



NLM

Fellows  
'14

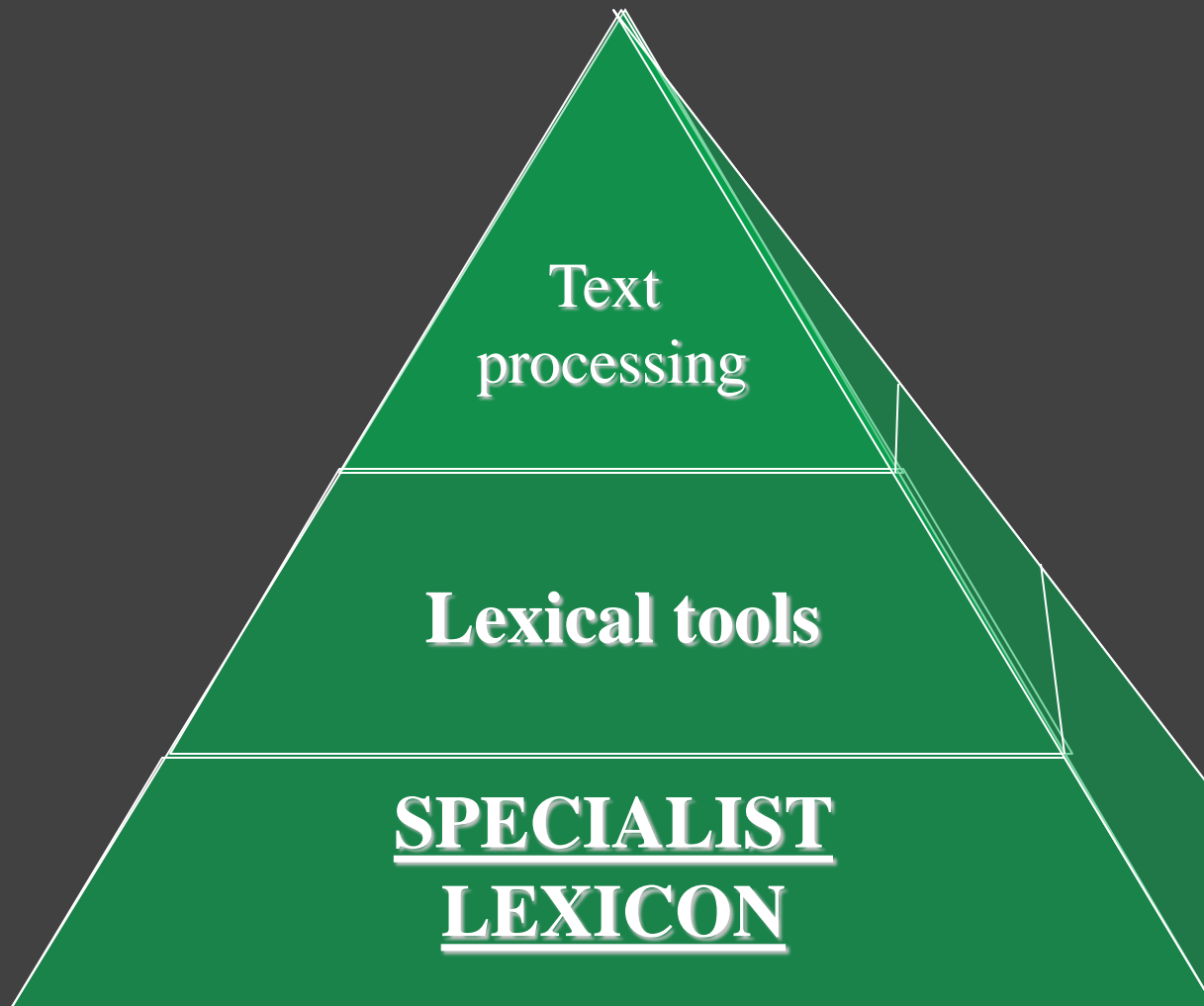
June 11, 2014



Allen Browne  
Chris Lu

The SPECIALIST Lexicon and Lexical Tools

June 11, 2014

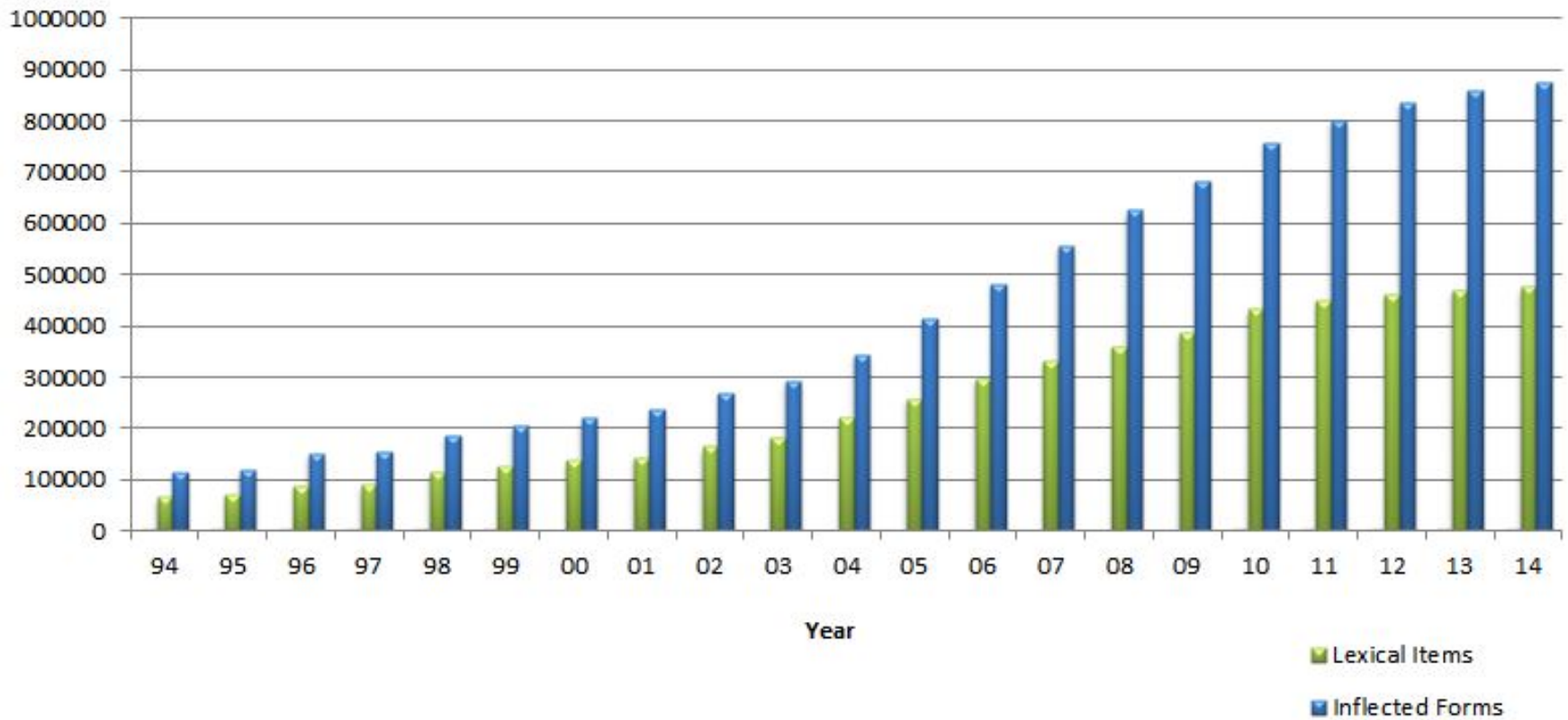


# The SPECIALIST lexicon

- A syntactic lexicon
- Biomedical and general English
- Over 476,000 records

# Lexicon Growth

## SPECIALIST LEXICON Growth

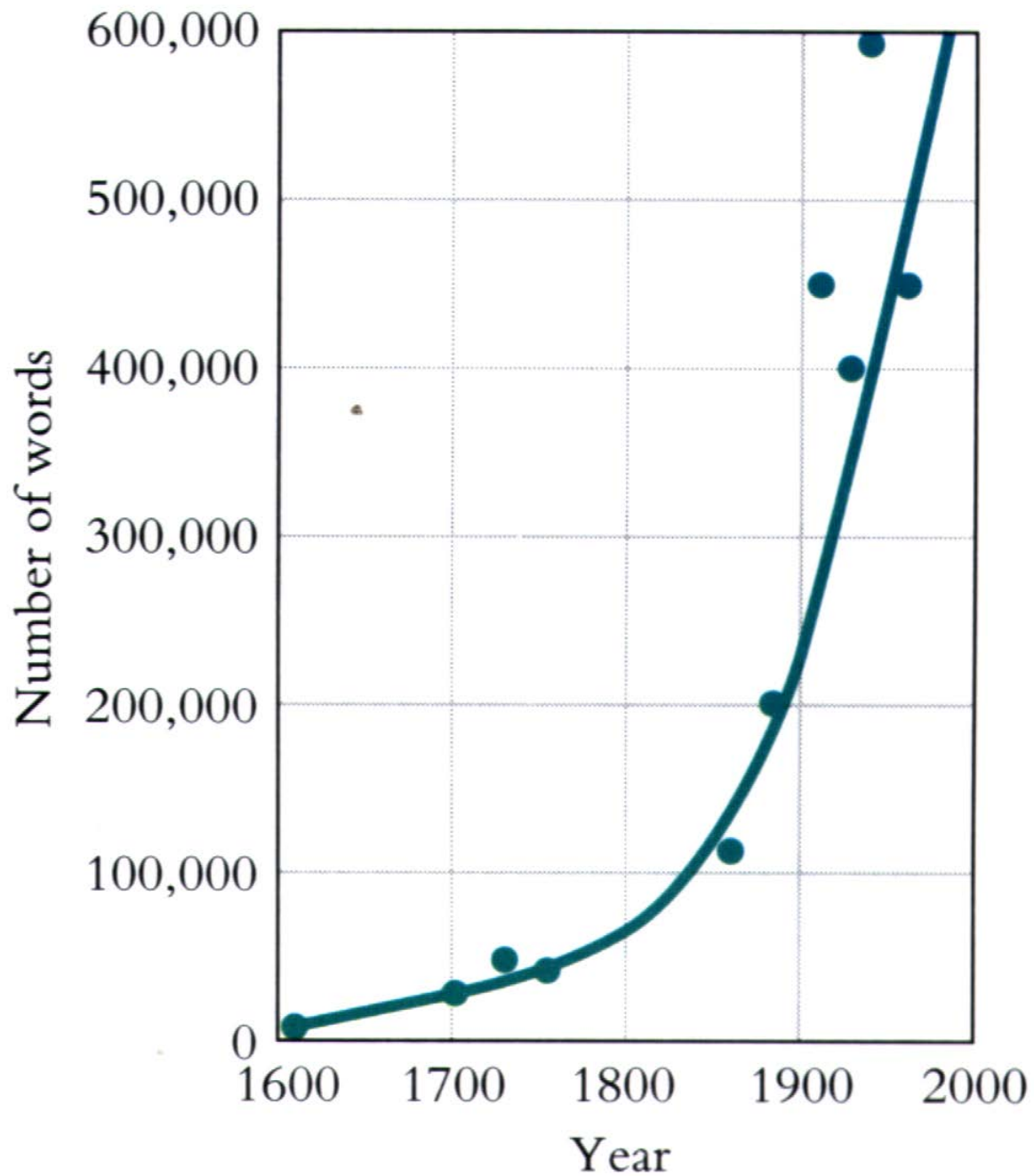


village  
square the circle  
square  
square  
fair and square root

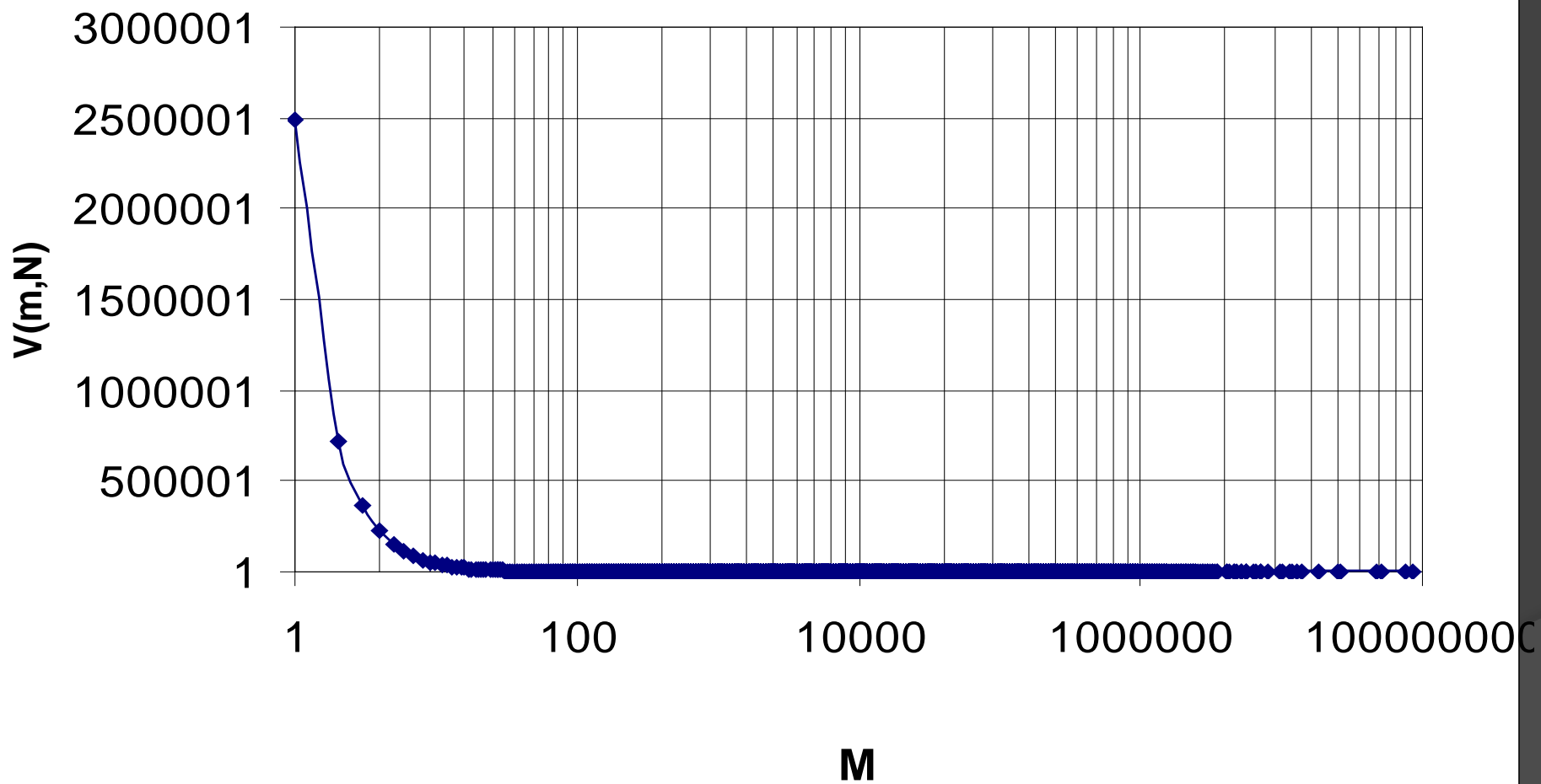
George A.  
Miller

The Science  
of Words

1991

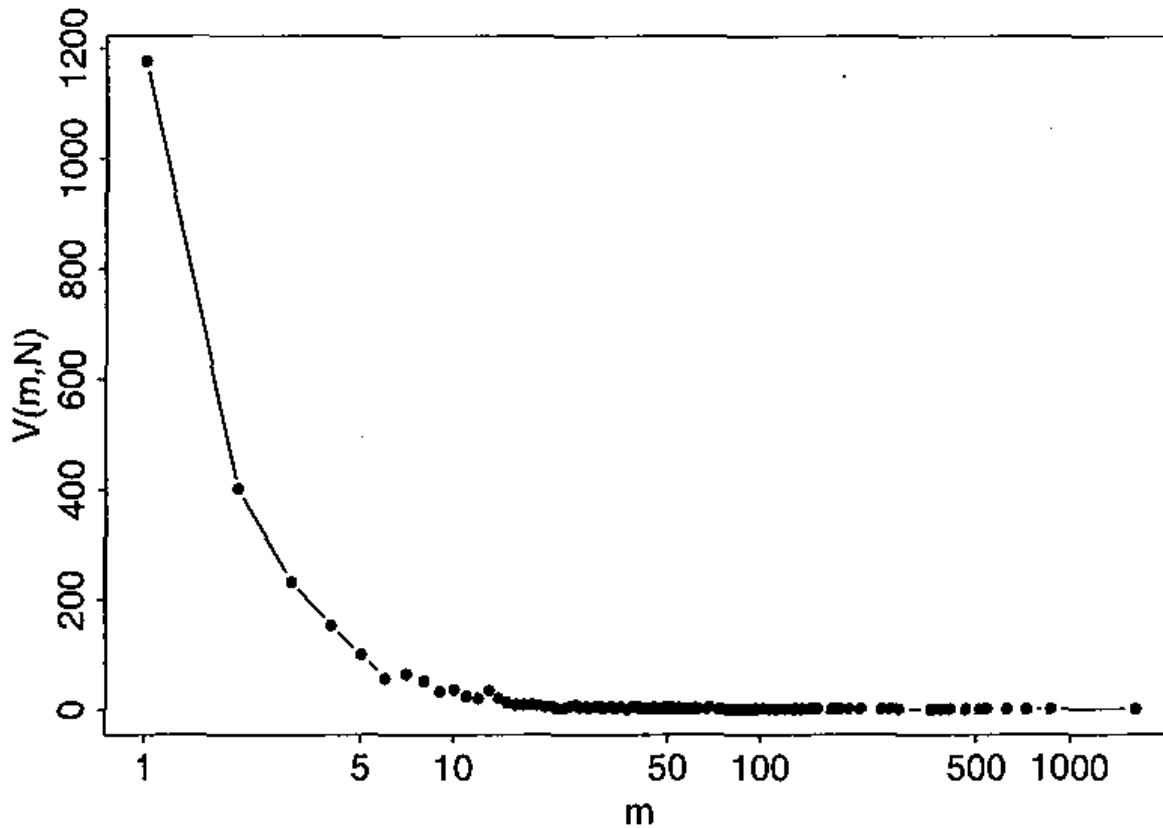


# Frequency Spectrum of Medline 2006

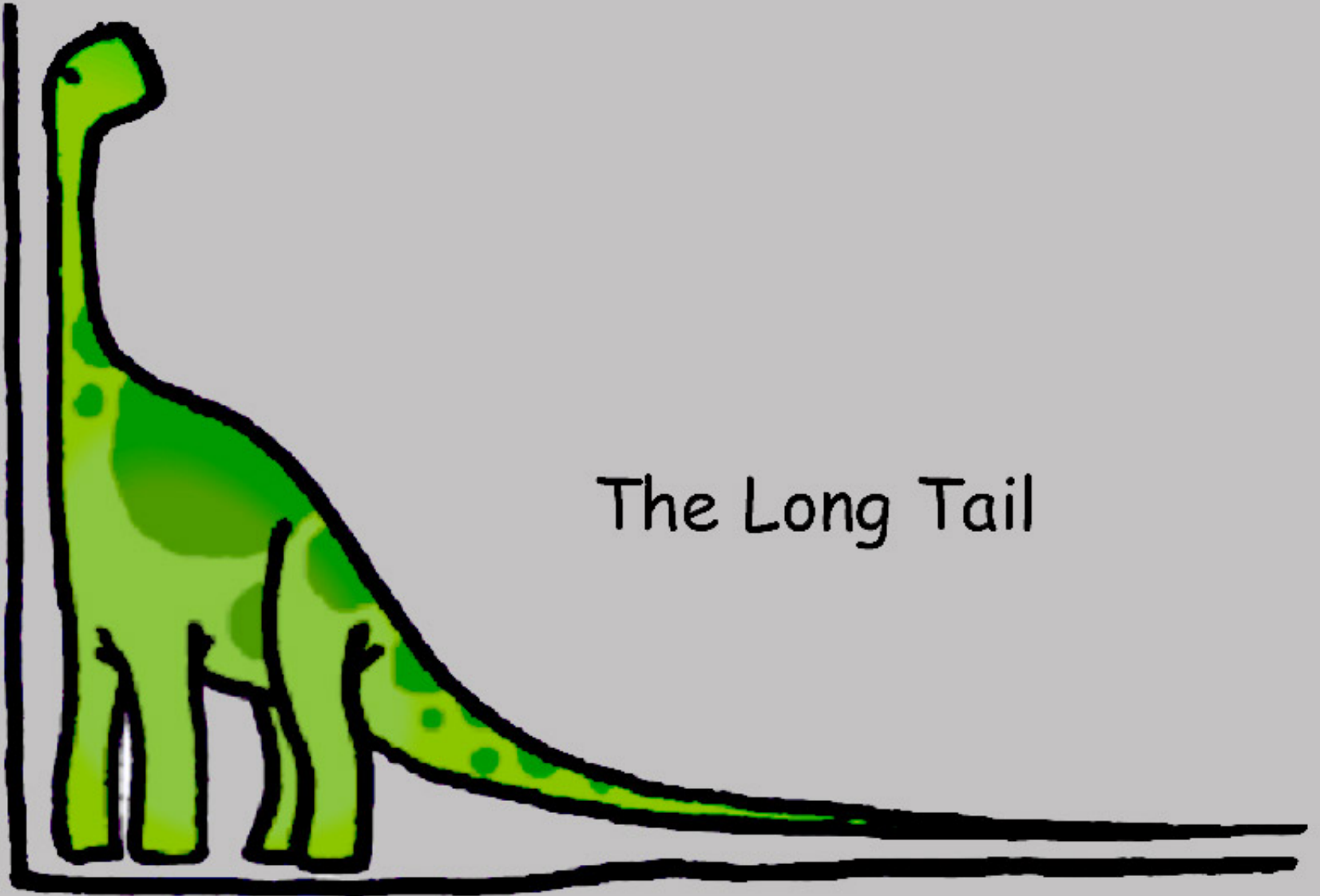




# Frequency Spectrum: Alice in Wonderland



Number of Words ->



Word Frequency ->

# The SPECIALIST LEXICON

- ⊙ Morphology
  - Inflection
  - Derivation
- ⊙ Orthography
  - Spelling variants
- ⊙ Syntax
  - Complementation for verbs, nouns, and adjectives

# Morphology

## ⦿ Inflectional

- nucleus, nuclei
- cauterize, cauterizes, cauterized, cauterizing
- red, redder reddest

## ⦿ Derivational

- laryngeal -- larynx
- transport -- transportation



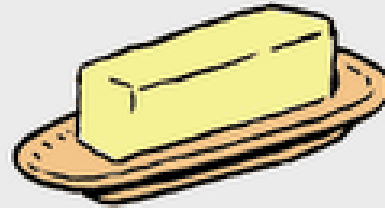
# Derivational Morphology



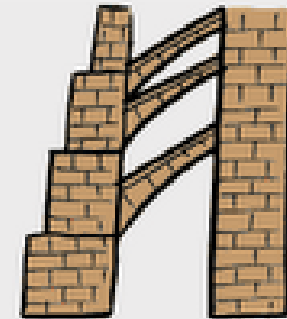
WAITER



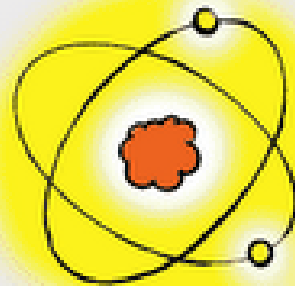
WAITRESS



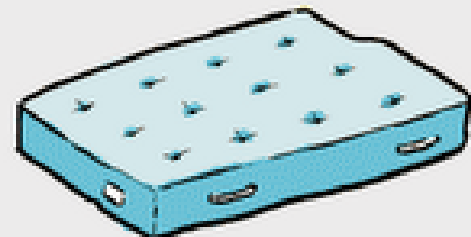
BUTTER



BUTTRESS



MATTER



MATTRESS

© 2006 Stivess

# Orthography

## Spelling Variation

- ⦿ align -- aline
- ⦿ Grave's disease -- Graves's disease -- Graves' disease
- ⦿ anesthetize -- anesthetise
- ⦿ Esophagus -- oesophagus
- ⦿ foetus – fetus
- ⦿ centre -- center

# Orthography

**HAGAR THE HORRIBLE** CHRIS BROWNE





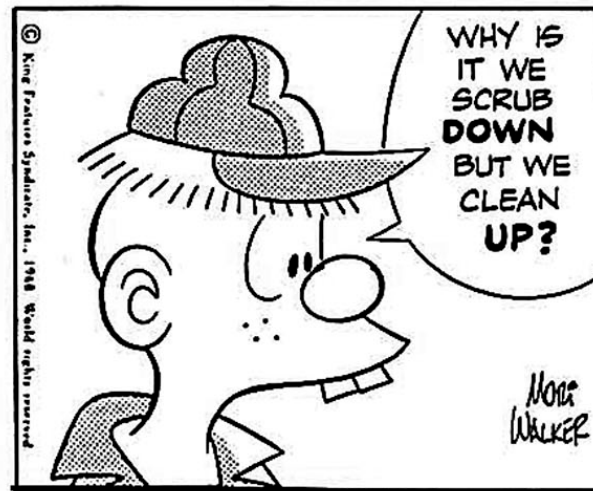
# Syntax -- Verb Complements

- ⊙ intran
  - I'll treat.
- ⊙ tran=np
  - He treated the patient.
- ⊙ ditran=np,pphr(with,np)
  - She treated the patient with the drug.

# Syntax -- Verb Complements

```
{base=treat
entry=E0061964
  cat=verb
  variants=reg
  intran
  tran=np
  tran=pphr(with,np)
  tran=pphr(of,np)
  ditran=np,pphr(to,np)
  ditran=np,pphr(with,np)
  ditran=np,pphr(for,np)
  cplxtran=np,advbl
  nominalization=treatment|noun|E0061968
}
```

# Syntax -- Verb Particle Constructions

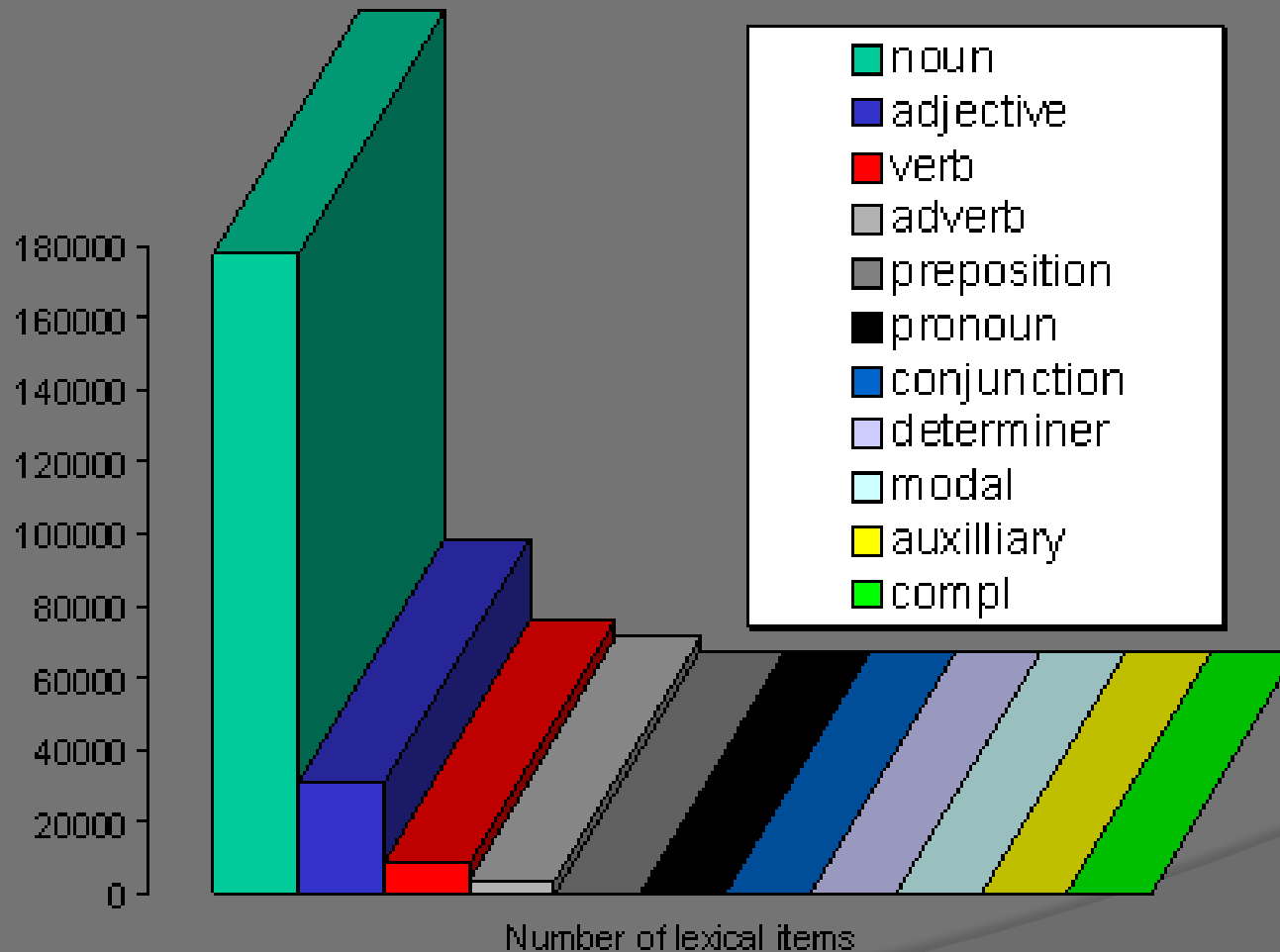


© King Features Syndicate.

clean up  
scrub down  
look up

```
{base=clean
  entry=E0017272
  cat=verb
  variants=reg
  intran
  intran;part(up)
  tran=np
  tran=np;part(up)
  nominalization=clean|noun|E0017273
  nominalization=cleanup|noun|E0319808}
```

# Categories – Parts of Speech



# Lexicon Unit Records

```
{ base=Kaposi's sarcoma
spelling_variant=Kaposi sarcoma
entry=E0003576
    cat=noun
    variants=uncount
    variants=reg
    variants=glreg
}
```

```
{ base=chronic
entry=E0016869
    cat=adj
    variants=inv
    position=attrib(1)
    position=pred
    stative
}
```

```
{ base=aspirate
entry=E0010803
    cat=verb
    variants=reg
    tran=np
    nominalization=aspiration|noun|E0010804
}
```

```
{ base=in
entry=E0033870
    cat=prep
}
```

# Orthographic vs. Lexicographic Word:

Why, for instance, if a two-word boy scout feels chilly on his one-word campground, does he pull up a two-word camp chair in front of his one-word campfire? Anyone who seeks a strictly logical answer to such questions is chasing will-o'-the-wisps (chargeable in telegrams as a single word, because of the hyphens) in a semantic bog.

# UTF-8

```
{base=resume  
spelling_variant=résumé  
spelling_variant=resumé  
entry=E0053099  
    cat=noun  
    variants=reg  
}
```

```
{base=role  
spelling_variant=rôle  
entry=E0053757  
    cat=noun  
    variants=reg  
}
```

```
{base=deja vu  
spelling_variant=deja-vu  
spelling_variant=déjà vu  
entry=E0021340  
    cat=noun  
    variants=uncount  
}
```

```
{base=cafe  
spelling_variant=café  
entry=E0420690  
    cat=noun  
    variants=reg  
}
```



# Noun Variants

```
{base=Kaposi's sarcoma  
spelling_variant=Kaposi sarcoma  
entry=E0003576  
  cat=noun  
  variants=uncount  
  variants=reg  
  variants=glreg  
}
```

- ⦿ Kaposi's sarcoma
- ⦿ Kaposi's sarcomas
- ⦿ Kaposi's sarcomata
- ⦿ Kaposi sarcoma
- ⦿ Kaposi sarcomas
- ⦿ Kaposi sarcomata

# Regular Nouns

**The plural suffix is *s*.**

*y* becomes *ie* following a consonant before *s*.

*e* is inserted before *s* if the base ends in *s*, *z*, *x*, *ch*, or *s*

Leach – Leaches

Stomach – Stomachs ← irregular

# Greco-latin Regular nouns

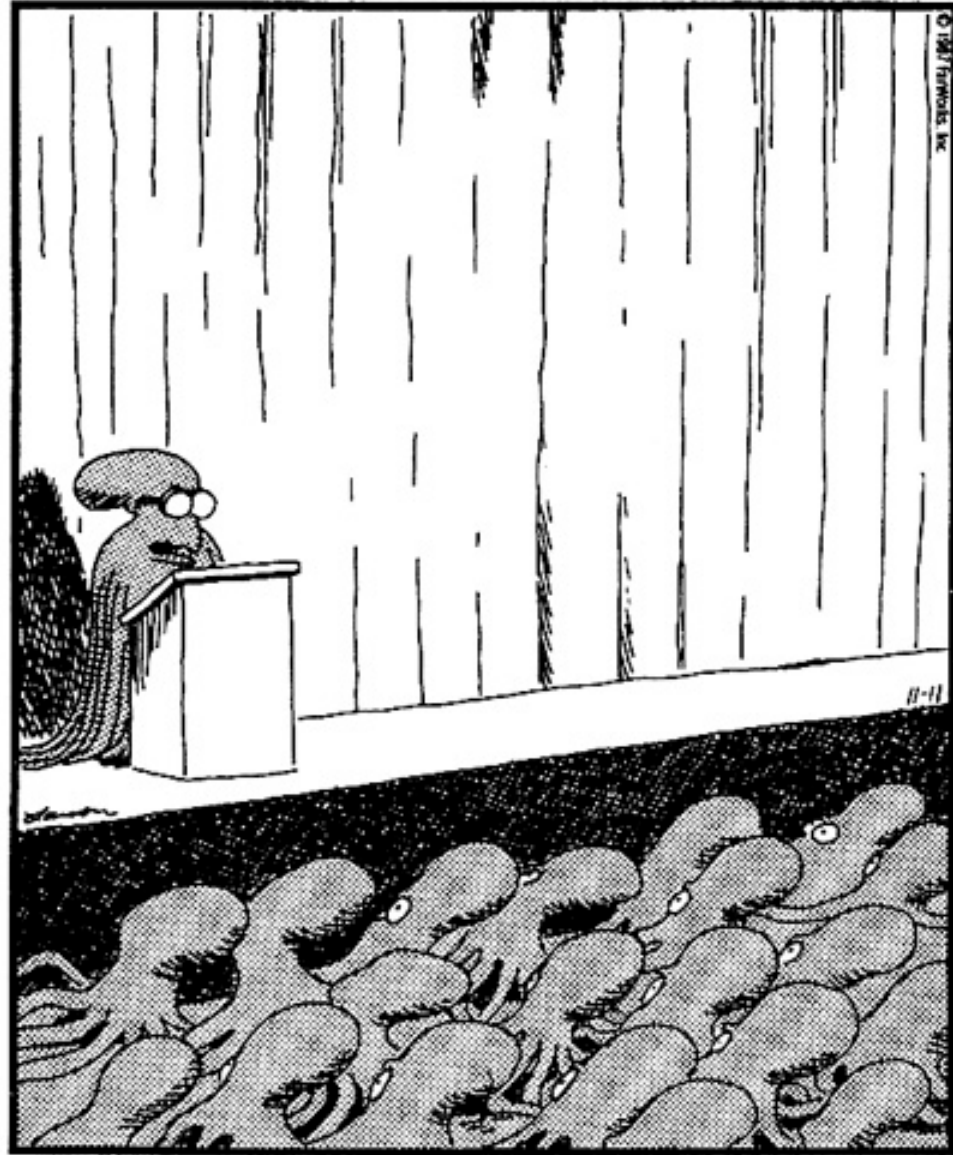
<b>singular ends with:</b>	<b>plural ends with:</b>	<b>Examples</b>
-us	-i	focus/foci
-ma	-mata	trauma/traumata
-a	-ae	larva/larvae
-um	-a	ilium/ilia
-on	-a	taxon/taxa
-sis	-ses	analysis/analyses
-is	-ides	cystis/cystides
-men	-mina	foramen/foramina
-ex	-ices	index/indices
-x	-ces	matrix/matrices

# Octopuses

```
{base=octopus  
entry=E0204527  
cat=noun  
variants=reg  
variants=greg  
}
```

## THE FAR SIDE

By GARY LARSON



"Fellow octopi, or octopuses ... octopi? ... Dang, it's hard to start a speech with this crowd."

All rights reserved.

THE FAR SIDE © 1987 FARWORKS, INC.

# Uncount Nouns

(abstract or mass)

```
{base=smallpox  
entry=E0056359  
  cat=noun  
  variants=uncount  
}
```

```
{base=potassium  
entry=E0049387  
  cat=noun  
  variants=uncount  
}
```

- ⦿ \* a smallpox
- ⦿ \* two smallpoxes
- ⦿ much smallpox
- ⦿ \* a potassium
- ⦿ \* two potassiums
- ⦿ much potassium

\* This form does not occur

# Countability

## ● Mail

\* A mail

much mail

\* many mails

## ● E-Mail

An e-mail

much e-mail

many e-mails

\* This form does not occur

# Uncount Nouns

Singulars



Spaghetti



Broccolo



Jacuzzi

EARLY  
+ 1/2

# Fixed Plural Nouns

```
{base=police  
entry=E0048616  
  cat=noun  
  variants=plur  
}
```

```
{base=scissors  
entry=E0054633  
  cat=noun  
  variants=plur  
}
```



# Irregular Nouns

```
{base=corpus  
entry=E0019113  
  cat=noun  
  variants=irreg|corpora|  
  variants=reg  
}
```

```
{base=larynx  
entry=E0036919  
  cat=noun  
  variants=irreg|larynges|  
  variants=reg  
}
```

# Regular Verbs

- ◎ The third person present tense suffix is *s*.
  - *y* becomes *ie* following a consonant before *s*.
  - *e* is inserted between *z*, *x*, *ch*, or *sh* and *s*.
- ◎ The past tense suffix is *ed*.
  - *y* becomes *ie* following a consonant before *ed*.
  - Final *e* is deleted before *ed*.
- The past participle is the same as the past tense.
- The present participle suffix is *ing*.
  - *y* becomes *ie* following a consonant before *ing*.
  - Final *e* is deleted before *ing*  
unless preceded by *e*, *y* or *o*.

# Regular Verbs

- ⦿ dismiss: dismisses, dismissed, dismissing
- ⦿ agree: agrees; agreed; agreeing
- ⦿ dry: dries, dried, drying

# Regular Doubling Verbs

- End in a CVC pattern
- Double the final consonant before *ed* and *ing*.
- Are otherwise regular
- variants=regd

control: controls, controlled, controlling

# Irregular Verbs



Bite: bite, bites, bit, bitten

# Irregular Verbs

**{base=bite**

**entry=E0013219**

**cat=verb**

**variants=irreg|bite|bites|bit|bitten|biting|**

**intran**

**tran=np**

**cplxtran=np,advbl**

**}**

# Ancillary Data Bases

- ◎ Synonymy
  - sm.db
- ◎ Derivation
  - dm.db, dm.rules
- ◎ Inflection
  - im.rules
- ◎ Neoclassical compounds
  - nc.db



# Derivational Facts and Rules

**dm.facts**

**treatment|noun|treat|verb**

**prohibition|noun|prohibitive|adj**

**cell lineage|noun|cell line|noun**

**photochemotherapeutic|adj|photochemotherapy|noun**

**pharmacotherapeutic|adj|pharmacotherapy|noun**



# Derivational Facts and Rules

**dm.rules**

**# e.g. alienation|alienate**

**ation\$|noun|ate|verb**

**ration|rate; station|state;**

# Inflectional Facts and Rules

**im.rules**

**# Noun rules (ggreg)**

**us\$|noun|singular|i\$|noun|plural**

**antus|anti;**

**ma\$|noun|singular|mata\$|noun|plural**

**a\$|noun|singular|ae\$|noun|plural**

**um\$|noun|singular|a\$|noun|plural**

**on\$|noun|singular|a\$|noun|plural**

**sis\$|noun|singular|ses\$|noun|plural**

**is\$|noun|singular|ides\$|noun|plural**

**men\$|noun|singular|mina\$|noun|plural**

**ex\$|noun|singular|ices\$|noun|plural**

**x\$|noun|singular|ces\$|noun|plural**

# Neoclassical compounds

## **nc.db**

abdomin(o)|abdomen|root

ab|away from|prefix

acanth(o)|prickle|root

acar(o)|mite|root

acetabul(o)|acetabulum|root

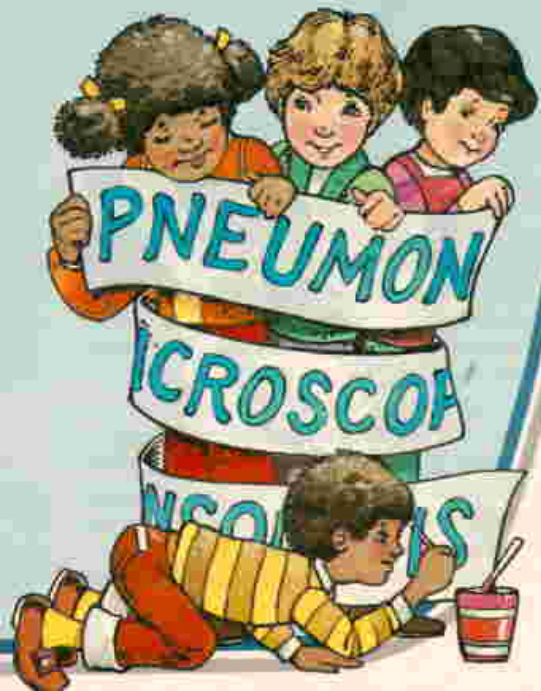
ad|towards|prefix

agogue|inducing|terminal

albumin(o)|albumin|root

sis|condition|terminal

stomy|surgical opening|terminal



# PNEUMONOUltrAMICROSCOPICS ILICOVOLCANOCONIOSIS

pneu.mo.no.ul.tra.mi.cro.scop.ic.sil.i.co.vol.ca.no.co.ni.o.sis  
\  
'n(y)u:-m\*-(.)no--.\*l-tr\*-.mi--kr\*-'ska:p-ik-'sil-i-(.)ko--(.)v\  
n [NL, fr. Gk pneumo-n + ISV ultramicroscopic + NL  
silicon +]a:l-'ka--no--.ko--ne--'o--s\*s ISV volcano + Gk  
konis dust : **a pneumoconiosis caused by the inhalation of  
very fine silicate or quartz dust**

-- Merriam Webster's 3rd International Dictionary, page 1747.

## The Protein of a tobacco mosaic virus, Dahlemense strain

acetylseryltyrosylserylisoleucylthreonylserylproylserylglutami  
nylphenylalanylvalylphenylalanylleucylserylserylvalyltryptoph  
ylalanylasparylproylisoleucylglutamylleucylleucylasparaginy  
lvalylcysteinylthreonylserylserylleucylglycylasparaginyglutami  
nylphenylalanylglutaminythreonylglutaminyglutaminyalanyla  
rginythreonythreonylglutaminyvalylglutaminyglutaminyphe  
nylalanylserylglutaminyvalyltryptophyllsylproylphenylalanyl  
proylglutaminyserylthreonylvalylarginylphenylalanylproylgly  
cylasparylvalyltyrosylsystylvalyltyrosylarginyltyrosylasparaginy  
lalanylvalylleucylasparylproylleucylisoleucylthreonylalanylleu  
cylleucylglycylthryonylphenylalanylasparylthreonylarginylasp  
araginylarginylisoleucylisoleucylglutamylvalylglutamylasparag  
inylglutaminyglutaminyserylproylthreonythreonylalanylgluta  
mylthreonylleucylasparylalanylthreonylarginylarginylvalylasp  
artylasparylalanylthreonylvalylalanylisoleucylarginylserylalan  
ylasparaginylisoleucylasparaginylleucylvallasparaginylglutam  
ylleucylvalylarginylglycylthreonyllycylleucultyrosylasparaginy  
lglutaminyasparaginythreonylphenylalanylglutamylserylmeth  
ionylserylglycylleucylvalyltryptophylthreonylserylalanylproylal  
anylserine

# Synonyms

**sm.db**

alar|adj|wing|noun

amygdaline|adj|tonsil|noun

articular|adj|joint|noun

bulbar|adj|medulla oblongata|noun

fununcular|adj|boil|noun

genicular|adj|knee|noun

hepatocellular|adj|liver cells|noun

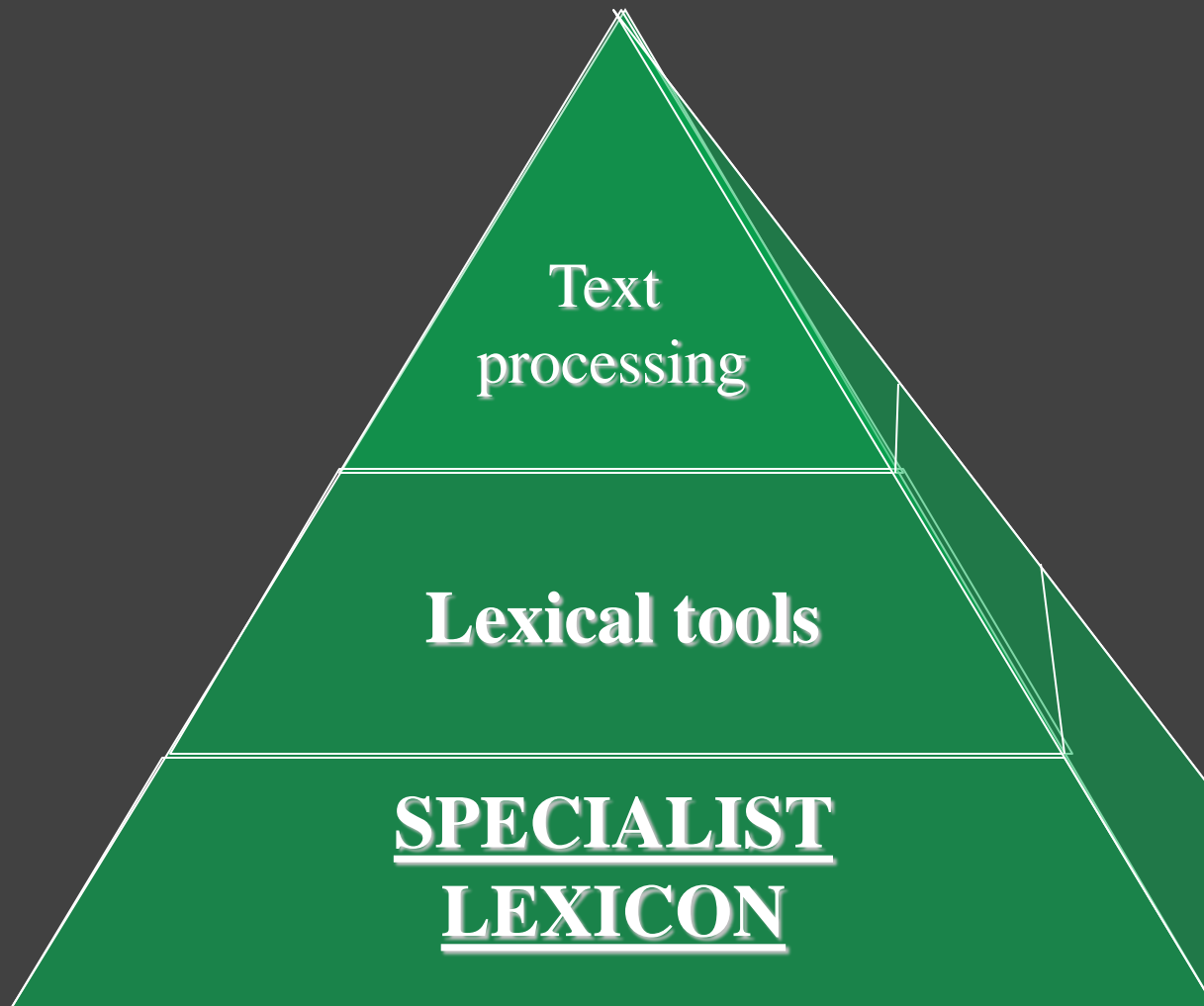
lazar|adj|leprosy|noun

lenticular|adj|crystalline lens|noun

ypsiform|adj|upsiloid|adj

wolfram|noun|tungsten|noun

double vision|noun|diplopia|noun



# Lexical Tools

- ◎ Wordind -- breaks strings into words
  - Produces the Metathesaurus word indexes (MRXW)
- ◎ LVG -- performs various lexical transformations
- ◎ NORM -- a selection of LVG transformations,
  - Used for Metathesaurus indexing
  - Produces the Metathesaurus Normalized word and string indexes (MRXNW & MRXNS)
  - Used to access those indexes



# Normalization

- Hodgkin Disease
- HODGKINS DISEASE
- Hodgkin's Disease
- Disease, Hodgkin's
- HODGKIN'S DISEASE
- Hodgkin's disease
- Hodgkins Disease
- Hodgkin's disease NOS
- Hodgkin's disease, NOS
- Disease, Hodgkins
- Diseases, Hodgkins
- Hodgkins Diseases
- Hodgkins disease
- hodgkin's disease
- Disease;Hodgkins
- Disease, Hodgkin
- disease hodgkin

# The Lexical Systems Group

- Allen Browne: [browne@nlm.nih.gov](mailto:browne@nlm.nih.gov)
- Chris Lu: [lu@nlm.nih.gov](mailto:lu@nlm.nih.gov)