INTEX and the processing of natural languages

Max Silberztein
silberz@bestweb.net

Contents

1. Introduction

INTEX is a linguistic development environment that allows users to build large-coverage Finite State descriptions of Natural Languages and apply them to large texts (several dozen million words in real time).

Several modules of INTEX have been available since 1992 under NextStep; INTEX has been fully integrated in a graphical interface since 1996 (release 3.0), at which point it began to be distributed to research centers as a linguistic development tool. INTEX has just been ported to Windows 95-NT, as INTEX 4.0.

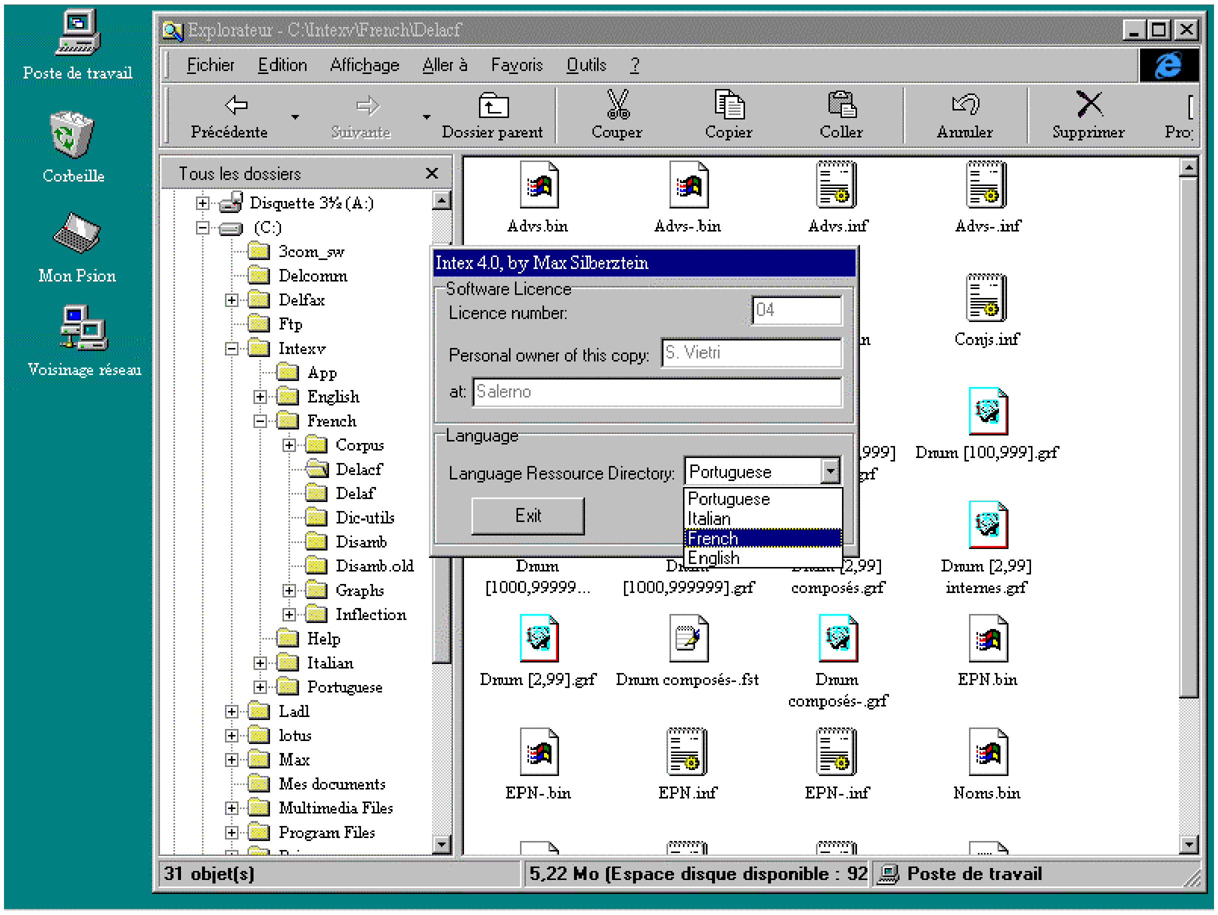
INTEX uses the work performed at the Laboratoire d'Automatique Documentaire et Linguistique (LADL), founded in 1967 by Prof. Maurice Gross. The goal of the LADL is to build a large-coverage description of Natural Languages by using 3 sets of tools:

- *Electronic dictionaries* and Finite State descriptions of the vocabulary and the morphology of Natural languages;
- *Local grammars* to identify frozen, semi-frozen and phrases in texts;
- Transformational rules described in a *Lexicon-Grammar* to extract (or generate) elementary sentences from (into) complex sentences.

One important aspect of INTEX is that Texts, Dictionaries and Grammars are all represented by **Finite State Transducers** (FSTs). Therefore, all the operations the user performs via the graphical interface are translated into a small number (about 30) of elementary operations on FSTs. For instance, applying a set of dictionaries to a text is performed by constructing a union of the dictionaries' FSTs, then applying the resulting FST to the text FST; removing lexical ambiguities in the text is performed by computing the intersection between a grammar FST and the text FST, etc.

2. Launching INTEX

The first operation consists of selecting the directory where the linguistic data is stored: alphabet of the language, preprocessing dictionaries and Finite State Transducers (FSTs), dictionaries for simple and compound words, FSTs representing the inflectional and derivational morphology of the language, FSTs used to remove lexical ambiguities, utilities for the maintenance of the dictionaries and the grammars.



Intex 4.0, by Max Silberztein

Software Licence

Licence number: 04

Personal owner of this copy: S. Vietri

at: Salerno

Language

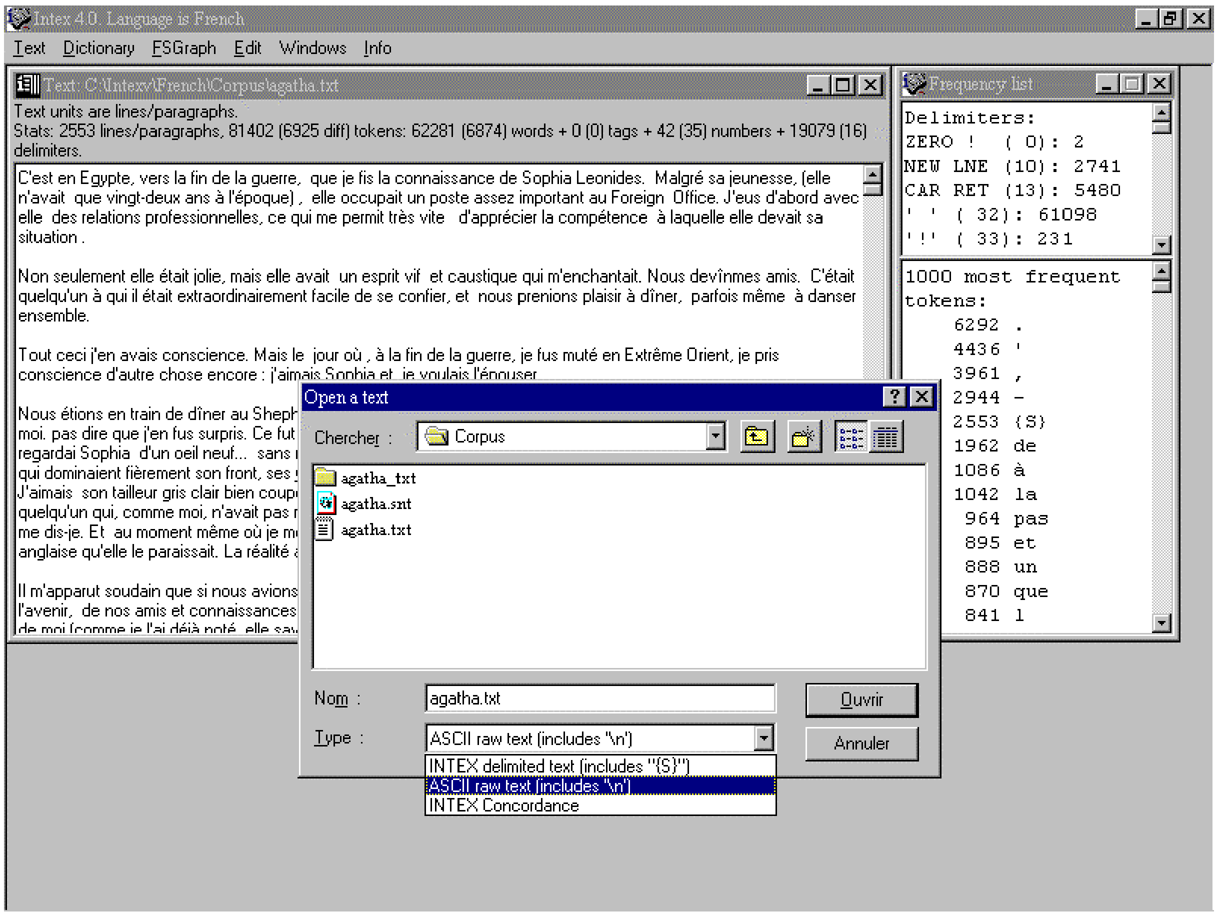
Language Ressource Directory: Portuguese

3. Opening a text (**Text->Open**)

INTEX processes three types of texts:

- ASCII files are considered as raw texts; they are considered as sequences of units delimited by the NEWLINE character; they do not contain any linguistic data;
- INTEX files are enriched texts where linguistic units (usually, sentences) have been delimited by the mark `{S}`; these files may contain linguistic tags;
- Concordance files are sequences of lines; each line consists of 3 columns: a left context, a sequence, and a right context. Concordances may have been created by previously indexing a FST in a text.

Below, a raw ASCII text is loaded and then indexed.


Intex 4.0. Language is French
Text Dictionary FSGraph Edit Windows Info
Text: C:\Intex\French\Corpus\agatha.txt
Frequency list
Open a text
Chercher : Corpus
agatha_txt
agatha.snt
agatha.txt
Ouvrir
Annuler

Text units are lines/paragraphs.
Stats: 2553 lines/paragraphs, 81402 (6925 diff) tokens: 62281 (6874) words + 0 (0) tags + 42 (35) numbers + 19079 (16) delimiters.

C'est en Egypte, vers la fin de la guerre, que je fis la connaissance de Sophia Leonides. Malgré sa jeunesse, (elle n'avait que vingt-deux ans à l'époque) , elle occupait un poste assez important au Foreign Office. J'eus d'abord avec elle des relations professionnelles, ce qui me permit très vite d'apprécier la compétence à laquelle elle devait sa situation .

Non seulement elle était jolie, mais elle avait un esprit vif et caustique qui m'enchantait. Nous devînmes amis. C'était quelqu'un à qui il était extraordinairement facile de se confier, et nous prenions plaisir à dîner, parfois même à danser ensemble.

Tout ceci j'en avais conscience. Mais le jour où , à la fin de la guerre, je fus muté en Extrême Orient, je pris conscience d'autre chose encore : j'aimais Sophia et je voulais l'épouser.

Nous étions en train de dîner au Sheph
moi. pas dire que j'en fus surpris. Ce fut
regardai Sophia d'un oeil neuf... sans
qui dominaient fièrement son front, ses
J'aimais son tailleur gris clair bien coup
quelqu'un qui, comme moi, n'avait pas
me dis-je. Et au moment même où je m
anglaise qu'elle le paraissait. La réalité

Il m'apparut soudain que si nous avions
l'avenir, de nos amis et connaissances
de moi (comme je l'ai déjà noté, elle sav

```
Delimiters:
ZERO !   ( 0): 2
NEW LNE (10): 2741
CAR RET (13): 5480
' '  ( 32): 61098
'!'  ( 33): 231
```

```
1000 most frequent
tokens:
    6292 .
    4436 '
    3961 ,
    2944 -
    2553 {S}
    1962 de
    1086 à
    1042 la
     964 pas
     895 et
     888 un
     870 que
     841 l
```

Nom : agatha.txt

Type : ASCII raw text (includes '\n')

- INTEX delimited text (includes "{S}")
- ASCII raw text (includes '\n')
- INTEX Concordance

4. Finite State Transducers (FSTs) in INTEX

First of all, a few definitions:

- an FST is a device that recognizes some sequences in the input, and associate them with some outputs. Typically, sequences are sequences of characters or sequences of words in the text written in a natural language; outputs are some linguistic information;

- an FST has the form of a graph that starts with an initial state, and ends with a terminal state. Recognized sequences are the ones that can be spelled by a path that goes from the initial state to the terminal state; Outputs of the FST are produced when a sequence has been recognized.

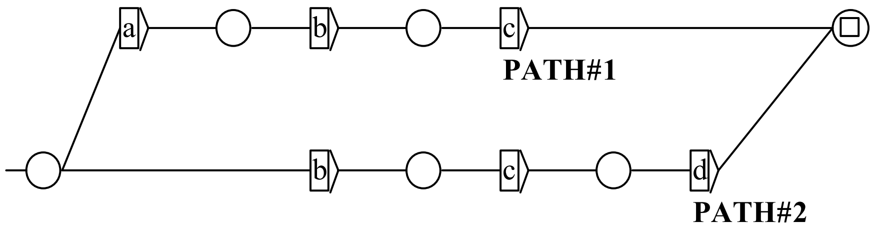
Now, let's explain how FSTs are applied to texts by the INTEX system:

(1) FSTs are applied from left to right

When the FST has matched one sequence of the text, it is reapplied **after** the end of the matching sequence. For instance, consider the following text:

```
z a b c d z
```

if we apply the following FST[1] to this text in REPLACE mode (*FST outputs replace matching sequences*):



we produce the resulting text:

```
z PATH#1 d z
```

The sequence `a b c` matched and was replaced by the output of the transducer, i.e. **`PATH#1`**. If we apply the same FST in MERGE mode (*FST outputs are inserted in the text*), we produce the following result:

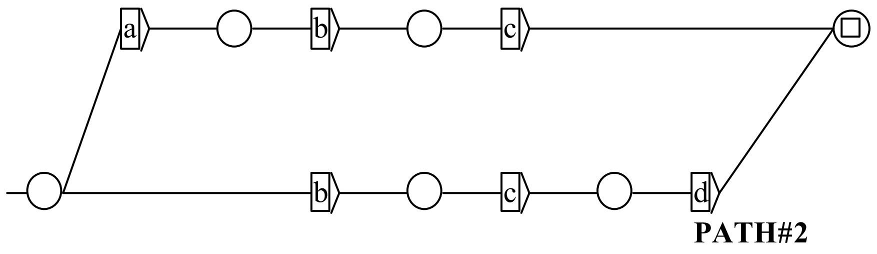
```
z a b PATH#1 c d z
```

The sequence `a b c` matched, and the output **`PATH#1`** was inserted before the character `c`.

[1]. FST inputs are displayed inside boxes, FST outputs are displayed below boxes.

In these two examples, the sequence **`b c d`** was not even 'seen' by the system. The sequence **`a b c`** has priority over the sequence **`b c d`** because it has matched *before*.

Note that the output of the FST may be the empty string. The same rule applies, even though the text was not modified. For instance, the following FST:



produces the following, unmodified text:

`z a b c d z`

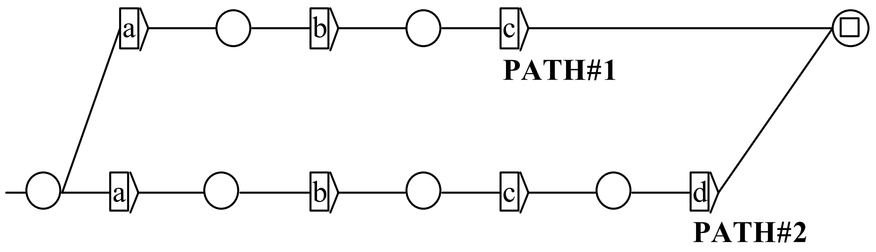
(**`a b c`** matched, then the FST is being applied at the position at **`d z`**).

(2) Longest matches have priority over shorter ones.

For instance, consider the following text:

`z a b c d z`

if we apply the following FST to it:



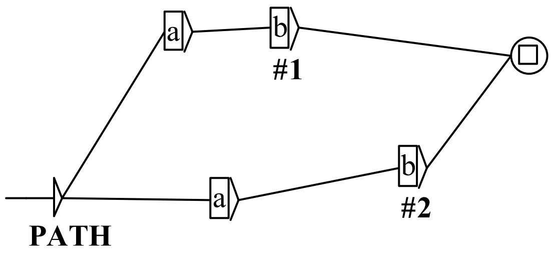
we get the resulting text:

(in REPLACE mode)	**`z PATH#2 z`**
(in MERGE mode)	**`z a b c PATH#2 d z`**

In other words, the matching sequence **`a b c d`** has priority over the sequence **`a b c`** because it is longer.

