



LABORATOIRE D'INFORMATIQUE  
GASPARD-MONGE

Sous la co-tutelle de :  
CNRS  
ÉCOLE DES PONTS PARISTECH  
ESIEE PARIS  
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

# Unitex Chinois

ZHANG Yue ,WANG Qi ,ZHANG Jiangting et ZHOU Shenyi  
2018.05

# VOICI LES DEUX LOGICIELS QUE NOUS UTILISONS

- Unitex
- Un logiciel de traitement de langues. Il permet de traiter des corpus de textes et de les analyser pour en extraire des informations. Il est utilisé dans le domaine de la linguistique et de l'information.
- Ses sites :
  - <http://unitex.org/>
  - <http://unitex.org/~unitex/>

- BRAT
- Un logiciel de annotation de textes. Il permet de créer et d'éditer des annotations sur des textes. Il est utilisé dans le domaine de la linguistique et de l'information.
- Ses sites :
  - <http://brat.nlplab.org/>

# L'environnement Unitex

- Environnement open-source développé principalement par Sébastien Paumier (écrit en C et interface en Java)
- Multiplateforme : Windows, Mac, Linux
- Multilingue : allemand, anglais, arabe, coréen, finnois, français, espagnol, géorgien ancien, grec, grec ancien, italien, norvégien bokmal, norvégien nynorsk, polonais, portugais du Portugal, portugais du Brésil, russe, serbo-croate latin, serbo-croate cyrillique, thaï, **et bientôt le chinois.** (une langue à la fois)
- Unicode 3.0 (UTF16, UTF8)

## Présentation du projet

■ Notre projet consiste à adapter et développer l'environnement Unitex à la langue chinoise. Ce développement a commencé depuis 2 ans avec la traduction du site et du manuel et la création d'un corpus (libre de droit pour qu'il puisse être diffusé librement). L'année dernière les ressources ont été validées et enrichies. Cette année nous avons intégré les entités nommées dans le but d'enrichir les ressources mais aussi tester l'outil d'évaluation.

## Description de l'existant

- Ressources existantes : alphabet, dictionnaire, codage des parties du discours, explication du système des tons.
- Quelques grammaires locales.
- Corpus *libre de droits* représentatif (littérature, presse, sciences, etc...) de la langue chinoise.

# Enrichissement des ressources

- Amélioration du dictionnaire et création de grammaires locales
  - vérification et la fusion des dictionnaires
  - entités nommés

# Entités Nommées

Grammaires locales et graphes dictionnaires concernant les :

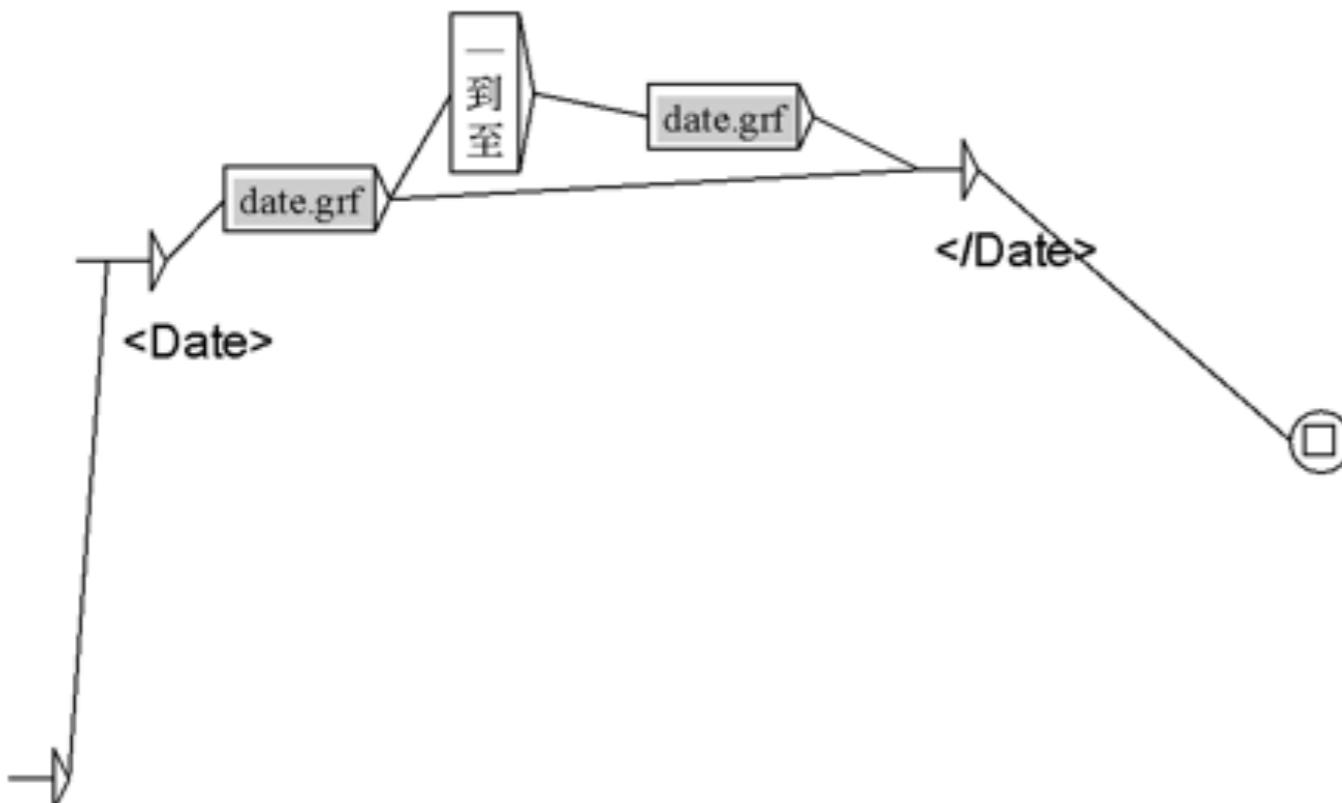
- **Dates(Expressions de durée)**
  
- **Personnes:** l'exemple de "ministre"

# Entités Nommées\_DATES

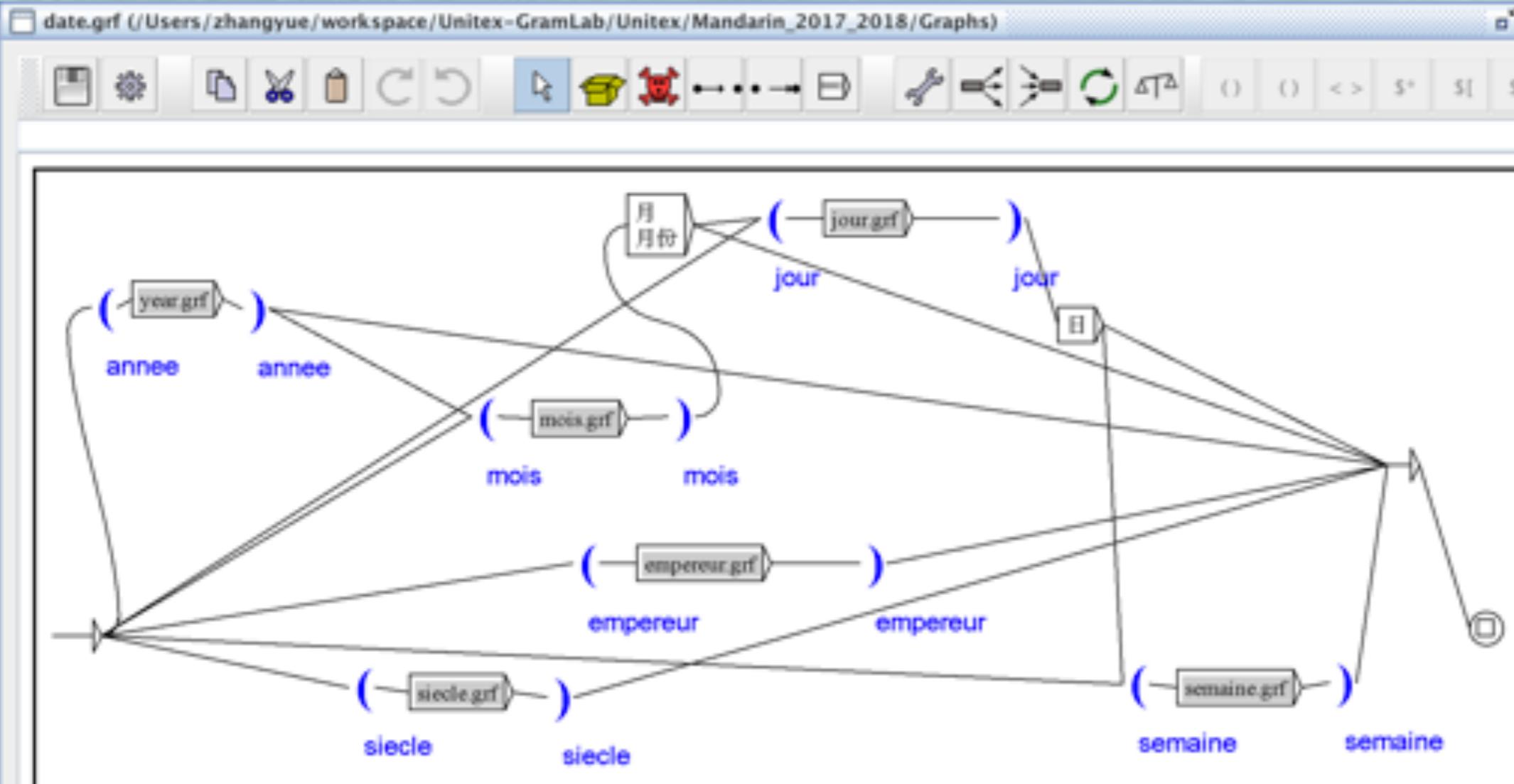
## Ce que l'on a fait:

- Etudier la grammaire fondamentale
- Analyser les méthodes d'expression du temps chinoises
- Concevoir les graphes et dictionnaires sur expressions de la date en fonction de la grammaire chinoise
- Chercher un article sur l'histoire de la Chine moderne
- Noter les expressions chinoises sur cet article avec Brat
- Modéliser et améliorer les graphes
- Comparer les résultats entre ce qui est traité par Brat et Unitex

# Graphe Totale



- Voici notre image générale: nous avons toutes nos expressions de date fondamental dans le diagramme *date.grf*.
- Dans cette graphe générale (*dateComplet.grf*), nous appelons deux fois *date.grf* pour exprimer une période de temps.



