NLTK (I installed with sudo pip install nltk) can download some corpora for you. This seems to be a mixture of whole corpora, partial corpora and stubs for corpora.

- http://nltk.org
- http://nltk.org/nltk_data/
- http://freecode.com/articles/processing-corpora-with-python-and-the-natural-language-toolkit

In [120]:
```
import nltk
nltk.download('propbank')
```

```
[nltk_data] Downloading package 'propbank' to /Users/alex/nltk_data...
[nltk_data]   Package propbank is already up-to-date!
```

Out[120]: True

In [95]:
```
from nltk.corpus import propbank
def vprt(nm,v): print '%-14s %s' % (nm+':',v)
def ctr(msg): print msg.center(80,'_')
def ctrs(msg): print msg.center(50,'_')
```

All the files available in the downloaded NLTK version of propbank. Not sure what 'propbank_ptr' corpus is or how it is different.

In [2]:
```
print propbank.fileids()[:10]
```

```
['prop.txt', 'verbs.txt', 'frames/abandon.xml', 'frames/abate.xml', 'frames/abdicate.xml',
'frames/abduct.xml', 'frames/abet.xml', 'frames/abide.xml', 'frames/abolish.xml',
'frames/abominate.xml']
```

They give us a list of the verbs. Not sure how useful this is to us.

In [32]:
```
v = propbank.verbs()
print v
print v.fileid
```

```
['abandon', 'abate', 'abdicate', 'abet', 'abide', ...]
/Users/alex/nltk_data/corpora/propbank/verbs.txt
```

We'll get general information about the roles for each predicate from the rolesets. Here's the frame for 'acquire.'

In [65]:
```
f = propbank.raw('frames/acquire.xml')
print f
```

```
<!DOCTYPE frameset SYSTEM "frameset.dtd">
<frameset>
<predicate lemma="acquire">
  <note>
    based on survey of initial sentences of big corpus
    and comparison with 'gain' and 'buy'
  </note>
<roleset id="acquire.01" name="get, acquire" vncls="13.5.2-1 14">
<roles>
  <role descr="agent, entity acquiring something" n="0">

  <vnrole vncls="14" vntheta="Agent"/><vnrole vncls="13.5.2-1" vntheta="Agent"/></role>

  <role descr="thing acquired" n="1">

  <vnrole vncls="14" vntheta="Topic"/><vnrole vncls="13.5.2-1" vntheta="Theme"/></role>

  <role descr="seller" n="2">

  <vnrole vncls="14" vntheta="Source"/><vnrole vncls="13.5.2-1" vntheta="Source"/></role>
```

```
  <role descr="price paid" n="3">

  <vnrole vncls="13.5.2-1" vntheta="Asset"/></role>

  <role descr="benefactive" n="4"/>
</roles>

<example>
  <text>
    New England Electric will acquire PS of New Hampshire.
  </text>
  <arg n="0">New England Electric</arg>
  <rel>acquire</rel>
  <arg n="1">PS of New Hampshire.</arg>
</example>

<example>
  <text>
    Its Moleculon affiliate acquired Kalipharma Inc for $23 million.
  </text>
  <arg n="0">Its Moleculon affiliate</arg>
  <rel>acquired</rel>
  <arg n="1">Kalipharma Inc</arg>
  <arg f="for" n="3">for $23 million.</arg>
</example>
</roleset>
</predicate>
</frameset>
```

Now let's get out the argument names.

In [113]:
```
for id in ['acquire.01','purchase.01','belly-flop.01']:
    ctrs(id)
    rs = propbank.roleset(id)
    vprt('vncls', rs.attrib['vncls'])
    vprt('id', rs.attrib['id'])
    vprt('name', rs.attrib['name'])
    roles = rs[0]
    for i, role in enumerate(roles.findall('role')):
        print role.attrib['n'], role.attrib['descr']
```

```
____________________acquire.01____________________
vncls:          13.5.2-1 14
id:             acquire.01
name:           get, acquire
0 agent, entity acquiring something
1 thing acquired
2 seller
3 price paid
4 benefactive
____________________purchase.01____________________
vncls:          13.5.2-1
id:             purchase.01
name:           buy
0 purchaser
1 thing purchased
2 seller
3 price paid
4 benefactive
____________________belly-flop.01____________________
vncls:          -
id:             belly-flop.01
```

```
name:          fail spectacularly
1 thing failing
```

Let's make a function to get descriptions of args.