(3) INTEX doesn't handle ambiguous FSTs

If one sequence in the text is associated with 2 or more different outputs, INTEX performs an undefined action. For instance, if the following FST is applied to the text **`z a b z`** in REPLACE mode:



we get one of the two results:

```
     z PATH#1 z
or   z PATH#2 z
```

(4) INTEX doesn't handle FSTs that recognize the empty string

INTEX produces an error message if one attempts to apply an FST that recognizes the empty string.

Attention:

- INTEX **can** apply ambiguous FSTs to texts that are represented by FSTs;
- Therefore, ambiguous FSTs **can** also be used to disambiguate texts (because the disambiguated text is internally represented by an FST);
- When applying Finite State Automata[2] to texts, users **can** choose to index only shortest matching sequences, only the longest matches or all matches.

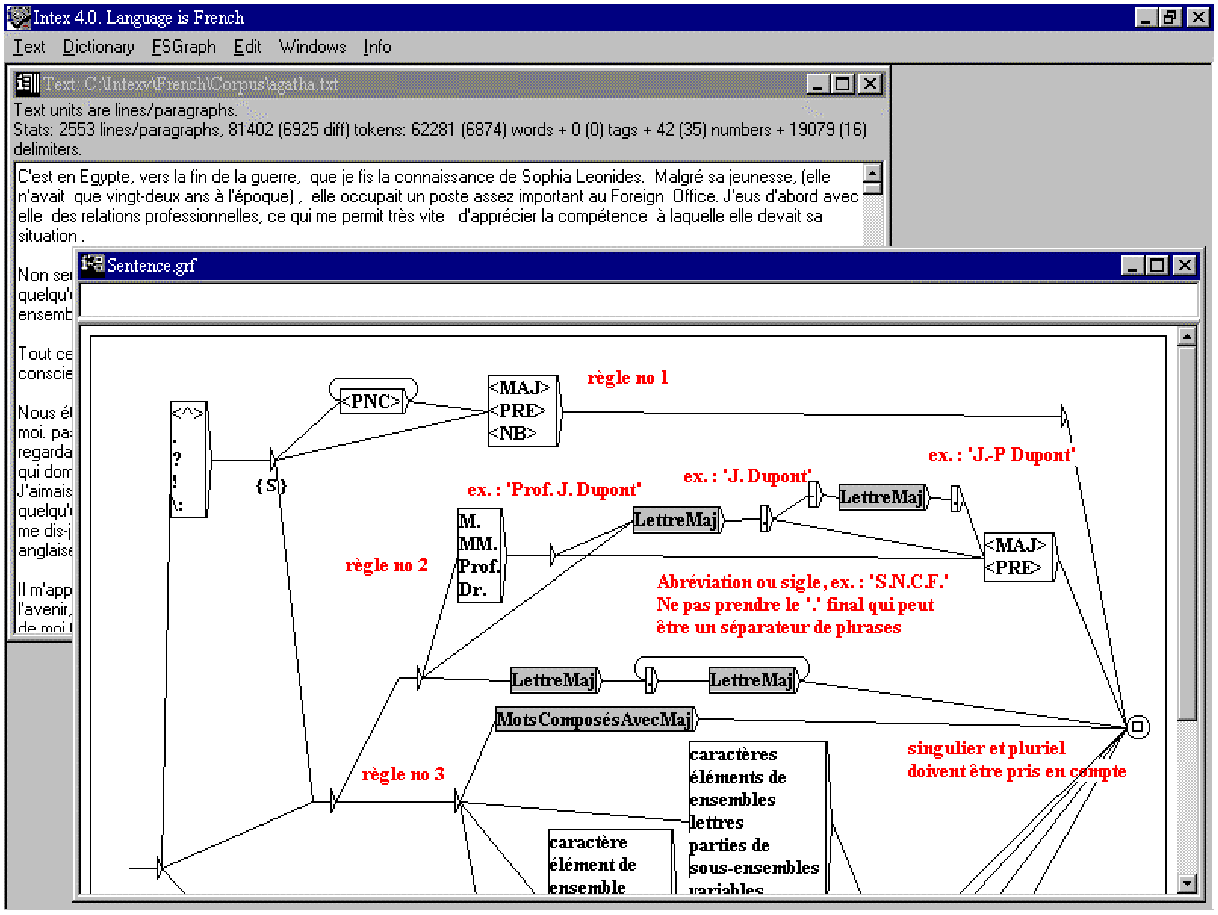
[2] . In INTEX, Finite State Automata are FSTs that produce the empty string.

5. Preprocessing the text (**Text->Preprocessing**)

Let us go back to the parsing of an ASCII file. After having loaded the file, users can preprocess the text, i.e. prepare the text for the linguistic analyses. The preprocessing consists of three operations: identification of sentences, of unambiguous compounds, and of special tokens.

2.1. Identifying sentences

The standard FST `Sentence.fst` (stored in the current language directory) is applied to the text in MERGE mode, i.e. the output of the FST is inserted in the text. Generally, this FST is used to insert the sentence delimiter `{S}` between consecutive sentences. Further INTEX processing will take this mark into account to process and index every linguistic unit.



Text units are lines/paragraphs.
Stats: 2553 lines/paragraphs, 81402 (6925 diff) tokens: 62281 (6874) words + 0 (0) tags + 42 (35) numbers + 19079 (16) delimiters.

C'est en Egypte, vers la fin de la guerre, que je fis la connaissance de Sophia Leonides. Malgré sa jeunesse, (elle n'avait que vingt-deux ans à l'époque) , elle occupait un poste assez important au Foreign Office. J'eus d'abord avec elle des relations professionnelles, ce qui me permit très vite d'apprécier la compétence à laquelle elle devait sa situation .

Gray nodes refer to embedded FSTs; for instance, LettreMaj is the name of an FST that identifies the 26 capital letters A…Z; MotsComposésAvecMaj is the name of an FST that lists all the French compound words that end with a capital letter (e.g. *Vitamine C*).

The FST **`Sentence.fst`** must be read in the following manner:

- if a period is followed by a word in capital letters, INTEX inserts the sentence delimiter between the period and the word (see on the top of the FST);

- if an single uppercase letter is followed by a period, followed by a word in uppercase (e.g. *J. Dupont*), INTEX does not insert any sentence delimiter;

- compound words that end with an uppercase letter (e.g. *Vitamine C*) may occur at the end of a sentence; the uppercase letter followed by a period must not be mistaken for an abbreviated firstname.

Thanks to the *Left to Right* priority seen previously, the last processing gets priority over the second processing, which has priority over the first processing. We then get the correct result:

```
J. Dupont comes. {S} Paul eats some vitamine C. {S} Luc also.
```

(*C. Luc* is not processed as *J. Dupont*). Although this FST is not perfect (some systematic errors are due to the use of some English abbreviations in French texts), it process usual French novels and journalistic texts with a high rate of success (>99.5%).

2.2. Identifying unambiguous compounds

The second step of the preprocessing will consist of identifying and tagging the unambiguous compound words in the text, this operation corresponds to a look-up of the dictionary **`Norm.dic`** (stored in the current language directory).

Let us first define INTEX units of processing.

INTEX units of processing

INTEX users define the **alphabet** of the language in the file `Alphabet` stored in the current language directory.

All the characters that are listed in this file are **letters**; the other characters are **delimiters**.

The alphabet file consists of a sequence of lines ordered alphabetically; each line contains two or three 'equivalent' letters; these equivalence classes are used by the sort and the lookup routines. For instance, here are the first 7 lines of the French alphabet:

```
Aa
AÀà
Aââ
Bb
Cc
CÇç
```

Linguistic units are classified in two main types:

- simple words are sequences of letters, e.g. *table*
- compound words are sequences of simple words, e.g. *red tape*

Since the apostrophe, the hyphen and the blank are not listed in the alphabet, INTEX treats the following words as compounds (even though their constituents cannot appear alone):

aujourd'hui, attaché-case, parce que

Generally, simple and compound words are identified by consulting the dictionaries of the system. In certain cases, they may be identified by applying morphological FSTs.

Ambiguity

Ambiguous words are words that correspond to more than one lexical entry in the dictionaries and the FSTs of the system.

Ambiguous compound words are sequences that correspond either to more than one lexical compound entry, such as:

`pied noir` (*Blackfoot*, Frenchman born in Algeria)
`pied noir` (*Blackfoot*, American Indian)

or to more than one sequence of lexical (simple or compound) entries, e.g.:

`red tape` : **`{red tape,.N} + {red,.A} {tape,.N}`**

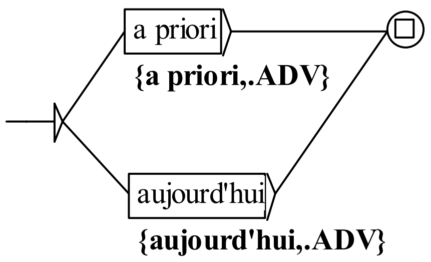
Unambiguous compound words correspond to only one lexical compound entry, e.g.:

`a priori` : **{a priori,.ADV}**

Tagging a text consists of replacing its forms by the corresponding lexical entry, written between curly brackets.

One can only replace unambiguous, or disambiguated forms

It is desirable to identify and tag unambiguous compound words as soon as possible, in order not to treat their constituents (e.g. '*a*' and '*priori*') as simple words. This operation can be performed during the preprocessing analysis by means of the special dictionary **`Norm.dic`** stored in the current language directory, or by FSTs applied in REPLACE mode. For instance, the following FST:



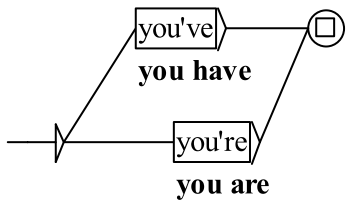
would tag the two French unambiguous adverbs.

The French dictionary **`Norm.dic`** lists over one thousand entries.

2.3. Identifying special tokens

The last stage of the preprocessing consists of identifying and tagging special tokens, such as elided and contracted words, unambiguous abbreviations, etc. This step is performed by applying the FST `Norm.fst` (stored in the current language directory) in REPLACE mode.

For instance, by applying the following FST to English texts:

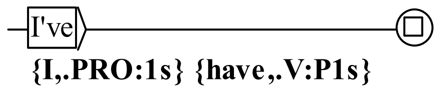


one replaces all the utterances of the sequences *you've* and *you're* by the more explicit sequences *you have, you are*. Such replacements are performed before indexing the text and before consulting the dictionaries.

One could perform a similar substitution to replace the sequence *I've* by the form *I have*, but that would lead to artificially add an ambiguity in the text: the sequence *I've* is not ambiguous (it is the pronoun *I*, followed by the verb *have*), while *I have* **is** ambiguous, as shown in the sentence:

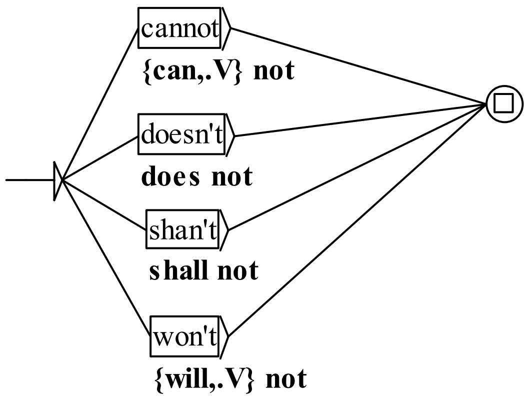
Nelson and Napoleon I have met in ...

(*I* is used as a roman numeral). Thus, it is better to use the following FST:



I is the pronoun, first person singular; *have* is the verb conjugated in the present, first person singular.

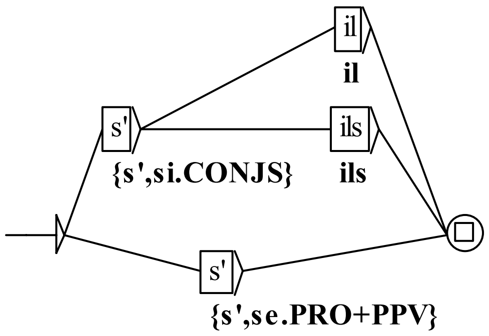
It is also possible to replace one word by a more complex sequence, such as in the following FST:



(the forms *can* and *will* are ambiguous in general; it is better to tag them here).

Unambiguous sequences of simple words

During this early stage of preprocessing, it is also possible to disambiguate a number of grammatical words by replacing them with a tag when they occur in certain contexts. For instance, the following FST:



replaces on the one hand all the utterances of the sequence s'*il* by the following sequence:

… {s',si.CONJS} il …

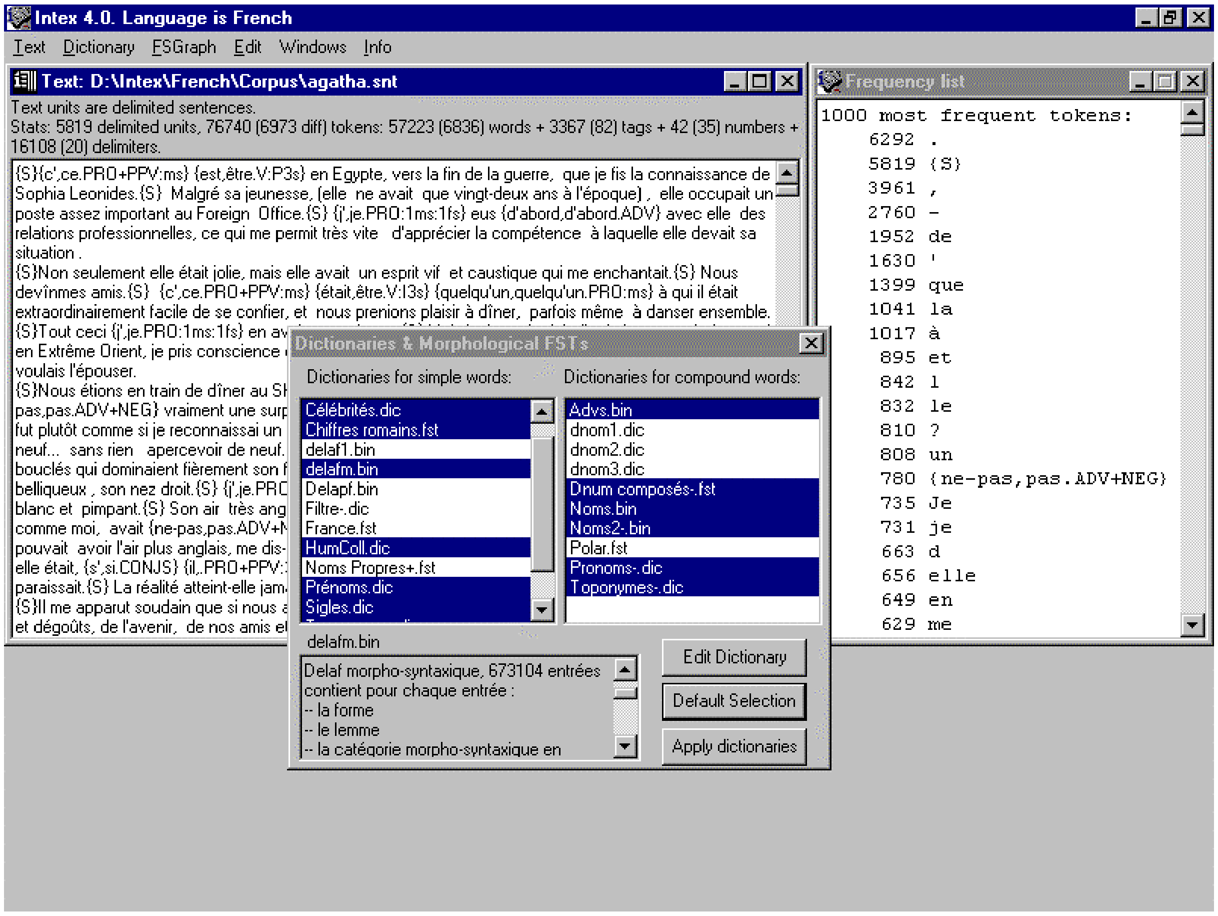
On the other hand, a sequence like *s'aide* would be replaced by:

… {s',se.PRO+PPV} aide …

Thus, the ambiguous form (*s'* corresponds either to the conjunction *si*, or the pronoun *se*) has been disambiguated.

6. Apply dictionaries and lexical FSTs (**Text->Apply dictionaries**)

After the completion of the preprocessing stage, the user selects the dictionaries and FSTs used to identify simple words (left column) and compound words (right column) in the text.


Intex 4.0. Language is French
Text Dictionary FSGraph Edit Windows Info
Text: D:\Intex\French\Corpus\agatha.snt
Frequency list
Dictionaries & Morphological FSTs
Dictionaries for simple words:
Dictionaries for compound words:
Edit Dictionary
Default Selection
Apply dictionaries

Text units are delimited sentences.
Stats: 5819 delimited units, 76740 (6973 diff) tokens: 57223 (6836) words + 3367 (82) tags + 42 (35) numbers + 16108 (20) delimiters.

{S}{c',ce.PRO+PPV:ms} {est,être.V:P3s} en Egypte, vers la fin de la guerre, que je fis la connaissance de
Sophia Leonides.{S} Malgré sa jeunesse, (elle ne avait que vingt-deux ans à l'époque) , elle occupait un
poste assez important au Foreign Office.{S} {j',je.PRO:1ms:1fs} eus {d'abord,d'abord.ADV} avec elle des
relations professionnelles, ce qui me permit très vite   d'apprécier la compétence à laquelle elle devait sa
situation .
{S}Non seulement elle était jolie, mais elle avait un esprit vif et caustique qui me enchantait.{S} Nous
devînmes amis.{S} {c',ce.PRO+PPV:ms} {était,être.V:I3s} {quelqu'un,quelqu'un.PRO:ms} à qui il était
extraordinairement facile de se confier, et nous prenions plaisir à dîner, parfois même à danser ensemble.
{S}Tout ceci {j',je.PRO:1ms:1fs} en av
en Extrême Orient, je pris conscience
voulais l'épouser.
{S}Nous étions en train de dîner au Sh
pas,pas.ADV+NEG} vraiment une surp
fut plutôt comme si je reconnaissai un
neuf... sans rien   apercevoir de neuf.
bouclés qui dominaient fièrement son f
belliqueux , son nez droit.{S} {j',je.PRO
blanc et pimpant.{S} Son air très ang
comme moi, avait {ne-pas,pas.ADV+N
pouvait avoir l'air plus anglais, me dis-
elle était, {s',si.CONJS} {il,.PRO+PPV:
paraissait.{S} La réalité atteint-elle jam
{S}Il me apparut soudain que si nous a
et dégoûts, de l'avenir, de nos amis et

Dictionaries for simple words:	Dictionaries for compound words:
Célébrités.dic	Advs.bin
Chiffres romains.fst	dnom1.dic
delaf1.bin	dnom2.dic
delafm.bin	dnom3.dic
Delapf.bin	Dnum composés-.fst
Filtre-.dic	Noms.bin
France.fst	Noms2-.bin
HumColl.dic	Polar.fst
Noms Propres+.fst	Pronoms-.dic
Prénoms.dic	Toponymes-.dic
Sigles.dic	

delafm.bin

Delaf morpho-syntaxique, 673104 entrées
contient pour chaque entrée :
-- la forme
-- le lemme
-- la catégorie morpho-syntaxique en

```
1000 most frequent tokens:
    6292 .
    5819 {S}
    3961 ,
    2760 -
    1952 de
    1630 '
    1399 que
    1041 la
    1017 à
     895 et
     842 l
     832 le
     810 ?
     808 un
     780 {ne-pas,pas.ADV+NEG}
     735 Je
     731 je
     663 d
     656 elle
     649 en
     629 me
```

The three types of files are:

■ **`.dic`**	an ASCII file that corresponds to a DELAF-type dictionary;
■ **`.fst`**	an FST graph
■ **`.bin`**	a dictionary compacted into a FST represented in a binary file.

7. Priority levels

Dictionaries and FSTs are associated with a 3-level priority system used to hide or impose some lexical data:

- if a (simple or a compound) form in the text matches a high priority dictionary or FST, the consultation stops;
- if not, all 'regular' dictionaries and FSTs are consulted;
- if no lexical entry corresponds to the form, low level priority dictionaries and FSTs are consulted.

Giving a high priority to a dictionary or a FST for simple words is generally useful in order to remove 'unpleasant' ambiguities from the general-purpose dictionaries: if one knows that a certain use of a word never occurs in the current text, one can insert the word in a high priority dictionary, without the irrelevant analysis. For instance, the word *la* in French is three-time ambiguous:

```
{la,le.DET:fs} + {la,le.PRO+PPV} + {la,.N:ms}
```

la can be either a determiner, a pronoun or the masculine noun (musical note). These three analyses are represented in the general-purpose `Delaf.bin` dictionary. By entering the only first two entries in a dictionary that has priority over the Delaf, one gets rid of the noun (which is not frequent).

Giving a high priority to a dictionary or a FST for compounds is useful to get rid of systematic structural ambiguities when compound words can be shortened. For instance, the following regular expression represents two synonymous variants of the same adverb:

```
Dans l'intimité (la plus stricte + <E>)
```

(*in the intimacy*). If the longest variant occurs in a text, INTEX must not parse it as ambiguous, as seen in the following two tagged results:

- either the adverb: `{dans l'intimité la plus stricte,ADV}`
- or the sequence: `{dans l'intimité,ADV} la plus stricte`

Obviously, the longest variant has to have priority over the shortest one. Giving priority to dictionaries and FSTs for compounds gives priority to the longest matches.

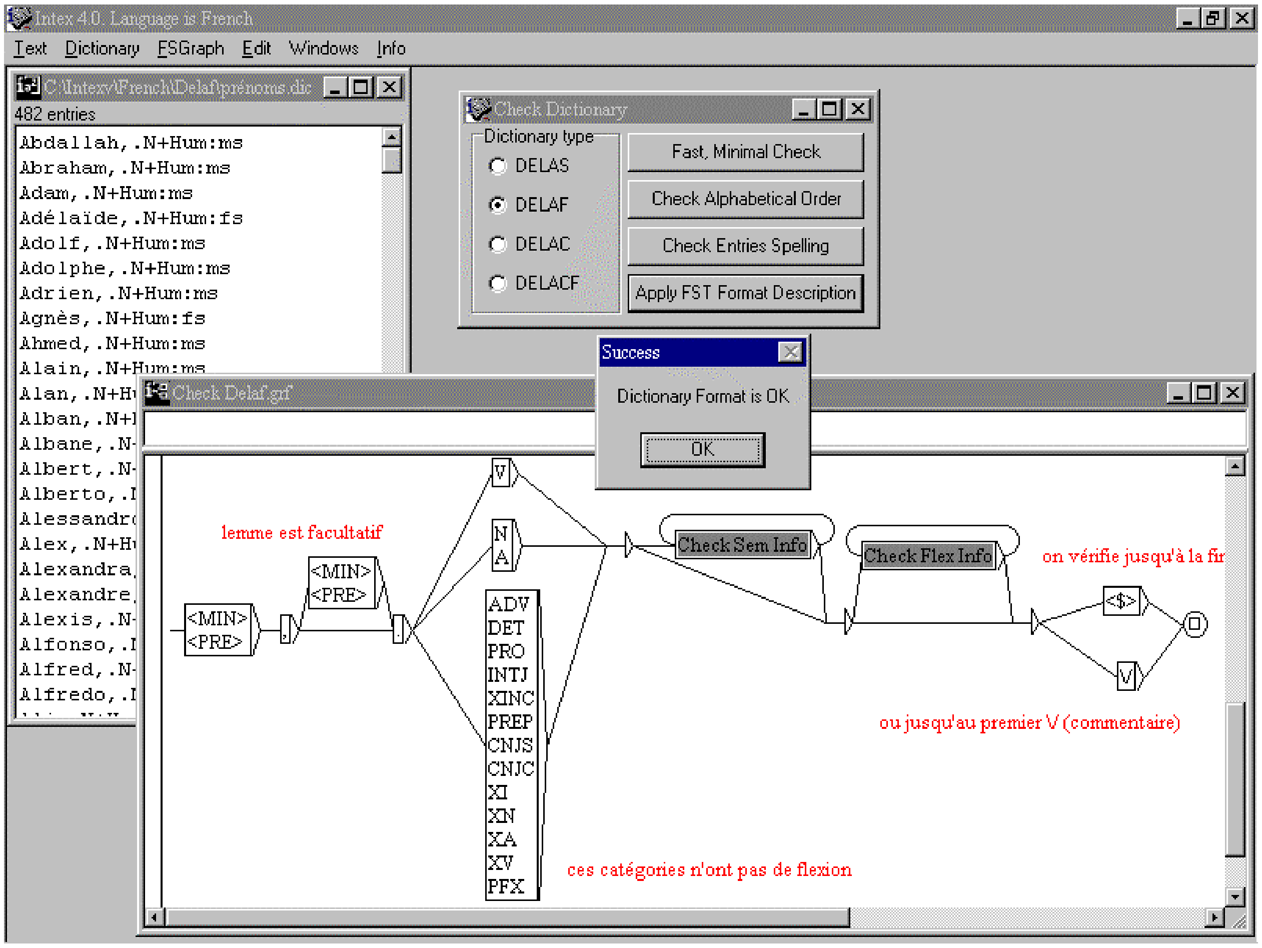
8. INTEX Dictionaries (**Dictionary->Open**)

Dictionaries applied by INTEX to identify simple forms in texts must be in the format of a **DELAF**; dictionaries used to identify compounds must be in the format of a **DELACF**. These dictionaries associate text utterances with one lemma and some linguistic information. Here is for instance one entry of the French DELAF:

```
amuserions,amuser.V+t+4:C1p
```

the form *amuserions* is associated with the lemma *amuser* which is a transitive (`+t`) verb (`V`) of the syntactic class #4 (`+4`); the form is conjugated in the Conditional, first person plural (`C1p`).

The format of user-defined dictionaries must be checked via the command: Check format[3]:



```
Abdallah,.N+Hum:ms
Abraham,.N+Hum:ms
Adam,.N+Hum:ms
Adélaïde,.N+Hum:fs
Adolf,.N+Hum:ms
Adolphe,.N+Hum:ms
Adrien,.N+Hum:ms
Agnès,.N+Hum:fs
Ahmed,.N+Hum:ms
Alain,.N+Hum:ms
Alan,.N+H
Alban,.N+
Albane,.N-
Albert,.N-
Alberto,.
Alessandr
Alex,.N+H
Alexandra
Alexandre
Alexis,.N-
Alfonso,.
Alfred,.N-
Alfredo,.
```

CheckSemInfo describes the syntactic and semantic features associated with the entry (e.g. `+t` for transitive); CheckFlexInfo describes the inflectional codes (e.g.`:ms` for masculine singular).

[3]. Of course, users may edit the FSTs that describe INTEX dictionaries' format in order to add their own linguistic information.

9. From a DELAS to a DELAF (Dictionary->Inflect)

Some dictionaries can be entered by the users directly in the form of a DELAF; typically, dictionaries of first names, abbreviations, etc.

However, in most languages, the dictionary of all the inflected forms would be too big to be entered 'manually'. It is then better to manually build a DELAS-type dictionary, and automatically generate the corresponding DELAF.

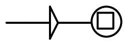
In a DELAS dictionary, lexical entries are lemmas; each entry is associated with one FST that represents all the corresponding inflected forms. For instance, here are five entries of the French DELAS:

```
amuser,V3+t+4
cousin,N1+Anim
cousin,N32+Hum
de,PREP
```

V3, **A31**, **N1**, **N32** and **PREP** are names of inflectional FST files.

- All the words in a language that have the same set of suffixes, associated with the same inflectional information are associated with the same inflectional FST. For instance, in French, all the verbs that conjugate like *amuser* (*aider*, *voler*, etc.) are associated with the FST **V3**; all the nouns that take an 'e' in the feminine and an 's' in the plural are associated with the FST **N32**, etc.;

- if a word has no inflection, it must be associated with an FST that has no input and produces no output. For instance, in French, the FST **PREP** is the following (prepositions do not inflect):



10. Multiple entries in the DELAS

If one word is associated with more than one inflectional class, it must be represented by more than one DELAS entry. For instance:

```
cousin,N32+Hum
cousin,N1+Anim
```

The first entry corresponds to a person (= cousin); the noun gets the feminine form *cousine*; the second entry corresponds to an animal (= mosquito); this noun is masculine only. In the same manner, the French DELAS includes the two following entries:

```
voile,N1+Conc
voile,N21+Conc
```

the first entry corresponds to a masculine noun (= veil); the second entry corresponds to a feminine noun (= sail). Although both entries are associated with the same distributional class (**`+Conc`** stands for concrete nouns), and to the same inflection (they both take an 's' in the plural) they must be associated with two different FSTs: one generates the code **`m`** (masculine) whereas the other generates the code **`f`** (feminine).

Sometimes, different sets of syntactic or semantic properties may lead to different inflections. For instance, the verb *voler* in the meaning of 'to steal' accepts a passive construction; therefore, the four past participle forms *volé, volée, volés, volées* must be represented, as we see in:

Ces fleurs ont été volées (= these flowers have been stolen)

The verb *voler* in the meaning of 'to fly' is intransitive; therefore, only the invariable form *volé* must be represented. This leads to at least two entries in the DELAS:

```
voler,V3U+i
voler,V3+t
```

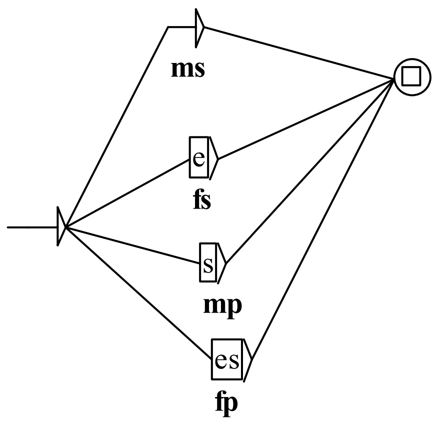
`V3U` represents only one invariable participle form whereas **`V3`** represents the four forms.

First entries of the French DELAS dictionary, 'basic' vocabulary

```
a,N2
a,XI+PMCO
à,PREP
ab,XI+PMCO
abaissé,A32
abaisser,V3+t+11
abaisser,V3+t+32RA
abaisser,V3+t+38L
abaisser,V3E+pi
abandon,N1
abandonné,A32
abandonné,N32+Hum
abandonner,V3+t+32H
abandonner,V3+t+36DT
abandonner,V3+t+38L1
abandonner,V3+t+38LR
abandonner,V3+t+6
abandonner,V3+t+9
abandonner,V3E+pi+7
abandonner,V3U+i+31H
abandonner,V3U+i+35R
abasourdi,A32
abasourdissant,A32
abatage,N1
abats,N2P+Conc
abattage,N1
abattement,N1
abattoir,N1+Conc
abattre,V68+t+32H
abattre,V68+t+32R3
abattre,V68+t+37E
abattre,V68+t+38LD
abattre,V68+t+4
abattre,V68E+pi+35R
abattu,A32
abattu,N1
abbaye,N21+Conc
abc,N2
abcès,N2+Conc
abdication,N21
abdiquer,V3+t+32R3
abdomen,N1+Conc
abdominal,A76
abdominaux,N2P
abeille,N21+Anl
aberrant,A32
aberration,N21
abêtir,V18+t+11
abêtir,V18+t+4
abêtissant,A32
abêtissement,N1
abhorrer,V3+t+12
abîme,N1+Conc
abîmé,A32
abîmer,V3+t+32C
abîmer,V3+t+38LD
abîmer,V3+t+4
abject,A32
abjection,N21
ablation,N21
ablution,N21
abnégation,N21
aboiement,N1
abois,N2P
abolir,V18+t+10
abolir,V18+t+32R2
abolition,N21
abolitionniste,A31
abolitionniste,N31+Hum
abominable,A31
abominablement,ADV
abomination,N21
abominer,V3+t+12
abondamment,ADV+Dadv
abondamment,ADV+PADV
abondance,N1
abondance,N21
abondant,A32
abonder,V3U+i+34L0
abonné,A32
abonné,N32+Hum
abonnement,N1
abonner,V3+t+11
abord,N1
abord,XI+PMCO+Préd
abordable,A31
abordage,N1
aborder,V3+t+32H
aborder,V3+t+32R2
aborder,V3+t+38L1
aborder,V3U+i+35L
abords,N2P
abouti,A32
aboutir,V18U+i+14
aboutir,V18U+i+31R
aboutir,V18U+i+35L
aboutissant,A32
aboutissement,N1
aboyer,V13+t+32R3
aboyer,V13+t+9
abracadabrant,A32
abrégé,N1+Abst
abrégé,N1+Conc
abrégement,N1
abrègement,N1
abréger,V10+t+32RA
abreuver,V3+t+37M1
abréviation,N21
abri,N1+Conc
abricot,A80+Conc
abricot,N1+Conc
abricotier,N1+Conc
abrité,A32
abriter,V3+t+10
abriter,V3+t+38LD
abriter,V3+t+38R
abriter,V3E+pi+35R
abrogation,N21
abrupt,A32
abrupt,N1+Conc
abruptement,ADV
abrupto,XI+PMCO
abruti,A32
abruti,N32+Hum
abrutir,V18+t+11
abrutir,V18+t+37M1
abrutissant,A32
abrutissement,N1
abscons,A61
absence,N21
absent,A32
absent,N32+Hum
absentéisme,N1
absentéiste,A31
absentéiste,N31+Hum
absenter,V3E+pi+2
absenter,V3E+pi+31H
absinthe,N21+Conc
absolu,A32
…
```

11. Inflectional FSTs

Inflectional FSTs associate sets of suffixes to the corresponding inflectional information. For instance, the FST **N32** (used to describe the inflection of the noun cousin) is displayed below (this file is stored in the **Inflection** subdirectory of the current language directory):



The FST associates each suffix of the DELAS entry *cousin* to some inflectional codes:

- if nothing is concatenated to the DELAS entry, one gets *cousin*; this form is associated with the inflectional codes **ms** (masculine singular);
- if 'e' is concatenated to the DELAS entry, one gets *cousine*; this form is associated with the inflectional codes **fs** (feminine singular);
- if 's' is concatenated to the DELAS entry, one gets *cousins*; this form is associated with the inflectional codes **mp** (masculine plural);
- if 'es' is concatenated to the DELAS entry, one gets *cousines*; this form is associated with the inflectional codes **fp** (feminine plural).

A simple process of concatenating the FST **N32** to the lexical entry, and then exploring the FST to generate all the paths will produce the resulting DELAF entries:

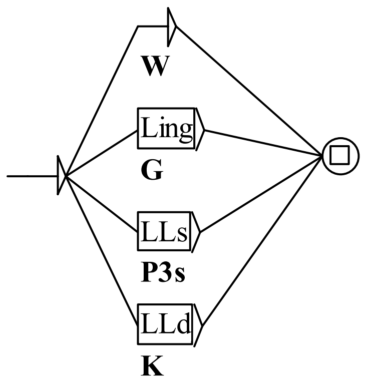
```
cousin,cousin.N32+Hum:ms
cousine,cousin.N32+Hum:fs
cousins,cousin.N32+Hum:mp
cousines,cousin.N32+Hum:fp
```

'Delete' operator

Lemmas are not always proper prefixes of their inflected forms. For instance, the verb 'have' gets the following forms:

have, having, has, had

In order to produce the last three forms, one needs to be able to 'delete' characters from the lemma. This operation is performed by adding the 'delete' character **`L`** (*Go Left*) to the alphabet of the DELAS dictionary. The FST used to produce the four inflected forms is then:



From the DELAS entry: **`have,V1`**, the process of constructing a DELAF dictionary becomes the following: (1) first we get the four following lines by exploring the FST V1:

```
have,have.V1:W
haveLing,have.V1:G
haveLLs,have.V1:P3s
haveLLd,have.V1:K
```

(2) then we spell each inflected form, interpreting the operator **`L`** as a 'delete last character' command. The final result is then:

```
have,have.V1:W
having,have.V1:G
has,have.V1:P3s
had,have.V1:K
```

Stack operators

In theory, the two operators *insertion*, *deletion* are sufficient to represent all kinds of inflection, even highly irregular ones:

```
avoir LLLLLont => ont
```

Basically, one can always connect any form to any lemma, by deleting all the letters, and then spelling the form.

The problem I discuss now is a practical one. Consider for instance the following German nouns:

Anrand, Balg, Blatt, Dach, Daus, Fach, Gehalt, Gemach, Geschmack, Gewand, Haus, Inland, Kaff, Kalb, Kraut, Lamm, Land, Mahd, Mann, Mark, Maul, Pfand, Rand, Salband, Wald, Wams…

These nouns have a similar inflected form:

Anränder, Bälger, Blätter, Dächer, Däuser, Fächer, Gehälter, Gemächer, Geschmäcker, Gewänder, Häuser, Inländer, Käffer, Kälber, Kräuter, Lämmer, Länder, Mähder, Männer, Märker, Mäuler, Pfänder, Ränder, Salbänder, Wälder, Wämser…

Basically, the inflection of all these forms is the same: add an ¨ to the *a* two letters before the end, and add the suffix *er*.

Unfortunately, with the only two operators of insertion and deletion, each of these forms would have to be associated with a different FST:

```
Anrand LLLänder => Anränder
  Balg LLLälger => Bälger
 Blatt LLLätter => Blätter
```

⇨ Over 40 almost identical inflectional FSTs would have to be added to the system!

I added two operators to the inflection process: `C` (*Copy*) and `R` (*Go Right*); all the previous inflections are then performed with the same command:

`LLLäRCCer`
go left three times, insert an 'ä', go right one time, copy the two letters, insert 'er'

The string produced by the FST is processed by four stack operators that take a constant time:

c	insert character 'c' at the end of the form
`L`	delete last character; push it onto the stack
`R`	pop the stack
`C`	copy the character at the top of the stack to the end of the form; pop the stack

Here is how the string **`BalgLLLälger`** is processed:

Operator		***Resulting form***	***Stack***
B	Insert B	B^	-
a	Insert a	Ba^	-
l	Insert l	Bal^	-
g	Insert g	Balg^	-
L	Push g	Bal^	g
L	Push l	Ba^	l g
L	Push a	B^	a l g
ä	Insert ä	Bä^	a l g
R	Pop	Bä^	l g
C	Pop & Insert	Bäl	g
C	Pop & Insert	Bälg	-
e	Insert e	Bälge	-
r	Insert r	Bälger	-

The result is **`Bälger`**. In the same manner, the following French verbs are associated with a unique FST **V6**

acheter, amener, beder, écarteler, élever, peser, semer

The conjugated forms: achète, amène, bède, écartèle, élève, pèse, and sème are produced by the command: **`LLLLèCC`**.

- 200 FSTs are required for the description of English,
- 250 FSTs for Italian,
- 300 for French and Spanish,
- 450 for Bulgarian,
- 670 for German.

⇨ The construction of DELAF dictionaries takes a time proportional to their length (linear time).

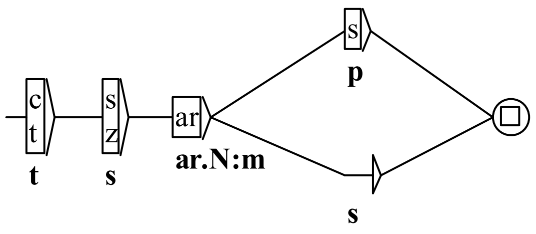
Resulting DELAF dictionary

```
a,.N:ms:mp
a,.XI+PMCO
à,.PREP
ab,.XI+PMCO
abaissé,.A:ms
abaissée,abaissé.A:fs
abaissés,abaissé.A:mp
abaissées,abaissé.A:fp
abaisser,.V+t+11:W
abaissant,abaisser.V+t+11:G
abaissé,abaisser.V+t+11:Kms
abaissée,abaisser.V+t+11:Kfs
abaissés,abaisser.V+t+11:Kmp
abaissées,abaisser.V+t+11:Kfp
abaisse,abaisser.V+t+11:P1s:P3s:S1s:S3s:Y2s
abaisses,abaisser.V+t+11:P2s:S2s
abaissons,abaisser.V+t+11:P1p:Y1p
abaissez,abaisser.V+t+11:P2p:Y2p
abaissent,abaisser.V+t+11:P3p:S3p
abaissais,abaisser.V+t+11:I1s:I2s
abaissait,abaisser.V+t+11:I3s
abaissions,abaisser.V+t+11:I1p:S1p
abaissiez,abaisser.V+t+11:I2p:S2p
abaissaient,abaisser.V+t+11:I3p
abaissai,abaisser.V+t+11:J1s
abaissas,abaisser.V+t+11:J2s
abaissa,abaisser.V+t+11:J3s
abaissâmes,abaisser.V+t+11:J1p
abaissâtes,abaisser.V+t+11:J2p
abaissèrent,abaisser.V+t+11:J3p
abaisserai,abaisser.V+t+11:F1s
abaisseras,abaisser.V+t+11:F2s
abaissera,abaisser.V+t+11:F3s
abaisserons,abaisser.V+t+11:F1p
abaisserez,abaisser.V+t+11:F2p
abaisseront,abaisser.V+t+11:F3p
abaissasse,abaisser.V+t+11:T1s
abaissasses,abaisser.V+t+11:T2s
abaissât,abaisser.V+t+11:T3s
abaissassions,abaisser.V+t+11:T1p
abaissassiez,abaisser.V+t+11:T2p
abaissassent,abaisser.V+t+11:T3p
…
```

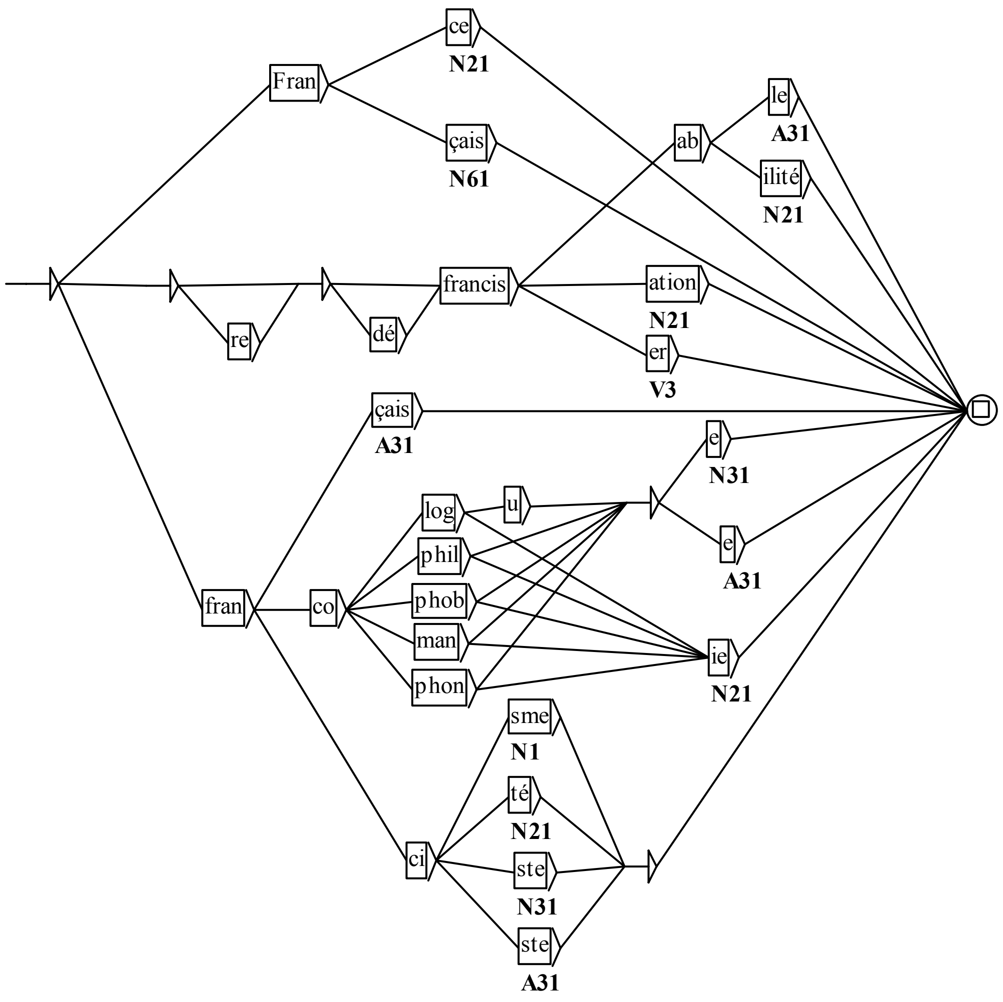
DELAF dictionaries are stored in minimal deterministic Finite State Transducers.

12. Lexical FSTs

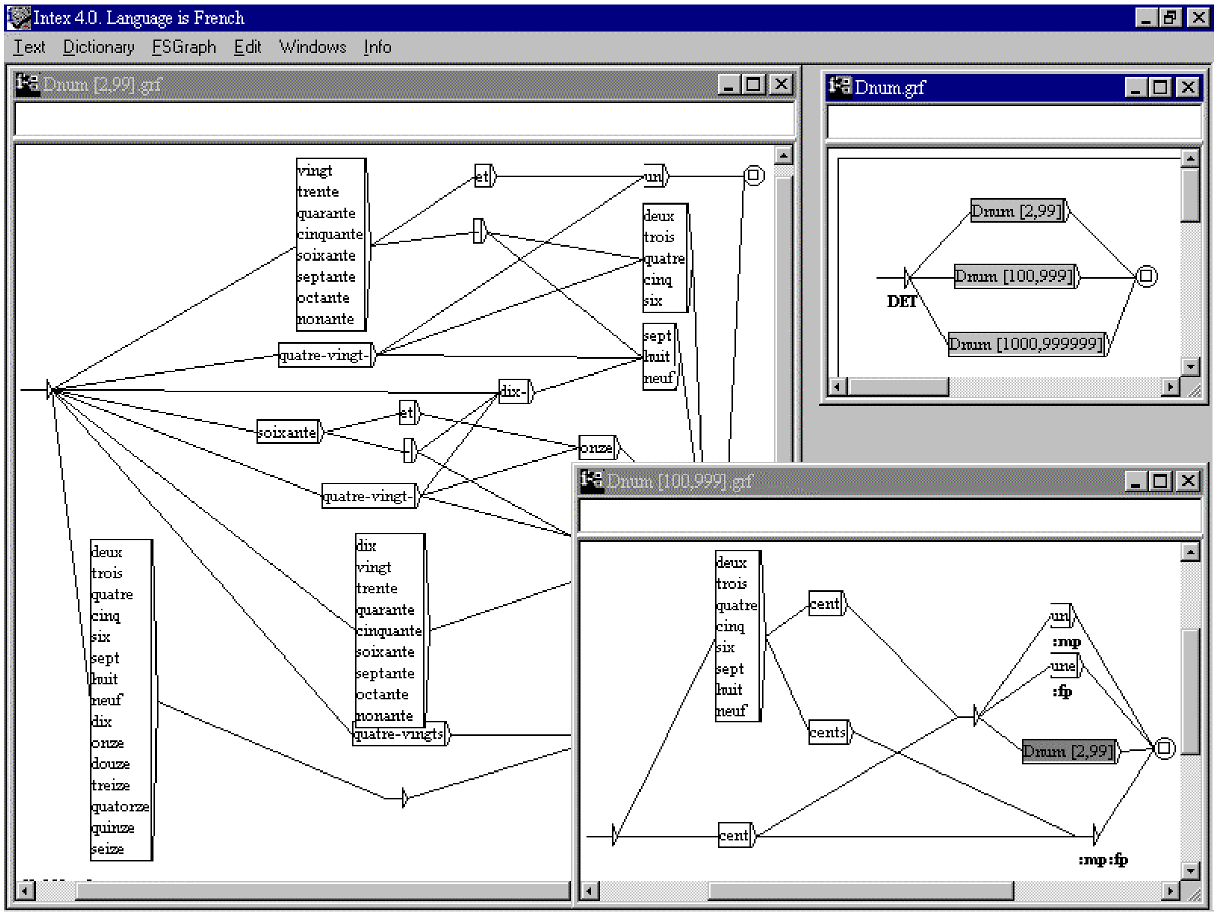
Generally, FSTs must be used to represent infinite sets of lexical entries (e.g. numerical determiners); they are used as well to put together members of a 'natural' family, such as derived forms of a lemma, orthographic variants, synonymous expressions, etc. For instance, here is a FST used to tag the orthographic variants of the noun *tsar* in French:



Here is a FST that represents all the derived forms of the noun *France*:



The following FST **Dnum** identifies and tag all the numerical determiners from 2 to 999999 written in French:



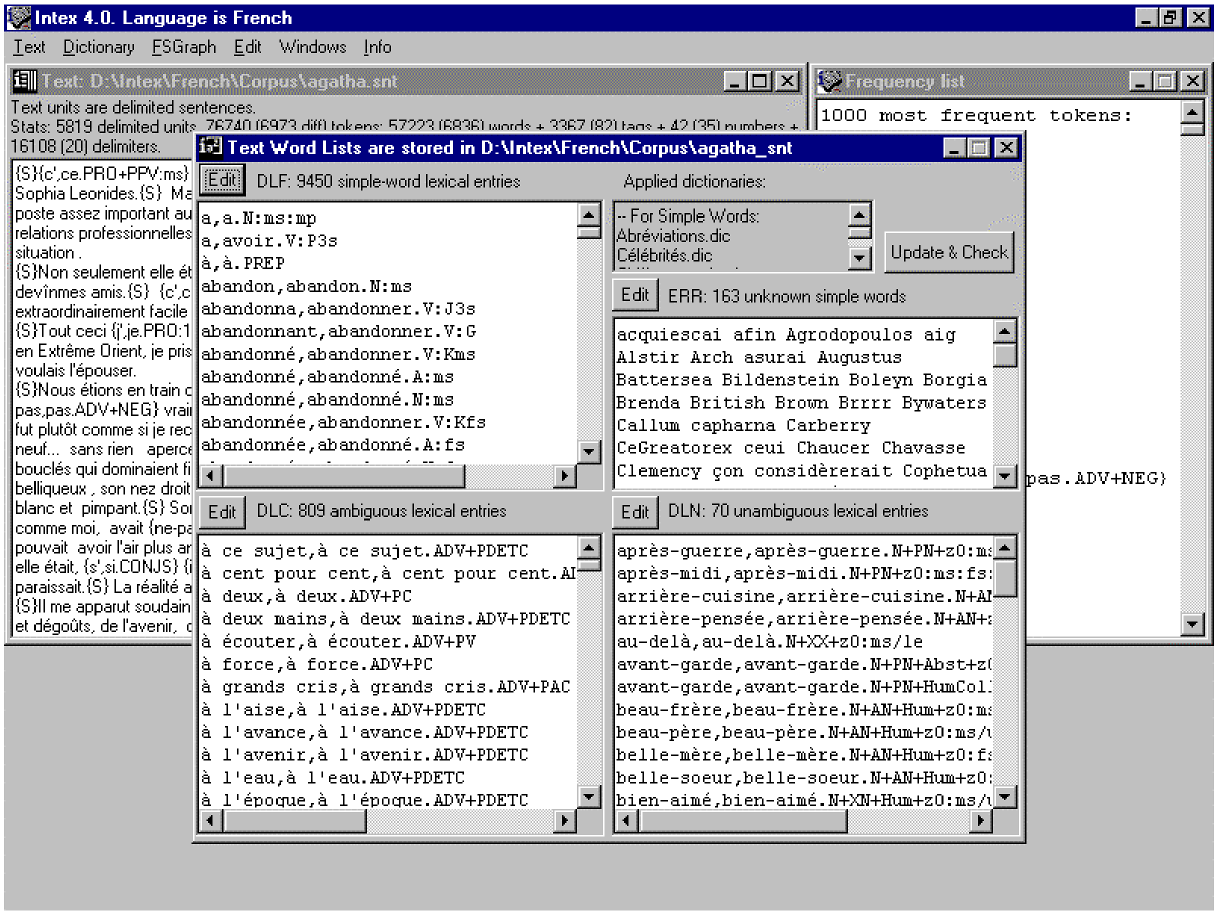
Since it is possible to embed FSTs in others, one can construct a large library of FSTs that will be re-used in other bigger projects. More than 3,000 FSTs have already been built for various descriptions of French:

- technical expressions (e.g. stock market, weather reports, medical statements, etc.),
- semi-frozen expressions (e.g. expressions of feelings),
- complements of dates, duration, addresses,
- etc.

13. Text dictionaries

The result of the application of the system dictionaries and FSTs is displayed below:

- the list of all the simple forms associated with their lemma and some linguistic data
- the list of all the simple forms that have not been found in the selected dictionaries
- the list of all 'ambiguous' compounds, such as 'red tape' (either a person, or a tape)
- the list of all 'unambiguous' compounds, such as 'grand parent'


Intex 4.0. Language is French
Text Dictionary FSGraph Edit Windows Info
Text: D:\Intex\French\Corpus\agatha.snt
Frequency list
Text Word Lists are stored in D:\Intex\French\Corpus\agatha_snt
Edit
Applied dictionaries:
Update & Check

```
Text units are delimited sentences.
Stats: 5819 delimited units, 76740 (6973 diff) tokens: 57223 (6836) words + 3367 (82) tags + 42 (35) numbers +
16108 (20) delimiters.

{S}{c',ce.PRO+PPV:ms}
Sophia Leonides.{S} Ma
poste assez important au
relations professionnelles
situation .
{S}Non seulement elle ét
devînmes amis.{S} {c',c
extraordinairement facile
{S}Tout ceci {j',je.PRO:1
en Extrême Orient, je pris
voulais l'épouser.
{S}Nous étions en train c
pas,pas.ADV+NEG} vrai
fut plutôt comme si je rec
neuf...  sans rien  aperce
bouclés qui dominaient fi
belliqueux , son nez droit
blanc et  pimpant.{S} So
comme moi,  avait {ne-pa
pouvait  avoir l'air plus ar
elle était, {s',si.CONJS} {i
paraissait.{S} La réalité a
{S}Il me apparut soudain
et dégoûts, de l'avenir,  c
```

```
1000 most frequent tokens:

pas.ADV+NEG}
```

DLF: 9450 simple-word lexical entries

```
a,a.N:ms:mp
a,avoir.V:P3s
à,à.PREP
abandon,abandon.N:ms
abandonna,abandonner.V:J3s
abandonnant,abandonner.V:G
abandonné,abandonner.V:Kms
abandonné,abandonné.A:ms
abandonné,abandonné.N:ms
abandonnée,abandonner.V:Kfs
abandonnée,abandonné.A:fs
```

```
-- For Simple Words:
Abréviations.dic
Célébrités.dic
```

ERR: 163 unknown simple words

```
acquiescai afin Agrodopoulos aig
Alstir Arch asurai Augustus
Battersea Bildenstein Boleyn Borgia
Brenda British Brown Brrrr Bywaters
Callum capharna Carberry
CeGreatorex ceui Chaucer Chavasse
Clemency çon considèrerait Cophetua
```

DLC: 809 ambiguous lexical entries

```
à ce sujet,à ce sujet.ADV+PDETC
à cent pour cent,à cent pour cent.AI
à deux,à deux.ADV+PC
à deux mains,à deux mains.ADV+PDETC
à écouter,à écouter.ADV+PV
à force,à force.ADV+PC
à grands cris,à grands cris.ADV+PAC
à l'aise,à l'aise.ADV+PDETC
à l'avance,à l'avance.ADV+PDETC
à l'avenir,à l'avenir.ADV+PDETC
à l'eau,à l'eau.ADV+PDETC
à l'époque,à l'époque.ADV+PDETC
```

DLN: 70 unambiguous lexical entries

```
après-guerre,après-guerre.N+PN+z0:ms
après-midi,après-midi.N+PN+z0:ms:fs:
arrière-cuisine,arrière-cuisine.N+AN
arrière-pensée,arrière-pensée.N+AN+
au-delà,au-delà.N+XX+z0:ms/le
avant-garde,avant-garde.N+PN+Abst+z0
avant-garde,avant-garde.N+PN+HumCol
beau-frère,beau-frère.N+AN+Hum+z0:ms
beau-père,beau-père.N+AN+Hum+z0:ms/
belle-mère,belle-mère.N+AN+Hum+z0:fs
belle-soeur,belle-soeur.N+AN+Hum+z0:
bien-aimé,bien-aimé.N+XN+Hum+z0:ms/
```

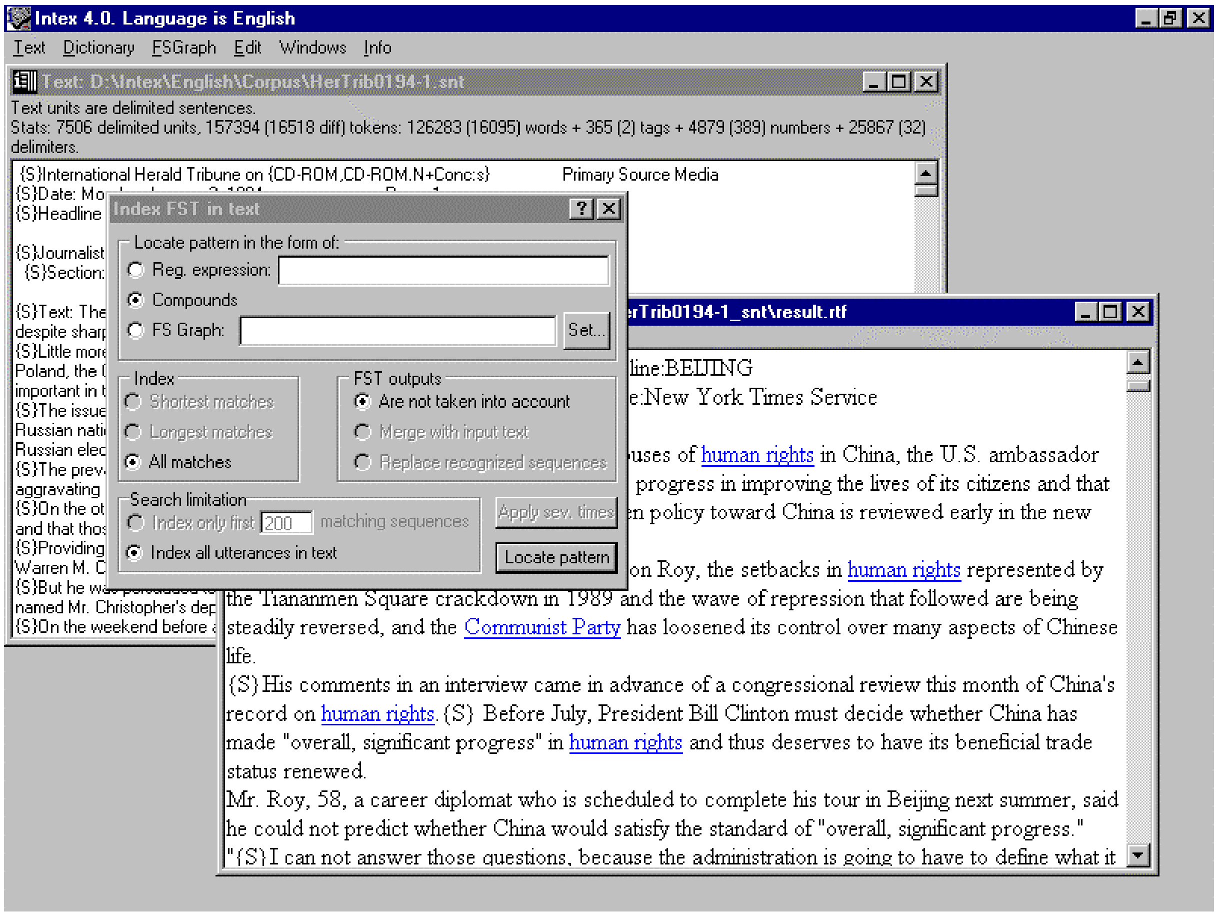
If a form is ambiguous, it appears on more than one line. For instance, *abandonné* is ambiguous: either the adjective masculine singular (**`A:ms`**), or the noun (**`N:ms`**), or the participle of the verb *abandonner* (**`V:Kms`**). The sequence *à ce sujet* (= *speaking of this*) can be an adverb; it is considered as an 'ambiguous' compound because the sequence is not obligatorily the adverb, as for instance in:

Je pense à ce sujet bien défini
(= I think of this well defined topic)

These four files can be edited, so that the dictionaries used by further processings are closely adapted to the text (one can easily get rid of ambiguities that do not occur in the specific text).

14. Highlight Compounds in the text (Text->Locate)

All the compounds that have been identified by the consultation of the selected dictionaries are indexed; it is then possible to highlight them in the text (via the *Display indexed sequences* panel).


Intex 4.0. Language is English
Text Dictionary FSGraph Edit Windows Info
Text: D:\Intex\English\Corpus\HerTrib0194-1.snt
Index FST in text
Set...
Apply sev. times
Locate pattern
rTrib0194-1_snt\result.rtf

Text units are delimited sentences.
Stats: 7506 delimited units, 157394 (16518 diff) tokens: 126283 (16095) words + 365 (2) tags + 4879 (389) numbers + 25867 (32) delimiters.

{S}International Herald Tribune on {CD-ROM,CD-ROM.N+Conc:s} Primary Source Media
{S}Date: Mo
{S}Headline
{S}Journalist
{S}Section:
{S}Text: The
despite sharp
{S}Little more
Poland, the (
important in t
{S}The issue
Russian nati
Russian elec
{S}The prev
aggravating
{S}On the ot
and that thos
{S}Providing
Warren M. C
{S}But he wa
named Mr. Christopher's dep
{S}On the weekend before

Locate pattern in the form of:
Reg. expression: [✓ Compounds] FS Graph:

Index: Shortest matches / Longest matches / [✓ All matches]
FST outputs: [✓ Are not taken into account] / Merge with input text / Replace recognized sequences
Search limitation: Index only first 200 matching sequences / [✓ Index all utterances in text]

line:BEIJING
e:New York Times Service
uses of human rights in China, the U.S. ambassador
progress in improving the lives of its citizens and that
n policy toward China is reviewed early in the new
on Roy, the setbacks in human rights represented by
the Tiananmen Square crackdown in 1989 and the wave of repression that followed are being steadily reversed, and the Communist Party has loosened its control over many aspects of Chinese life.
{S} His comments in an interview came in advance of a congressional review this month of China's record on human rights. {S} Before July, President Bill Clinton must decide whether China has made "overall, significant progress" in human rights and thus deserves to have its beneficial trade status renewed.
Mr. Roy, 58, a career diplomat who is scheduled to complete his tour in Beijing next summer, said he could not predict whether China would satisfy the standard of "overall, significant progress."
"{S}I can not answer those questions, because the administration is going to have to define what it

Generally, the indexation of compound nouns produces a very precious corpus to be used by information retrieval systems, because compound nouns are almost never ambiguous (compared to simple nouns).

15. Locate a regular expression in the text (Text->Locate)

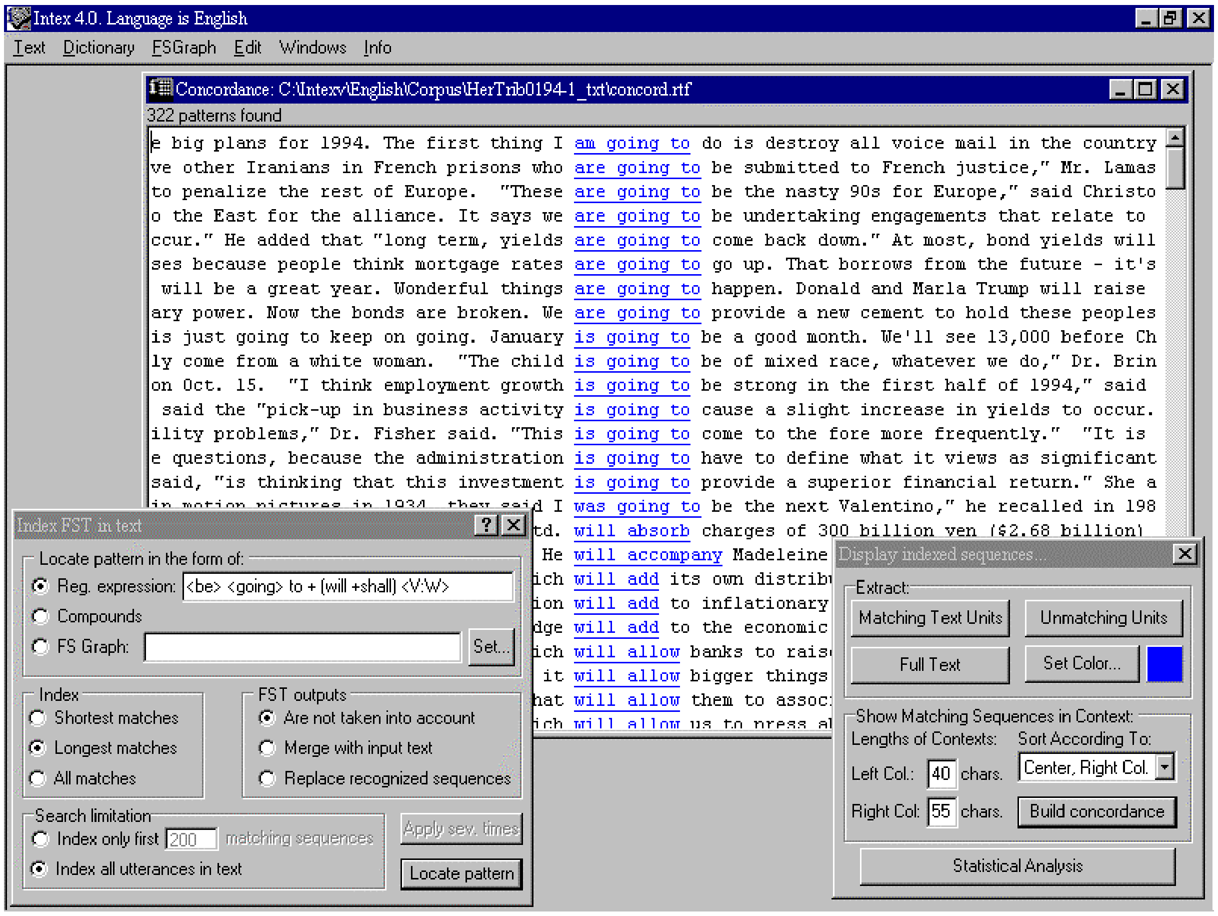
One can index all the sequences of the text that match a particular regular expression. For instance:

`<be> going to + (will + shall ) <V:W>`

`<be>` stands for any inflected form associated with the lemma *be* (that is, any conjugated form of the verb *to be*); `<V:W>` stands for any form associated with the category **V** and the inflectional code **W** (that is, any verb in the infinitive). Any code present in any of the dictionaries is instanteneously useable, for instance:

`<N+Hum>`	Noun, associated with the code Hum (Human)
`<N-Hum>`	Noun, not Human
`<N:fp>`	feminine plural Noun
`<!PREP>`	any word that is not a Preposition

The resulting index can be displayed in the form of a concordance:


Intex 4.0. Language is English
Text Dictionary FSGraph Edit Windows Info
Concordance: C:\Intexv\English\Corpus\HerTrib0194-1_txt\concord.rtf

322 patterns found

```
e big plans for 1994. The first thing I am going to do is destroy all voice mail in the country
ve other Iranians in French prisons who are going to be submitted to French justice," Mr. Lamas
to penalize the rest of Europe.  "These are going to be the nasty 90s for Europe," said Christo
o the East for the alliance. It says we are going to be undertaking engagements that relate to
ccur." He added that "long term, yields are going to come back down." At most, bond yields will
ses because people think mortgage rates are going to go up. That borrows from the future - it's
 will be a great year. Wonderful things are going to happen. Donald and Marla Trump will raise
ary power. Now the bonds are broken. We are going to provide a new cement to hold these peoples
is just going to keep on going. January is going to be a good month. We'll see 13,000 before Ch
ly come from a white woman.  "The child is going to be of mixed race, whatever we do," Dr. Brin
on Oct. 15.  "I think employment growth is going to be strong in the first half of 1994," said
 said the "pick-up in business activity is going to cause a slight increase in yields to occur.
ility problems," Dr. Fisher said. "This is going to come to the fore more frequently."  "It is
e questions, because the administration is going to have to define what it views as significant
said, "is thinking that this investment is going to provide a superior financial return." She a
in motion pictures in 1934, they said I was going to be the next Valentino," he recalled in 198
                                    td. will absorb charges of 300 billion yen ($2.68 billion)
                                     He will accompany Madeleine
                                    ich will add its own distrib
                                    ion will add to inflationary
                                    dge will add to the economic
                                    ich will allow banks to rais
                                     it will allow bigger things
                                    hat will allow them to assoc
                                    ich will allow us to press al
```

Locate pattern in the form of:
- [✓ Reg. expression:] <be> <going> to + (will +shall) <V:W>
- Compounds
- FS Graph: (Set...)

Index:
- Shortest matches
- [✓ Longest matches]
- All matches

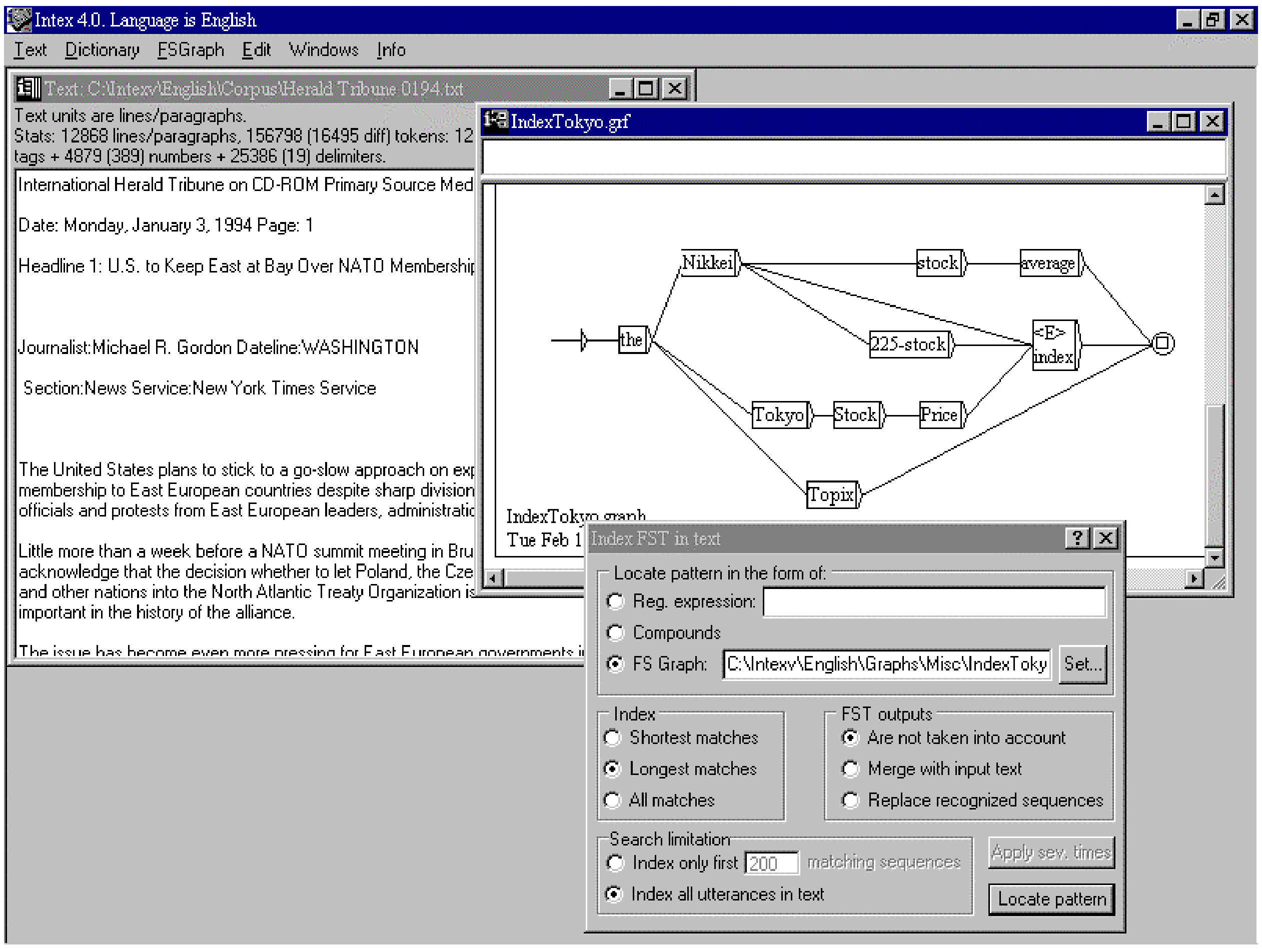
FST outputs:
- [✓ Are not taken into account]
- Merge with input text
- Replace recognized sequences

Search limitation:
- Index only first 200 matching sequences
- [✓ Index all utterances in text]

Show Matching Sequences in Context:
Lengths of Contexts: Left Col.: 40 chars. Right Col: 55 chars.
Sort According To: Center, Right Col.

16. Index a FST in the text (**Text->Locate**)

One can index all the sequences of the text that match a particular FST.



Text units are lines/paragraphs.
Stats: 12868 lines/paragraphs, 156798 (16495 diff) tokens: 12
tags + 4879 (389) numbers + 25386 (19) delimiters.

International Herald Tribune on CD-ROM Primary Source Med

Date: Monday, January 3, 1994 Page: 1

Headline 1: U.S. to Keep East at Bay Over NATO Membershi

Journalist:Michael R. Gordon Dateline:WASHINGTON

Section:News Service:New York Times Service

The United States plans to stick to a go-slow approach on exp
membership to East European countries despite sharp division
officials and protests from East European leaders, administratic

Little more than a week before a NATO summit meeting in Bru
acknowledge that the decision whether to let Poland, the Cze
and other nations into the North Atlantic Treaty Organization is
important in the history of the alliance.

The issue has become even more pressing for East European governments i

Locate pattern in the form of:
Reg. expression:
Compounds
[✓ FS Graph:] C:\Intexv\English\Graphs\Misc\IndexToky

Index
Shortest matches
[✓ Longest matches]
All matches

FST outputs
[✓ Are not taken into account]
Merge with input text
Replace recognized sequences

Search limitation
Index only first 200 matching sequences
[✓ Index all utterances in text]

The user can choose to index:

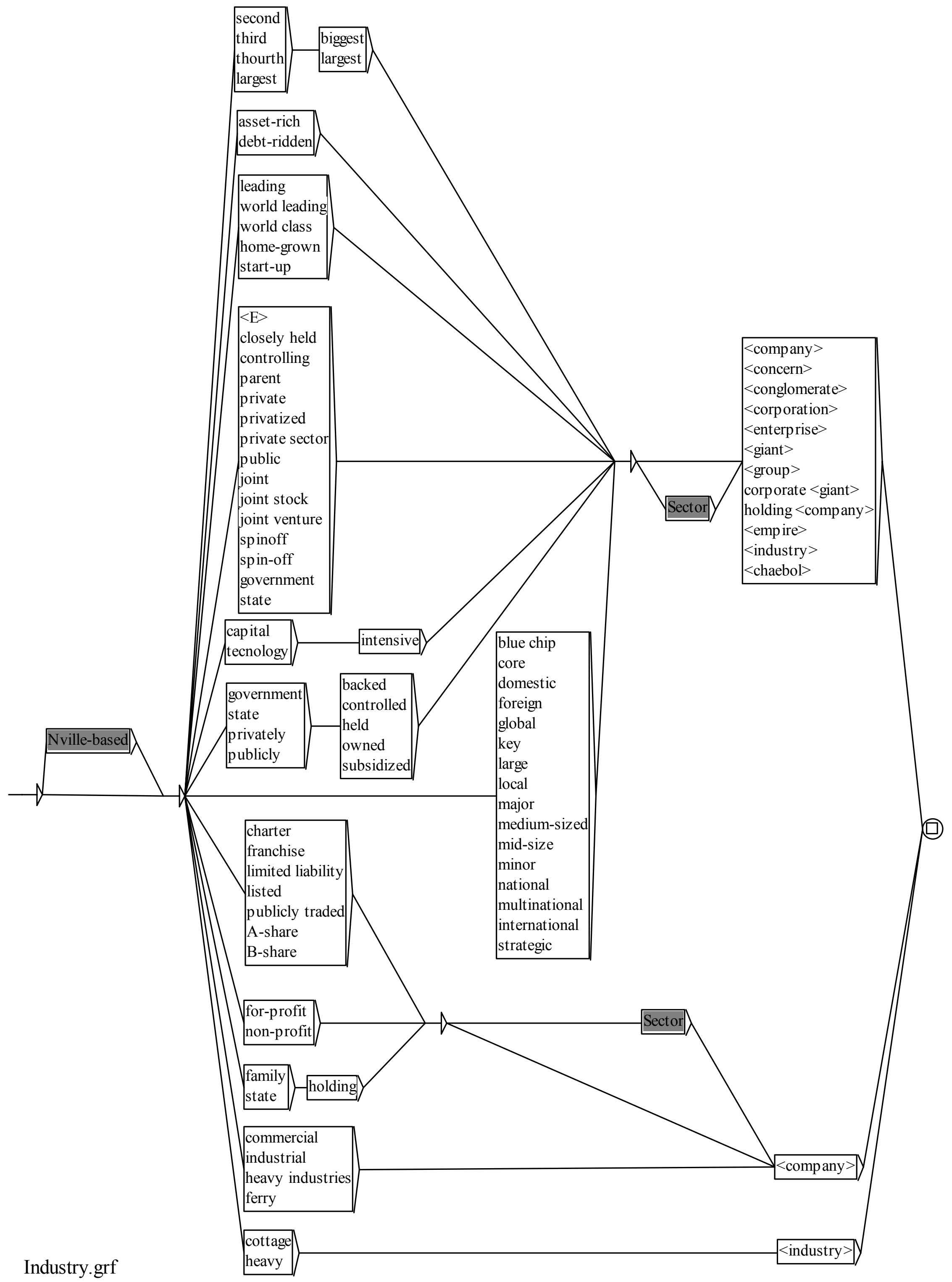
- shortest matches;
- longuest matches;
- all matches.

FST outputs can:

- be ignored;
- be merged into the text;
- replace recognized sequences.

With the two last options, the FST can be reapplied several times to the text, for instance until no modification has been performed.

One of the grammars used to locate semi-frozen noun phrases that represent companies; when applied to newspapers, the resulting index can be used by information retrieval systems:

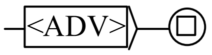


Industry.grf

17. Various Text Transformations

FSTs can also be applied to texts in order to modify the text itself.

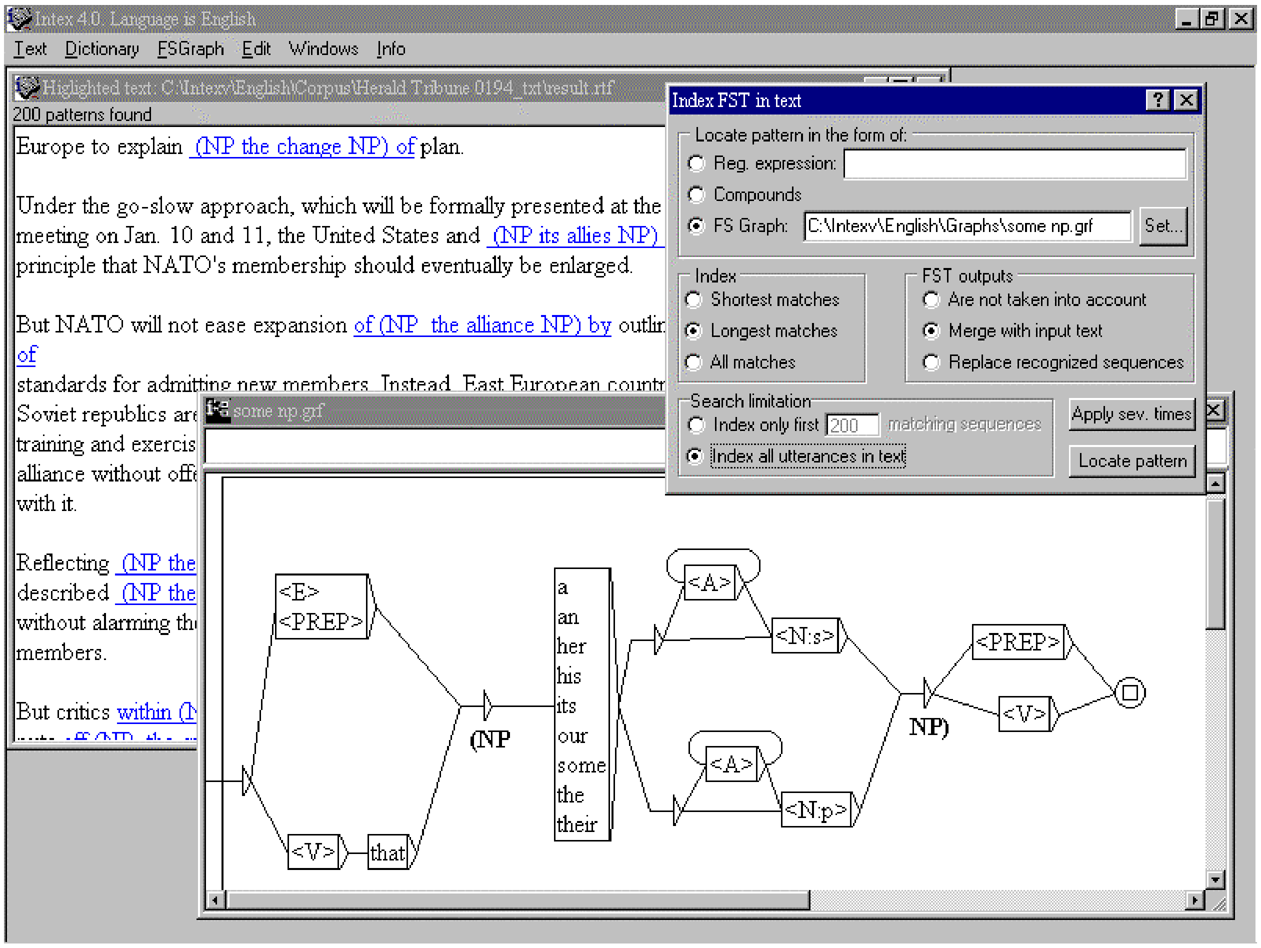
For instance, the following FST, if used in REPLACE mode, could be used to remove adverbs from the text (useful for information retrieval systems):



In the same manner, one could design FSTs that would delete expressions like:

> *I think that, In the same manner, as far as I am concerned,*
> *officials said that, he argued that, it seems doubtfull that,* etc.

The following FST can be used in MERGE mode to insert parentheses around some noun phrases:


Intex 4.0. Language is English
Text Dictionary FSGraph Edit Windows Info
Higlighted text: C:\Intexv\English\Corpus\Herald Tribune 0194_txt\result.rtf
some np.grf
<E>
<PREP>
<V>
that
(NP
a
an
her
his
its
our
some
the
their
<A>
<N:s>
<N:p>
NP)
<PREP>
<V>

200 patterns found

Europe to explain (NP the change NP) of plan.

Under the go-slow approach, which will be formally presented at the meeting on Jan. 10 and 11, the United States and (NP its allies NP)
principle that NATO's membership should eventually be enlarged.

But NATO will not ease expansion of (NP the alliance NP) by outli
of
standards for admitting new members. Instead, East European countr
Soviet republics are
training and exercis
alliance without offe
with it.

Reflecting (NP the
described (NP the
without alarming th
members.

But critics within (N

Index FST in text

Locate pattern in the form of:
- Reg. expression:
- Compounds
- [✓ FS Graph:] C:\Intexv\English\Graphs\some np.grf

Index
- Shortest matches
- [✓ Longest matches]
- All matches

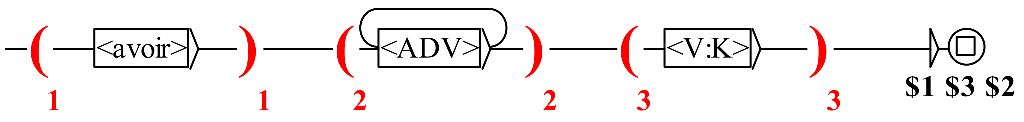
FST outputs
- Are not taken into account
- [✓ Merge with input text]
- Replace recognized sequences

Search limitation
- Index only first 200 matching sequences
- [✓ Index all utterances in text]

Enhanced FSTs

In INTEX FSTs, outputs and inputs are *synchronized*, that is, the input sequences that match are indexed in the same time as the corresponding output. That feature is essential to the application in MERGE mode (see for instance how parentheses are inserted by the previous FST), or by disambiguation FSTs, where lexical constraints have to be applied to the correct matching forms.

INTEX users are allowed to purposely modify the conditions of this synchronization, by using internal variables to store parts of the matching input sequence. This feature is the same as in the UNIX tool **`sed`**. Here is one example. The following FST:



is used to extract the adverbs that may be inserted between the auxiliary verb *avoir* and the following past participle. For instance, when the FST is applied to the following text:

`Luc m'avait souvent amusé`

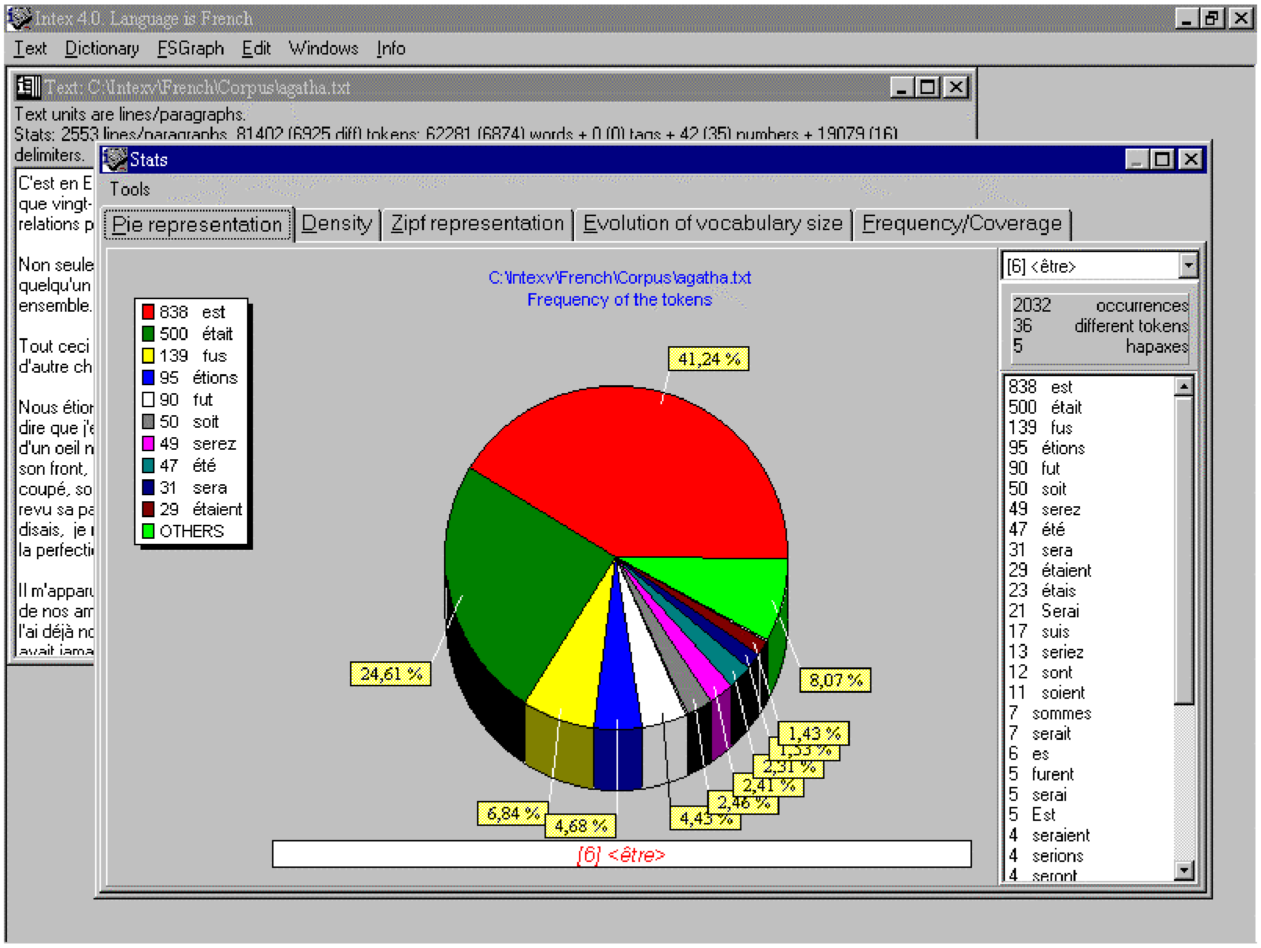
the sequence *avait* is stored in memory slot $1; the sequence souvent is stored in memory slot $2, and the sequence amusé is stored in memory slot $3. Performing the replacement produces the following result:

`Luc m'avait amusé souvent`

Look at the FST that identifies some French negations in the Norm directory (on the form *ne <V> Neg*).

18. Statistical Analyses

Any index (the index of the compounds, of the sequences that match a regular expression, or a FST) can be studied with a number of statistical measurements. Here for instance, the frequencies of the forms matching the regular expression **`<être>`** (all the conjugated forms of the verb *être*) are represented in a pie:



Text: C:\Intexv\French\Corpus\agatha.txt

Text units are lines/paragraphs.
Stats: 2553 lines/paragraphs. 81402 (6925 diff) tokens: 62281 (6874) words + 0 (0) tags + 42 (35) numbers + 19079 (16) delimiters.

C'est en E
que vingt-
relations p

Non seule
quelqu'un
ensemble.

Tout ceci
d'autre ch

Nous étior
dire que j'e
d'un oeil n
son front,
coupé, so
revu sa pa
disais, je i
la perfecti

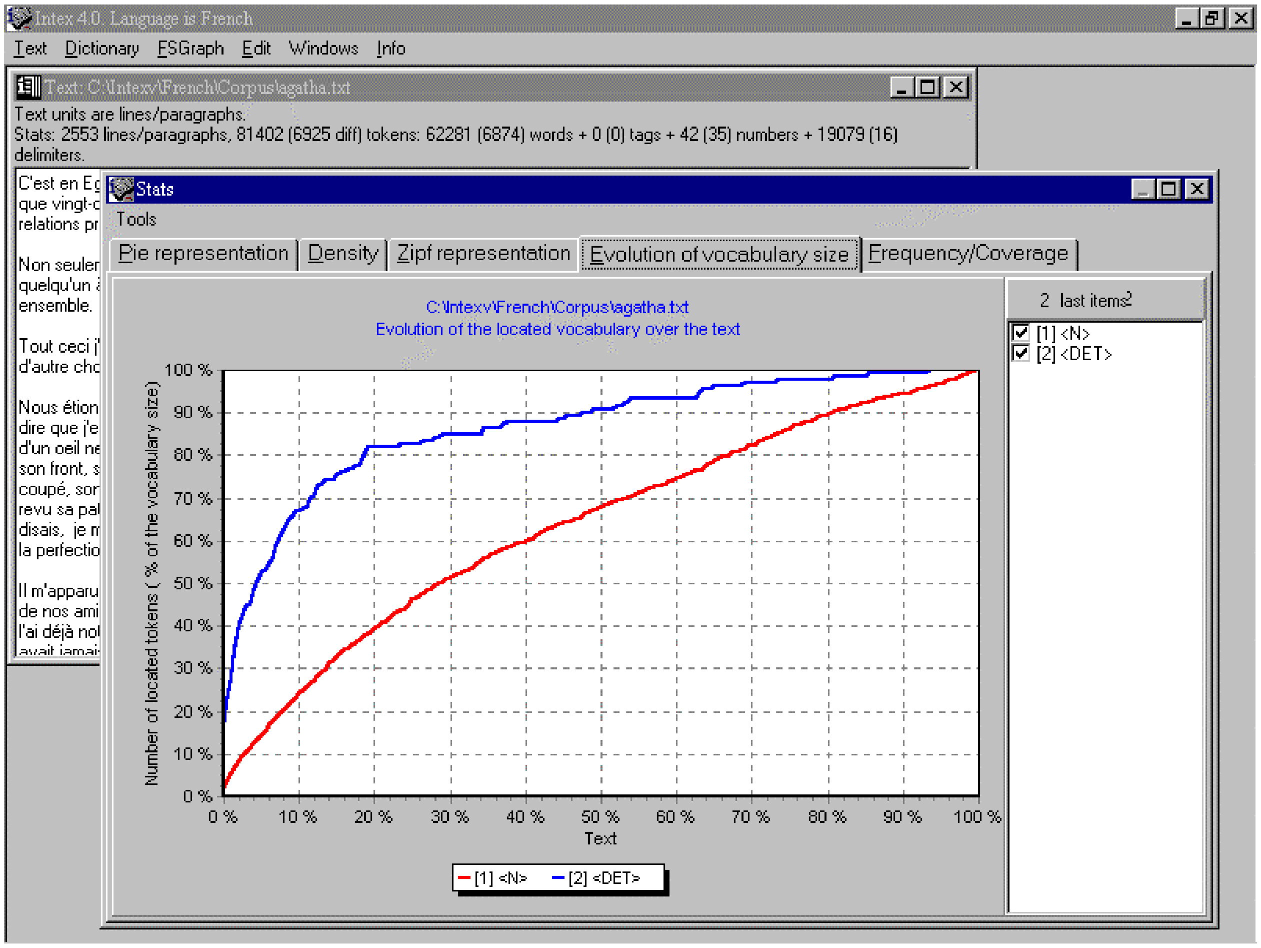
Il m'appari
de nos an
l'ai déjà nc

🟥	838	est
🟩	500	était
🟨	139	fus
🟦	95	étions
⬜	90	fut
▨	50	soit
🟪	49	serez
▨	47	été
▨	31	sera
▨	29	étaient
🟩		OTHERS

2032	occurrences
36	different tokens
5	hapaxes

838	est
500	était
139	fus
95	étions
90	fut
50	soit
49	serez
47	été
31	sera
29	étaient
23	étais
21	Serai
17	suis
13	seriez
12	sont
11	soient
7	sommes
7	serait
6	es
5	furent
5	serai
5	Est
4	seraient
4	serions
4	seront

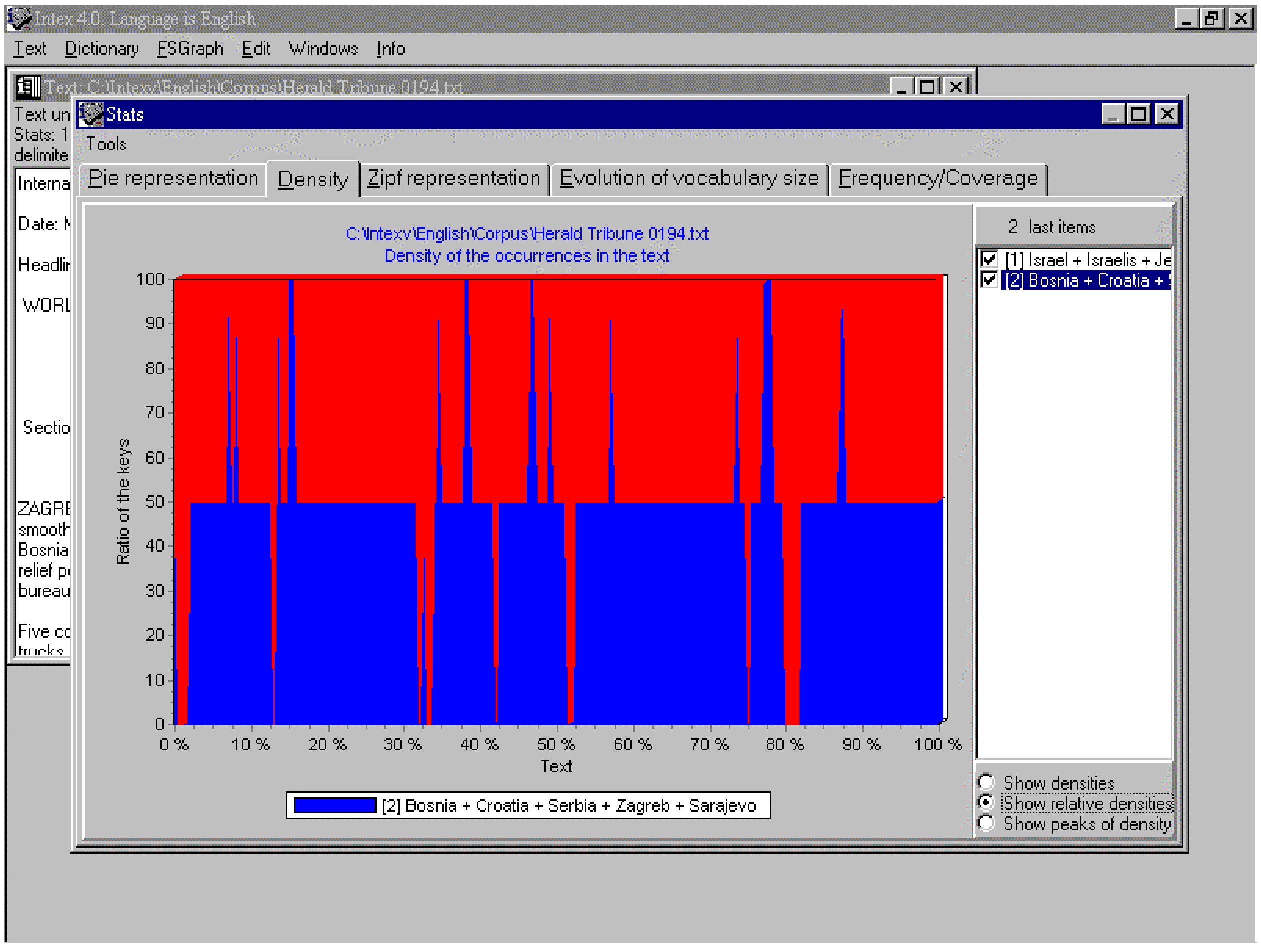
Below, we study the evolution of the number of Nouns (**`<N>`**) and Determiners (**`<DET>`**). The proportion of determiners grows much faster than the one of the nouns (which is almost linear).


Intex 4.0. Language is French
Text Dictionary FSGraph Edit Windows Info
Text: C:\Intexv\French\Corpus\agatha.txt

Text units are lines/paragraphs.
Stats: 2553 lines/paragraphs, 81402 (6925 diff) tokens: 62281 (6874) words + 0 (0) tags + 42 (35) numbers + 19079 (16) delimiters.

Stats
Tools
Pie representation
Density
Zipf representation
Evolution of vocabulary size
Frequency/Coverage
C:\Intexv\French\Corpus\agatha.txt
Evolution of the located vocabulary over the text
Number of located tokens (% of the vocabulary size)
Text
0 %
10 %
20 %
30 %
40 %
50 %
60 %
70 %
80 %
90 %
100 %
[1] <N>
[2] <DET>
2 last items
[1] <N>
[2] <DET>

Below we study the density of two regular expressions in the *Herald Tribune*, january 1994:

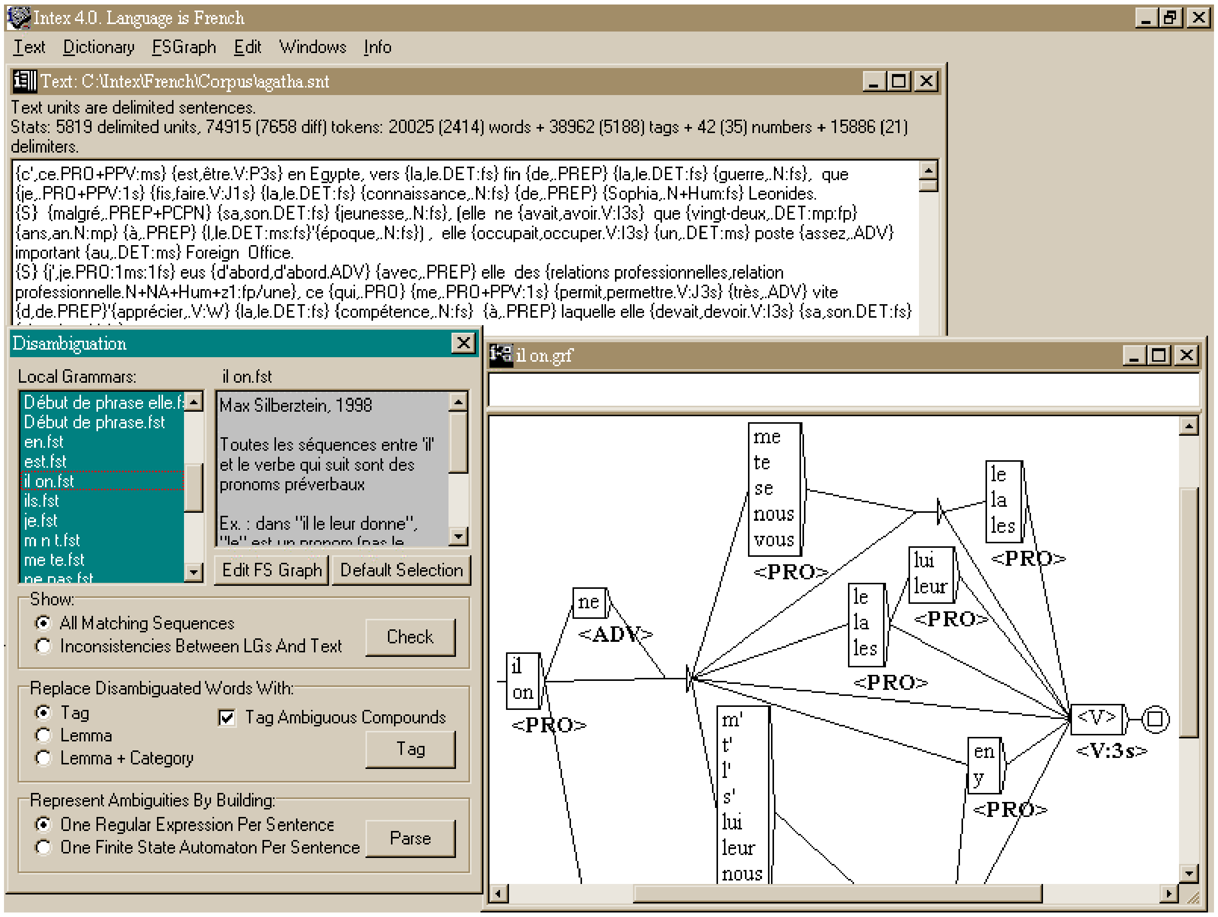

Intex 4.0. Language is English
Text Dictionary FSGraph Edit Windows Info
Text: C:\Intexv\English\Corpus\Herald Tribune 0194.txt
Stats
Tools
Pie representation
Density
Zipf representation
Evolution of vocabulary size
Frequency/Coverage
C:\Intexv\English\Corpus\Herald Tribune 0194.txt
Density of the occurrences in the text
Ratio of the keys
Text
[2] Bosnia + Croatia + Serbia + Zagreb + Sarajevo
2 last items
[1] Israel + Israelis + Je
[2] Bosnia + Croatia +
Show densities
Show relative densities
Show peaks of density

Peaks correspond to an important local density of the forms that match the indexed regular expression.

Various other statistical measures are available.

19. Disambiguation with Local Grammars

Local grammars are FSTs that recognize text sequences (e.g. *il le la donne*), then applies corresponding lexical constraints (e.g. `<PRO> <PRO> <PRO> <V:3s>`) to remove lexical hypotheses.


Intex 4.0. Language is French
Text Dictionary FSGraph Edit Windows Info
Text: C:\Intex\French\Corpus\agatha.snt
Disambiguation
il on.grf
me te se nous vous
<PRO>
le la les
<PRO>
lui leur
<PRO>
ne
<ADV>
le la les
<PRO>
il on
<PRO>
m' t' l' s' lui leur nous
en y
<PRO>
<V>
<V:3s>

Text units are delimited sentences.
Stats: 5819 delimited units, 74915 (7658 diff) tokens: 20025 (2414) words + 38962 (5188) tags + 42 (35) numbers + 15886 (21) delimiters.

```
{c',ce.PRO+PPV:ms} {est,être.V:P3s} en Egypte, vers {la,le.DET:fs} fin {de,.PREP} {la,le.DET:fs} {guerre,.N:fs},  que
{je,.PRO+PPV:1s} {fis,faire.V:J1s} {la,le.DET:fs} {connaissance,.N:fs} {de,.PREP} {Sophia,.N+Hum:fs} Leonides.
{S}  {malgré,.PREP+PCPN} {sa,son.DET:fs} {jeunesse,.N:fs}, (elle  ne {avait,avoir.V:I3s}  que {vingt-deux,.DET:mp:fp}
{ans,an.N:mp} {à,.PREP} {l,le.DET:ms:fs}'{époque,.N:fs}) ,  elle {occupait,occuper.V:I3s} {un,.DET:ms} poste {assez,.ADV}
important {au,.DET:ms} Foreign  Office.
{S} {j',je.PRO:1ms:1fs} eus {d'abord,d'abord.ADV} {avec,.PREP} elle  des {relations professionnelles,relation
professionnelle.N+NA+Hum+z1:fp/une}, ce {qui,.PRO} {me,.PRO+PPV:1s} {permit,permettre.V:J3s} {très,.ADV} vite
{d,de.PREP}'{apprécier,.V:W} {la,le.DET:fs} {compétence,.N:fs}  {à,.PREP} laquelle elle {devait,devoir.V:I3s} {sa,son.DET:fs}
```

Local Grammars:
Début de phrase elle.f
Début de phrase.fst
en.fst
est.fst
il on.fst
ils.fst
je.fst
m n t.fst
me te.fst
ne pas.fst

il on.fst
Max Silberztein, 1998

Toutes les séquences entre 'il' et le verbe qui suit sont des pronoms préverbaux

Ex. : dans "il le leur donne",

Edit FS Graph | Default Selection

Show:
[✓ All Matching Sequences]
Inconsistencies Between LGs And Text
Check

Replace Disambiguated Words With:
[✓ Tag]
Lemma
Lemma + Category
[✓ Tag Ambiguous Compounds]
Tag

Represent Ambiguities By Building:
[✓ One Regular Expression Per Sentence]
One Finite State Automaton Per Sentence
Parse

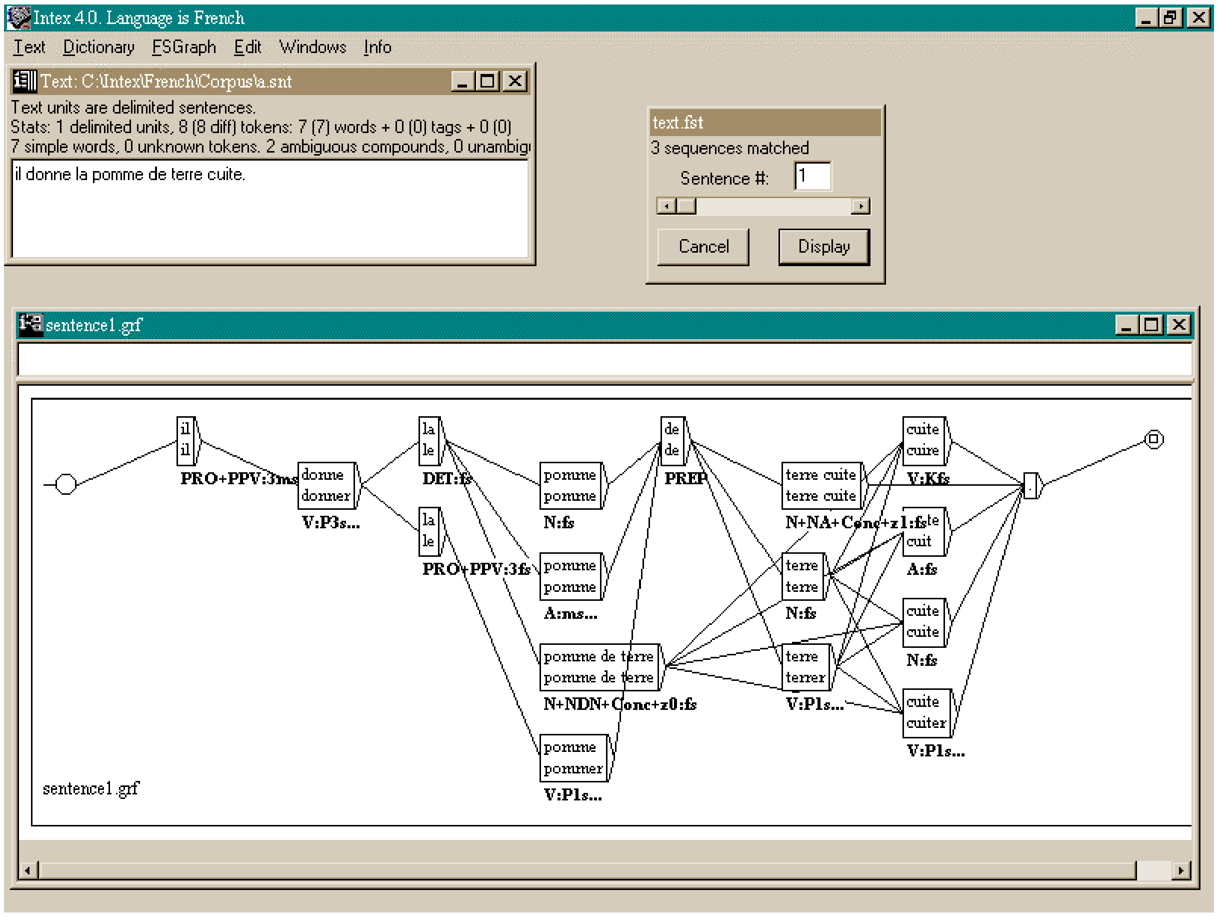
The result of the application of one or more disambiguating FSTs can be displayed in several ways. INTEX can:

- index all matching sequences (i.e. to study the coverage of the local grammars)
- index all inconsistencies between selected local grammars and the text (to locate errors in the grammars, or agreement errors in the text)
- tag the text, i.e. replace all disambiguated forms by the corresponding lexical entry;
- lemmatize the text, i.e. replace all disambiguated forms by the corresponding lemma;
- build a regular expression that represents all the ambiguities of the text;
- build and display the text in the form of a FST.

Internally, each text sentence is represented by a FST. INTEX can display it. For instance, the text:

```
il donne la pomme de terre cuite
```

is represented below with all the ambiguities that remain after having applied selected local grammars. Notice how the two ambiguous compound nouns *pomme de terre* (= *potato*) and *terre cuite* (= *clay*) are represented in the FST.



Text units are delimited sentences.
Stats: 1 delimited units, 8 (8 diff) tokens: 7 (7) words + 0 (0) tags + 0 (0)
7 simple words, 0 unknown tokens. 2 ambiguous compounds, 0 unambigi

il donne la pomme de terre cuite.

3 sequences matched
Sentence #: 1

Conclusion

Research Tool

Over 30 research centers are presently using INTEX as a research tool in various domains: computational linguistics, corpus-based linguistics, information retrieval, terminology, literature studies, teaching of a second-language. At the First INTEX users' Workhop (March 1993, Université Paris 7), 60 participants came from 10 countries; the 15 papers demonstrated the various applications of the system.

I would like to emphasize the fact that thanks to INTEX, about 60 researchers from all over Europe have started to work within a shared framework, using the same tools and the same methodology to construct large coverage descriptions of a dozen natural languages. I believe this situation is unique in the field of Linguistics.

Large-coverage INTEX descriptions have already been built for Bulgarian (contact Prof. Elena Paskaleva, University of Sofia), English, French and Spanish (Contact Prof. Maurice Gross, Université Paris 7), German (Prof. Franz Günthner, University Maximilian, Münich), Greek (Prof. Ana Symeonides, University of Salonique), Italian (Prof. Annibale Elia, University of Salerne), Polish (Prof. Zygmunt Vetulani), Portuguese (Prof. Elisabete Ranchodd, University of Lisbon), Russian (Alex Kolesov, Academy of Sciences, Moscow). INTEX Dictionaries are being constructed for Korean (Jeesun Nam, University of Seoul) and Old French (Prof. Hava Bat-Zeev, University of Tel Aviv).

Teaching Tool for Linguistics, Corpus Linguistics and Computational Linguistics

INTEX is a great tool to teach Corpus Linguistics and Computational Linguistics.

Describing the linguistic data included in INTEX (Dictionaries for Simple Words, Description of the Morphology, Dictionaries for Compound words, Dictionaries for frozen expressions, Local grammars for disambiguating texts, etc.) as well as the technology used to handle this data (Finite State Automata and Transducers) corresponds to a full year course (50 Hours) for Masters students. There are a plethora of projects that remain to be proposed to students, as the description of Natural Languages is not nearly complete!

Teaching Tool for Second Languages

INTEX is also used to teach French as a second-language; it allows teachers to ask students to locate morpho-syntactic patterns in wide texts (such as 5 years of the newspaper *Le Monde*), to build local grammars for semi-frozen expressions (e.g. how to express a date in French) and to ‘play’ (edit, correct, apply and test) with some linguistic descriptions.