Dans cette graphe nous avons quelques expressions de date fondamentaux

date.grf

Il existe quatre types d'expressions de date: chaque sous-graphes est associée aux unités de temps chinoises nécessaires pour former l'expression de date courante (date.grf).

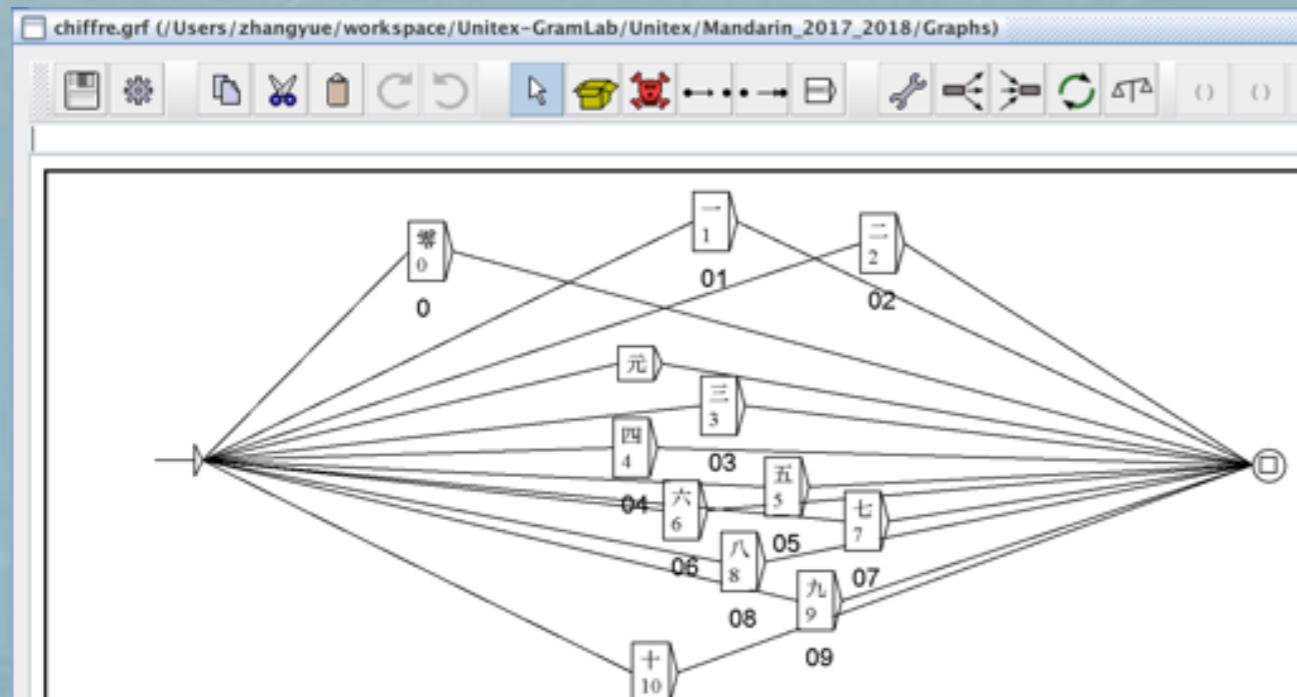
Nous décrivons ci-dessous ces quatre expressions fondamentale en détail.

La première forme est l'expression la plus fréquemment utilisée dans la vie quotidienne.

- ◆ Date simple utilisant année, mois, un jour du mois, un jour de la semaine
  - (chiffres arabes)
- ◆ Forme : xxxx**年** (année) xx**月**(mois) xx**日**(un jour du mois) **星期**x(un jour de la semaine)
- ◆ Exemple: la date : **1895年4月17日星期一**
  - ◆ égale le **lundi 17 avril 1895**

- (chiffres caractères chinoises)
- Exemple: la date : 一八九五年四月十七日星期一
  - égale le lundi 17 avril 1895

Dans la première expression, Nous avons utilisées cinq sous-graphes :  
**chiffre.grf, year.grf, mois.grf, jour.grf, semaine.grf**

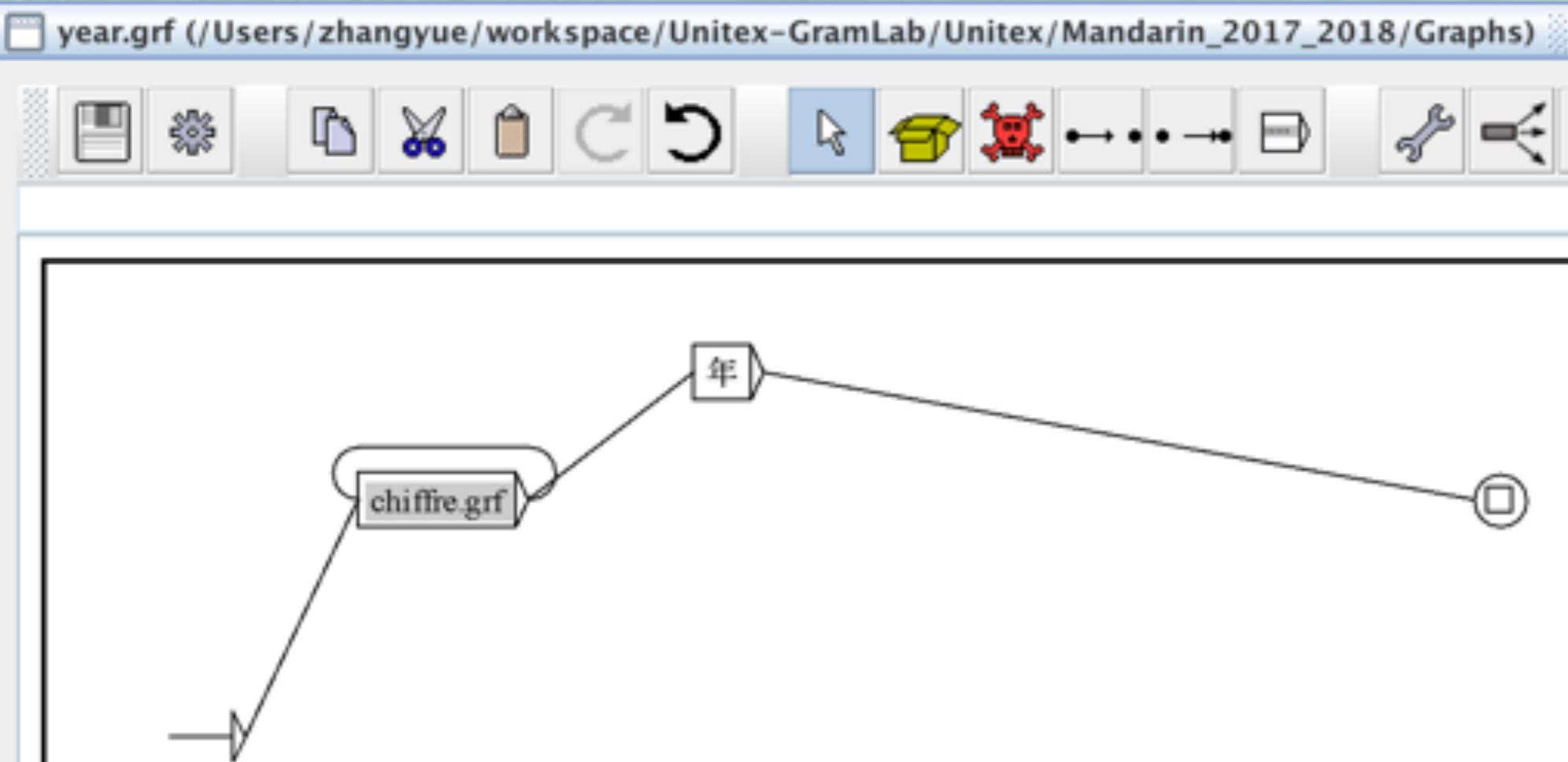


## 1.chiffre.grf

La Chine utilise deux méthodes quand il s'agit de chiffres, l'une se sert des chiffres arabes et l'autre des caractères chinois correspondants.

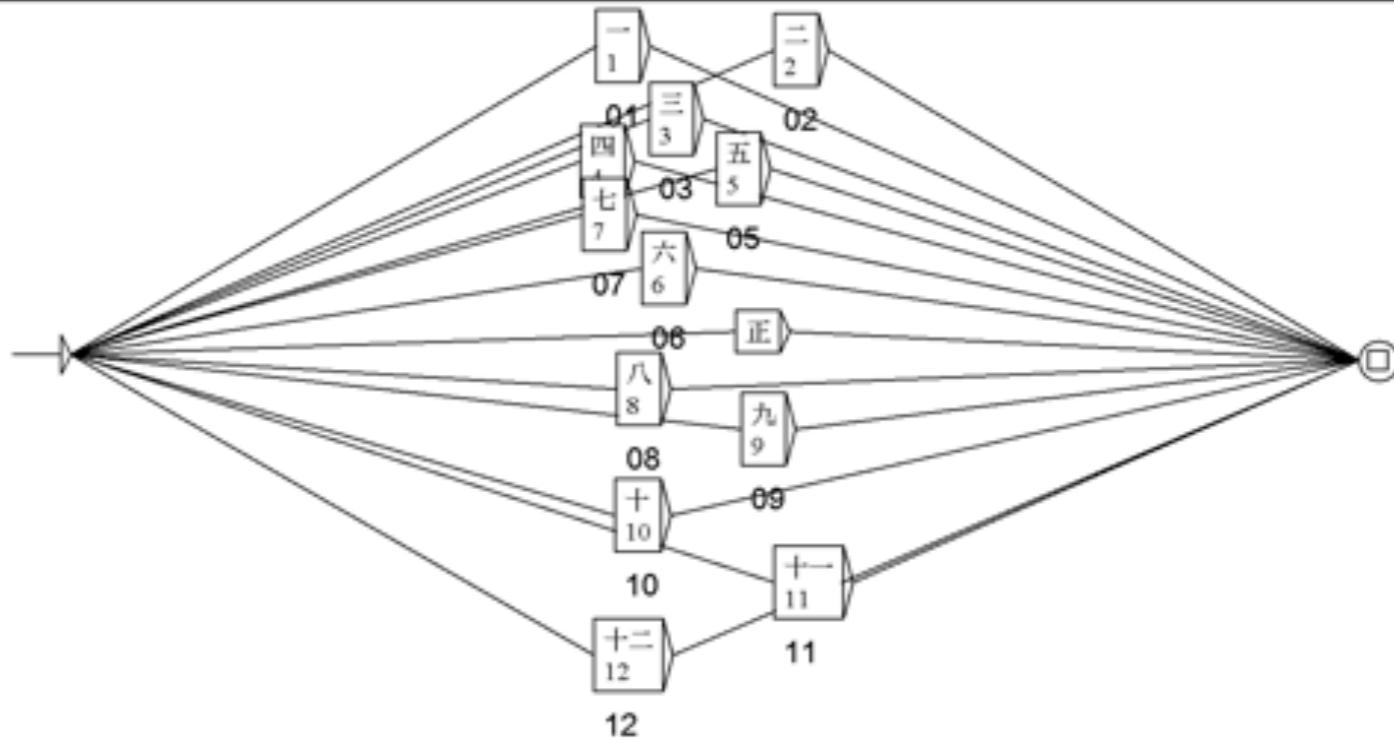
En outre, nous utilisons parfois des caractères chinois spéciaux supplémentaires.