In [106]:
```
def rsargs(id):
    rs = propbank.roleset(id)
    args = {}
    roles = rs[0]
    for i, role in enumerate(roles.findall('role')):
        args['ARG'+role.attrib['n']] = role.attrib['descr']
    return args
rsargs('acquire.01')
```

Out[106]:
```
{'ARG0': 'agent, entity acquiring something',
 'ARG1': 'thing acquired',
 'ARG2': 'seller',
 'ARG3': 'price paid',
 'ARG4': 'benefactive'}
```

The instances are the individual verb annotations.

In [38]:
```
print propbank.instances()[:2]
```

```
[<PropbankInstance: wsj_0001.mrg, sent 0, word 8>, <PropbankInstance: wsj_0001.mrg, sent 1,
word 10>]
```

Let's loop through some instances of the predicate 'acquire'. Note that in the arguments, we are often getting an entire tree but if we want to pull out say, just the proper nouns (NNP) we could attempt to do that. The proper nouns might be a better starting point if we wanted to match the entities to another data source, e.g. Freebase.

In [117]:
```
for baseform in ['acquire','purchase']:
    for i in [i for i in propbank.instances()[:2000] if i.baseform == baseform][:3]:
        ctr(i.baseform)
        vprt('fileid',i.fileid)
        vprt('sentnum',i.sentnum)
        vprt('wordnum',i.wordnum)
        vprt('roleset',i.roleset)
        args = rsargs(i.roleset)
        vprt('inflection',i.inflection)
        vprt('tagger',i.tagger)

        ctrs('sentence')
        print ' '.join(i.tree.leaves())

        ctrs('predicate')
        vprt('wordnum',i.predicate.wordnum)
        vprt('height',i.predicate.height)
        vprt('word', ''.join(i.predicate.select(i.tree)))
        print i.predicate.select(i.tree)
        #vprt('parse_corpus',i.parse_corpus)

        for a in i.arguments:
            id = a[1]
            ctrs(id)
            if id in args:
                vprt('descr', args[a[1]])
                vprt('loc',a[0])
                t = a[0].select(i.tree)
                vprt('arg', ' '.join(t.leaves()))
                print t

        ctrs('tree')
```

```
        #vprt('tree leaves', i.tree.leaves())
        print i.tree
        print ''
```

```
_____________________________________acquire_____________________________________
fileid:         wsj_0013.mrg
sentnum:        2
wordnum:        20
roleset:        acquire.01
inflection:     i---a
tagger:         gold
_____________________sentence_____________________
New England Electric , based * in Westborough , Mass. , had offered $ 2 billion *U* *ICH*-1
*-4 to acquire PS of New Hampshire , well below the $ 2.29 billion *U* value 0 United
Illuminating places *T*-2 on its bid and the $ 2.25 billion *U* 0 Northeast says 0 its bid
is worth *T*-3 .
_____________________predicate_____________________
wordnum:        20
height:         0
word:           acquire
(VB acquire)
_____________________ARG0_____________________
descr:          agent, entity acquiring something
loc:            0:2*18:0
arg:            New England Electric , based * in Westborough , Mass. , *-4
(*CHAIN*
  (NP-SBJ-4
