D'aquest capítol disposes dels següents arxius:

- El capítol en pdf: 07-Wordnet-cat.pdf
- Els arxius i programes necessaris: programes8-cat.zip

WordNet és una base de dades lèxica de l'anglès. En aquesta base de dades les paraules pertanyents a categories obertes (substantius, verbs, adjectius i adverbis) s'agrupen en conjunts de sinònims anomenats *synsets*. Cada synset representa un concepte lexicalitzat en anglès. Els synsets de WordNet estam relacionats per les següents relacions semàntiques:

- homonímia: és una relació d'especificitat entre un mot (l'hiònim) i un altre de significat més genèric (l'hiperònim).
- meronímia: és la relació entre una part i el tot
- antonímia: o significat oposat
- troponímia o implicació lèxica: és una relació que es dóna entre els verbs i que en certa manera és equivalent a la relació d'hiponímia en els substantius

NLTK proporciona una accés fàcil a WordNet. De fet es tracta simplement d'un lector de corpus específic per a WordNet i que es pot importar fent:

```
>>> from nltk.corpus import wordnet
```

o bé, d'una manera més compacta

```
>>> from nltk.corpus import wordnet as wn
```

En el programa-8-1.py podem veure com recórrer tot el WordNet i obtenir certa informació:

```
from nltk.corpus import wordnet as wn

def completa(offset):
    offset=str(offset)
    if len(offset)==1: offset="0000000"+offset
    elif len(offset)==2: offset="000000"+offset
    elif len(offset)==3: offset="00000"+offset
    elif len(offset)==4: offset="0000"+offset
    elif len(offset)==5: offset="000"+offset
    elif len(offset)==6: offset="00"+offset
    elif len(offset)==7: offset="0"+offset
    elif len(offset)==8: offset=str(offset)
    return(offset)


for synset in wn.all_synsets():
    print("SYNSET:",synset)
    offset=synset.offset()
    pos=synset.pos()
    offsetpos=completa(offset)+"-"+pos
    print("OFFSET POS:",offsetpos)
```

```
    print("LEMES:",synset.lemmas())
    for lemma in synset.lemmas():
        print(lemma.name())
    print("DEFINITION:",synset.definition())
    print("EXEMPLE:",synset.examples())
    print("-------------------------------")
```

El programa s'estarà una bona estona executant-se i el podeu parar quan vulgueu amb Ctrl+C. Veiem una mostra de la informació:

```
-------------------------------
SYNSET: Synset('inside_job.n.01')
OFFSET POS: 00767633-n
LEMES: [Lemma('inside_job.n.01.inside_job')]
inside_job
DEFINITION: some transgression committed with the assistance of someone trusted by the victim
EXEMPLE: ['the police decided that the crime was an inside job']
-------------------------------
```

Aprofitarem ara una sessió interactiva per explicar alguns aspectes d'aquest programa

El primer que fem és importar wordnet, que no és més que un lector de corpus específic. Tal i com l'importem, a wn ara tenim imporat el lector.

```
>>> from nltk.corpus import wordnet as wn
```

Podem definir un synset determinat, com per exemple el que es defineix com a car.n.1.

```
>>> synset=wn.synset("car.n.01")
```

Ara a synset tenim un objecte determinat de la classe synset. Podem accedir a la definició d'aquest synset, fent:

```
>>> synset.definition()
'a motor vehicle with four wheels; usually propelled by an internal combustion engine'
```

Cada synset té una sèrie de lemes que també són objectes determinats de la classe lemma.

```
>>> synset.lemmas()
[Lemma('car.n.01.car'), Lemma('car.n.01.auto'), Lemma('car.n.01.automobile'), Lemma('car.n.01.machine'), Lemma('car.n.01.motorcar')]
```

Per accedir a la paraula concreta podem fer servir el mètode name() de la classe lemma. Amb el bucle for recorrem tots aquests lemes i escrivim la paraula associada:

```
>>> for lemma in synset.lemmas():
```

```
...     print(lemma.name())
...
car
auto
automobile
machine
motorcar
```

També podem accedir directament al nom dels lemes fent:

```
>>> synset.lemma_names()
['car', 'auto', 'automobile', 'machine', 'motorcar']
```

Alguns synsets també tenen associats uns exemples:

```
>>> synset.examples()
['he needs a car to get to work']
```

Una manera molt habitual de referir-nos als synsets de wordnets és mitjançant el seu offset (un número concret associat al synset) i la seva categoria gramatical (pos).

Podem accedir a l'offset d'un synset amb el mètode offset:

```
>>> synset.offset()
2958343
```

i a la seva categoria gramatical amb el mètode pos()

```
>>> synset.pos()
'n'
```

És habitual donar l'offset amb un número de 8 xifres, omplint de 0 la part esquerra si és necessari. Per això hem creat la funció completa que afegeix tants zeros a l'esquerra com és necessari.

Per al nostre exemple, l'offset-pos quedaria 02958343-n

Per veure les relacions, podem fer per exemple:

```
>>> synset.hypernyms()
[Synset('motor_vehicle.n.01')]
>>> synset.hyponyms()
[Synset('ambulance.n.01'), Synset('beach_wagon.n.01'), Synset('bus.n.04'), Synset('cab.n.03'),
Synset('compact.n.03'), Synset('convertible.n.01'), Synset('coupe.n.01'), Synset('cruiser.n.01'),
Synset('electric.n.01'), Synset('gas_guzzler.n.01'), Synset('hardtop.n.01'), Synset('hatchback.n.01'),
Synset('horseless_carriage.n.01'), Synset('hot_rod.n.01'), Synset('jeep.n.01'), Synset('limousine.n.01'),
Synset('loaner.n.02'), Synset('minicar.n.01'), Synset('minivan.n.01'), Synset('model_t.n.01'),
Synset('pace_car.n.01'), Synset('racer.n.02'), Synset('roadster.n.01'), Synset('sedan.n.01'),
```

Synset('sport_utility.n.01'), Synset('sports_car.n.01'), Synset('stanley_steamer.n.01'), Synset('stock_car.n.01'), Synset('subcompact.n.01'), Synset('touring_car.n.01'), Synset('used-car.n.01')]

Recordeu que es pot accedir a la documentació d'una determinada classe fent

help(wn)

Consulteu l'ajuda per veure altres mètodes associats.

Open Multilingual WorNet

El WordNet es va desenvolupar inicialment per a l'anglès i posteriorment s'han desenvolupat wordnets per a moltes altres llengües. Alguns d'aquests wordnets tenen llicències propietàries i són d'accés restringit. Tot i això, hi ha molts wordnets lliures, entre ells el del català i castellà. El projecte Open Multilingual WordNet recopila els wordnets lliures i els publica en un format comú. NLTK implementa també l'accés fàcil a aquests wordnets.

Podem accedir a la llista de llengües amb wordnets disponibles fent:

```
>>> sorted(wn.langs())
['als', 'arb', 'bul', 'cat', 'cmn', 'dan', 'ell', 'eng', 'eus', 'fas', 'fin', 'fra', 'fre', 'glg', 'heb', 'hrv', 'ind', 'ita', 'jpn', 'nno', 'nob', 'pol', 'por', 'qcn', 'slv', 'spa', 'swe', 'tha', 'zsm']
```

Podem accedir per exemple als noms dels lemes d'un determinat synset (recordem que més endavant hem definit la variable synset com a car.n.01:

```
['auto', 'automòbil', 'cotxe', 'turisme']
>>> synset.lemma_names('jpn')
['オートモビル', 'オートモービル', '乗用車', '四輪車', '自動車', '車']
```

No tots els wordnets de les diferents llengües tenen la mateixa mida. Podem veure el nombre de lemes dels wordnets d'algunes llengües determinades:

```
>>> len(wn.all_lemma_names(pos='n', lang='eng'))
119034
>>> len(wn.all_lemma_names(pos='n', lang='cat'))
38736
>>> len(wn.all_lemma_names(pos='n', lang='jpn'))
64797
>>> len(wn.all_lemma_names(pos='n', lang='hrv'))
19623
```