Par exemple: 元, représentant zéro; 千, représentant mille;  
百, représentant cent.



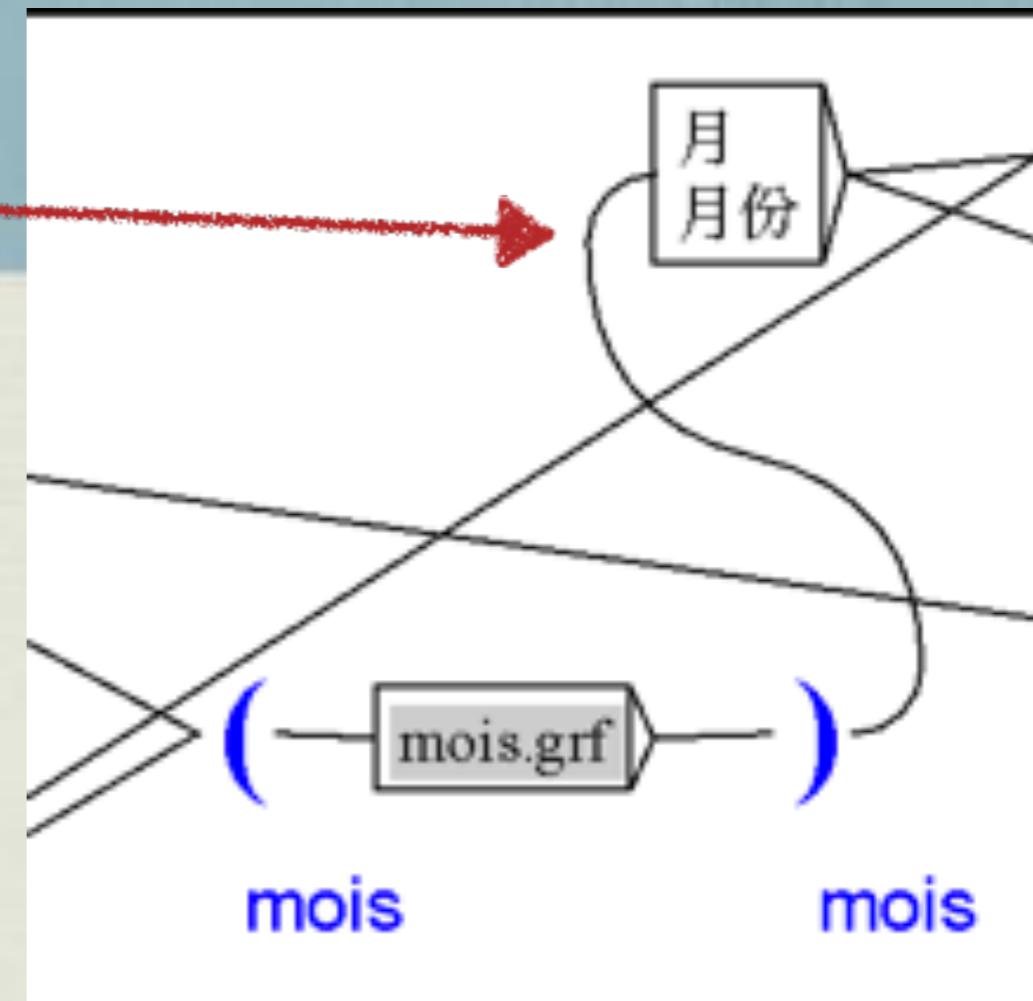
- 2.year.grf ( xxxx**年** (année) )
- Un nombre et un caractère chinois “年” forment une année. Un chiffre lui-même se boucle pour obtenir un nombre de l’année, en plus avec un mot “年” (année) ,ils composent une expression de l’année.

### 3. mois.grf

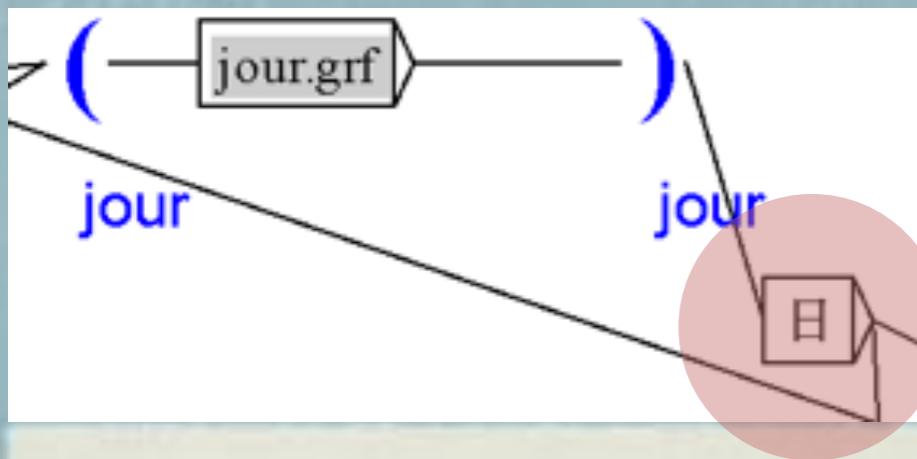


- Le mois se compose d'un nombre (1-12) et d'un caractères du mois chinois 月 ou 月份(mois).
- En outre, nous utilisons parfois des caractères chinois spéciaux supplémentaires. (Il y a deux façons d'exprimer pour certains mois)
- Par exemple: 正月, représentant janvier  
二月, représentant février etc.

- ◆ Nous avons essayé de mettre cette unité de mesure dans la sous-graphe des mois.grf, mais cela ne correspondait **pas** au bon mois à *Unitex*, et après avoir communiqué avec l'enseignant, nous n'avons pas résolu ce problème.

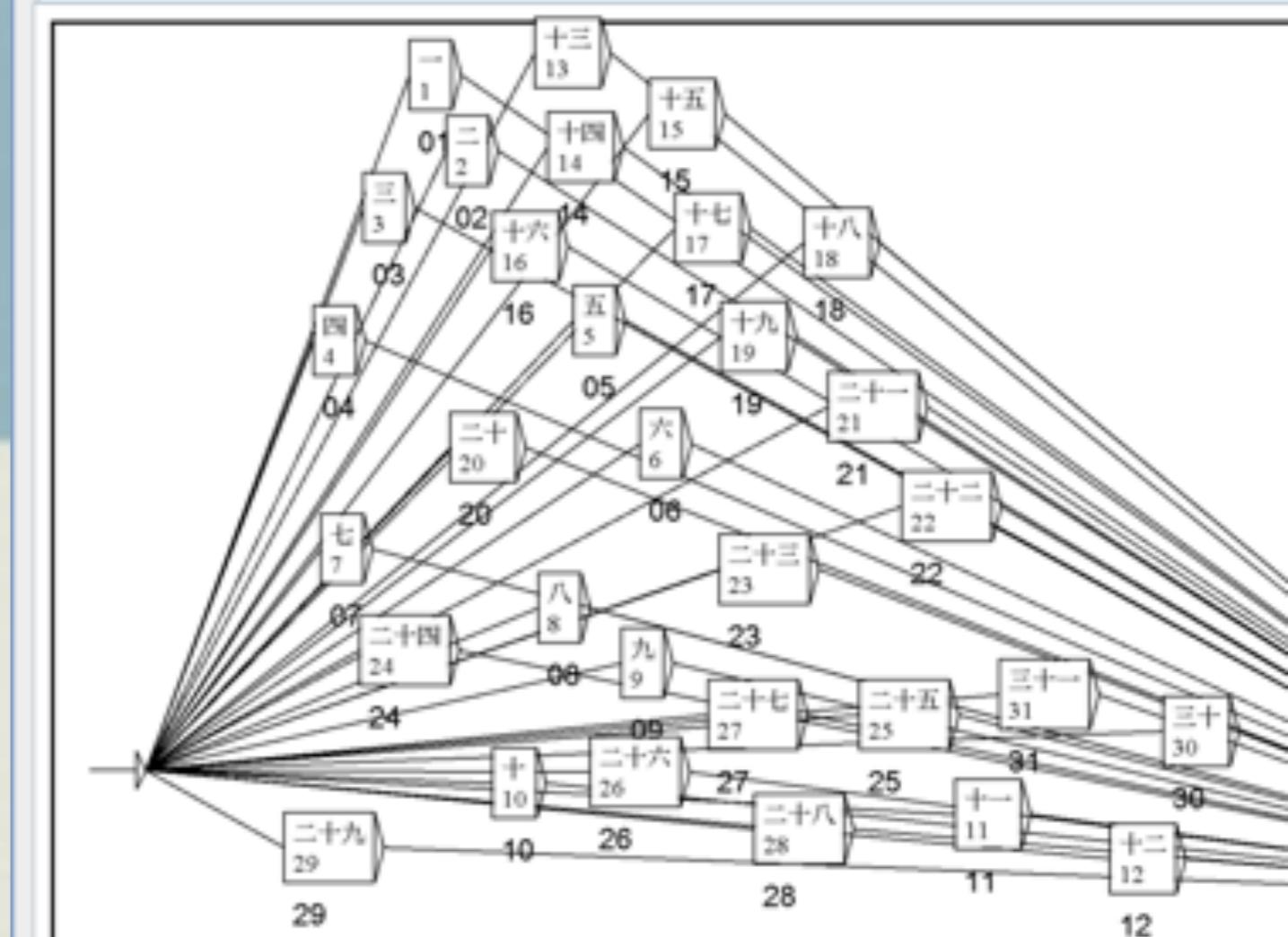


- ◆ Nous mettons donc l'unité de mesure du mois dehors du *mois.grf* et la graphe marche bien.



- ◆ *Et* nous avons rencontré le même problème que *mois.grf* dans *jour.grf* et nous avons traité de la même façon.

- ◆ 4. *jour.grf*
- ◆ Il a été réalisé par des chiffres de 1 à 31 ou des caractères chinois correspondants connecté avec un caractère 日 ou 号 (xx jour).

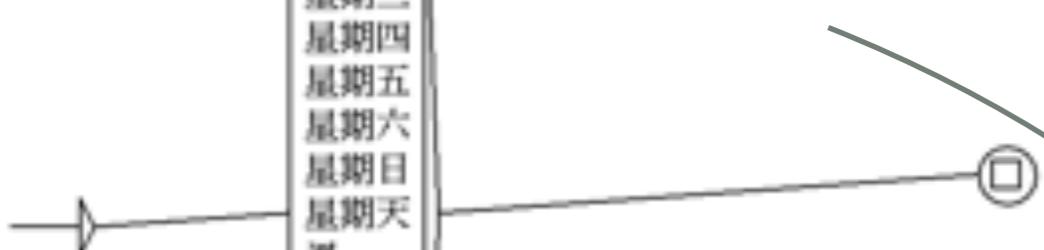




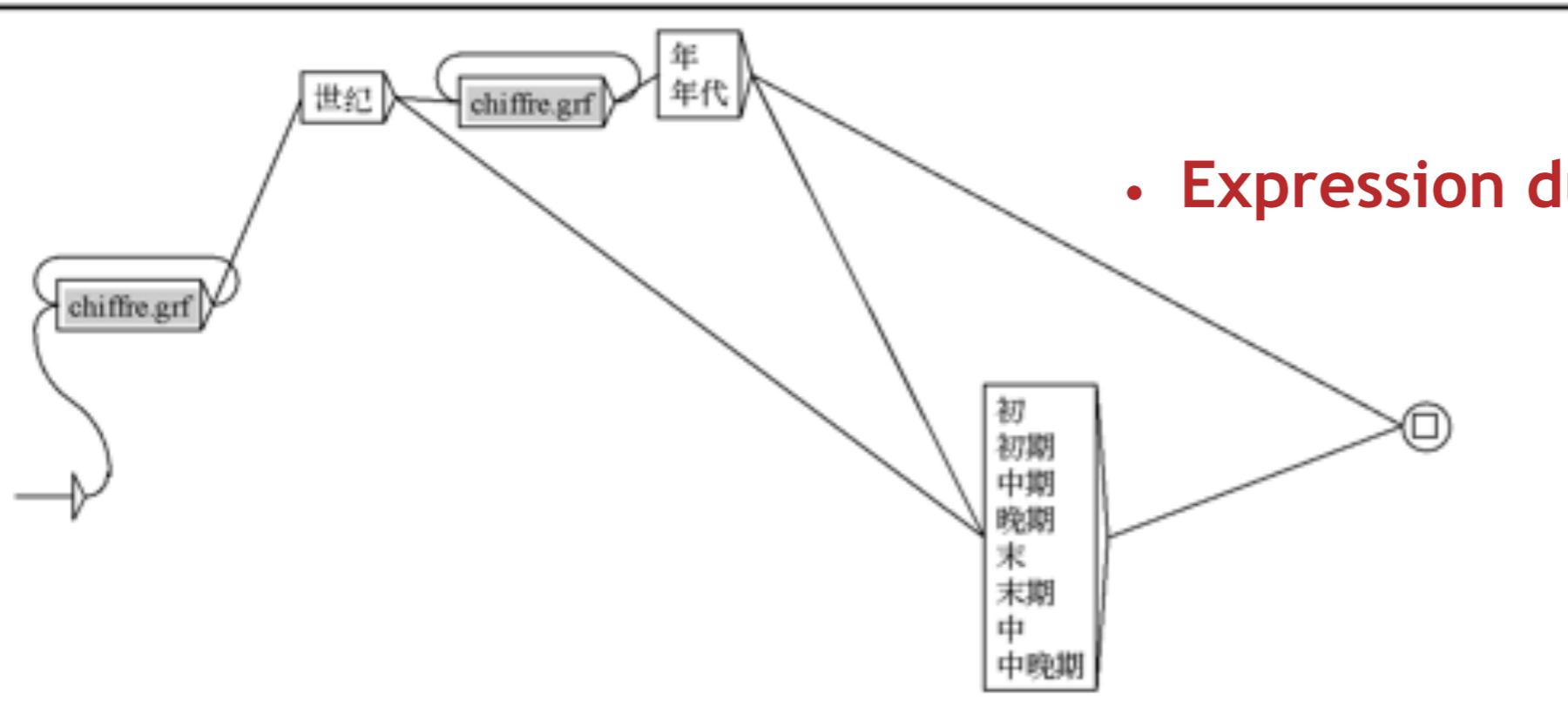
1

星期一  
星期二  
星期三  
星期四  
星期五  
星期六  
星期日  
星期天  
周一  
周二  
周三  
周四  
周五  
周六  
周日

14



- 5.semaine.grf
- Il y a **deux** façons d'exprimer **chaque** jour de la semaine en Chine ,donc on a 14 mots.



## • Expression du siècle et des années

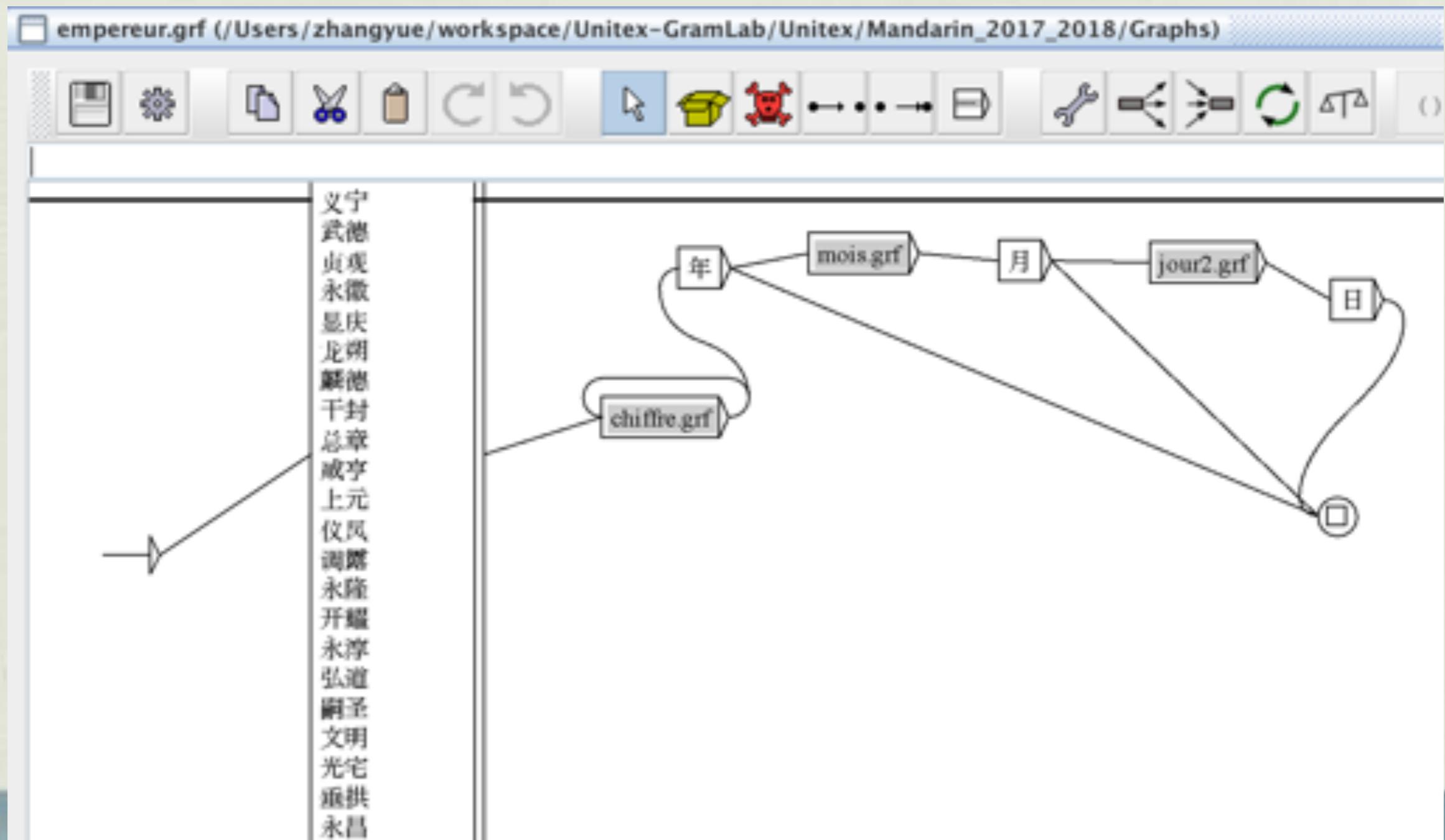
- Nous utilisons *chiffre.grf* à nouveau, puis on créer l'expression du siècle avec des caractères chinois liées. Le chiffre (qui se boucle )compose une nombre du siècle.
- De même pour l'expression :xx年(代)(des années xxxx).
- Il y a 8 façons de représenter la période du siècle, comme le stade précoce (早期), le stade intermédiaire (中晚期), etc.
- Forme : xx**世纪** (siècle)xx**年(代)**(des années xxxx)xx**期**(période)
- Exemple:20**世纪80年代初期** égale au **début des années** 80 du 20<sup>e</sup> **siècle**

- **Expression de la date ancienne utilisant les noms des empereurs**

**Forme :**

**le nom d'empereur** + xxxx年(année) xx月(mois) xx日(un jour du mois)

Nous avons utilisées une boîte contenant tous les nom des empereurs et trois sous-graphes : *chiffre.grf*, *mois.grf* et *jour2.grf*.



- Afin d'obtenir les noms des anciens empereurs, nous avons cherché une liste sur Internet. Nous utilisons un simple programme de python pour extraire les noms et ajouter un signe plus(+) entre les deux noms (afin de s'adapter à la syntaxe *Unitex*).
- C'est le programme de python que nous avons écrit et nous obtenir les noms formés spécifiques:

```

oksup.py ×
1 def sup(text):
2     liste=["|","_","-","_","_","+", "|",".", "B","C","(,") ","周"]
3     f=open(text,"r",encoding='UTF-8')
4     ll=""
5     for ligne in f:
6         for mot in ligne:
7             if mot in liste:
8                 continue
9             else:
10                ll+=mot
11    f.close()
12    print("eee"+ll)
13    with open('/Users/wangqi/Desktop/YooHoo/em.txt','a') as fa:
14        fa.write(ll)
15
16
17 def plus(text):
18     f=open(text,"r",encoding='UTF-8')
19     ll=""
20     for ligne in f:
21         for mot in ligne:
22             if mot =='\n':
23                 ll+= '+'
24             else:
25                 ll+=mot
26     f.close()
27     return ll
28 if __name__ == '__main__':
29     sup("empoire.txt")
30     ll=plus("em.txt")
31     print(ll)
32
33

```

empoire.txt

元光+元朔+元狩+元鼎+元封+太初+天汉+太始+征和+后元+始元+元凤+元平+本始+地节+元  
光+建昭+竟宁+建始+河平+阳朔+鸿嘉+永始+元延+绥和+建平+元寿+元始+居摄+初始+始建  
平+建初+元和+章和+永元+元兴+延平+永初+元初+永宁+建光+延光+永建+阳嘉+永和+汉安  
+永兴+永寿+延熹+永康+建宁+熹平+光和+中平+光熹+昭宁+永汉+中平+初平+兴平+建安+延  
平+正元+甘露+景元+咸熙+章武+建兴+延熙+景耀+炎兴+黄武+黄龙+嘉禾+赤乌+太元+神凤  
+宝鼎+建衡+凤凰+天玺+天纪+泰始+咸宁+太康+太熙+永熙+永平+元康+永康+永宁+太安+永  
兴+建武+太兴+永昌+太宁+咸和+咸康+建元+永和+升平+隆和+兴宁+太和+咸安+宁康+太元  
+元嘉+孝建+大明+永光+景和+泰始+泰豫+元徽+升明+建元+永明+隆昌+延兴+建武+永泰+永  
+大同+中大同+太清+大宝+天正+承圣+天成+绍泰+太平+永定+天嘉+天康+光大+太建+至德  
+神瑞+泰常+始光+神+廷和+太延+太平真君+正平+承平+兴安+兴光+太安+和平+天安+皇兴  
+延昌+熙平+神龟+正光+孝昌+武泰+建义+永安+建明+普泰+中兴+太昌+永兴+永熙+天平+永  
和+建德+南朝+宋+齐+梁+陈+西晋+东晋+宋+齐+梁+陈+南朝+北魏+西魏+东魏+北齐+北周+南

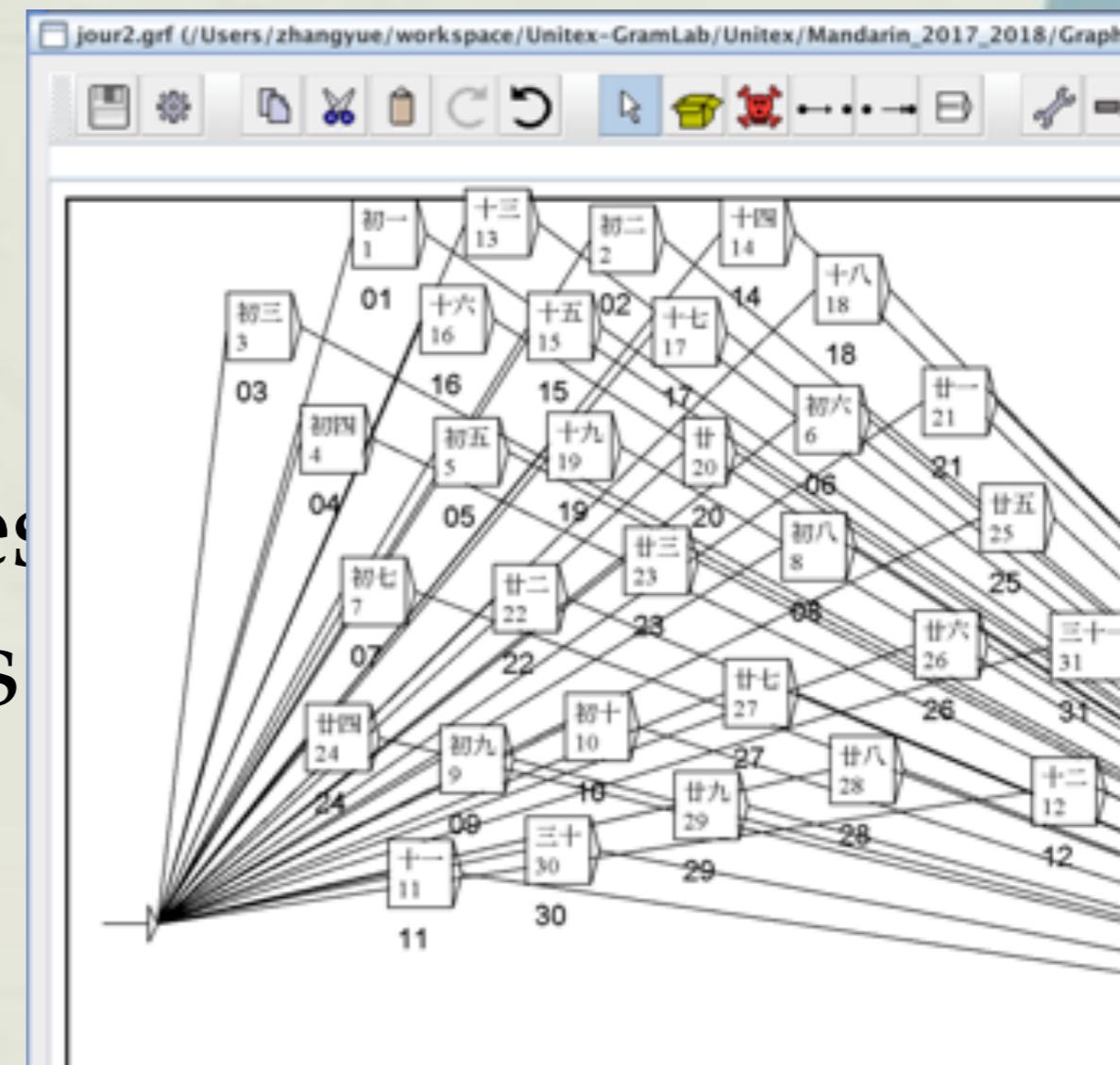
- Expression de la date de la période spéciale qui s'appelle
  - La République de Chine (1912-1949)

Forme : 民国/民 (La République de Chine) + xxxx年 (année) xx月 (mois) xx日 (un jour du mois)

Exemple : 民国23年5月 égale le mai 23<sup>e</sup> de La République de Chine

C'est presque le même que l'expression utilisant les noms des empereurs

Contrairement à *jour.grf*, nous en avons ajouté d'autres les expressions des jours du mois lunaire dans *jour2.grf*, on l'utilise pour une expression un peu ancienne.



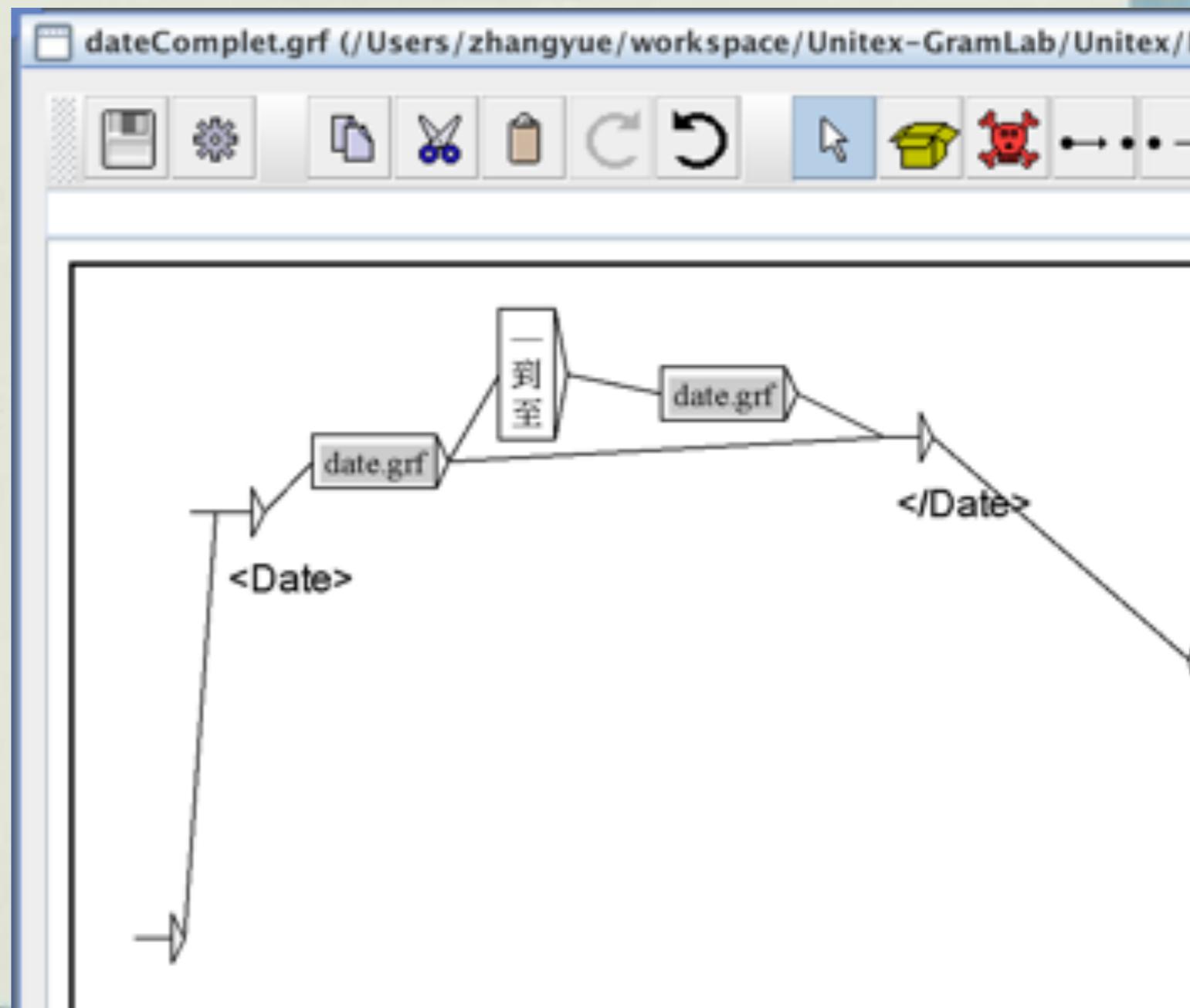
## • Expression de la durée

Forme : une date1 — / 到 / 至 une date2

→ (de date1 à date2)

Pour exprimer la durée, nous réutilisons la date.grf,  
les 3 caractères — ou 到 ou 至 (à)  
relie les deux dates de la durée.

Au début, nous ne pensons pas avoir besoin d'ajouter une durée. Car il n'était pas une expression de date exacte  
Après avoir discuté avec l'enseignant, nous avons décidé de composer une durée en appelant *date.grf*.



# Utilisation des graphes et *Brat*

- Nous avons téléchargé un article concerné l'histoire chinoise récente sur Internet, nous avons utilisé le graphe complété (*dateComplet.grf*) pour identifier ce document, et nous avons obtenu le résultat ci-dessous et il y a 189 dates est marquées avec balise `<Date></Date>`:



|   |                               |      |       |       |             |      |      |      |      |
|---|-------------------------------|------|-------|-------|-------------|------|------|------|------|
| Date                                      | Date                          | Date | Date  | Date  | Date        | Date | Date | Date | Date |
| 道光十九年正月 (1839年3月)                         | ，林则徐到达广                       |      |       |       |             |      |      |      |      |
| 第一次鸦片战争                                   |                               |      |       |       |             |      |      |      |      |
| 第一次鸦片战争(2张)                               |                               |      |       |       |             |      |      |      |      |
| Date                                      | Date                          | Date | Date  | Date  | Date        | Date | Date | Date | Date |
| 并严肃表示禁烟的决心：{S}“若鸦片一日未绝，本大臣一日不回，誓与此事相抗到底。” |                               |      |       |       |             |      |      |      |      |
| 四月二十二日至五月十三日 (6月3日至6月25日)                 | ，将缴获的鸦片全部在虎门滩当众销毁。{S}虎门销烟打击了外 |      |       |       |             |      |      |      |      |
| Date                                      | Date                          | Date | Date  | Date  | Date        | Date | Date | Date | Date |
| 157                                       | T149                          | Date | 28195 | 28205 | 20世纪80年代    |      |      |      |      |
| 158                                       | T151                          | Date | 28251 | 28259 | 20世纪80年代    |      |      |      |      |
| 159                                       | T158                          | Date | 28523 | 28533 | 1840–1919年  |      |      |      |      |
| 160                                       | T168                          | Date | 28993 | 29003 | 1840–1949年  |      |      |      |      |
| 161                                       | T169                          | Date | 29150 | 29157 | 19世纪中晚期     |      |      |      |      |
| 162                                       | T171                          | Date | 29329 | 29337 | 20世纪80年代    |      |      |      |      |
| 163                                       | T173                          | Date | 30378 | 30388 | (1840–1911) |      |      |      |      |
| 164                                       | T174                          | Date | 30488 | 30497 | 1912–1949   |      |      |      |      |
| 165                                       |                               |      |       |       |             |      |      |      |      |

img.1

道光十九年正月 (1839年3月)，林则徐到达广  
第一次鸦片战争  
第一次鸦片战争(2张)

并严肃表示禁烟的决心：{S}“若鸦片一日未绝，本大臣一日不回，誓与此事相抗到底。”

四月二十二日至五月十三日 (6月3日至6月25日)，将缴获的鸦片全部在虎门滩当众销毁。{S}虎门销烟打击了外

img.2

- Nous avons utilisé le logiciel Brat et on a marqué la date sur le fichier manuellement et on a exportée un fichier xx.ann. On a trouvée 164 dates.
- On peut comparer le fichier traité par Unitex et le fichier qui est bien été marqué avec Brat pour évaluation.(img3)

1931年9月18日，爆发了九一八事变。(S)日本与中国之间的矛盾进一步激化。而日本国内主战派日本军部地位上升，导致日本走上全面侵华的道路。为抗日战争的全面爆发埋下了导火索。(S)日本关东军占领者，并利用投靠日本的前清遗老博古在日本东北建立了满洲国傀儡政权，实行了14年之久的殖民统治。(S)

1934年10月，中央红军主力开始长征。(S)

同年11月至次年4月，在照搬苏革命根据地的红二十五军和照搬革命根据地的红四方面军分别开始长征。(S)

1935年11月，在照搬苏革命根据地的红二、六军团也离开根据地开始长征。(S)

1936年6月，第二、六军团组成第二方面军。(S)

同年10月，红军第一、二、四方面军在甘肃会宁胜利会合，结束了长征。(S)其中红一方面军长征历时一年，转战十一个省，翻越行军的二万五千里。(S)

1936年12月12日，发生了西安事变，确立了抗日民族统一战线的主张。(S)

### 抗日战争

(War of Resistance Against Japan)简称抗战。指 20世纪中期 第二次世界大战中，中国抵抗日本侵略的一场民族性的全面战争。(S)国际上称作第二次中日战争(Second Sino-Japanese War)。日本侵华战争(Attack on China)。(S)抗战时间从 1931年9月18日 九一八事变开始算起，至 1945年 结束，共 14 四年 抗战。(S)(1)

1931年，侵华日军发动九一八事变后，完全侵占中国东北，并成立伪满洲国。此后陆续在华北、上海等地制造事端，挑起战争。国民政府则采取妥协政策避免冲突扩大。(S)1937年7月7日，日军在上海卢沟桥附近制造卢沟桥事变，中日战争全面爆发。(S)1941年12月7日 日本发动太平洋战争后，12月9日 重庆国民政府正式对日宣战。(S)1945年8月15日，日本向同盟国无条件投降。(S)

中国战场是二战的主要战场之一。(S)中国人民的抗日战争是二战的主要组成部分。(S)

中国人的抗日战争，是中华民族历史上最伟大的卫国战争，是中国人民反抗日本帝国主义侵略的正义战争，是世界反法西斯战争的重要组成部分，也是中国近代以来抗击外来侵略第一次取得完全胜利的民族解放战争。

img.3

## Entités Nommées\_PERSONNES

### ■ Le travail que nous avons accompli

- ◆ La analyse des méthodes d'expression du nom+prénom chinois
- ◆ Noté tous les noms de famille chinois
- ◆ Enregistré tous les noms+prénoms chinois communs
- ◆ La Modélisation de dictionnaire et des graphiques
- ◆ La recherche et le test du corpus
- ◆ La comparaison les résultats UNITEX et BRAT
- ◆ Les problèmes qui existent

## Entités Nommées\_PERSONNES

### ■ La analyse des méthodes d'expression du nom+prénom

- ◆ Forme : nom(rouge)+appellation / profession(orange)
- ◆ Exemple: 张先生
  - ◆ égale le Monsieur Zhang

En Chine, si on veut appeler qn plus poli, on peut l'appeler comme monsieur, mademoiselle ou bien madame. De plus, on aussi l'appeler son profession comme docteur, directeur , president etc...

C'est le format plus formel et poli. Les gens les utilisent au lieu de leurs prénoms quand on est dans les cadres formels.

D'ailleur, il peut aussi être utilisé régulièrement à appeler qn qui n'est pas très familler.

## Entités Nommées\_PERSONNES

### ■ La analyse des méthodes d'expression du nom+prénom

- ◆ Forme : nom(rouge)+prénom(orange)
- ◆ Exemple: 张可
  - ◆ égale le **KE ZHANG**

C'est un nom complet, qui aussi peut être utilisé dans les cadres formels comme réunion. Il est une forme sérieuse mais pas très respectueuse comme le premier cas. Normalement, les gens qui sont supérieurs peuvent appeler leurs sous-ordres en ce cas.

## Entités Nommées\_PERSONNES

### ■ La analyse des méthodes d'expression du nom+prénom

En chine, il y a des prénoms bizarres qui contient des mots comme "fleure" ou "rouge", dans ce cas, l'unitex ne peut pas bien connaître s'il est un nom ou les autres choses comme les adjectifs, les adverbes, etc..

- Exemple: ...一张纸... égale le ...Un Zhi Zhang...

(numéro: 1+nom+prénom)

- Ou: ...一张纸... égale le ...un morceau de papier...

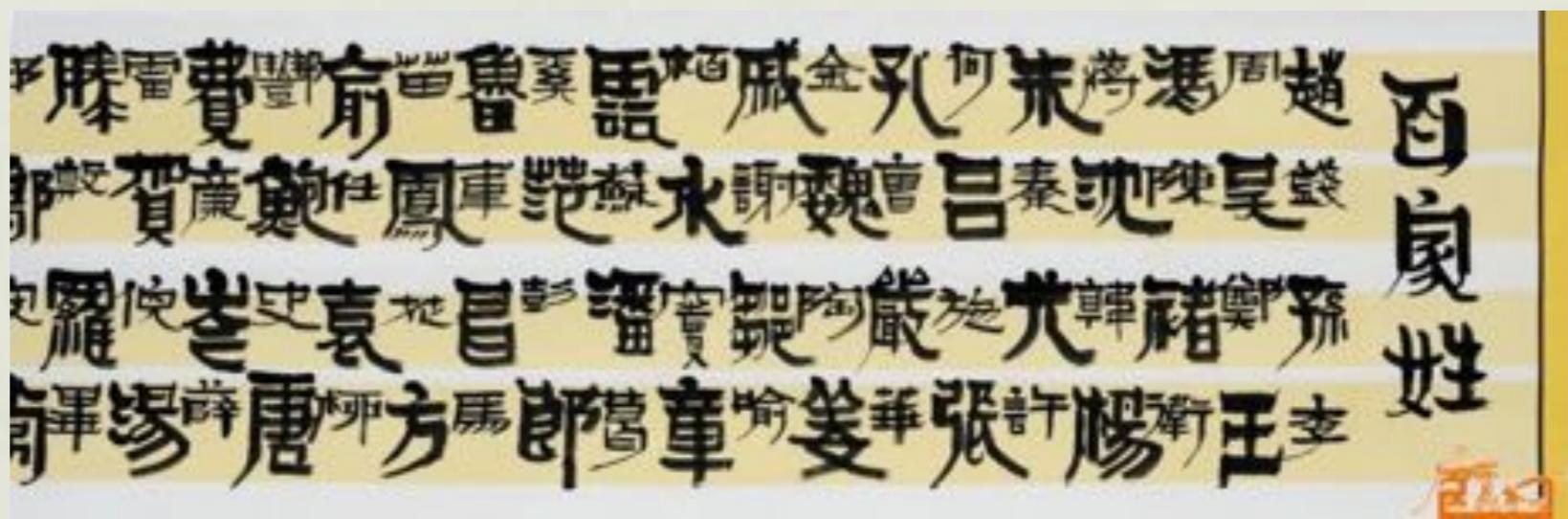
## Entités Nommées\_PERSONNES

■ Noté tous les noms de famille chinois

### Nom

En Chine, tous les noms sont certains depuis les temps anciens, chaque personne a le même nom que ses ancêtres, on ne les crée pas par soi-même. Les noms chinois sont collectés dans un livre des noms de famille(BaiJiaXing) qui est redigé depuis la dynastie son du nord.

D'ailleur, on peut utiliser le même nom même si on n'est pas famille.



## Entités Nommées\_PERSONNES

### ■ Noté tous les noms de famille chinois

Nous avons extrait tous les noms chinois et ajouté le signe plus (“+”) dans le programme python pour l'adapter à la syntaxe unitex.

```
def affiche_fichier(entree):
    f=open(entree,"r",encoding='UTF-8')
    text=f.read()
    f.close()
    return text

if __name__ == '__main__':
    text=affiche_fichier('tra')
    lst=text.split('\n')
    for i in range(len(lst)):
        if lst[i]=='\n' or lst[i]==' ':
            lst[i]='+'
    print(lst)
    text='+'.join(lst)
    f=open('sortir.txt','w',encoding='UTF-8')
    f.write(text)
    f.close()
```

赵+钱+孙+李+周+吴+郑+王+冯+陈+褚+卫+蒋+沈+韩+杨+朱+秦+尤+许+何+吕+施+张+孔+曹+严+华+金+魏+陶+姜+戚+谢+邹+喻+柏+水+窦+章+云+苏+潘+葛+奚+范+彭+郎+鲁+韦+昌+马+苗+凤+花+方+俞+任+袁+柳+酆+鲍+史+唐+费+廉+岑+薛+雷+贺+倪+汤+滕+殷+罗+毕+郝+邬+安+常+乐+于+时+傅+皮+卞+齐+康+伍+余+元+卜+顾+孟+平+黄+和+穆+萧+尹+姚+邵+湛+汪+祁+毛+禹+狄+米+贝+明+臧+计+伏+成+戴+谈+宋+茅+庞+熊+纪+舒+屈+项+祝+董+梁+杜+阮+蓝+闵+席+季+麻+强+贾+路+娄+危+江+童+颜+郭



## Entités Nommées\_PERSONNES

■ Enregistré tous les noms+prénoms chinois communs

### Prénom

Les prénoms chinois sont nommés par des mots(normalement un mot ou deux mots) sans ordre certain. On peut avoir les noms bizarres comme petite fleure, petit rouge, etc..

Donc, si on veut connaître le nom du seconde cas, soit on les connaît manuellement à l'aide de Brat, soit on récupère une partie de nom+prénom qui est déjà enregistré dans Internet. Et on les met dans une dictionnaire nouvelle et les ajoute une nature qui s'appelle <<Personne>>.

<http://www.resgain.net/> (Voici le site qui enregistre les ressources de nom+prénom)

## Entités Nommées\_PERSONNES

### ■ Enregistré tous les noms+prénoms chinois communs

Nous écrivons le programme python et récupérons tous les noms+prénom communs sur internet ajouté “,.N+Personne” dans le programme python pour l'adapter à la syntaxe unitex.

```

11 def affiche_fichier(entree):#从文件中导入并返回字符串
12     f=open(entree,'r',encoding='UTF-8')
13     text=f.read()
14     f.close()
15     return text
16
17 text=affiche_fichier('xinpinyin.txt')#导入所有姓氏(拼音)文件
18 xlst=text.split('+')
19 xinglst=[]
20 for x in range(0,len(xlst)-1):
21     flag=0
22     xing=xlst[x]
23     while(xing in xinglst):
24         flag+=1
25         xing=xlst[x]+str(flag)
26     xinglst.append(xing)
27 xinglst.sort()
28 print(xinglst)
29
30 for x in range(0,len(xinglst)-1):
31
32     Nom=xinglst[x]
33     print(Nom)
34     text=''
35     gender=['boys','girls']#用于分别读取男生姓名和女生姓名
36     igender=0#gender数组的指针
37
38     for i in range(1,11):#从网页的第一页到第十页
39         r = requests.get('http://'+Nom+'.resgain.net/name/'+gender[0]+'_'+str(i)+'.html')

```

```

38     for i in range(1,11):#从网页的第一页到第十页
39         r = requests.get('http://'+Nom+'.resgain.net/name/'+gender[0]+'_'+str(i)+'.html')
40         #print (r.text)
41         pattern = r"/name/(.+).html\" class=\"btn";#python爬虫匹配网页源代码，从而找出源代码
42         result = re.findall(pattern,r.text)
43
44         r = requests.get('http://'+Nom+'.resgain.net/name/'+gender[1]+'_'+str(i)+'.html')
45         #print (r.text)
46         pattern = r"/name/(.+).html\" class=\"btn";
47         result2 = re.findall(pattern,r.text)
48
49         if result==[]:
50             print(Nom+' null!!!')#如果网页找不到当前的姓氏，输出NULL
51             break
52         else:
53             for j in result:
54                 text+=j+',.N+Personne\n'#dictionnaire的格式，所以在每个名字后面加上,.N+Personne
55             for j in result2:
56                 text+=j+',.N+Personne\n'
57
58         #print (result)
59         if text!='':
60             f=open(Nom+'.txt','w',encoding='UTF-8')#导出到文件
61             f.write(text)
62             f.close()

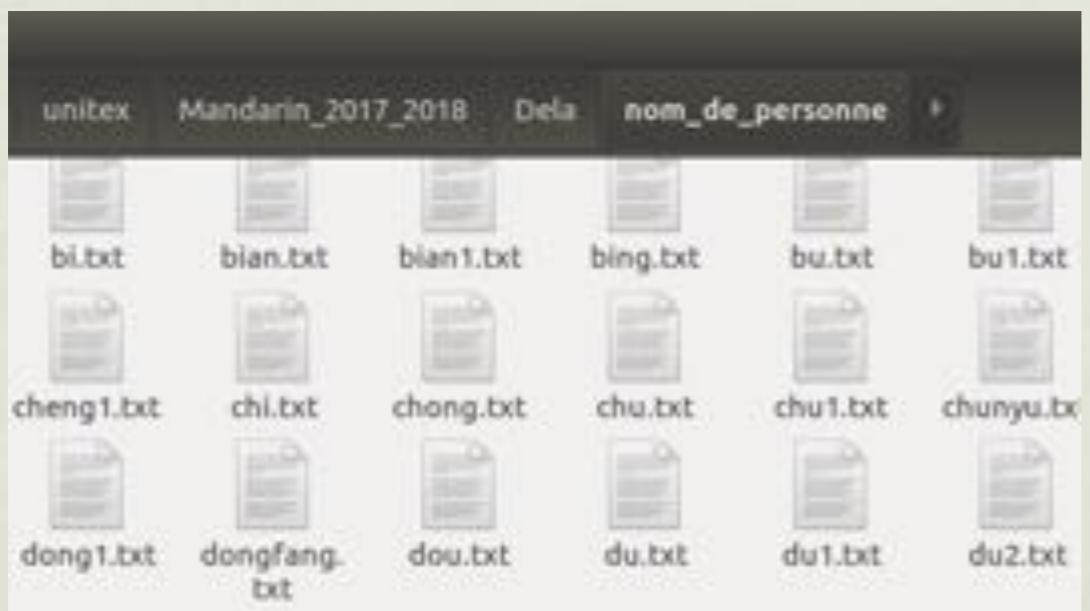
```

## Entités Nommées\_PERSONNES

### ■ La Modélisation de dictionnaire et des graphiques

#### Dictionnaire

Nous intégrons tous les fichiers qui contient les noms+prénom communs en tant que dictionnaire



```
艾宗铧,.N+Personne
艾芯锋,.N+Personne
艾鑫,.N+Personne
艾忻,.N+Personne
艾熙凡,.N+Personne
艾顺光,.N+Personne
艾春富,.N+Personne
艾继东,.N+Personne
艾建楠,.N+Personne
艾安琪,.N+Personne
艾华,.N+Personne
艾一鸣,.N+Personne
艾福旗,.N+Personne
艾水桥,.N+Personne
艾剑波,.N+Personne
艾欣达,.N+Personne
艾培培,.N+Personne
艾鑫尘,.N+Personne
```

```
Text DELA FSGraph Lexico
/house/szhos02/unitex/Na
//! nom_de_personne/ai.txt
//! nom_de_personne/an.txt
//! nom_de_personne/ao.txt
//! nom_de_personne/bai1.txt
//! nom_de_personne/bai.txt
//! nom_de_personne/ban.txt
//! nom_de_personne/bao1.txt
//! nom_de_personne/bao2.txt
//! nom_de_personne/bao.txt
//! nom_de_personne/ba.txt
//! nom_de_personne/bei.txt
//! nom_de_personne/ben.txt
//! nom_de_personne/bian1.txt
```

**Exemple de fichier Intégrer tous les nom+prenom fichiers en tant que commun dictionnaire**

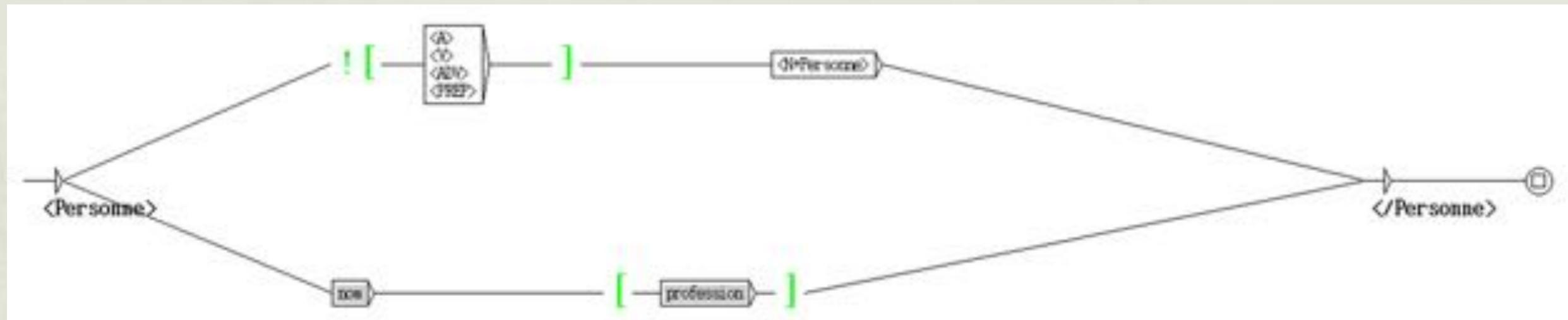
## Entités Nommées\_PERSONNES

### □ La Modélisation de dictionnaire et des graphiques

#### Graphiques

La route ci-dessus correspond à nom+prénom et le dictionnaire que nous avons fait était utilisé. Les crochets vertes avec le point d'exclamation indiquent: excluent les adjectifs, les adverbes, les verbes, les prépositions.

.

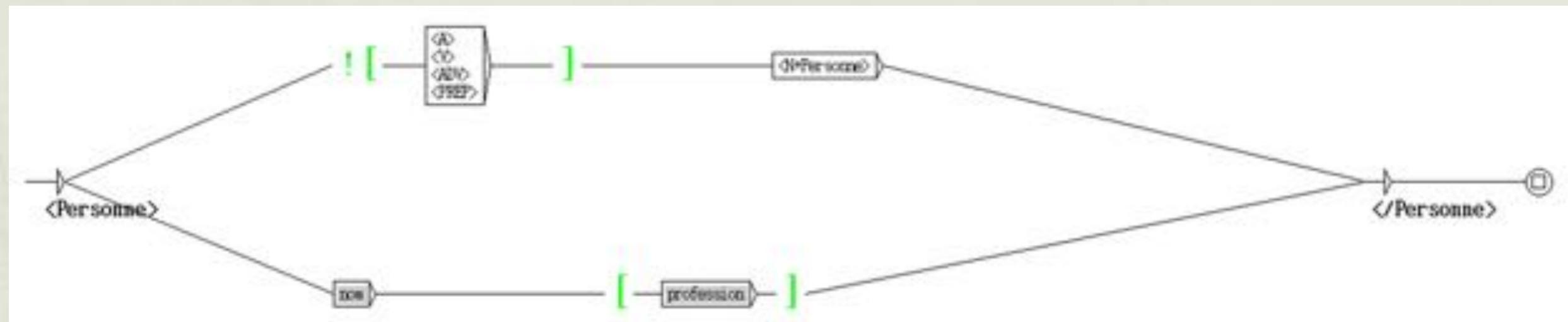


## Entités Nommées\_PERSONNES

### □ La Modélisation de dictionnaire et des graphiques

#### Graphiques

La route suivante correspond à nom+appellation/profession. Et on mettre le partie du nom et du prénom dans le sous graph. Les crochets vertes indiquent: ne pas mettre le profession dans le résultat du test unitex.



## Entités Nommées\_PERSONNES

### La recherche et le test du corpus

#### Partie Unitex

Unitex étiquette le nom en bleu selon le dictionnaire et les graphs. Et dans le fichier qui contient le résultat, les deux côtés du nom ont <Personne>et</Personne>.



résultat



le fichier qui contient le résultat

## Entités Nommées\_PERSONNES

### □ La recherche et le test du corpus

#### Partie Brat

Nous avons besoin d'une opération manuelle dans Brat. Et dans le fichier du résultat, il y a des numéros, position de départ et de fin.

| 7      | 对此，胡雷曾不止一次的吐槽，然而，他毕竟只是一个小小的打工仔，光明正大的质疑老板品味是一件很危险的事情，所以 | Personne |
|--------|--|----------|
| 9      | 「班主任确认，请选择系统定制主题。」 半个小时前，一个清脆的机械音忽然在胡雷脑海中响起，起初胡雷同学还以为最 | Personne |
| 11     | 一分钟后…… 「请选择系统定制主题。」                                    | Personne |
| 13     | 同样的声音，这次胡雷听得真切，更重要的是，明明是机械音，但他分明从中听出了一丝不耐？或许真的是幻听了，胡雷再 | Personne |
| 又一分钟…… |  | Personne |
| 15     | 「请选择系统定制主题。」 机械音第三次响起……                                | Personne |
| 17     | 直到这个声音以每分钟一次的频率循环播放了十几次的时候，胡雷才意识到，这并不是幻听或者是谁的恶作剧，而是……真 | Personne |

résultat

|    |          |     |     |    |
|----|----------|-----|-----|----|
| T1 | Personne | 286 | 288 | 胡雷 |
| T2 | Personne | 359 | 361 | 胡雷 |
| T3 | Personne | 435 | 437 | 胡雷 |
| T4 | Personne | 507 | 509 | 胡雷 |
| T5 | Personne | 553 | 555 | 胡雷 |
| T6 | Personne | 611 | 612 | 林  |
| T7 | Personne | 705 | 707 | 胡雷 |
| T8 | Personne | 800 | 802 | 胡雷 |

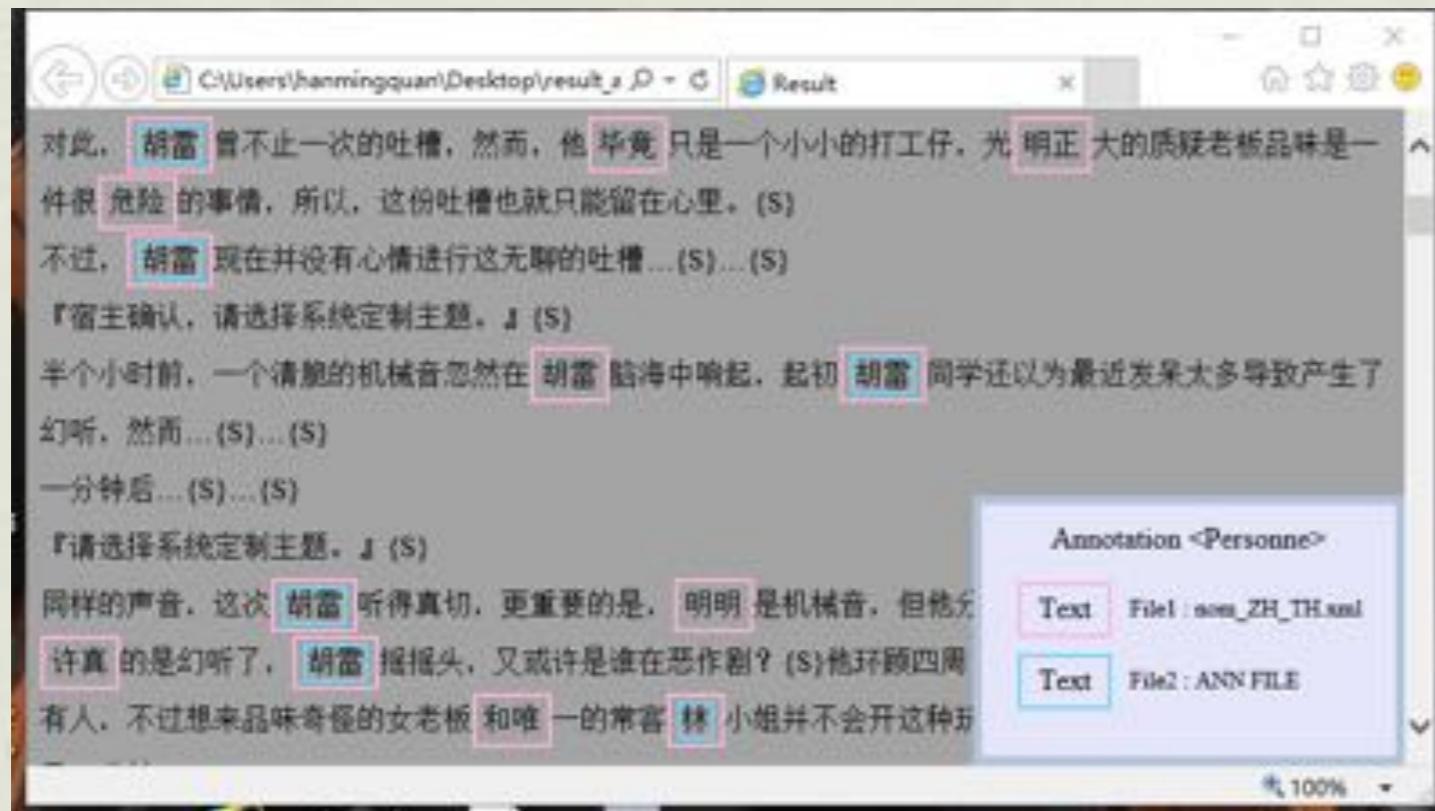
le fichier qui contient le résultat

## Entités Nommées\_PERSONNES

### □ La comparaison les résultats UNITEX et BRAT

#### Le résultat de la comparaison

Les boîtes bleue sont les résultats Unitex, et les boîtes rouge sont des Brat.



résultat

```

Similarity score (weak precision) : 0.3566879
Similarity score (strict precision) : 0.27388534
Similarity score (weighted precision) : 0.32484075
Similarity score (weak recall) : 0.6292135
Similarity score (strict recall) : 0.48314607
Similarity score (weighted recall) : 0.5730337
Similarity score (weak F-measure) : 0.45528454
Similarity score (strict F-measure) : 0.3495935
Similarity score (weighted F-measure) : 0.4146341

```

Taux de coïncidence dans différentes situations

## Entités Nommées\_PERSONNES

### □ Les problèmes qui existent

1. Dans les résultats d'UNITEX, nous avons trouvé qu'il y a des résultats qui n'est pas un nom.
2. Si le nom+prénom n'est pas dans le dictionnaire, on ne peut pas l'identifier (pour le nom+appellation/profession, ça marche).

没有了交谈，小庙瞬间安静了下来，只听得是说道，“她就这么安静的坐在那里。” (S) “嗯，你们安静。” (S) ·一番叫 [S] “便一点安全。” (S) ·杜聆对于小小来说人便些比较安全，但是安全，但是太做了似乎又不太安全了。 (S) 看着众多学妹回帖安慰自己的宁相宇随白不要，胡露自我安慰着在系统的催促下。“ (S) 巴掌大小的洁白小瓷盘上摆放的居然越弹越好，颇有专业级别的感觉透出。 (S) 下午下班的时候杜聆叫住自己包中，然后从包中拿出了一本杜聆直接收到了自己包中，然后从包中这不是晨跑时遇到的包子头妹子吗？琢的小女孩，肉嘟嘟的包子脸上长着一口气。 (S) 遮面跑来的包子头小美女笑了。 (S) 哟，这不是晨跑的包子头妹子吗？胡露呆呆的看着包子头妹子吃了一口。

# Perspectives

- Adaptation éventuelle de la mise en page du site et du manuel à la graphie chinoise.
- Tester les résultats avec l'outil d'évaluation (en cours).