    (NP (NNP New) (NNP England) (NNP Electric))
    (, ,)
    (VP
      (VBN based)
      (NP (-NONE- *))
      (PP-LOC-CLR
        (IN in)
        (NP (NP (NNP Westborough)) (, ,) (NP (NNP Mass.)))))
    (, ,))
  (-NONE- *-4))
_____________________ARG1_____________________
descr:          thing acquired
loc:            21:2
arg:            PS of New Hampshire
(NP (NP (NNP PS)) (PP (IN of) (NP (NNP New) (NNP Hampshire))))
_____________________tree_____________________
(S
  (NP-SBJ-4
    (NP (NNP New) (NNP England) (NNP Electric))
    (, ,)
    (VP
      (VBN based)
      (NP (-NONE- *))
      (PP-LOC-CLR
        (IN in)
        (NP (NP (NNP Westborough)) (, ,) (NP (NNP Mass.)))))
    (, ,))
  (VP
    (VBD had)
    (VP
      (VBN offered)
      (NP
        (NP (QP ($ $) (CD 2) (CD billion)) (-NONE- *U*))
        (PP (-NONE- *ICH*-1)))
      (S-PRP
        (NP-SBJ (-NONE- *-4))
```

```
        (VP
          (TO to)
          (VP
            (VB acquire)
            (NP
              (NP (NNP PS))
              (PP (IN of) (NP (NNP New) (NNP Hampshire)))))))
      (, ,)
      (PP-1
        (ADVP (RB well))
        (IN below)
        (NP
          (NP
            (NP
              (NP
                (DT the)
                (ADJP (QP ($ $) (CD 2.29) (CD billion)) (-NONE- *U*))
                (NN value))
              (SBAR
                (WHNP-2 (-NONE- 0))
                (S
                  (NP-SBJ (NNP United) (NNP Illuminating))
                  (VP
                    (NNS places)
                    (NP (-NONE- *T*-2))
                    (PP-DIR (IN on) (NP (PRP$ its) (NN bid)))))))
            (CC and)
            (NP
              (NP
                (DT the)
                (QP ($ $) (CD 2.25) (CD billion))
                (-NONE- *U*))
              (SBAR
                (WHNP-3 (-NONE- 0))
                (S
                  (NP-SBJ (NNP Northeast))
                  (VP
                    (VBZ says)
                    (SBAR
                      (-NONE- 0)
                      (S
                        (NP-SBJ (PRP$ its) (NN bid))
                        (VP
                          (VBZ is)
                          (ADJP-PRD (IN worth) (NP (-NONE- *T*-3)))))))))))))
  (. .))

_____________________________________acquire_____________________________________
fileid:        wsj_0023.mrg
sentnum:       0
wordnum:       16
roleset:       acquire.01
inflection:    vp--a
tagger:        gold
_____________________sentence_____________________
F.H. Faulding & Co. , an Australian pharmaceuticals company , said 0 its Moleculon Inc.
affiliate acquired Kalipharma Inc. for $ 23 million *U* .
_____________________predicate_____________________
wordnum:       16
height:        0
word:          acquired
(VBD acquired)
_____________________ARG0_____________________
descr:         agent, entity acquiring something
loc:           12:1
arg:           its Moleculon Inc. affiliate
(NP-SBJ (PRP$ its) (NNP Moleculon) (NNP Inc. ) (NN affiliate))
```

