# Etymachine

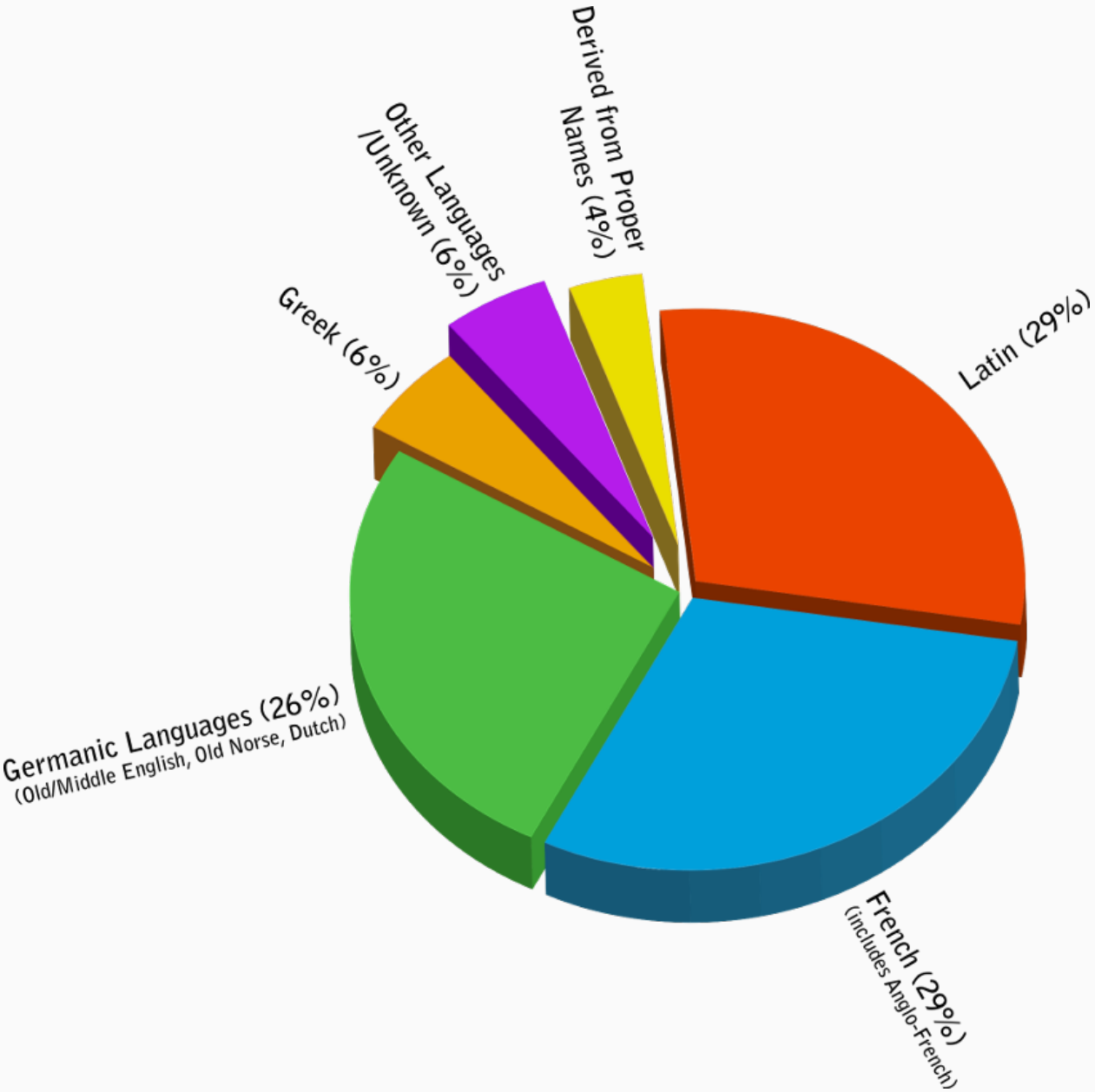
## The Etymological Machine

Trevor Sullivan

25 February 2017

University of Arizona

# The Problem


French (29%)
(includes Anglo-French)
Germanic Languages (26%)
(Old/Middle English, Old Norse, Dutch)
Greek (6%)
Other Languages
/Unknown (6%)
Derived from Proper
Names (4%)
Latin (29%)

# The Problem

- talaric

# The Problem

- talaric

# The Problem

- talaric (of the ankle)
- thallasocracy

# The Problem

- talaric (of the ankle)
- thallasocracy

# The Problem

- talaric (of the ankle)
- thallasocracy (reign of the water)
- lexiphanicism

# The Problem

- talaric (of the ankle)
- thallasocracy (reign of the water)
- lexiphanicism

# The Problem

- talaric (of the ankle)
- thallasocracy (reign of the water)
- lexiphanicism (using too many words)
- foudroyant

# The Problem

- talaric (of the ankle)
- thallasocracy (reign of the water)
- lexiphanicism (using too many words)
- foudroyant

# The Problem

- talaric (of the ankle)
- thallasocracy (reign of the water)
- lexiphanicism (using too many words)
- foudroyant (sudden)
- rynchosporous

# The Problem

- talaric (of the ankle)
- thallasocracy (reign of the water)
- lexiphanicism (using too many words)
- foudroyant (sudden)
- rynchosporous

# The Problem

- talaric (of the ankle)
- thallasocracy (reign of the water)
- lexiphanicism (using too many words)
- foudroyant (sudden)
- rynchosporous (having pointy fruit)
- nepheligenous

# The Problem

- talaric (of the ankle)
- thallasocracy (reign of the water)
- lexiphanicism (using too many words)
- foudroyant (sudden)
- rynchosporous (having pointy fruit)
- nepheligenous

# The Problem

- talaric (of the ankle)
- thallasocracy (reign of the water)
- lexiphanicism (using too many words)
- foudroyant (sudden)
- rynchosporous (having pointy fruit)
- nepheligenous (discharging smoke)
- tachyphylaxis

# The Problem

- talaric (of the ankle)
- thallasocracy (reign of the water)
- lexiphanicism (using too many words)
- foudroyant (sudden)
- rynchosporous (having pointy fruit)
- nepheligenous (discharging smoke)
- tachyphylaxis

# The Problem

- talaric (of the ankle)
- thallasocracy (reign of the water)
- lexiphanicism (using too many words)
- foudroyant (sudden)
- rynchosporous (having pointy fruit)
- nepheligenous (discharging smoke)
- tachyphylaxis (rapid immunity)
- calcoprapher

# The Problem

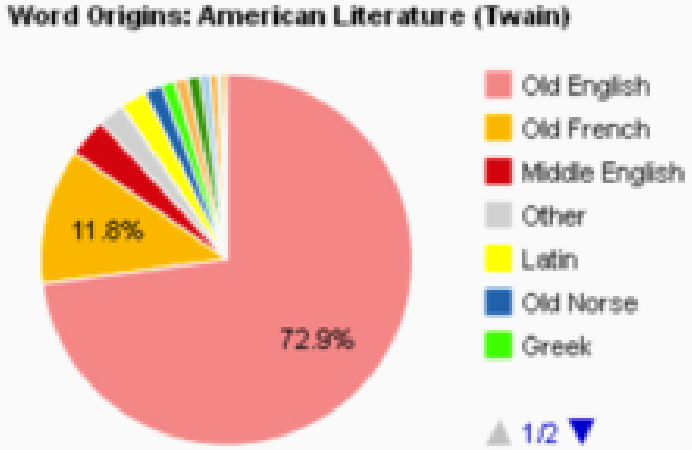
- talaric (of the ankle)
- thallasocracy (reign of the water)
- lexiphanicism (using too many words)
- foudroyant (sudden)
- rynchosporous (having pointy fruit)
- nepheligenous (discharging smoke)
- tachyphylaxis (rapid immunity)
- calcoprapher

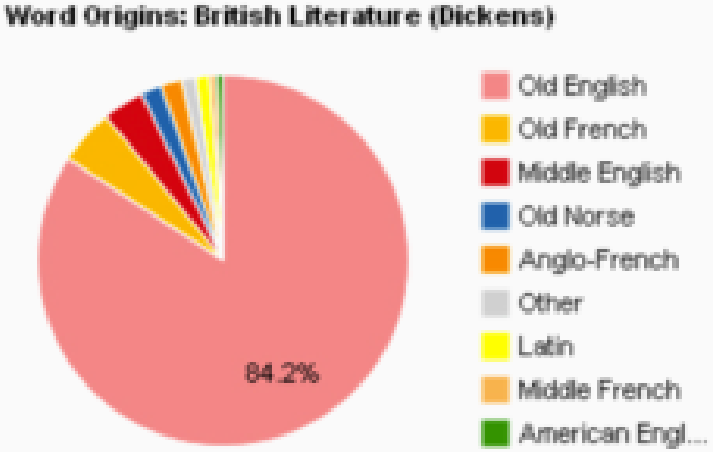
# The Problem

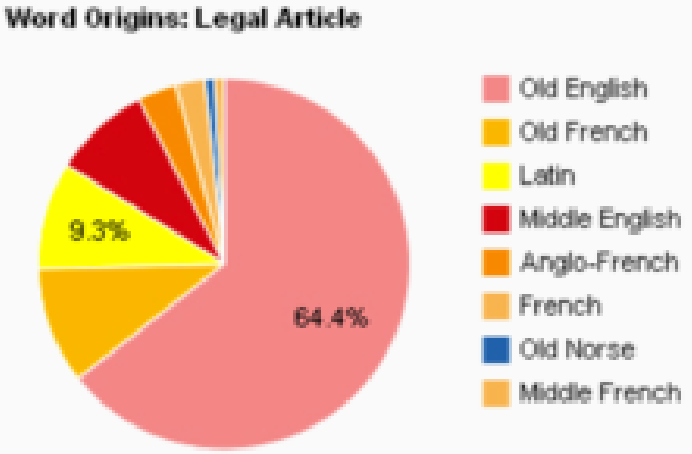
- talaric (of the ankle)
- thallasocracy (reign of the water)
- lexiphanicism (using too many words)
- foudroyant (sudden)
- rynchosporous (having pointy fruit)
- nepheligenous (discharging smoke)
- tachyphylaxis (rapid immunity)
- calcoprapher (someone who draws with crayons)

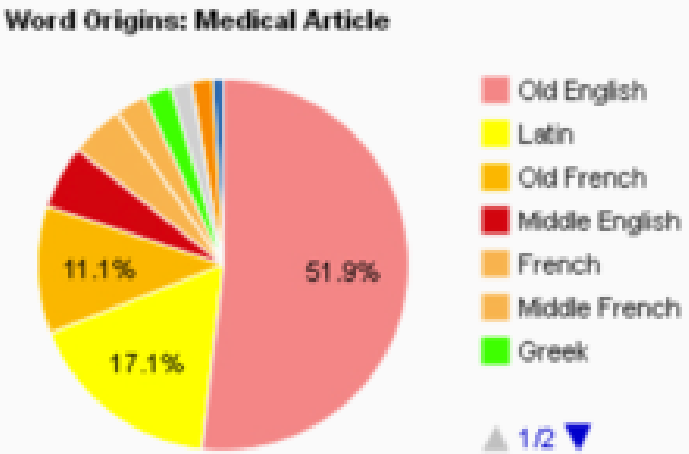
# The Problem

# The Problem




Based on a very small corpus, one paragraph each

# Goal

# Goal

To construct a machine that can make a chart of the etymological content of a text on the fly.

# Goal

To construct a machine that can make a chart of the etymological content of a text on the fly.

New problem: there is no database of etymological classifications

# Constructing a Database

# Constructing a Database

Source: Etymonline.com, the Online Etymological Dictionary

# Constructing a Database

Source: Etymonline.com, the Online Etymological Dictionary

ONLINE ETYMOLOGY DICTIONARY

**Search:** OK

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

This is a map of the wheel-ruts of modern English. Etymologies are not definitions; they're explanations of what our words meant and how they sounded 600 or 2,000 years ago.

The dates beside a word indicate the earliest year for which there is a surviving written record of that word (in English, unless otherwise indicated). This should be taken as approximate, especially before about 1700, since a word may have been used in conversation for hundreds of years before it turns up in a manuscript that has had the good fortune to survive the centuries.

The basic sources of this work are Weekley's "An Etymological Dictionary of Modern English," Klein's "A Comprehensive Etymological Dictionary of the English Language," "Oxford English Dictionary" (second edition), "Barnhart Dictionary of Etymology," Holthausen's "Etymologisches Wörterbuch der Englischen Sprache," and Kipfer and Chapman's "Dictionary of American Slang." **A full list of print sources used in this compilation can be found here**.

Since this dictionary went up, it has benefited from the suggestions of dozens of people I have never met, from around the world. Tremendous thanks and appreciation to all of you.

**培根词汇 - 英语词根词源词典 (App for iOS)**
The brand-new official Online Etymology Dictionary app for China, Taiwan, Hong Kong and Macau

**Like the Etymonline Page on Facebook**

**Ye Olde Swag Shoppe**
**Get yer Etym on**

**Why I probably haven't answered your e-mail**

Donate

Help keep etymonline free and open

**Introduction and abbreviations**
**Who did this?**
**Sources**
**Links**

© 2001-2016 **Douglas Harper**
**Custom logo design** by LogoBee.com
Web page design by **Dan McCormack**
**Sponsored Words**

# Constructing a Database

Etymonline has lots of etymological entries, but they are not neatly categorized into the source language of each.

This is categorization problem, but the categories are obscured in the source data. Rather than trying to do unsupervised learning to it, I chose to attempt to approximate a golden dataset using a deterministic process, and then use that for machine learning.

# Constructing a Database

Languages investigated

- Old English

# Constructing a Database

Languages investigated

- Old English
- French

# Constructing a Database

Languages investigated

- Old English
- French
- Old Norse

# Constructing a Database

Languages investigated

- Old English
- French
- Old Norse
- Latin

# Constructing a Database

Languages investigated

- Old English
- French
- Old Norse
- Latin
- Greek

# Constructing a Database

Languages investigated

- Old English
- French
- Old Norse
- Latin
- Greek
- Other

# Constructing a Database

Languages investigated

- Old English
- French
- Old Norse
- Latin
- Greek
- Other

# Constructing a Database

Languages investigated

- Old English
- French
- Old Norse
- Latin
- Greek
- Other

Some of these have a relationship to each other, as in the following definition

**polysemy** 1900, from French polysmie, from Medieval Latin polysemus, from Greek polysemos, from poly- + sema

# Chains

French > Middle French > Old French > Medieval Latin > Latin > Greek > PIE

Greek(modern, koine, classical) > PIE

Old English > Ingvaeonic (Old Frisian, Old Saxon) > Proto-West-Germanic > Proto-Germanic > PIE

Old Norse > Proto-North-Germanic > Proto-Germanic > PIE

# Problem: links

Problem: many dictionary entries take the form

**alumna** see **alumnus**

This means that our analysis will have to collapse these "two words that are really just one word" into a single element, and then be able to un-collapse them when the actual analysis of a real text is performed.

# Problem: links

Problem: many dictionary entries take the form

**alumna** see **alumnus**

This means that our analysis will have to collapse these "two words that are really just one word" into a single element, and then be able to un-collapse them when the actual analysis of a real text is performed.

This is fixed by using the data we have to find these "links" and then just filling the linked defintion into the linker's definition space. This happens roughly 260 times

# Problem: hidden links

The "other" category is large because of definitions like this:

# Problem: hidden links

The "other" category is large because of definitions like this:

```
Word:  curvy (adj.)  Category:  Other
Definition:  1902, from curve (n.) + -y (2). Related: Curviness.
```

# Problem: hidden links

The "other" category is large because of definitions like this:

```
Word:  curvy (adj.)  Category:  Other
Definition:  1902, from curve (n.) + -y (2). Related: Curviness.
```

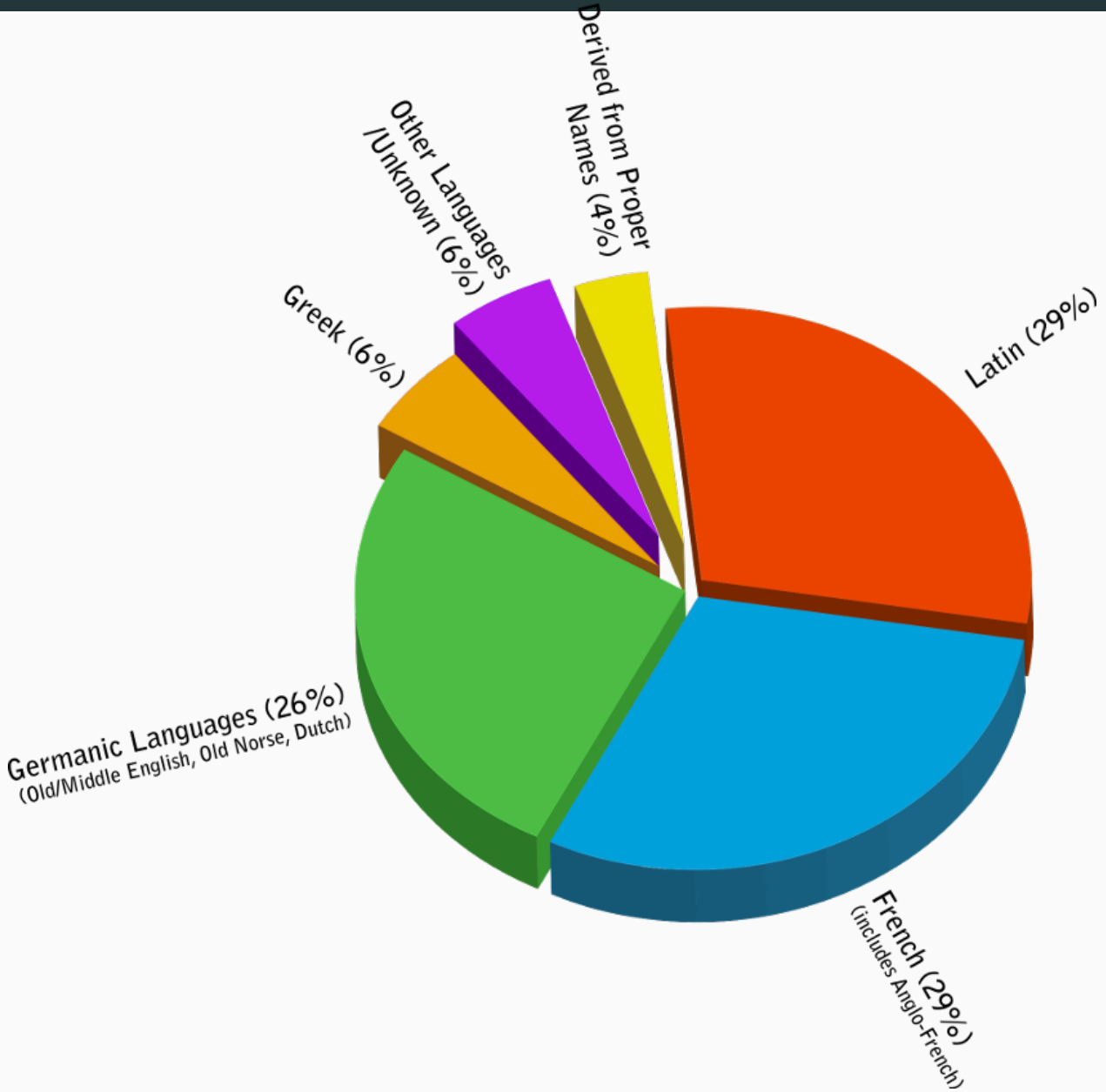
This might be fixable by searching for these "from" and "form of" type expressions and treating them like links recursively

# Second try

# Second try

```
English:  11300  percent of total:  24.71
French:   12712  percent of total:  27.80
Norse:      495  percent of total:   1.08
Latin:     6946  percent of total:  15.19
Greek:     1465  percent of total:   3.20
Other:    11131  percent of total:  24.34
Total:    45723
```

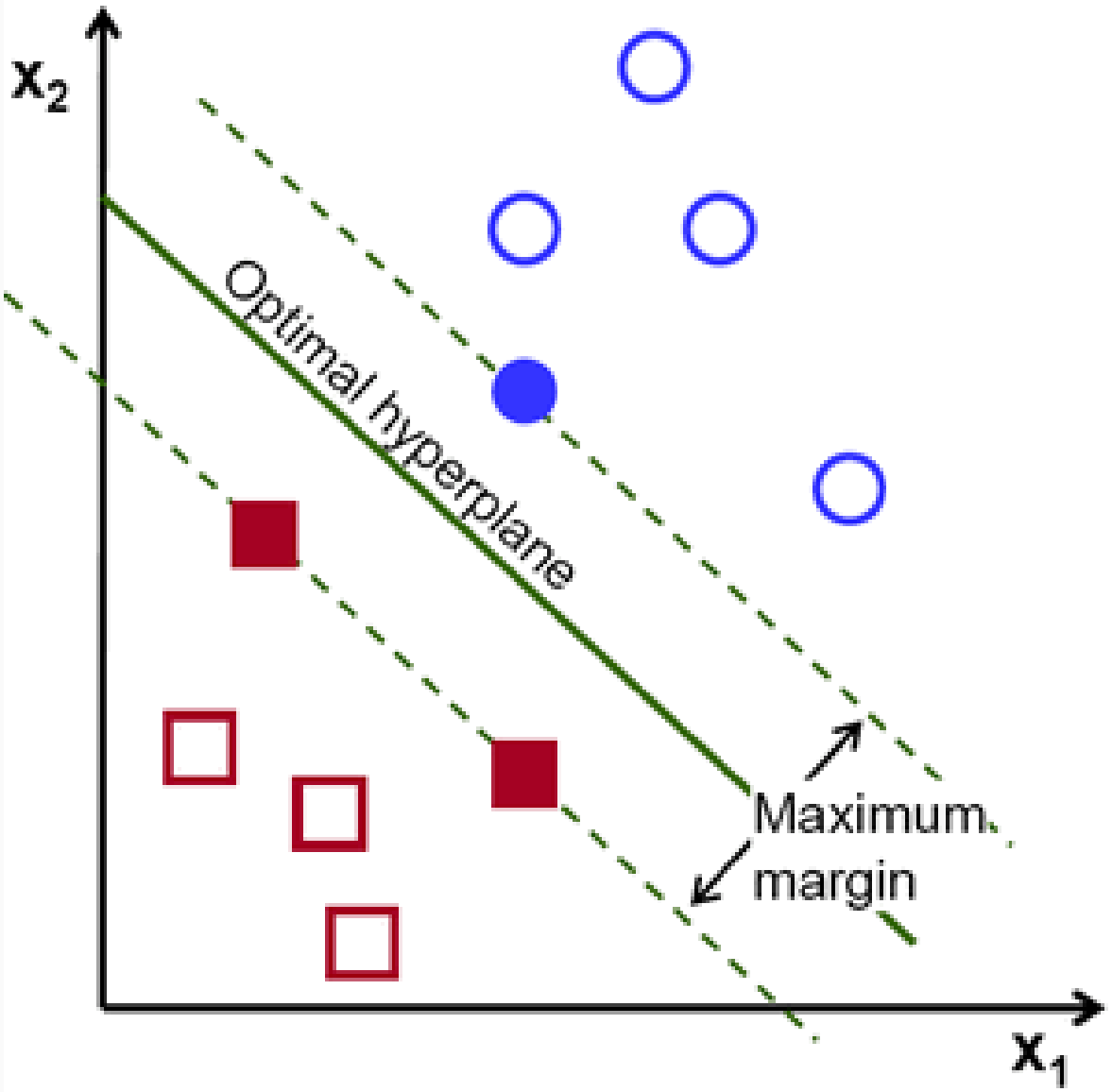
# Pattern match categorization results



# Machine Learning

# Machine Learning attempts

Support Vector Machine



# Selecting features

- Bag-of-words

# Selecting features

- Bag-of-words
- What characters are present

# Selecting features

- Bag-of-words
- What characters are present
- First attested century

# Selecting features

- Bag-of-words
- What characters are present
- First attested century
- Syllables present

# Selecting features

- Bag-of-words
- What characters are present
- First attested century
- Syllables present
- Bigrams

# Selecting features

- Bag-of-words
- What characters are present
- First attested century
- Syllables present
- Bigrams

# Extracting Features

(conscript)

# Extracting Features

(conscript)

**Bag-of-Words** 159644 dimensional sparse vector, where each dimension is a word count. The language names (English, French, Norse, Latin, Greek) are not counted.
eg, {simplicity: 1, participle: 2, literally: 1, genea: 0}

# Extracting Features

(conscript)

**Bag-of-Words** 159644 dimensional sparse vector, where each dimension is a word count. The language names (English, French, Norse, Latin, Greek) are not counted.
eg, {simplicity: 1, participle: 2, literally: 1, genea: 0}

**Characters** 80 dimensional sparse vector, where each dimension is a binary indicator of whether a character is present or not.
eg {i: 1, 1: 1, .: 1, a: 1, h: 1, : 0, : 0}

# Extracting Features

(conscript)

**Bag-of-Words** 159644 dimensional sparse vector, where each dimension is a word count. The language names (English, French, Norse, Latin, Greek) are not counted.
eg, {simplicity: 1, participle: 2, literally: 1, genea: 0}

**Characters** 80 dimensional sparse vector, where each dimension is a binary indicator of whether a character is present or not.
eg {i: 1, 1: 1, .: 1, a: 1, h: 1, : 0, : 0}

**Century** 1-dimensional vector, simply a number corresponding to the first two digits of the century, found by searching for the first 3 or 4-digit number, or "th/nd/st"
eg [15]

# Extracting Features

(conscript)

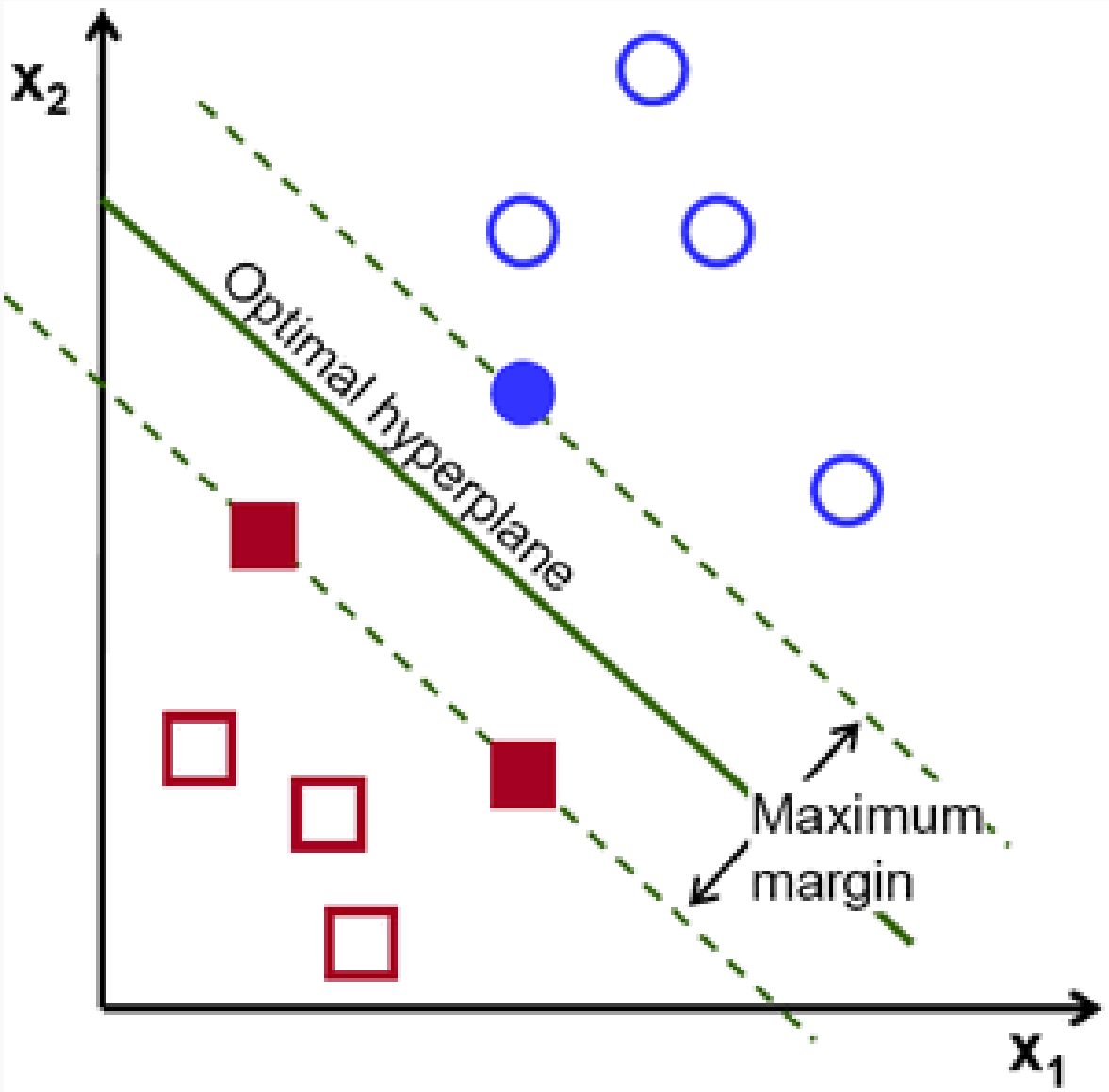
**Bag-of-Words** 159644 dimensional sparse vector, where each dimension is a word count. The language names (English, French, Norse, Latin, Greek) are not counted.
eg, {simplicity: 1, participle: 2, literally: 1, genea: 0}

**Characters** 80 dimensional sparse vector, where each dimension is a binary indicator of whether a character is present or not.
eg {i: 1, 1: 1, .: 1, a: 1, h: 1, : 0, : 0}

