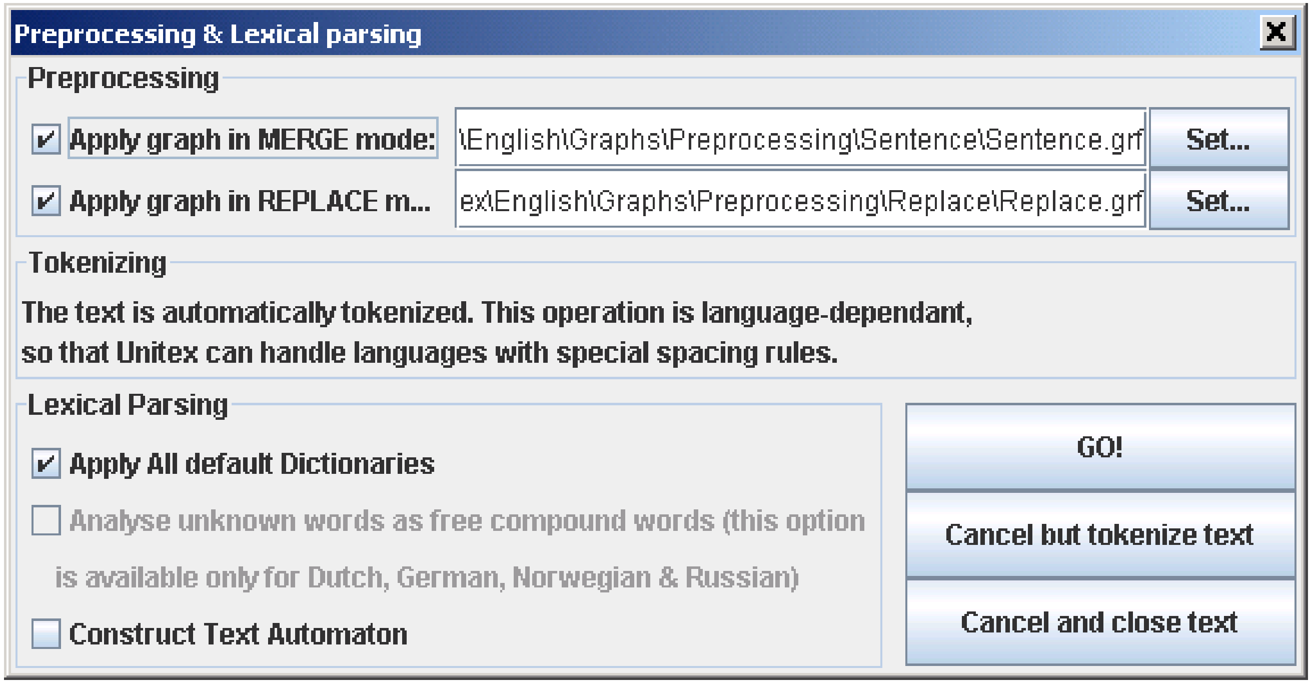
Unitex

- corpus processor based on linguistic resources (dictionaries, grammars)
- open source (LGPL + LGPLLR)
- portable (Windows, Linux, MacOS, ...)
- Unicode Little-Endian 16 bits
- programs in C/C++, GUI in Java
- handling many languages

http://www-igm.univ-mlv.fr/~unitex/

Opening a text

- the text must be a raw UTF16LE one
- it must be preprocessed:



Preprocessing & Lexical parsing

Preprocessing

☑ Apply graph in MERGE mode: \English\Graphs\Preprocessing\Sentence\Sentence.grf

☑ Apply graph in REPLACE m... ex\English\Graphs\Preprocessing\Replace\Replace.grf

Tokenizing

The text is automatically tokenized. This operation is language-dependant,
so that Unitex can handle languages with special spacing rules.

Lexical Parsing

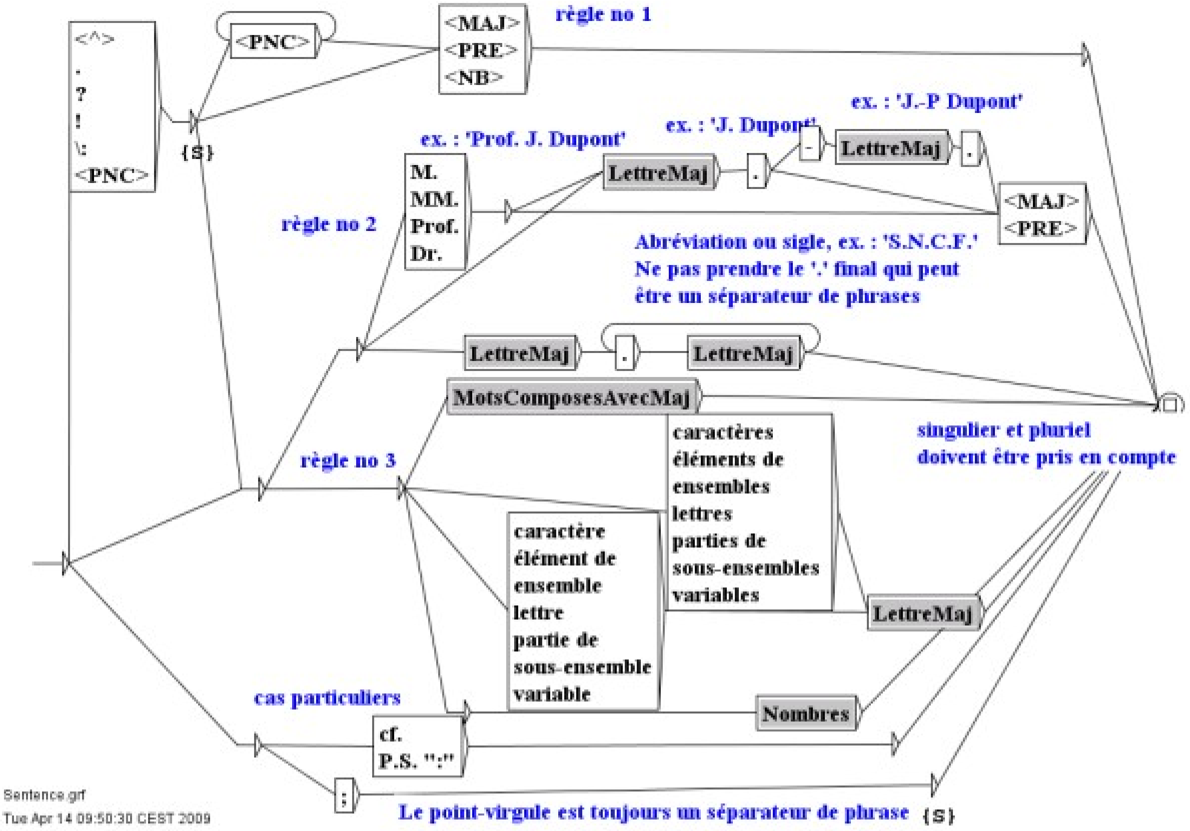
☑ Apply All default Dictionaries

□ Analyse unknown words as free compound words (this option is available only for Dutch, German, Norwegian & Russian)

□ Construct Text Automaton

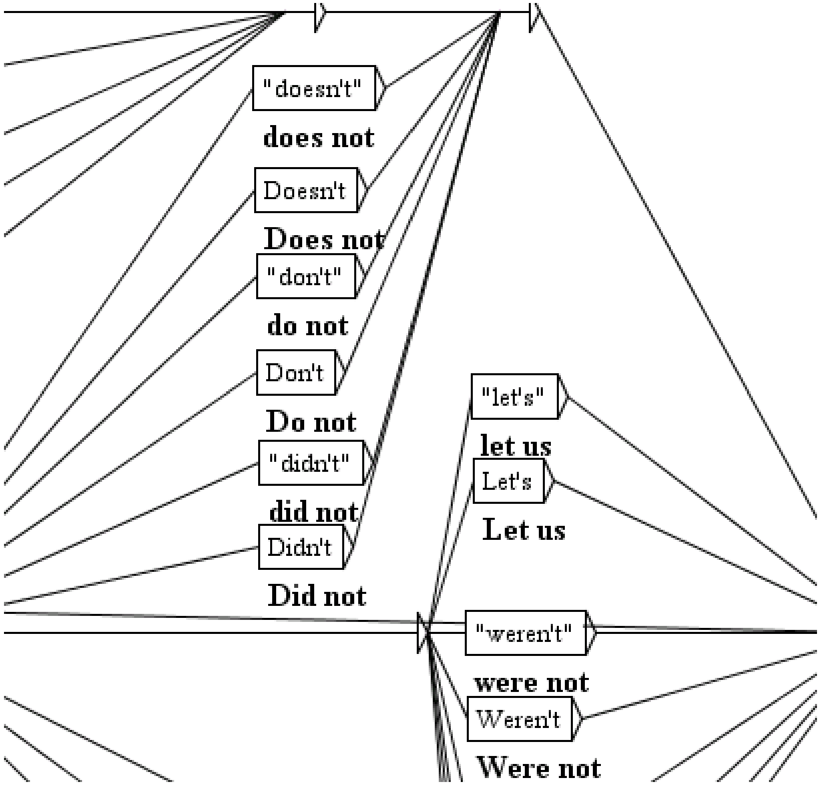
Sentence splitting

- {S} symbol=sentence delimiter
- inserted in the text according to a grammar:



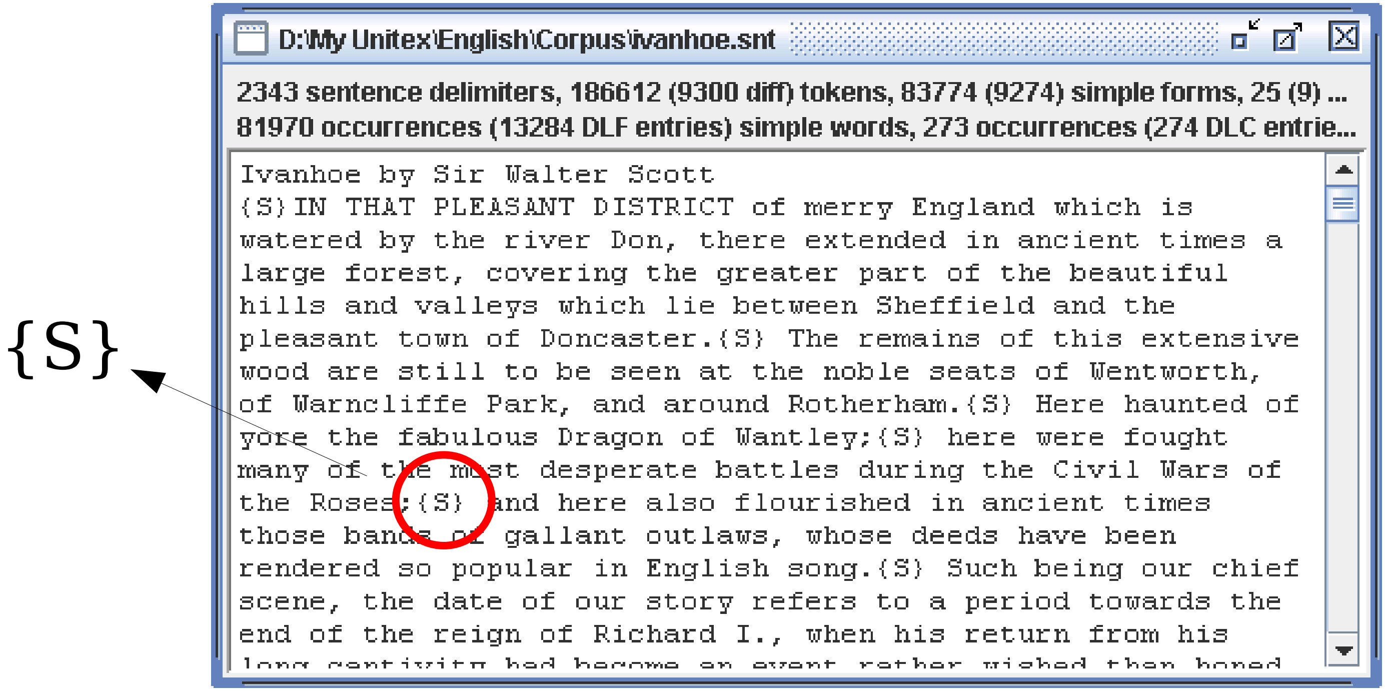
Normalization

- some unambiguous sequences can be normalized:



Preprocessed text

- the text, after preprocessing:



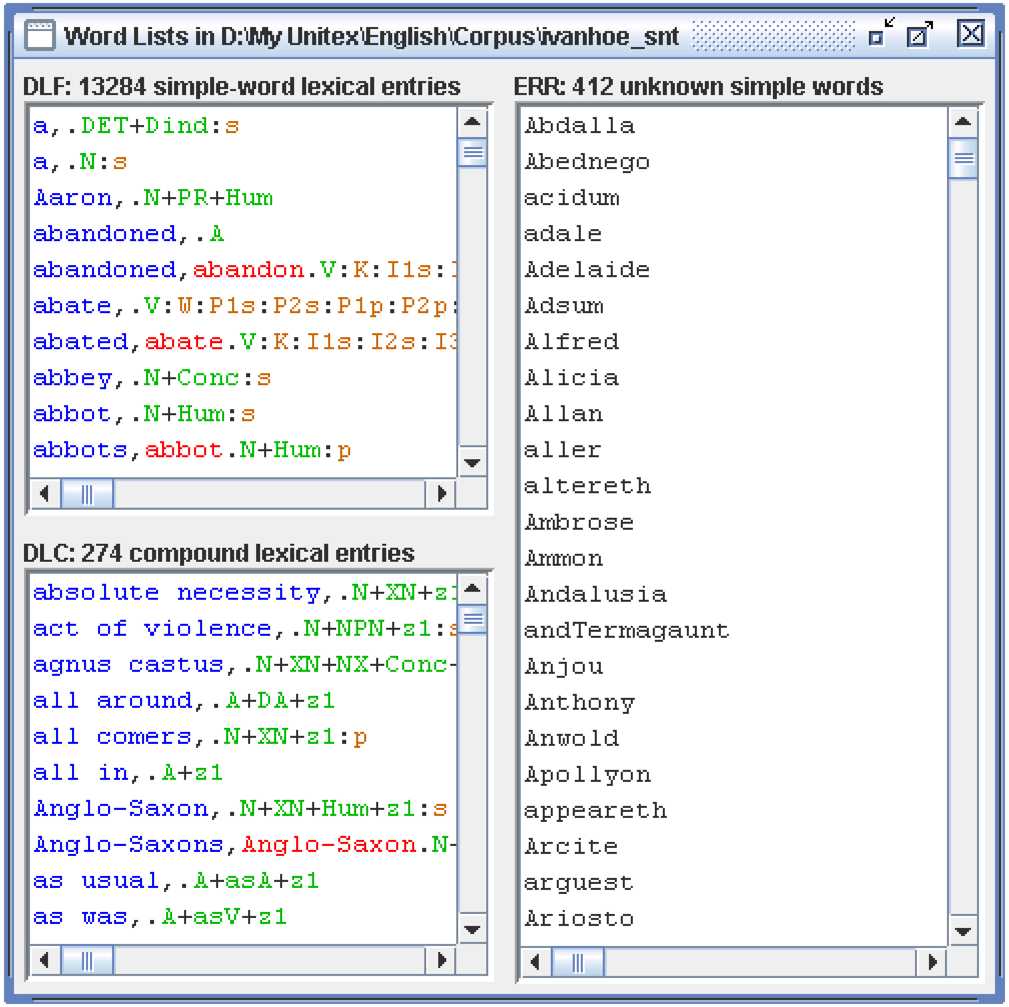
D:\My Unitex\English\Corpus\ivanhoe.snt

2343 sentence delimiters, 186612 (9300 diff) tokens, 83774 (9274) simple forms, 25 (9) ...
81970 occurrences (13284 DLF entries) simple words, 273 occurrences (274 DLC entrie...

```
Ivanhoe by Sir Walter Scott
{S}IN THAT PLEASANT DISTRICT of merry England which is
watered by the river Don, there extended in ancient times a
large forest, covering the greater part of the beautiful
hills and valleys which lie between Sheffield and the
pleasant town of Doncaster.{S} The remains of this extensive
wood are still to be seen at the noble seats of Wentworth,
of Warncliffe Park, and around Rotherham.{S} Here haunted of
yore the fabulous Dragon of Wantley;{S} here were fought
many of the most desperate battles during the Civil Wars of
the Roses;{S} and here also flourished in ancient times
those bands of gallant outlaws, whose deeds have been
rendered so popular in English song.{S} Such being our chief
scene, the date of our story refers to a period towards the
end of the reign of Richard I., when his return from his
long captivity had become an event rather wished than hoped
```

Applying dictionaries

- applying dictionaries means constructing subdictionaries containing all words of the text
- dlf=simple words
- dlc=compound words
- err=unknown words



DLF: 13284 simple-word lexical entries

```
a,.DET+Dind:s
a,.N:s
Aaron,.N+PR+Hum
abandoned,.A
abandoned,abandon.V:K:I1s:I
abate,.V:W:P1s:P2s:P1p:P2p:
abated,abate.V:K:I1s:I2s:I3
abbey,.N+Conc:s
abbot,.N+Hum:s
abbots,abbot.N+Hum:p
```

DLC: 274 compound lexical entries

```
absolute necessity,.N+XN+z1
act of violence,.N+NPN+z1:s
agnus castus,.N+XN+NX+Conc-
all around,.A+DA+z1
all comers,.N+XN+z1:p
all in,.A+z1
Anglo-Saxon,.N+XN+Hum+z1:s
Anglo-Saxons,Anglo-Saxon.N-
as usual,.A+asA+z1
as was,.A+asV+z1
```

ERR: 412 unknown simple words

```
Abdalla
Abednego
acidum
adale
Adelaide
Adsum
Alfred
Alicia
Allan
aller
altereth
Ambrose
Ammon
Andalusia
andTermagaunt
Anjou
Anthony
Anwold
Apollyon
appeareth
Arcite
arguest
Ariosto
```

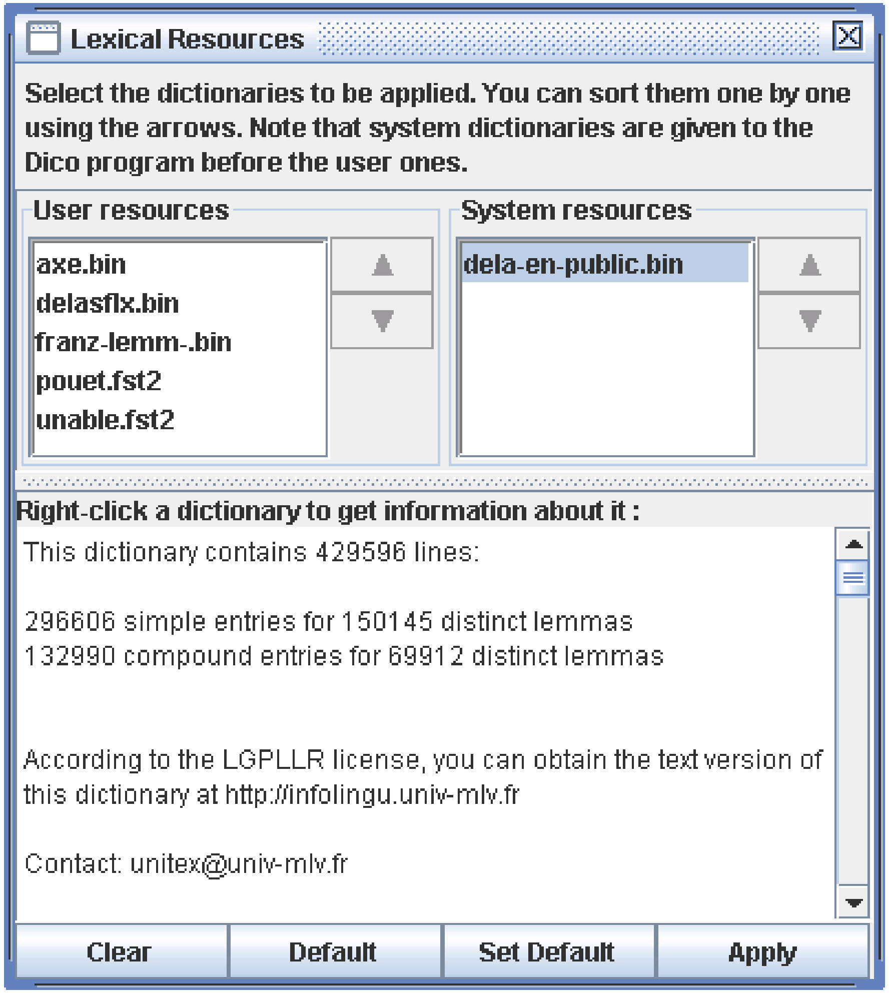
Selecting dictionaries

- one can select the dictionaries to be applied
- system ones are applied prior to user ones
- priority rules:

  foo-.bin prior to

  foo.bin prior to

  foo+.bin



Select the dictionaries to be applied. You can sort them one by one using the arrows. Note that system dictionaries are given to the Dico program before the user ones.

User resources

axe.bin
delasflx.bin
franz-lemm-.bin
pouet.fst2
unable.fst2

System resources

dela-en-public.bin

Right-click a dictionary to get information about it :

This dictionary contains 429596 lines:

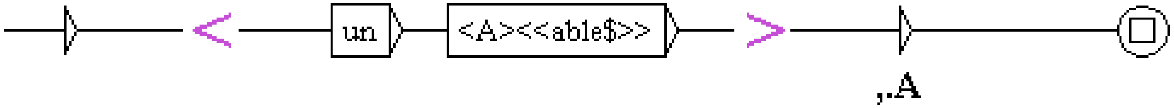
296606 simple entries for 150145 distinct lemmas
132990 compound entries for 69912 distinct lemmas

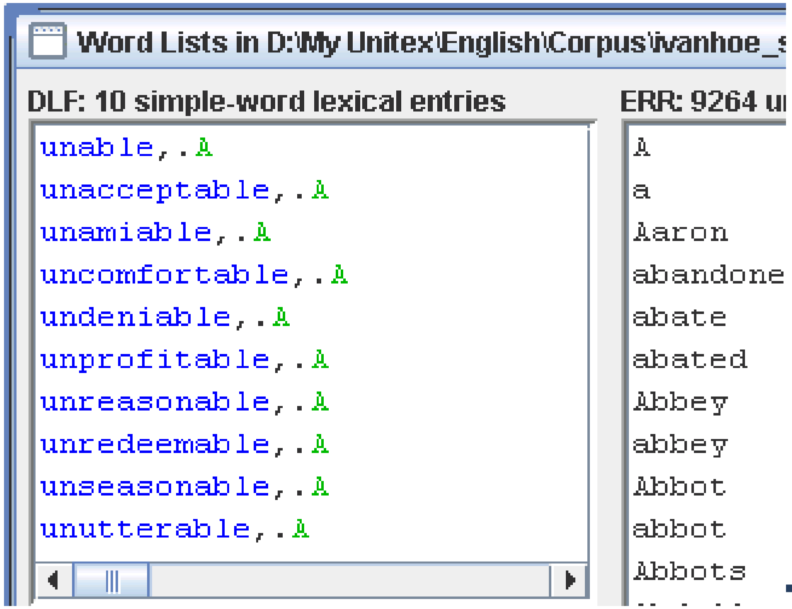
According to the LGPLLR license, you can obtain the text version of this dictionary at http://infolingu.univ-mlv.fr

Contact: unitex@univ-mlv.fr

Dictionary graphs

- one can design dictionary graphs
- such a graph must produce an output that is a valid dictionary line:




Word Lists in D:\My Unitex\English\Corpus\ivanhoe_s

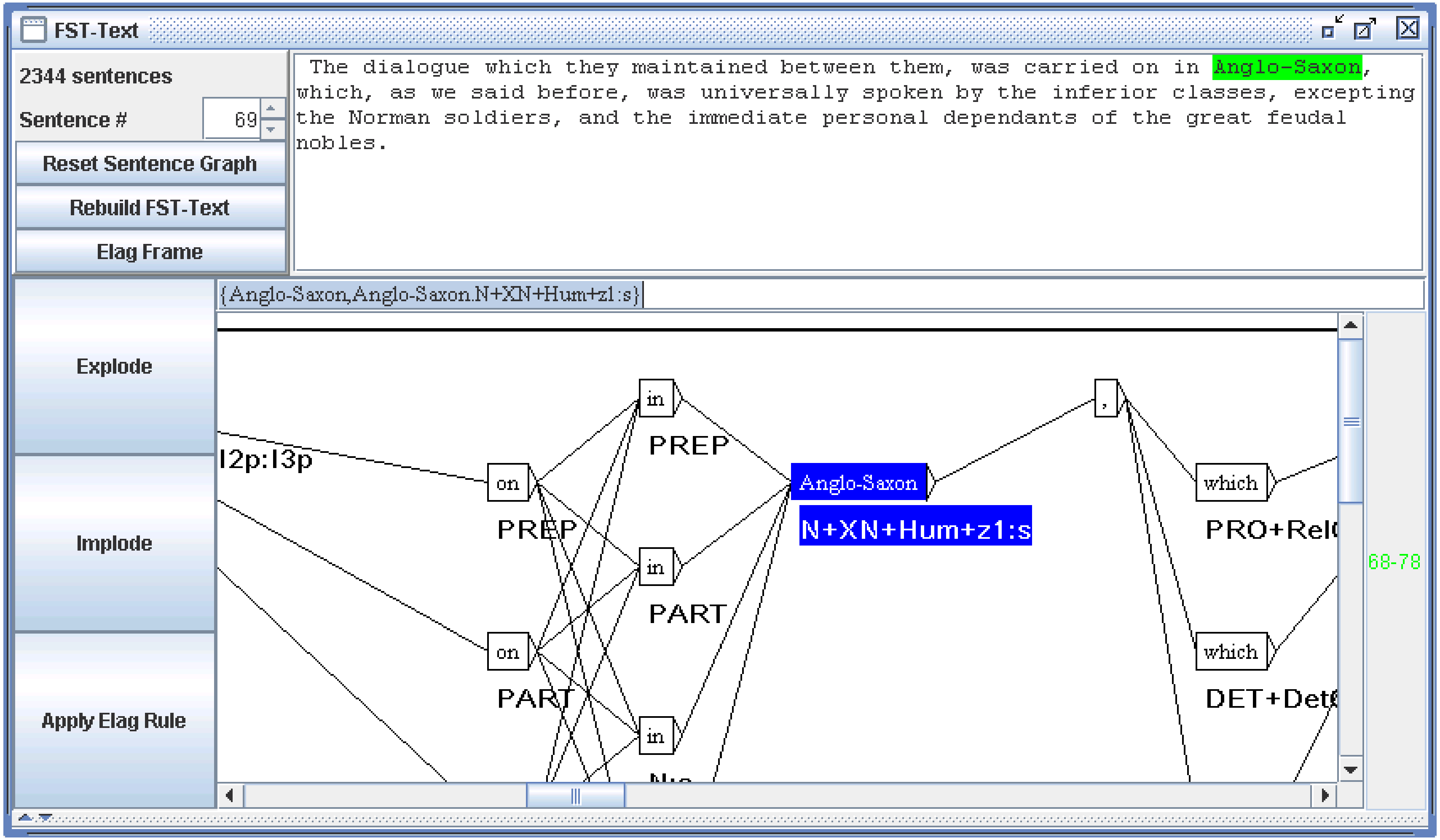
DLF: 10 simple-word lexical entries

```
unable,.A
unacceptable,.A
unamiable,.A
uncomfortable,.A
undeniable,.A
unprofitable,.A
unreasonable,.A
unredeemable,.A
unseasonable,.A
unutterable,.A
```

ERR: 9264 u

```
A
a
Aaron
abandone
abate
abated
Abbey
abbey
Abbot
abbot
Abbots
```

Text automaton



```
The dialogue which they maintained between them, was carried on in Anglo-Saxon,
which, as we said before, was universally spoken by the inferior classes, excepting
the Norman soldiers, and the immediate personal dependants of the great feudal
nobles.
```

{Anglo-Saxon,Anglo-Saxon.N+XN+Hum+z1:s}

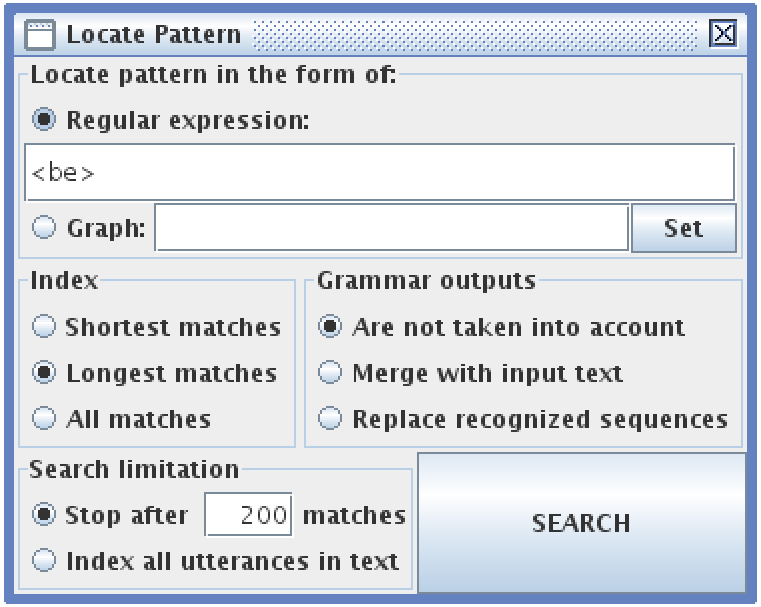
DELAF dictionaries

- text files containing one line per entry:

```
could,can.V+aux:I1s:I2s:I3s:I1p:I2p:I3p/foo
```

could=inflected form
can=lemma
V=grammatical code
aux=optional grammatical/semantic code
I1s ... I3p=inflectional codes
foo=optional comment

Locate patterns



Locate Pattern

Locate pattern in the form of:

● Regular expression:

<be>

○ Graph:

Index

○ Shortest matches

● Longest matches

○ All matches

Grammar outputs

● Are not taken into account

○ Merge with input text

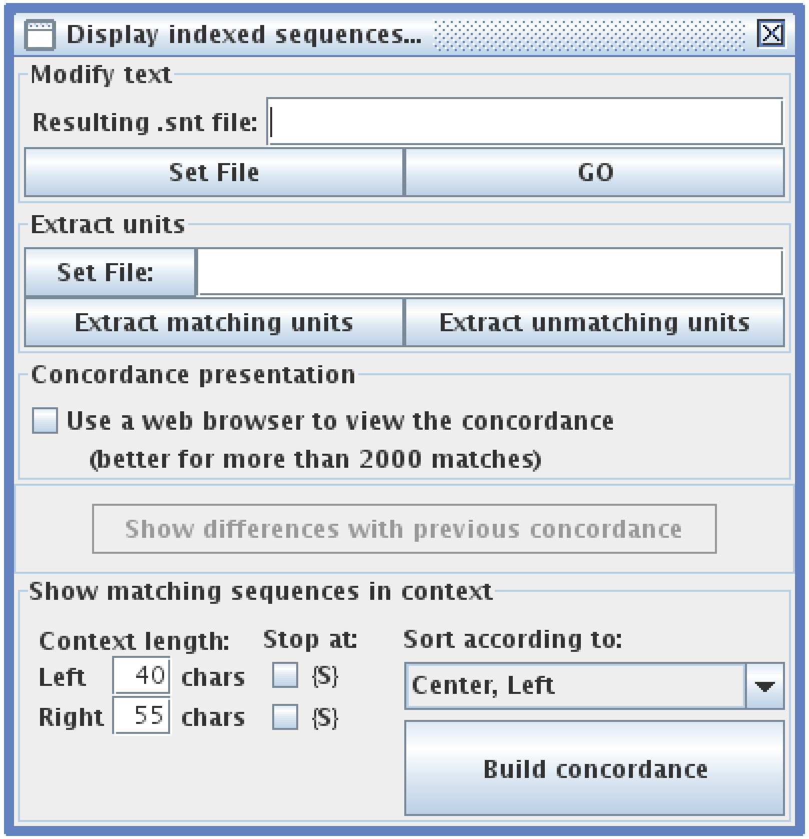
○ Replace recognized sequences

Search limitation

● Stop after 200 matches

○ Index all utterances in text

Displaying occurrences



Display indexed sequences...

Modify text

Resulting .snt file:

Extract units

Concordance presentation

□ Use a web browser to view the concordance
(better for more than 2000 matches)

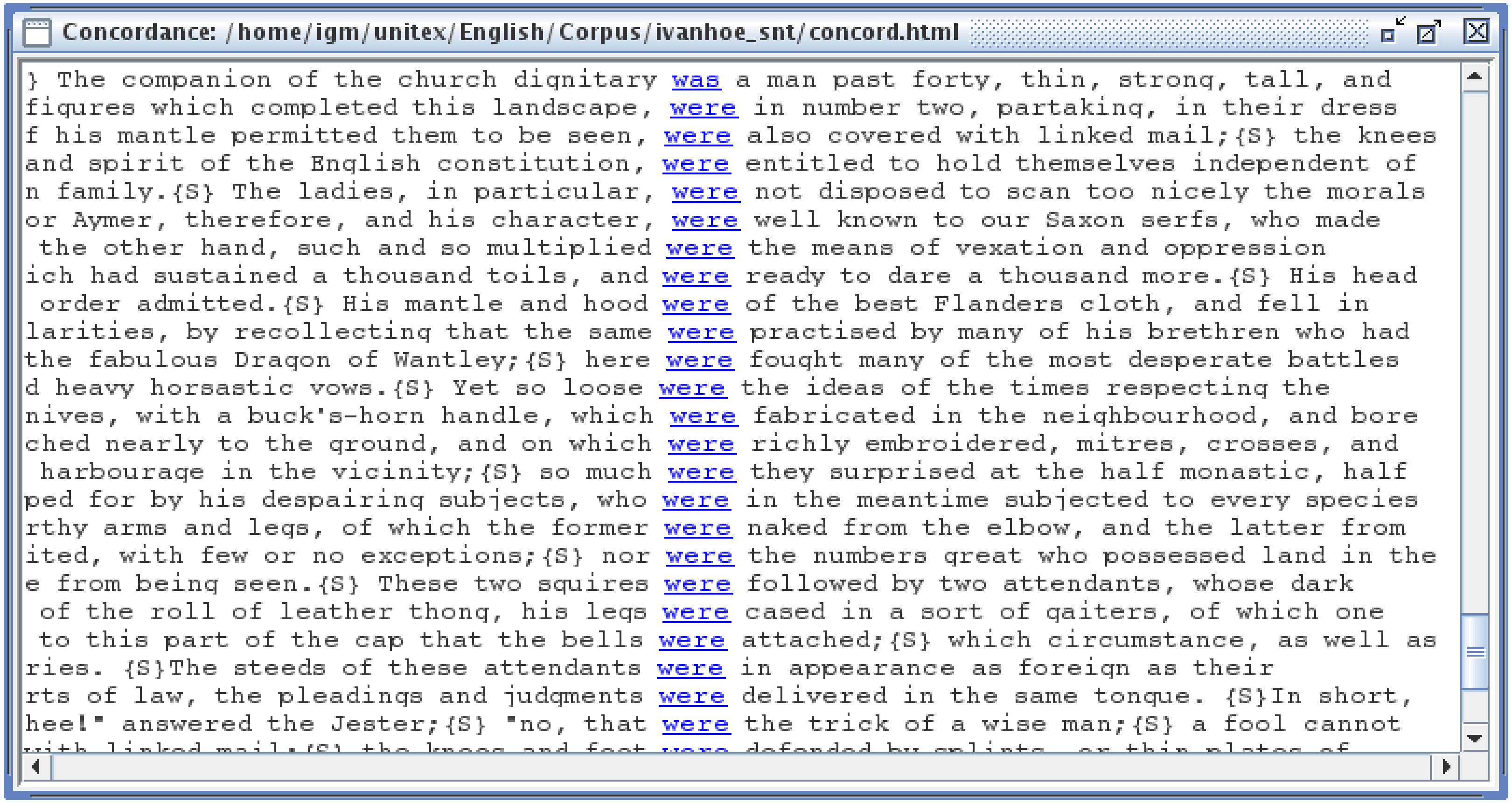
Show matching sequences in context

Context length:
Left 40 chars
Right 55 chars

Stop at:
□ {S}
□ {S}

Sort according to: Center, Left

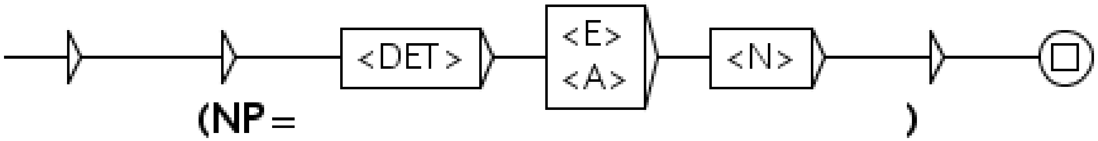
Concordance

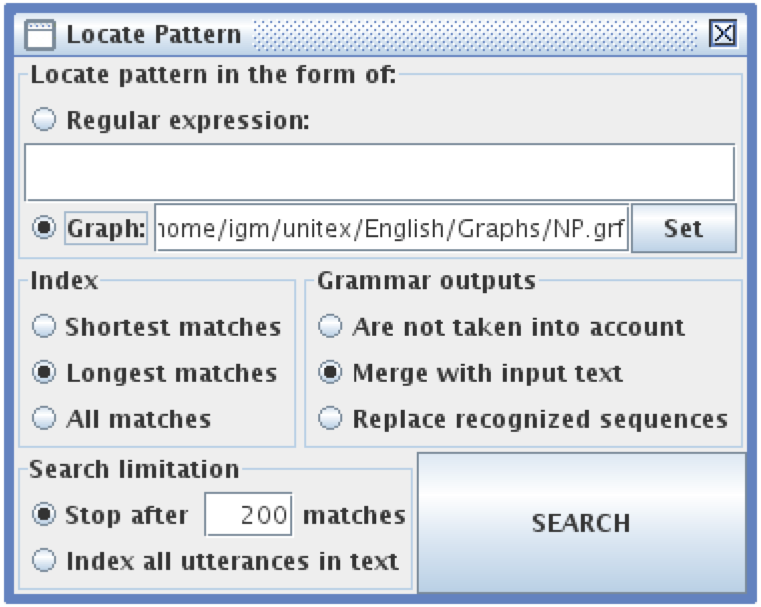


Concordance: /home/igm/unitex/English/Corpus/ivanhoe_snt/concord.html

```
} The companion of the church dignitary was a man past forty, thin, strong, tall, and
figures which completed this landscape, were in number two, partaking, in their dress
f his mantle permitted them to be seen, were also covered with linked mail;{S} the knees
and spirit of the English constitution, were entitled to hold themselves independent of
n family.{S} The ladies, in particular, were not disposed to scan too nicely the morals
or Aymer, therefore, and his character, were well known to our Saxon serfs, who made
 the other hand, such and so multiplied were the means of vexation and oppression
ich had sustained a thousand toils, and were ready to dare a thousand more.{S} His head
 order admitted.{S} His mantle and hood were of the best Flanders cloth, and fell in
larities, by recollecting that the same were practised by many of his brethren who had
the fabulous Dragon of Wantley;{S} here were fought many of the most desperate battles
d heavy horsastic vows.{S} Yet so loose were the ideas of the times respecting the
nives, with a buck's-horn handle, which were fabricated in the neighbourhood, and bore
ched nearly to the ground, and on which were richly embroidered, mitres, crosses, and
 harbourage in the vicinity;{S} so much were they surprised at the half monastic, half
ped for by his despairing subjects, who were in the meantime subjected to every species
rthy arms and legs, of which the former were naked from the elbow, and the latter from
ited, with few or no exceptions;{S} nor were the numbers great who possessed land in the
e from being seen.{S} These two squires were followed by two attendants, whose dark
 of the roll of leather thong, his legs were cased in a sort of gaiters, of which one
 to this part of the cap that the bells were attached;{S} which circumstance, as well as
ries. {S}The steeds of these attendants were in appearance as foreign as their
rts of law, the pleadings and judgments were delivered in the same tongue. {S}In short,
heel!" answered the Jester;{S} "no, that were the trick of a wise man;{S} a fool cannot
with linked mail;{S} the knees and feet were defended by splints, or thin plates of
```

A graph with outputs





Locate Pattern

Locate pattern in the form of:

○ Regular expression:

● Graph: home/igm/unitex/English/Graphs/NP.grf

Index

○ Shortest matches

● Longest matches

○ All matches

Grammar outputs

○ Are not taken into account

● Merge with input text

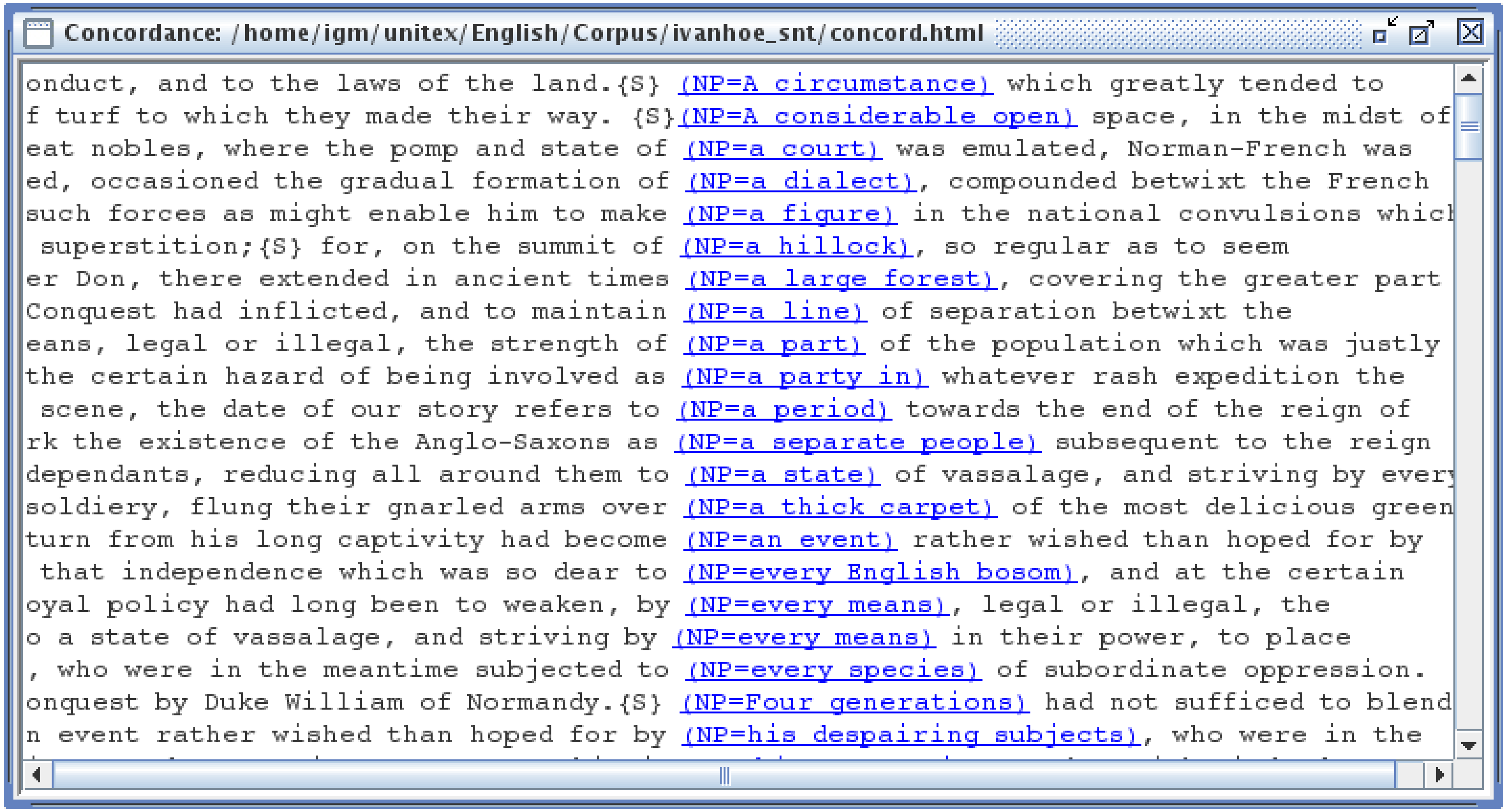
○ Replace recognized sequences

Search limitation

● Stop after 200 matches

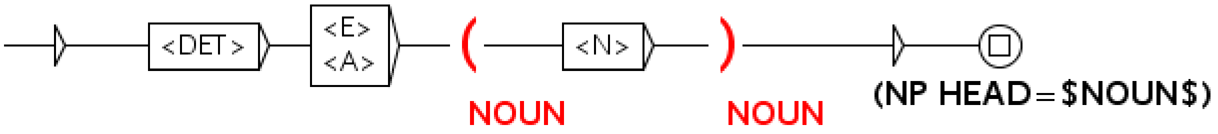
○ Index all utterances in text

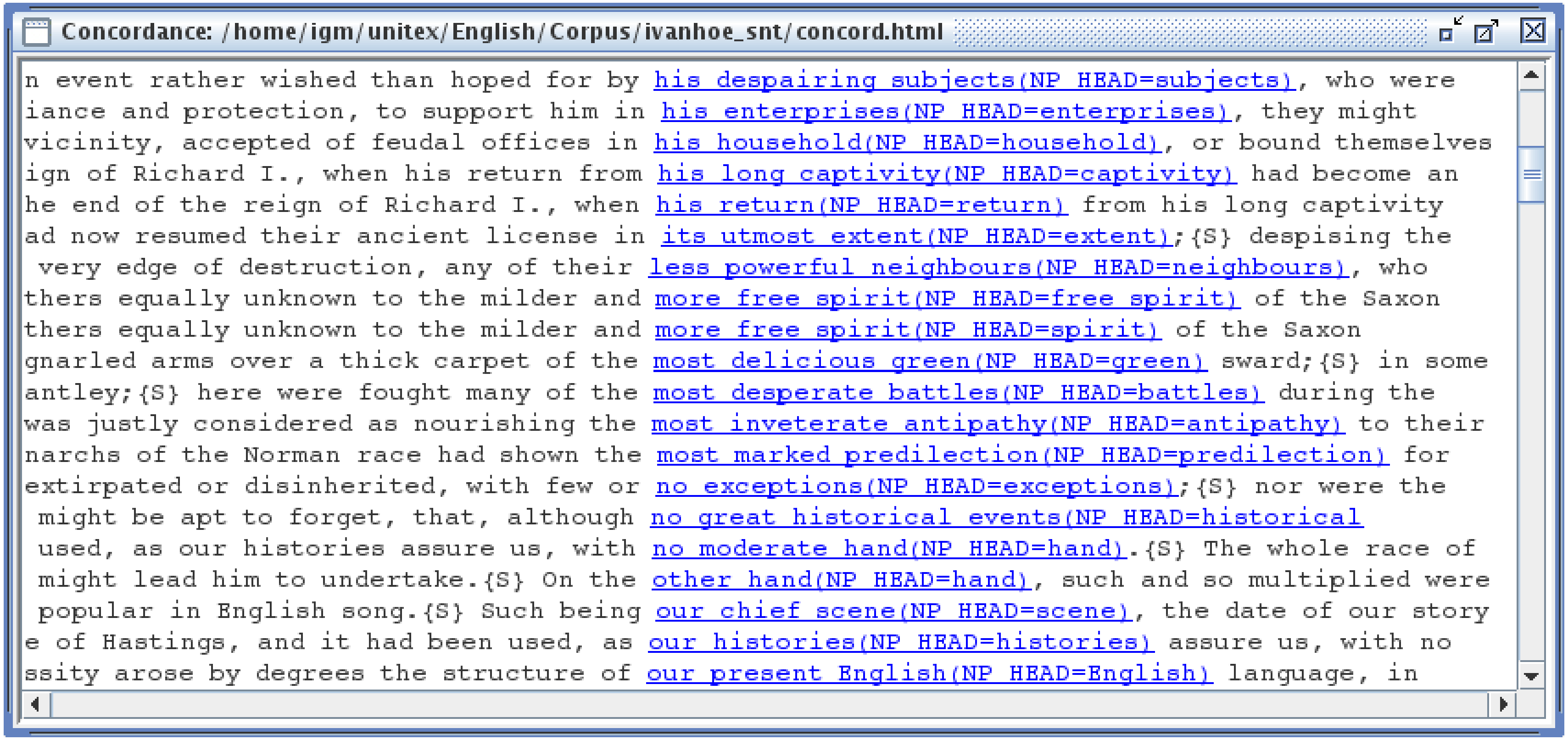
A graph with outputs


Concordance: /home/igm/unitex/English/Corpus/ivanhoe_snt/concord.html

```
onduct, and to the laws of the land.{S} (NP=A circumstance) which greatly tended to
f turf to which they made their way. {S}(NP=A considerable open) space, in the midst of
eat nobles, where the pomp and state of (NP=a court) was emulated, Norman-French was
ed, occasioned the gradual formation of (NP=a dialect), compounded betwixt the French
such forces as might enable him to make (NP=a figure) in the national convulsions which
 superstition;{S} for, on the summit of (NP=a hillock), so regular as to seem
er Don, there extended in ancient times (NP=a large forest), covering the greater part
Conquest had inflicted, and to maintain (NP=a line) of separation betwixt the
eans, legal or illegal, the strength of (NP=a part) of the population which was justly
the certain hazard of being involved as (NP=a party in) whatever rash expedition the
 scene, the date of our story refers to (NP=a period) towards the end of the reign of
rk the existence of the Anglo-Saxons as (NP=a separate people) subsequent to the reign
dependants, reducing all around them to (NP=a state) of vassalage, and striving by every
soldiery, flung their gnarled arms over (NP=a thick carpet) of the most delicious green
turn from his long captivity had become (NP=an event) rather wished than hoped for by
 that independence which was so dear to (NP=every English bosom), and at the certain
oyal policy had long been to weaken, by (NP=every means), legal or illegal, the
o a state of vassalage, and striving by (NP=every means) in their power, to place
, who were in the meantime subjected to (NP=every species) of subordinate oppression.
onquest by Duke William of Normandy.{S} (NP=Four generations) had not sufficed to blend
n event rather wished than hoped for by (NP=his despairing subjects), who were in the
```

A graph with variables




Concordance: /home/igm/unitex/English/Corpus/ivanhoe_snt/concord.html

```
n event rather wished than hoped for by his despairing subjects(NP HEAD=subjects), who were
iance and protection, to support him in his enterprises(NP HEAD=enterprises), they might
vicinity, accepted of feudal offices in his household(NP HEAD=household), or bound themselves
ign of Richard I., when his return from his long captivity(NP HEAD=captivity) had become an
he end of the reign of Richard I., when his return(NP HEAD=return) from his long captivity
ad now resumed their ancient license in its utmost extent(NP HEAD=extent);{S} despising the
 very edge of destruction, any of their less powerful neighbours(NP HEAD=neighbours), who
thers equally unknown to the milder and more free spirit(NP HEAD=free spirit) of the Saxon
thers equally unknown to the milder and more free spirit(NP HEAD=spirit) of the Saxon
gnarled arms over a thick carpet of the most delicious green(NP HEAD=green) sward;{S} in some
antley;{S} here were fought many of the most desperate battles(NP HEAD=battles) during the
was justly considered as nourishing the most inveterate antipathy(NP HEAD=antipathy) to their
narchs of the Norman race had shown the most marked predilection(NP HEAD=predilection) for
extirpated or disinherited, with few or no exceptions(NP HEAD=exceptions);{S} nor were the
 might be apt to forget, that, although no great historical events(NP HEAD=historical
 used, as our histories assure us, with no moderate hand(NP HEAD=hand).{S} The whole race of
 might lead him to undertake.{S} On the other hand(NP HEAD=hand), such and so multiplied were
 popular in English song.{S} Such being our chief scene(NP HEAD=scene), the date of our story
e of Hastings, and it had been used, as our histories(NP HEAD=histories) assure us, with no
ssity arose by degrees the structure of our present English(NP HEAD=English) language, in
```

Lexical masks

- inch: matches 'inch', no matter the case
- <inch>: matches any word whose lemma is 'inch'
- <inch.V>: matches any verbal form whose lemma is 'inch'
- <V:P1s:P3s>: matches any verb at present singular, 1st or 3rd form
- see the user manual for the code tables

Meta masks

- <MOT>: any sequence of letters
- <MIN>: any sequence of lower case letters
- <MAJ>: any sequence of upper case letters
- <PRE>: any sequence of letters starting with an upper case one

Meta masks

- <DIC>: any word in the text dictionaries
- <SDIC>: any simple word in the text dictionaries
- <CDIC>: any compound word in the text dictionaries
- <NB>: any contiguous sequence of digits ('1234' but not '1 234')

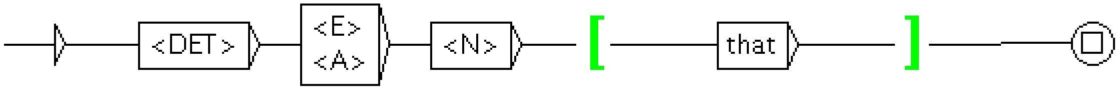
Morphological filters

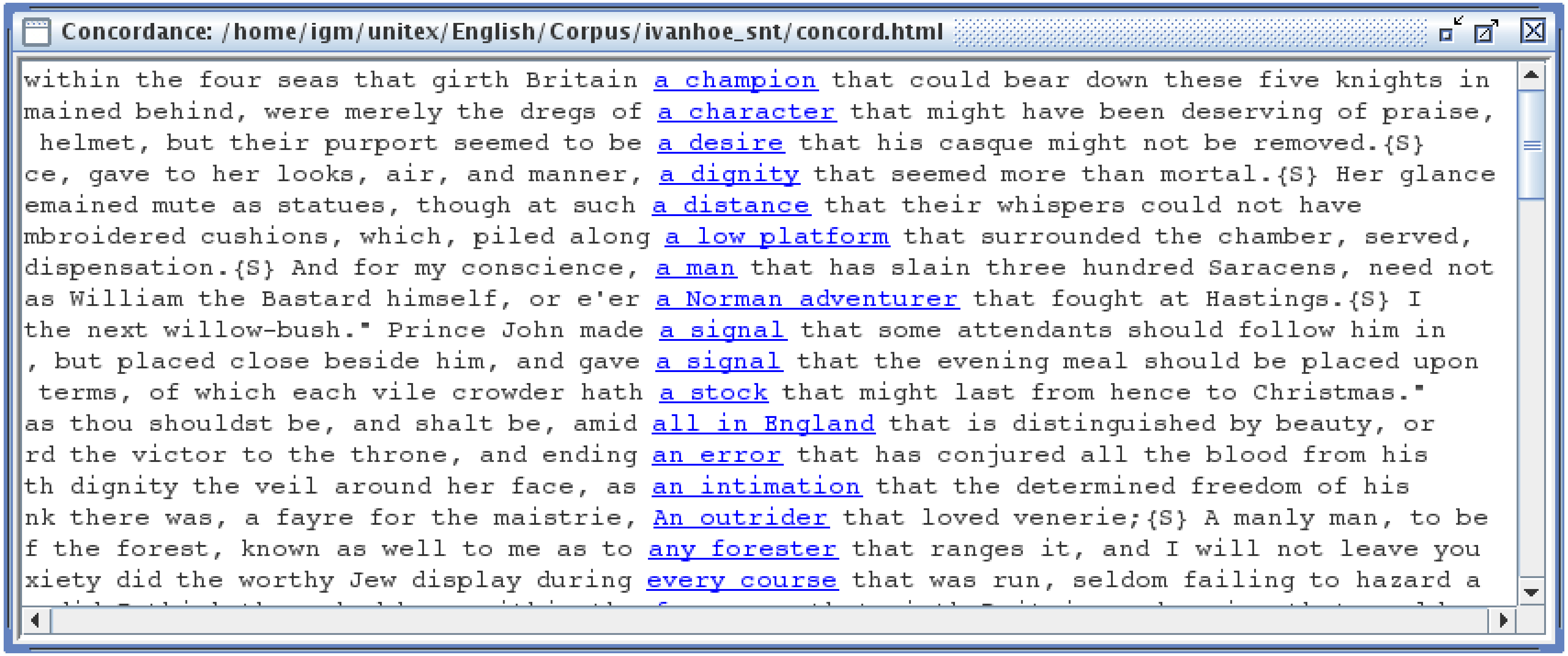
- <<ss>>: contains 'ss'
- <<^a>>: starts with 'a'
- <<s$>>: ends with 's'
- <<a.ss>>: contains 'a' followed by any character, followed by 'ss'
- <<a.*ss>>: contains 'a' followed by any sequence of characters, followed by 'ss'
- <<k|w>>: contains 'k' or 'w'
- <<es?>>: contains 'e' followed by an optional 's'

Combining filters and masks

- <V:K><<wn$>>: past participle ending with 'wn'
- <CDIC><<->>: compound word containing a dash
- <A:s><<^pro>>: feminine adjective starting with 'pro'
- <!DIC><<es$>>: a word that is not in the text dictionaries ending with 'es'

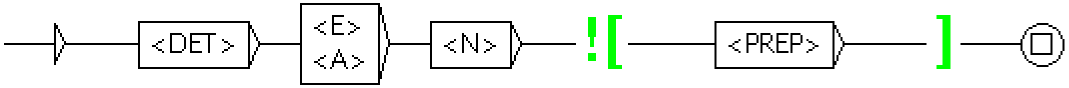
Positive contexts

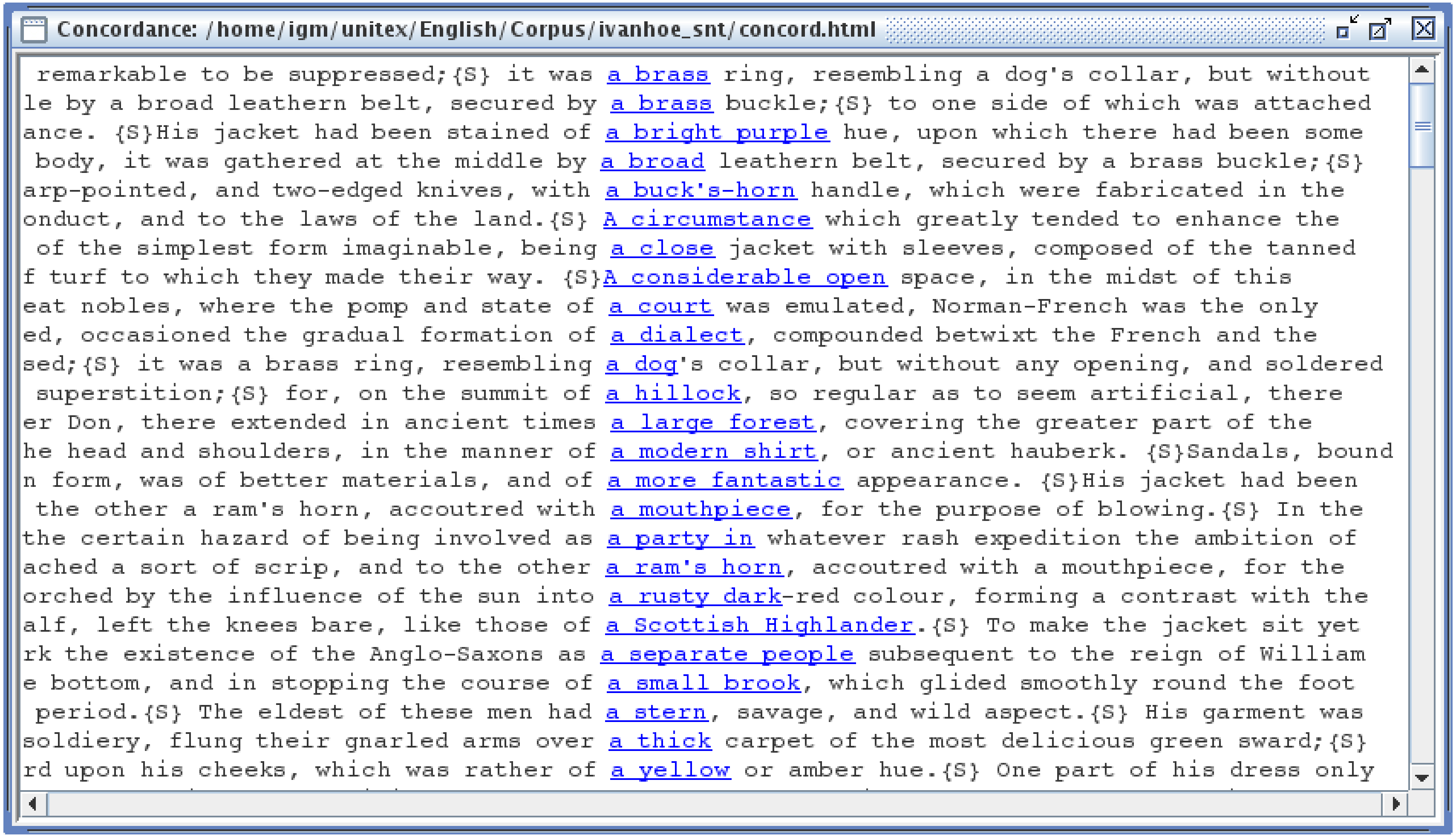



Concordance: /home/igm/unitex/English/Corpus/ivanhoe_snt/concord.html

```
within the four seas that girth Britain a champion that could bear down these five knights in
mained behind, were merely the dregs of a character that might have been deserving of praise,
 helmet, but their purport seemed to be a desire that his casque might not be removed.{S}
ce, gave to her looks, air, and manner, a dignity that seemed more than mortal.{S} Her glance
emained mute as statues, though at such a distance that their whispers could not have
mbroidered cushions, which, piled along a low platform that surrounded the chamber, served,
dispensation.{S} And for my conscience, a man that has slain three hundred Saracens, need not
as William the Bastard himself, or e'er a Norman adventurer that fought at Hastings.{S} I
the next willow-bush." Prince John made a signal that some attendants should follow him in
, but placed close beside him, and gave a signal that the evening meal should be placed upon
 terms, of which each vile crowder hath a stock that might last from hence to Christmas."
as thou shouldst be, and shalt be, amid all in England that is distinguished by beauty, or
rd the victor to the throne, and ending an error that has conjured all the blood from his
th dignity the veil around her face, as an intimation that the determined freedom of his
nk there was, a fayre for the maistrie, An outrider that loved venerie;{S} A manly man, to be
f the forest, known as well to me as to any forester that ranges it, and I will not leave you
xiety did the worthy Jew display during every course that was run, seldom failing to hazard a
```

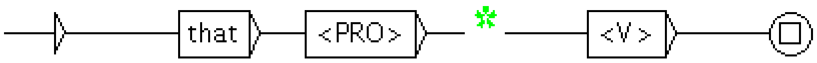
Negative contexts

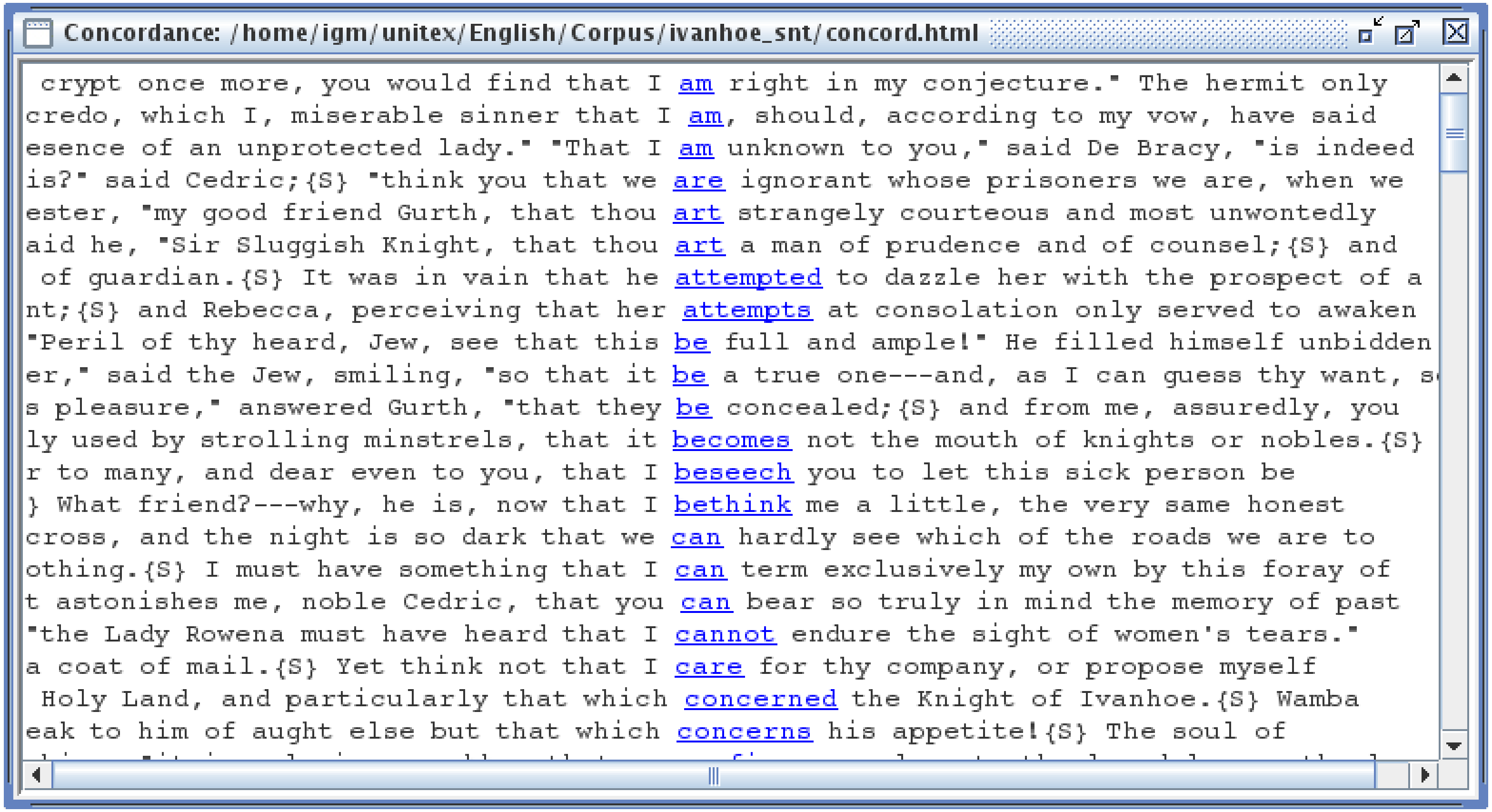



Concordance: /home/igm/unitex/English/Corpus/ivanhoe_snt/concord.html

```
 remarkable to be suppressed;{S} it was a brass ring, resembling a dog's collar, but without
le by a broad leathern belt, secured by a brass buckle;{S} to one side of which was attached
ance. {S}His jacket had been stained of a bright purple hue, upon which there had been some
 body, it was gathered at the middle by a broad leathern belt, secured by a brass buckle;{S}
arp-pointed, and two-edged knives, with a buck's-horn handle, which were fabricated in the
onduct, and to the laws of the land.{S} A circumstance which greatly tended to enhance the
 of the simplest form imaginable, being a close jacket with sleeves, composed of the tanned
f turf to which they made their way. {S}A considerable open space, in the midst of this
eat nobles, where the pomp and state of a court was emulated, Norman-French was the only
ed, occasioned the gradual formation of a dialect, compounded betwixt the French and the
sed;{S} it was a brass ring, resembling a dog's collar, but without any opening, and soldered
 superstition;{S} for, on the summit of a hillock, so regular as to seem artificial, there
er Don, there extended in ancient times a large forest, covering the greater part of the
he head and shoulders, in the manner of a modern shirt, or ancient hauberk. {S}Sandals, bound
n form, was of better materials, and of a more fantastic appearance. {S}His jacket had been
 the other a ram's horn, accoutred with a mouthpiece, for the purpose of blowing.{S} In the
the certain hazard of being involved as a party in whatever rash expedition the ambition of
ached a sort of scrip, and to the other a ram's horn, accoutred with a mouthpiece, for the
orched by the influence of the sun into a rusty dark-red colour, forming a contrast with the
alf, left the knees bare, like those of a Scottish Highlander.{S} To make the jacket sit yet
rk the existence of the Anglo-Saxons as a separate people subsequent to the reign of William
e bottom, and in stopping the course of a small brook, which glided smoothly round the foot
 period.{S} The eldest of these men had a stern, savage, and wild aspect.{S} His garment was
soldiery, flung their gnarled arms over a thick carpet of the most delicious green sward;{S}
rd upon his cheeks, which was rather of a yellow or amber hue.{S} One part of his dress only
```

Left contexts




Concordance: /home/igm/unitex/English/Corpus/ivanhoe_snt/concord.html

```
 crypt once more, you would find that I am right in my conjecture." The hermit only
credo, which I, miserable sinner that I am, should, according to my vow, have said
esence of an unprotected lady." "That I am unknown to you," said De Bracy, "is indeed
is?" said Cedric;{S} "think you that we are ignorant whose prisoners we are, when we
ester, "my good friend Gurth, that thou art strangely courteous and most unwontedly
aid he, "Sir Sluggish Knight, that thou art a man of prudence and of counsel;{S} and
 of guardian.{S} It was in vain that he attempted to dazzle her with the prospect of a
nt;{S} and Rebecca, perceiving that her attempts at consolation only served to awaken
"Peril of thy heard, Jew, see that this be full and ample!" He filled himself unbidden
er," said the Jew, smiling, "so that it be a true one---and, as I can guess thy want, s
s pleasure," answered Gurth, "that they be concealed;{S} and from me, assuredly, you
ly used by strolling minstrels, that it becomes not the mouth of knights or nobles.{S}
r to many, and dear even to you, that I beseech you to let this sick person be
} What friend?---why, he is, now that I bethink me a little, the very same honest
cross, and the night is so dark that we can hardly see which of the roads we are to
othing.{S} I must have something that I can term exclusively my own by this foray of
t astonishes me, noble Cedric, that you can bear so truly in mind the memory of past
"the Lady Rowena must have heard that I cannot endure the sight of women's tears."
a coat of mail.{S} Yet think not that I care for thy company, or propose myself
 Holy Land, and particularly that which concerned the Knight of Ivanhoe.{S} Wamba
eak to him of aught else but that which concerns his appetite!{S} The soul of
```

The history of Unitex, born as a clone of the non free software Intex, developed and used by the scientific community

Intex distribution policy

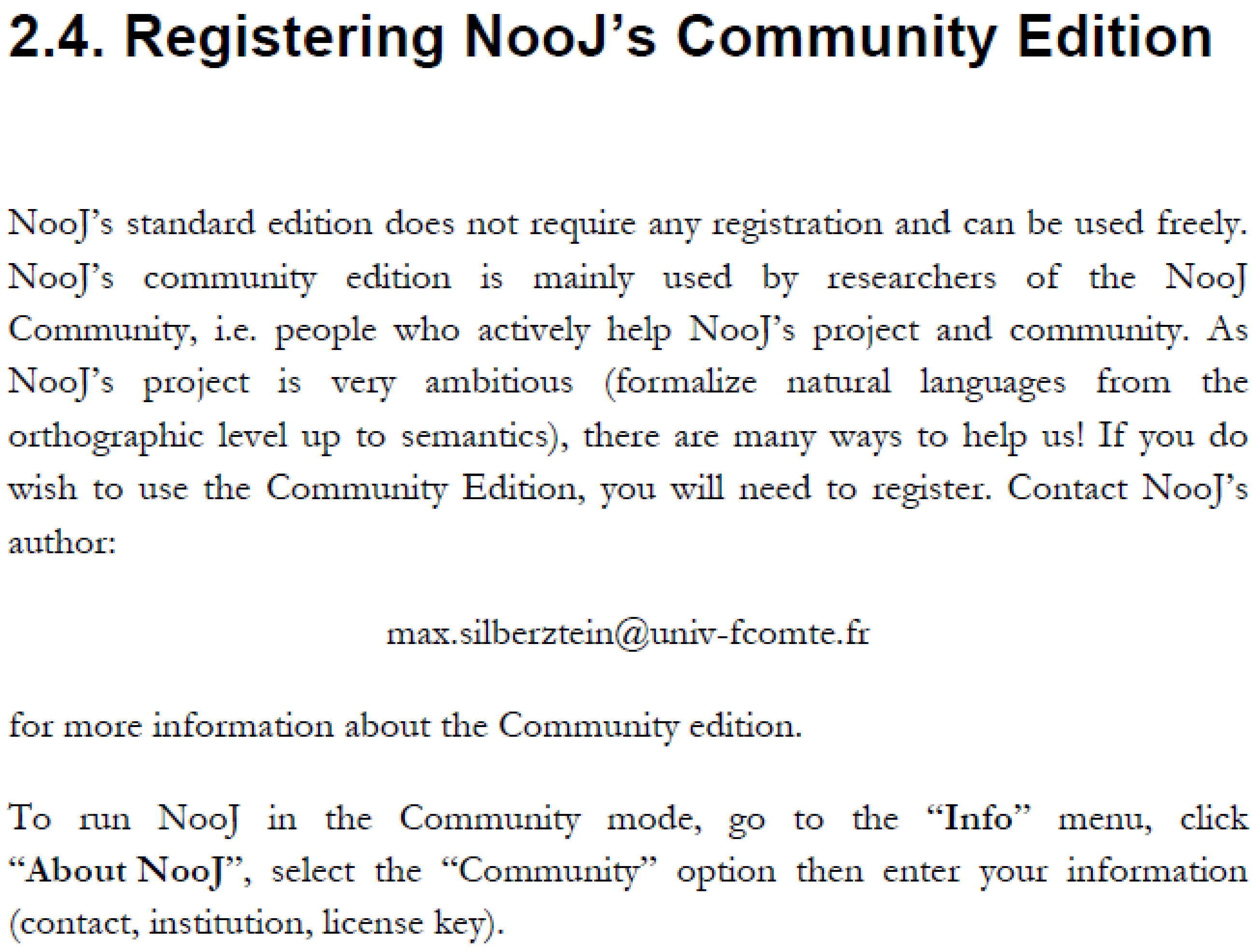
3. Launch INTEX; it will display a Machine Identification Number and ask for an Installation Key. To obtain this key:

- Send an email to ··max.silberztein@univ-fcomte.fr with the Subject: **Get INTEX**. Your email address should have a university suffix, such as ".edu", “univ-xxx.fr” or ".ac.uk"
- In the body of your email, enter the name of the person responsible for your work (typically the head of your laboratory, or your PhD adviser), your faculty or university, and your Machine Identification Number (displayed above the Installation Key in the Registration window), as follows:

  Contact: **John Smith**
  Institution: **University of North Texas**
  Machine ID: **12345**
- You should receive an automatic answer by email that will look like this:

  License number: **123**
  Contact: **John Smith**
  Institution: **University of North Texas**
  Machine ID: **12345**
  => Installation key: **ab1234cde**

NooJ distribution policy



2.4. Registering NooJ's Community Edition

NooJ's standard edition does not require any registration and can be used freely. NooJ's community edition is mainly used by researchers of the NooJ Community, i.e. people who actively help NooJ's project and community. As NooJ's project is very ambitious (formalize natural languages from the orthographic level up to semantics), there are many ways to help us! If you do wish to use the Community Edition, you will need to register. Contact NooJ's author:

max.silberztein@univ-fcomte.fr

for more information about the Community edition.

To run NooJ in the Community mode, go to the "**Info**" menu, click "**About NooJ**", select the "Community" option then enter your information (contact, institution, license key).

Problems

- if the software is modified, you may not be able to reproduce experiments
- as you can't read the code, you don't even know if there is a theoretical possibility of optimization
- if there are bugs, you cannot make complete benchmarks against other systems

Problems

- if a feature does not fit your needs, you have to design an awful home-made patch (usually in perl or python)
- if you want to test an idea, you have to recode quite the whole thing
- because of bugs and uncontrolled evolutions, you may have to keep several versions of the software, one per use

Open Source is good for science

- external contributions to Unitex:
  - ELAG: a disambiguation module
  - morphological filters, based on the TRE library
  - tools to generate Korean dictionaries
  - MultiFlex: MWU inflection module
  - PolyLex adapted for Russian and German
  - speed and memory optimizations
  - dictionary graphs

Open Source is good for science

  - statistic module
  - XAlign: text alignment module
  - linguistic resources for: English, Finnish, French (France, Quebec), Georgian (Ancient), German, Greek (Modern and Ancient), Italian, Korean, Norwegian, Polish, Portuguese (Brazil, Portugal), Russian, Serbian (Cyrillic and Latin alphabets), Spanish and Thai

- July 2009: more than 200 works that use or cite Unitex (source: Google scholar)

Conclusion

Every software produced by public research should be free software.