(NP-SBJ (PRP$ its) (NNP Moleculon) (NNP Inc.) (NN affiliate))

_____ARG1_____
descr: thing acquired
loc: 17:1
arg: Kalipharma Inc.
(NP (NNP Kalipharma) (NNP Inc.))
_____ARG3-for_____
_____tree_____
```
(S
  (NP-SBJ
    (NP (NNP F.H.) (NNP Faulding) (CC &) (NNP Co.))
    (, ,)
    (NP (DT an) (JJ Australian) (NNS pharmaceuticals) (NN company))
    (, ,))
  (VP
    (VBD said)
    (SBAR
      (-NONE- 0)
      (S
        (NP-SBJ (PRP$ its) (NNP Moleculon) (NNP Inc.) (NN affiliate))
        (VP
          (VBD acquired)
          (NP (NNP Kalipharma) (NNP Inc.))
          (PP-CLR
            (IN for)
            (NP (QP ($ $) (CD 23) (CD million)) (-NONE- *U*)))))))
  (. .))
```

_____acquire_____
fileid: wsj_0023.mrg
sentnum: 2
wordnum: 18
roleset: acquire.01
inflection: i---a
tagger: gold
_____sentence_____
Faulding said 0 it owns 33 % of Moleculon 's voting stock and has an agreement * to acquire an additional 19 % .
_____predicate_____
wordnum: 18
height: 0
word: acquire
(VB acquire)
_____ARG0_____
descr: agent, entity acquiring something
loc: 3:1*16:0
arg: it *
(*CHAIN* (NP-SBJ (PRP it)) (-NONE- *))
_____ARG1_____
descr: thing acquired
loc: 19:1
arg: an additional 19 %
(NP (DT an) (JJ additional) (CD 19) (NN %))
_____tree_____
```
(S
  (NP-SBJ (NNP Faulding))
  (VP
    (VBD said)
    (SBAR
      (-NONE- 0)
      (S
        (NP-SBJ (PRP it))
        (VP
          (VP
            (VBZ owns)
```

```
            (NP
              (NP (CD 33) (NN %))
              (PP
                (IN of)
                (NP
                  (NP (NNP Moleculon) (POS 's))
                  (NN voting)
                  (NN stock)))))
          (CC and)
          (VP
            (VBZ has)
            (NP
              (DT an)
              (NN agreement)
              (S
                (NP-SBJ (-NONE- *))
                (VP
                  (TO to)
                  (VP
                    (VB acquire)
                    (NP (DT an) (JJ additional) (CD 19) (NN %)))))))))
  (. .))

_____________________________________purchase_____________________________________
fileid:        wsj_0036.mrg
sentnum:       4
wordnum:       1
roleset:       purchase.01
inflection:    p---a
tagger:        gold
_____________________sentence_____________________
The purchasing managers , however , also said that orders turned up in October after four
months of decline .
_____________________predicate_____________________
wordnum:       1
height:        0
word:          purchasing
(VBG purchasing)
_____________________ARG0_____________________
descr:         purchaser
loc:           2:0
arg:           managers
(NNS managers)
_____________________tree_____________________
(S
  (NP-SBJ (DT The) (VBG purchasing) (NNS managers))
  (, ,)
  (ADVP (RB however))
  (, ,)
  (ADVP (RB also))
  (VP
    (VBD said)
    (SBAR
      (IN that)
      (S
        (NP-SBJ (NNS orders))
        (VP
          (VBD turned)
          (ADVP-CLR (RB up))
          (PP-TMP (IN in) (NP (NNP October)))
          (PP-TMP
            (IN after)
            (NP
              (NP (CD four) (NNS months))
              (PP (IN of) (NP (NN decline)))))))))
  (. .))
```

```
  (. .))

_____________________________________purchase_____________________________________
fileid:        wsj_0036.mrg
sentnum:       45
wordnum:       5
roleset:       purchase.01
inflection:    p---a
tagger:        gold
_____________________sentence_____________________
Only 19 % of the purchasing managers reported better export orders in October , down from
27 % in September .
_____________________predicate____________________
wordnum:       5
height:        0
word:          purchasing
(VBG purchasing)
_____________________ARG0_____________________
descr:         purchaser
loc:           6:0
arg:           managers
(NNS managers)
_____________________tree_____________________
(S
  (NP-SBJ
    (NP
      (NP (QP (RB Only) (CD 19)) (NN %))
      (PP (IN of) (NP (DT the) (VBG purchasing) (NNS managers)))))
  (VP
    (VBD reported)
    (NP (JJR better) (NN export) (NNS orders))
    (PP-TMP (IN in) (NP (NNP October)))
    (, ,)
    (ADVP
      (RB down)
      (PP
        (IN from)
        (NP (CD 27) (NN %))
        (PP-TMP (IN in) (NP (NNP September))))))
  (. .))

_____________________________________purchase_____________________________________
fileid:        wsj_0036.mrg
sentnum:       48
wordnum:       6
roleset:       purchase.01
inflection:    p---a
tagger:        gold
_____________________sentence_____________________
For the fifth consecutive month , purchasing managers said 0 prices for the goods 0 they
purchased *T*-1 fell .
_____________________predicate____________________
wordnum:       6
height:        0
word:          purchasing
(VBG purchasing)
_____________________ARG0_____________________
descr:         purchaser
loc:           7:0
arg:           managers
(NNS managers)
_____________________tree_____________________
(S
  (PP-TMP
    (IN For)
    (NP (DT the) (JJ fifth) (JJ consecutive) (NN month)))
```

```
  (NP (DT the) (JJ fifth) (JJ consecutive) (NN month)))
(, ,)
(NP-SBJ (VBG purchasing) (NNS managers))
(VP
  (VBD said)
  (SBAR
    (-NONE- 0)
    (S
      (NP-SBJ
        (NP (NNS prices))
        (PP
          (IN for)
          (NP
            (NP (DT the) (NNS goods))
            (SBAR
              (WHNP-1 (-NONE- 0))
              (S
                (NP-SBJ (PRP they))
                (VP (VBD purchased) (NP (-NONE- *T*-1))))))))
      (VP (VBD fell)))))
(. .))
```

In [ ]: