

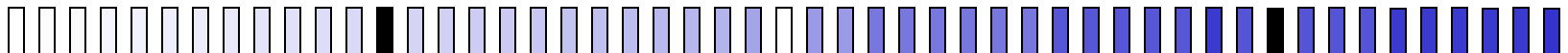


T25

Lexical Tools for UMLS Developers

11/10/2002

8:00 AM



Semantic Network

Metathesaurus

SPECIALIST Lexicon



Lexical Tools for UMLS Developers

Documentation

Resource

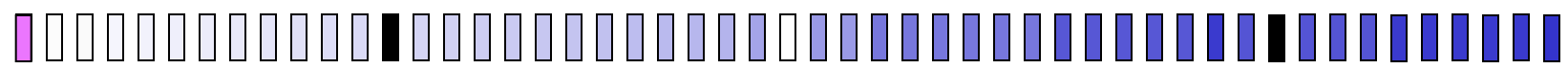
Resource

November 10, 2002

Allen C. Browne, Guy Divita,
Chris Lu

Lister Hill National Center for Biomedical Communications

National Library of Medicine





UMLS Knowledge Source Server (UMLSKS)

UMLSKS Version 2.1 UMLS Releases: 2002 2002AB

Metathesaurus

Semantic Network

SPECIALIST Lexicon

[Search](#)

[Advanced Search](#)

[Documentation](#)

[Resources](#)

[Views/Profiles](#)

[Logout](#)

Metathesaurus Focused Search:

1) Select UMLS Release:

2002AB ▼

2) Enter a term or a concept unique identifier (CUI):

- Exclude suppressible synonyms
 Include suppressible synonyms

3) [Restrict source vocabulary to:](#)

- All Source Vocabularies Restrict to selected sources:

AI/RHEUM	Normalized string index
Alternative	Normalized word index
Alcohol an	Approximate matching
Beth Israel	Word index
Classificat	Left truncation
Clinical Cl	Right truncation
Clinical Co	

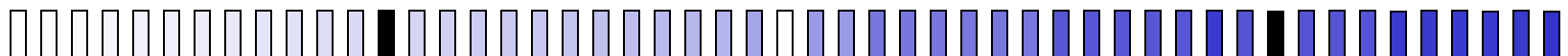
4) [String Matching Criteria](#)

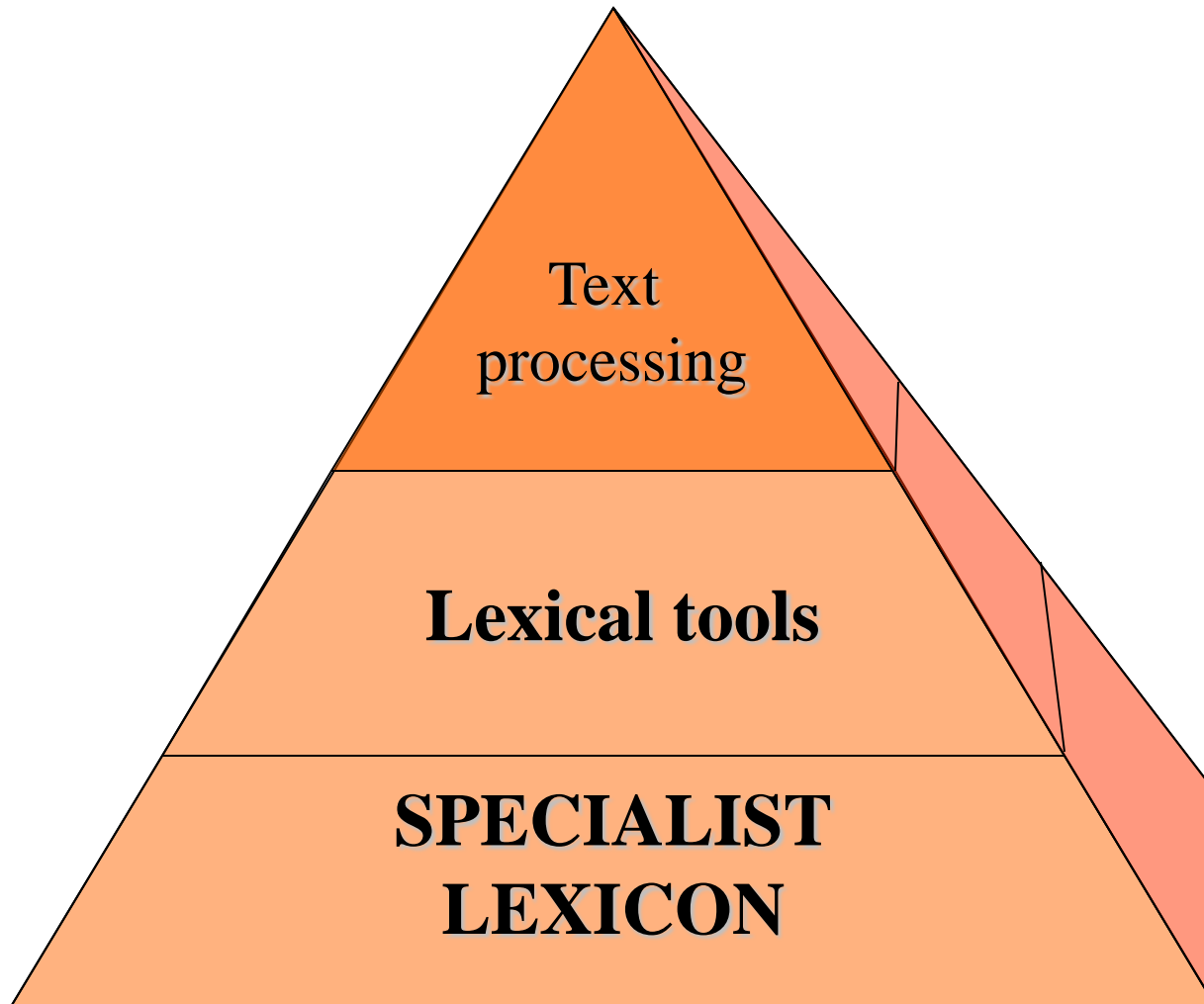
**Perform
Concept
Search**

**Perform
Term
Search**

Lexical Tools for UMLS Developers

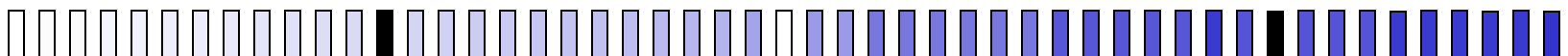
- The SPECIALIST lexicon – Browne
- The lexical tools – Divita/Lu
- Coffee Break– 10:00 - 10:30
- Lexical tools cont. – Divita/Lu





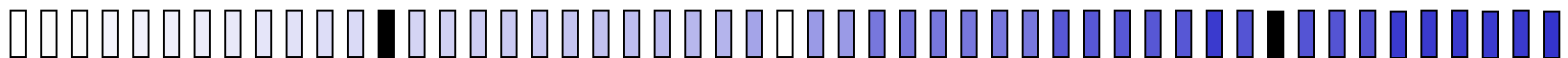
The SPECIALIST Lexicon

- A syntactic lexicon
- Biomedical and general English
- Over 180,000 records

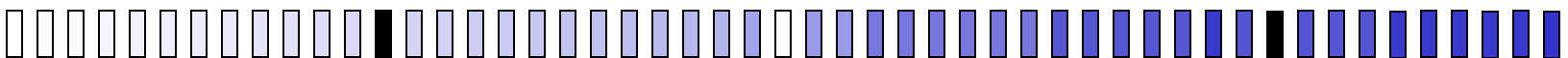
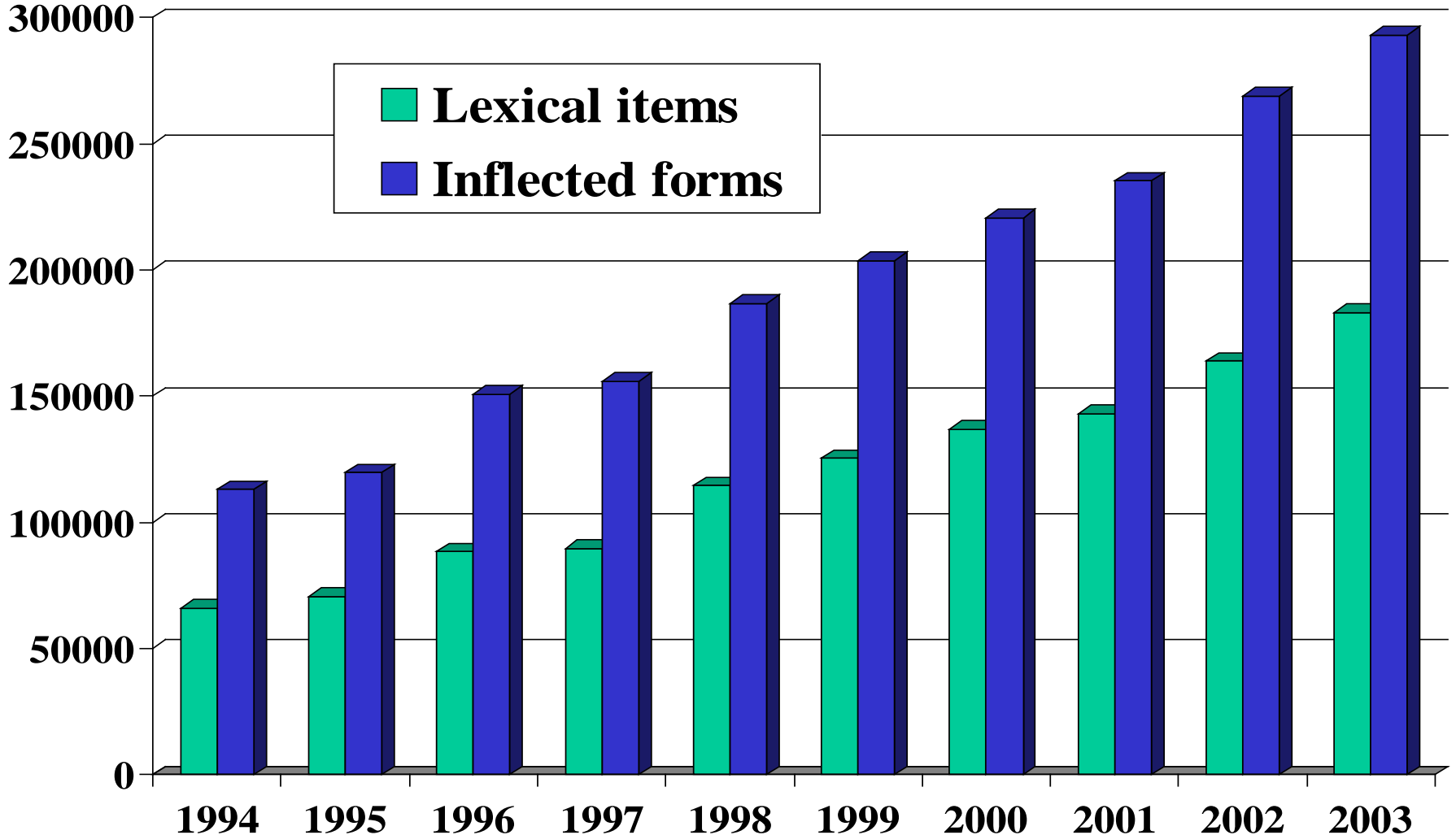


The SPECIALIST Lexicon

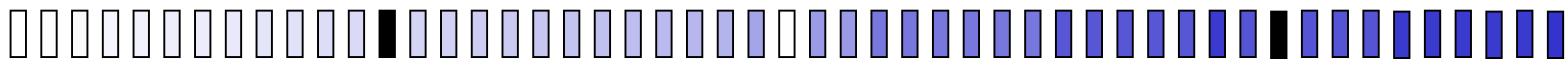
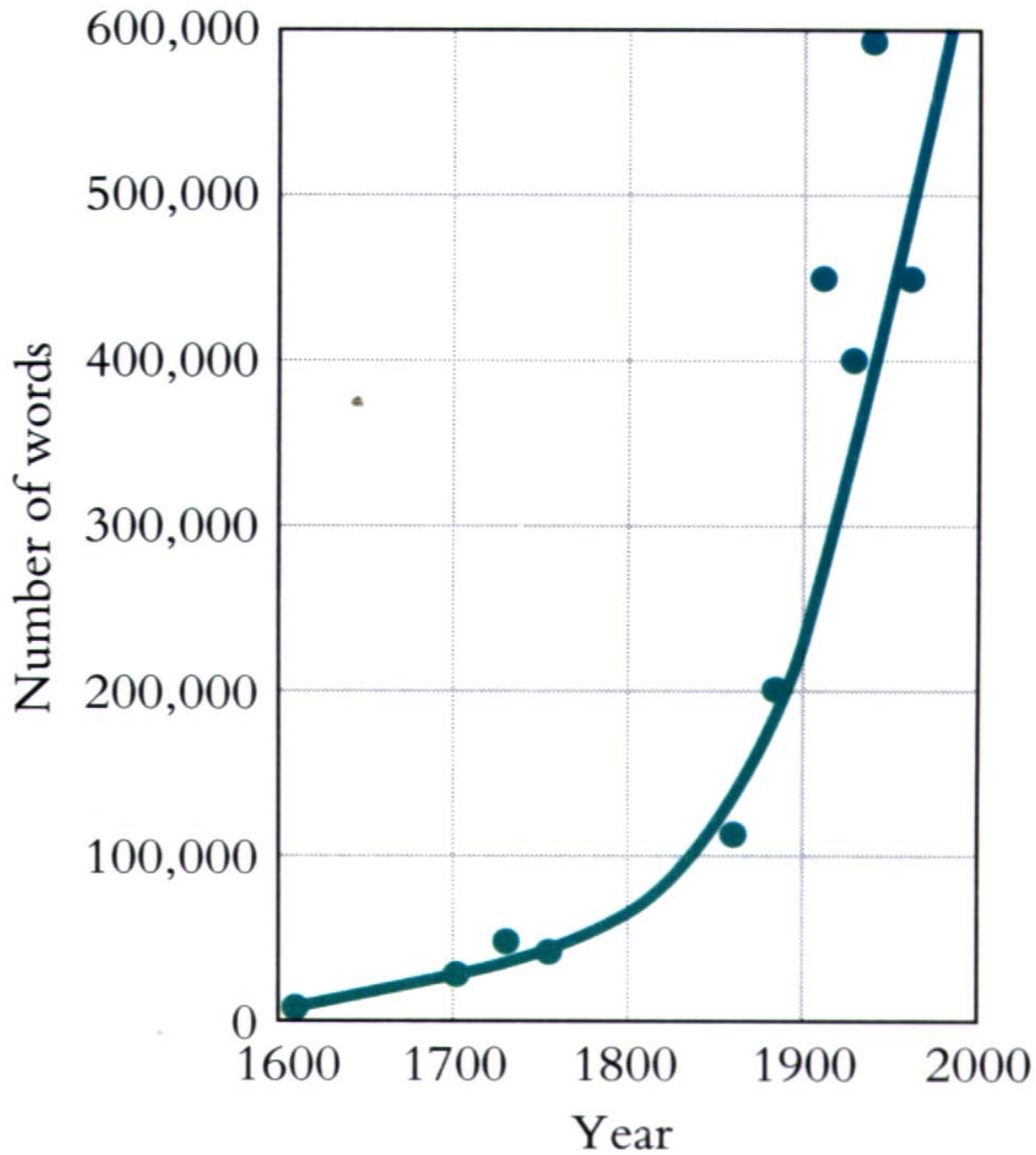
- General English:
- 10,000 most frequent words from the American Heritage word frequency list
- 2,000 words used by Longman's Dictionary of Contemporary English
- Verbs and adjectives identified by heuristics



Lexicon Growth

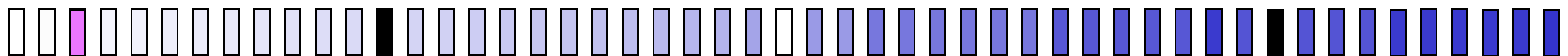


George A.
Miller
The Science
of Words
1991



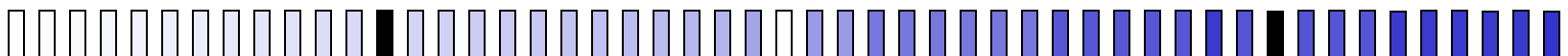
The SPECIALIST Lexicon

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



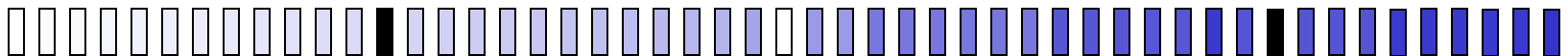
Morphology

- Inflectional
 - nucleus -- nuclei
 - cauterize, cauterizes, cauterized, cauterizing
 - red, redder reddest
- Derivational
 - laryngeal -- larynx
 - transport -- transportation



Orthography

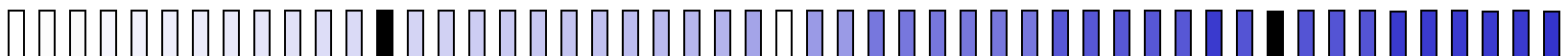
HAGAR THE HORRIBLE CHRIS BROWNE



Orthography

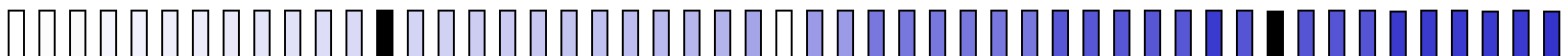
Spelling Variation

- **align -- aline**
- **Grave's disease -- Graves's disease -- Graves' disease**
- **anesthetize -- anaesthetise**
- **esophagus -- oesophagus**



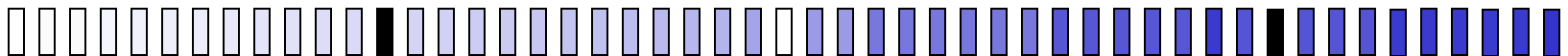
Syntax -- Verb Complements

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

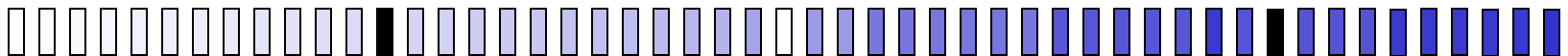
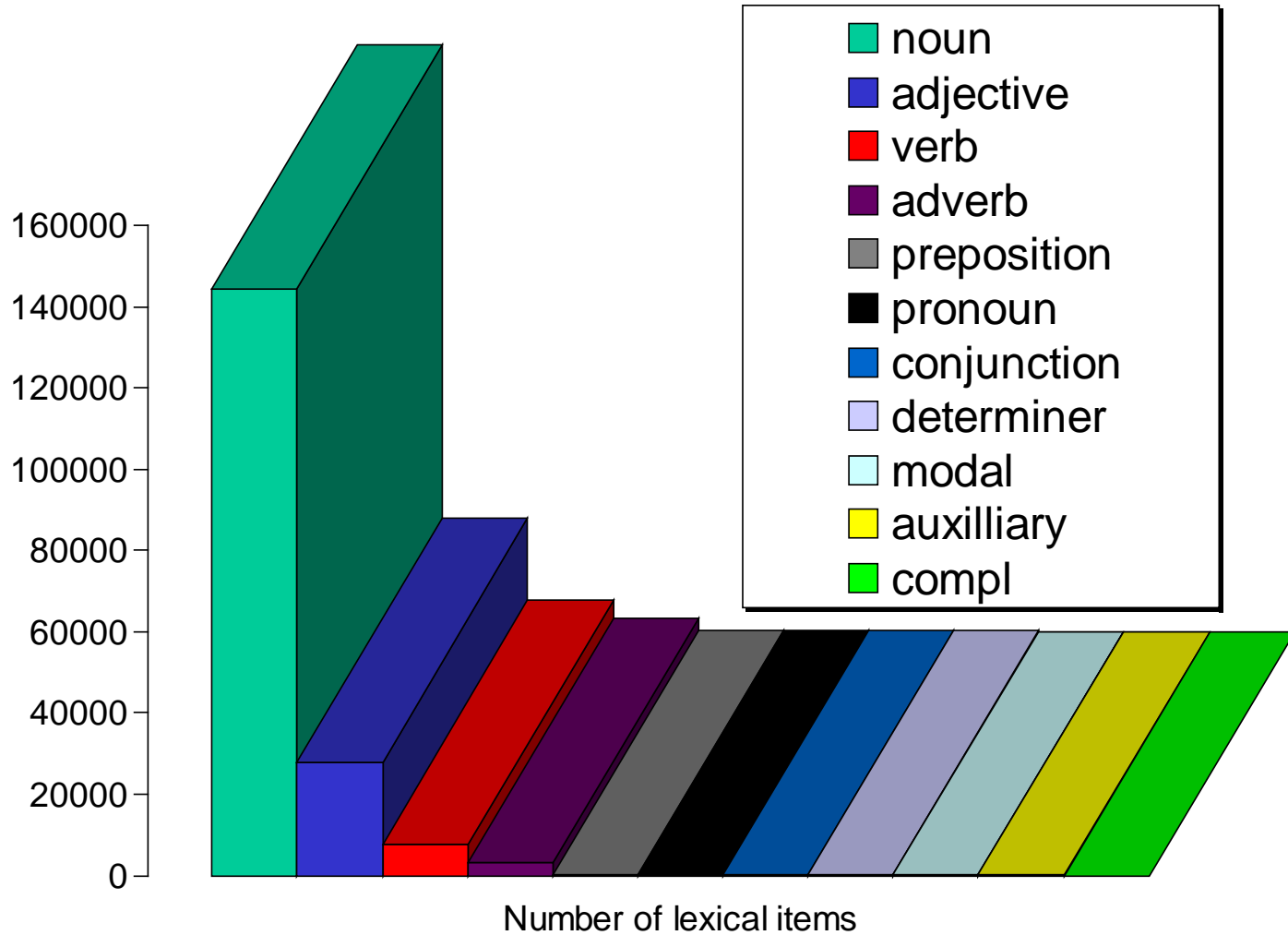


Syntax -- Verb Complements

```
{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
}
```



The 2003 SPECIALIST Lexicon



village

square

the circle

square

square

fair and

square

root

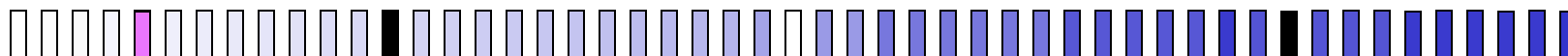
Lexicon Unit Records

{ **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
}

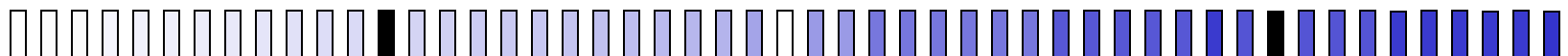
{ **base**=in
entry=E0033870
 cat=prep
}



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

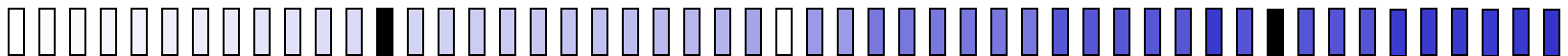


Regular Nouns

The plural suffix is *s*.

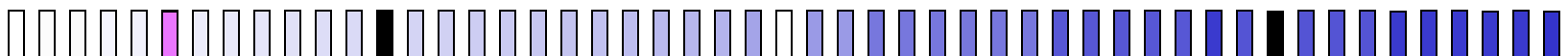
y becomes *ie* following a consonant before *s*.

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



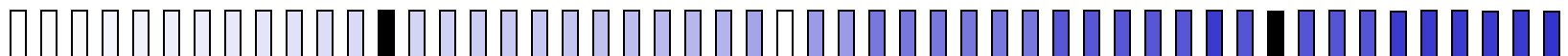
Regular Nouns

Base ends with	Plural ends with	Examples
Cy	Cies	fly: flies
-s	-ses	illness: illnesses
-z	-zes	waltz: waltzes
-x	-xes	box: boxes
-ch	-ches	match: matches
-sh	-shes	splash: splashes
X	Xs	book: books



Greco-latin Regular nouns

singular ends with:	plural ends with:	Examples
-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



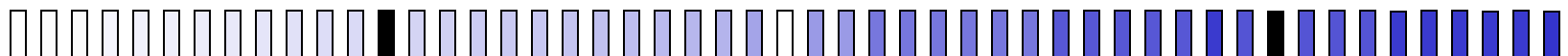
Uncount Nouns

(abstract or mass)

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

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

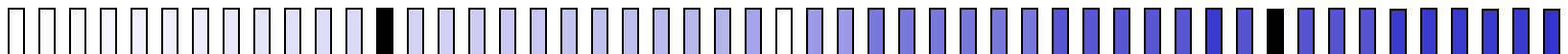
- * a smallpox
- * two smallpoxes
- much smallpox
- * a potassium
- * two potassiums
- much potassium



Fixed Plural Nouns

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

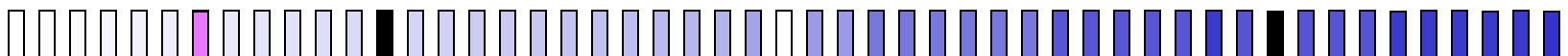
```
{base=scissors  
entry=E0054633  
  cat=noun  
  variants=plur  
}
```



Irregular Nouns

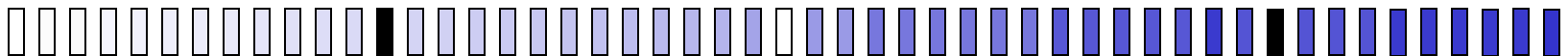
```
{base=corpus  
entry=E0019113  
  cat=noun  
  variants=irreg|corpora|  
  variants=reg  
}
```

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



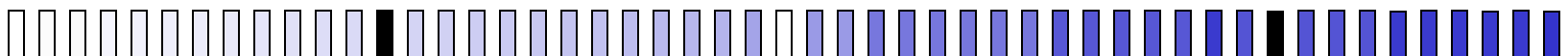
Regular Verbs

- 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*.



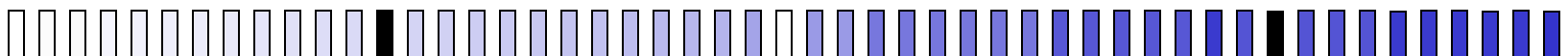
Regular Verbs

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



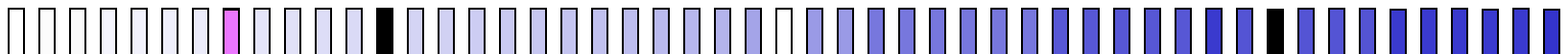
Regular Doubling Verbs

- End in a CVC pattern
- Double the final consonant before *ed* and *ing*.
- Are otherwise regular
- variants=regd
- e.g. control: controls, controlled, controlling

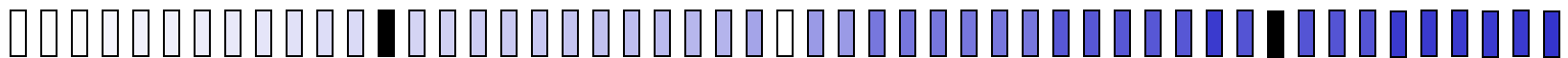
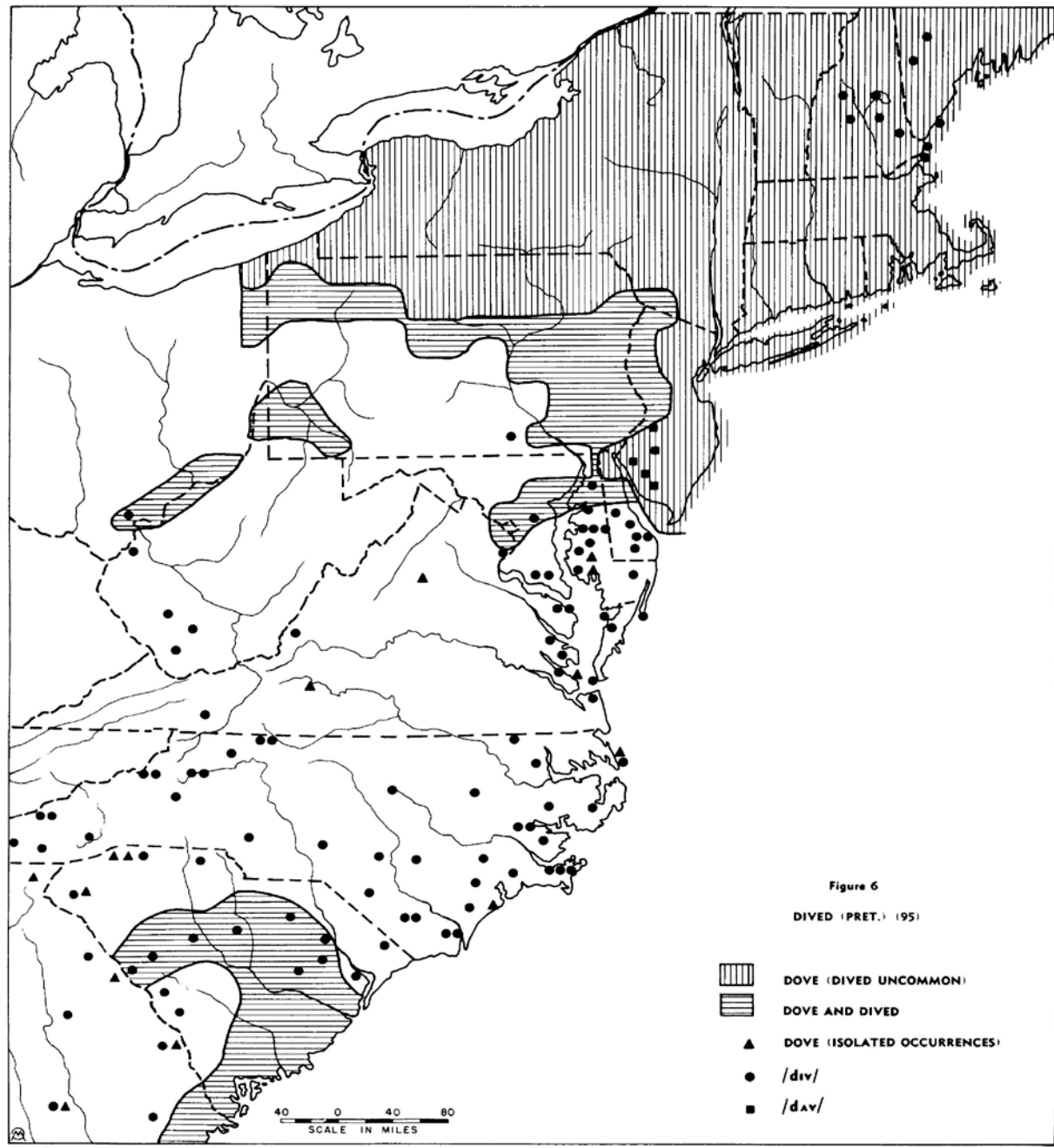


Irregular Verbs

```
{base=dive
cat=verb
  variants=reg
  variants=irreg|dives|dove|dove|diving|
  intrans
  intrans;part(in)
  ...
}
```

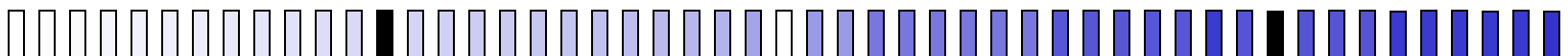


Dive vs. Dove



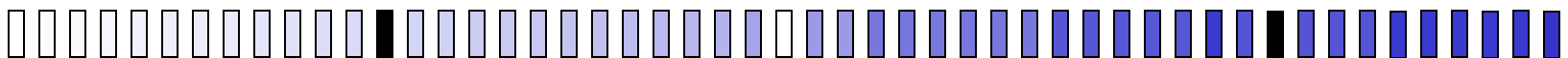
Regular Adjectives and Adverbs

- The comparative suffix is *er*.
- The superlative suffix is *est*.
 - *y* become *ie* after a consonant before *er* or *est*.
 - Final *e* is deleted before *er* or *est*.
- e.g. green: greener, greenest



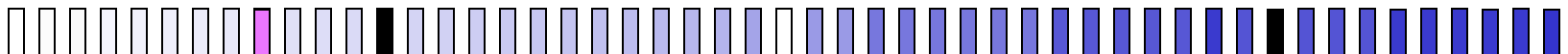
Regular Doubling Adjectives and Adverbs

- CVC final pattern
- Final consonant is doubled before ed or est.
- Otherwise regular
- e.g. red: redder, reddest



Ancillary Data Bases

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



Derivational Facts and Rules

dm.facts

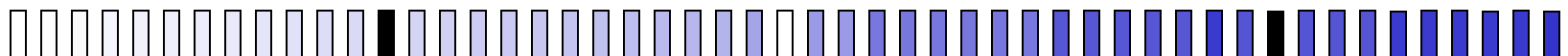
treatment|noun|treat|verb

prohibition|noun|prohibitive|adj

cell lineage|noun|cell line|noun

photochemotherapeutic|adj|photochemotherapy|noun

pharmacotherapeutic|adj|pharmacotherapy|noun



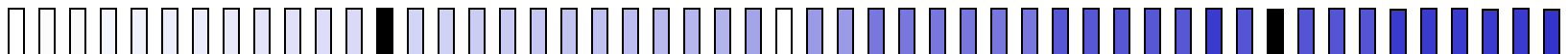
Derivational Facts and Rules

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

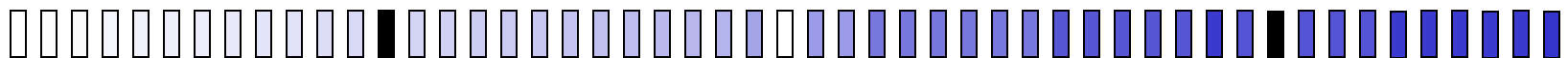
sis\$|noun|singular|ses\$|noun|plural

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

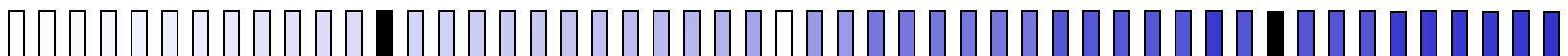
ad|towards|prefix

agogue|inducing|terminal

albumin(o)|albumin|root

sis|condition|terminal

stomy|surgical opening|terminal



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

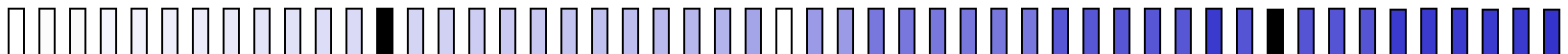
lazar|adj|leprosy|noun

lenticular|adj|crystalline lens|noun

ypsiform|adj|upsiloid|adj

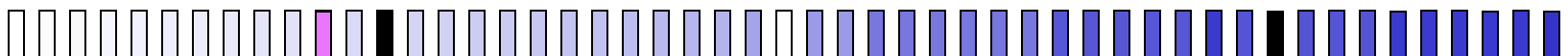
wolfram|noun|tungsten|noun

double vision|noun|diplopia|noun



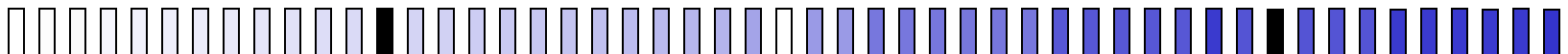
Relational Tables

- One line records
- Pipe separated Fields -- “|”
- Keyed to EUI
- LRAGR matches forms to EUIs
- Word index: LRWD



Relational Tables

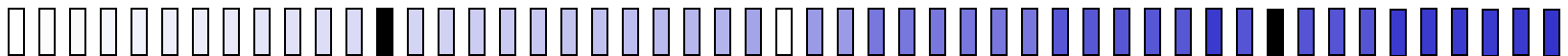
- LRAGR - Agreement
- LRCMP - Complements
- LRFIL - Files
- LRFLD - Fields
- LRMOD - Modification
- LRNOM - Nominalization
- LRPRN - Pronouns
- LRPRP - Properties
- LRSPL - Spelling
- LRTRM - Trademarks
- LRWD - Word index



LRAGR

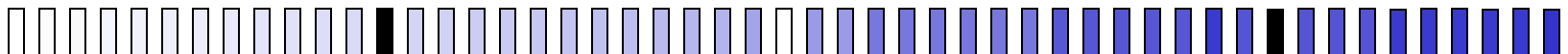
Agreement and Inflection

- EUI - Entry ID
- STR - Inflected form
- SCA - Syntactic category
- AGR - agreement information
- BAS - Base form (morphological)
- CIT - Citation form (base=)



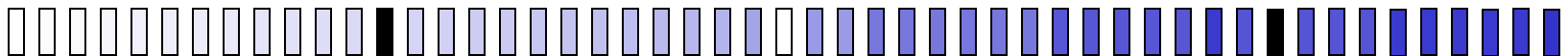
LRAGR

E0003576| Kaposi sarcomas| noun| count(thr_plur)| Kaposi sarcoma| Kaposi's sarcoma|
E0003576| Kaposi sarcomata| noun| count(thr_plur)| Kaposi sarcoma| Kaposi's sarcoma|
E0003576| Kaposi sarcoma| noun| count(thr_sing)| Kaposi sarcoma| Kaposi's sarcoma|
E0003576| Kaposi sarcoma| noun| uncount(thr_sing)| Kaposi sarcoma| Kaposi's sarcoma|
E0003576| Kaposi's sarcomas| noun| count(thr_plur)| Kaposi's sarcoma| Kaposi's sarcoma|
E0003576| Kaposi's sarcomata| noun| count(thr_plur)| Kaposi's sarcoma| Kaposi's sarcoma|
E0003576| Kaposi's sarcoma| noun| count(thr_sing)| Kaposi's sarcoma| Kaposi's sarcoma|
E0003576| Kaposi's sarcoma| noun| uncount(thr_sing)| Kaposi's sarcoma| Kaposi's
sarcoma|

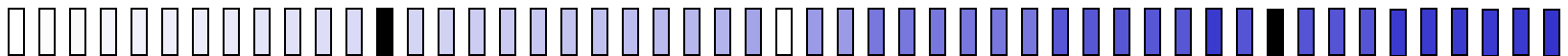


Number Words

- ‘one’, ‘thirteen’ ‘fifty’, ‘thousand’, ‘million’
- Not in the lexicon.
 - No part of speech
 - Used to construct number expressions:
“Three thousand eight hundred and five”
- To be released in the 2003 lexicon.
- Accompanying number tools.



```
{base=two
  cat=number_word
  entry=N0000003
  variant=second;ordinal
  variant=second;denominator,singular;part_denominator
  variant=second;denominator,plural;part_denominator
  variant=half;denominator,singular;full_denominator
  variant=halves;denominator,plural;full_denominator
  number_type=unit
  value=2
  digit=2
}
```



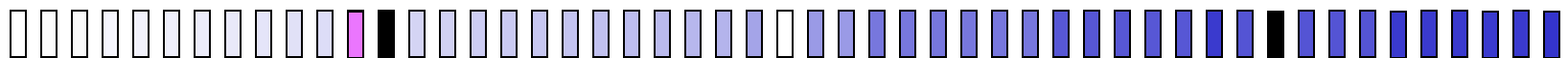
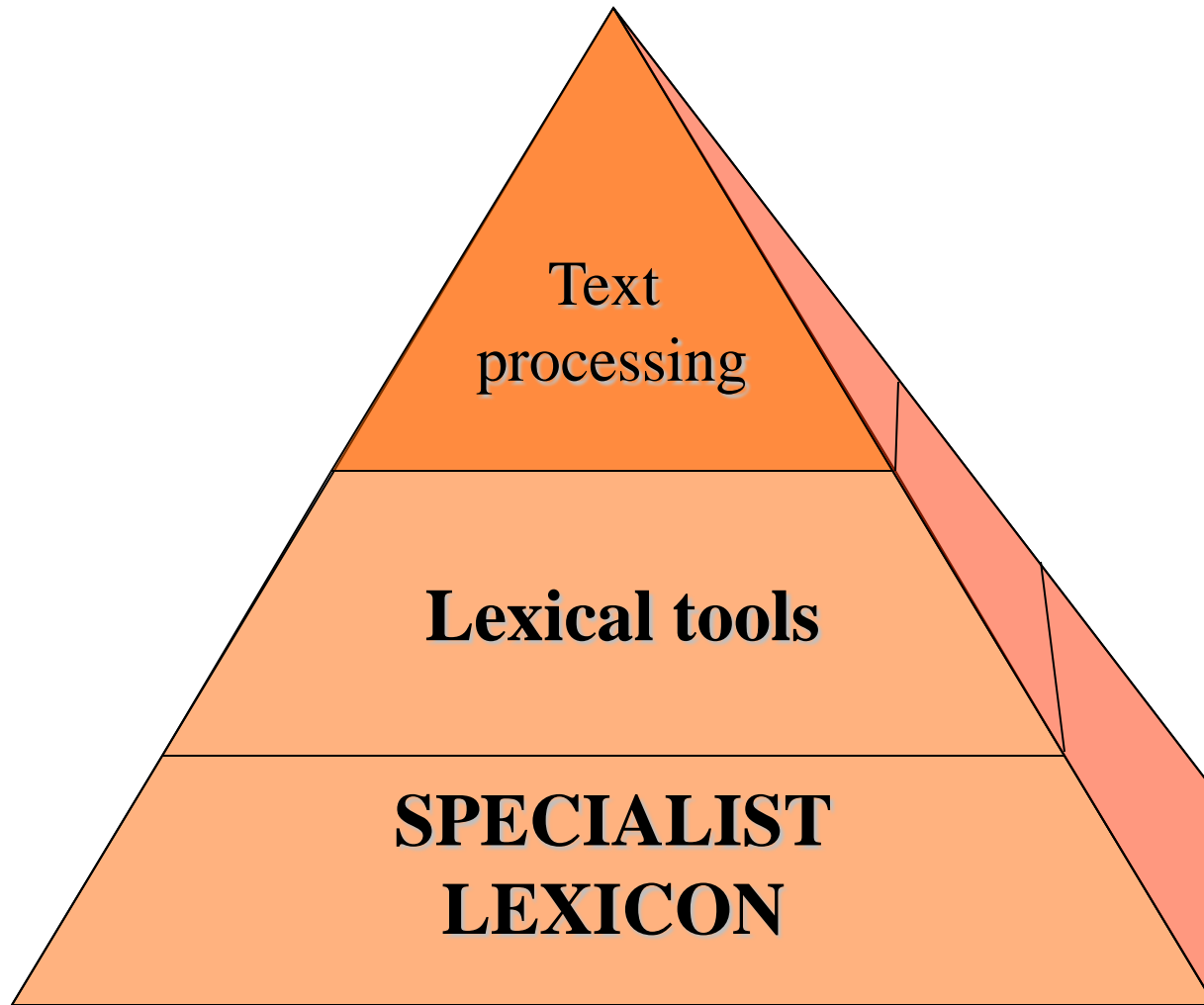
```
{base=twelve
  cat=number_word
  entry=N0000013
  variants=reg
  number_type=teen
  value=12
}
```

```
{base=twenty
  cat=number_word
  entry=N0000021
  variants=reg
  number_type=decade
  value=20
  digit=2
}
```

```
{base=billion
  cat=number_word
  entry=N0000032
  variants=reg
  number_type=magnitude
  power=3
}
```

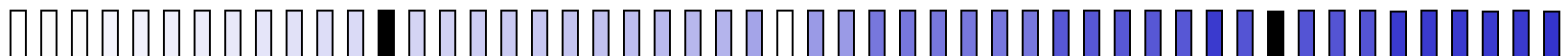
```
{base=sexdecillion
  cat=number_word
  entry=N0000046
  variants=reg
  number_type=magnitude
  power=17
}
```





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**