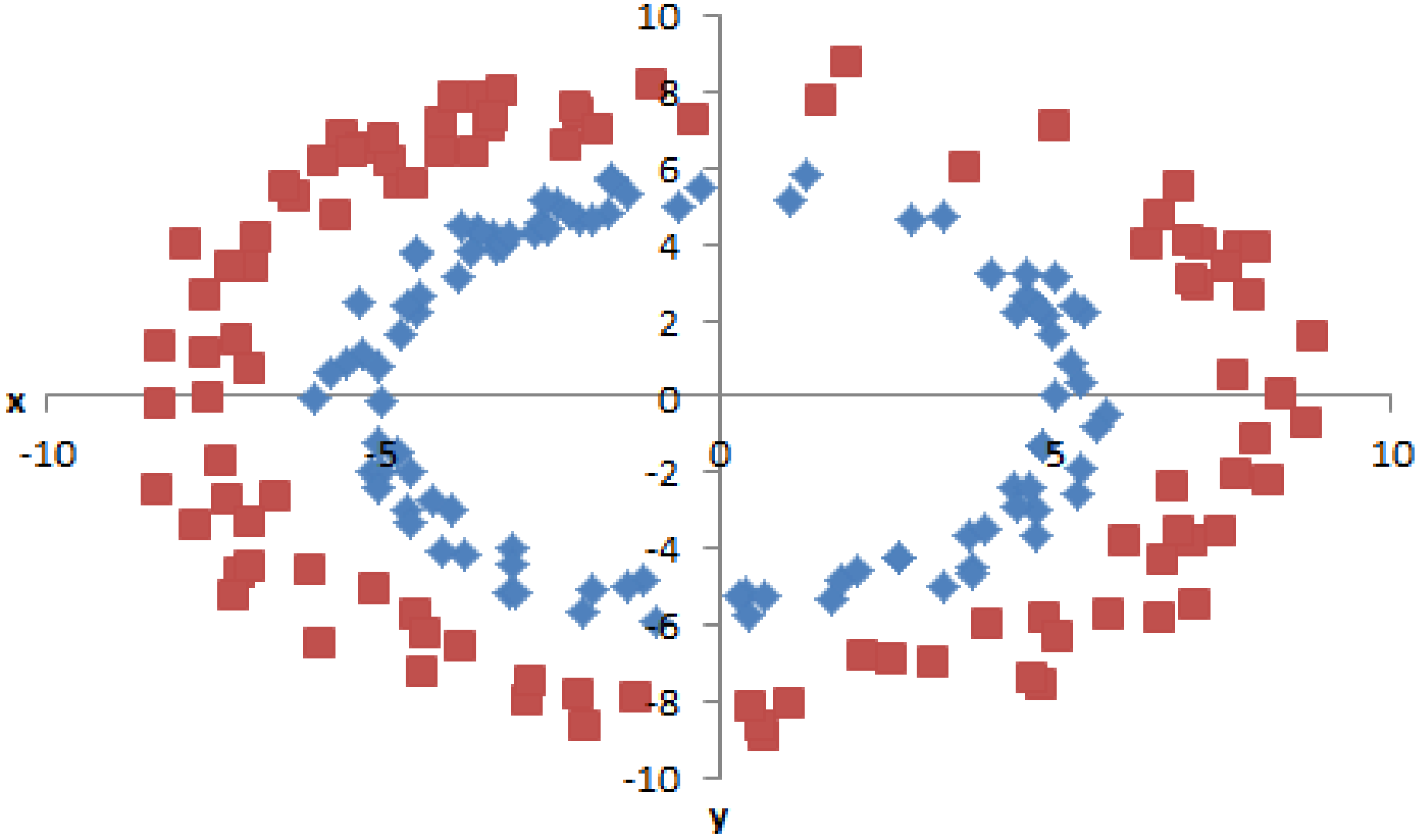
**Century** 1-dimensional vector, simply a number corresponding to the first two digits of the century, found by searching for the first 3 or 4-digit number, or "th/nd/st"
eg [15]

**Syllable** 7588 dimensional vector, Moses statistical machine-translation module based on CELEX, courtesy of Jungyeul Park.
eg {script: 1, ress: 0, thaw: 0, con: 1}

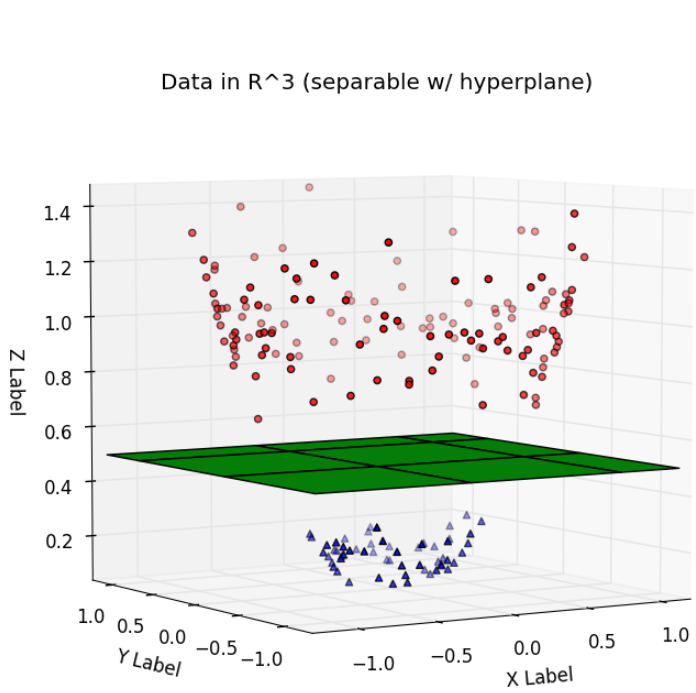
# Selecting Model Parameters

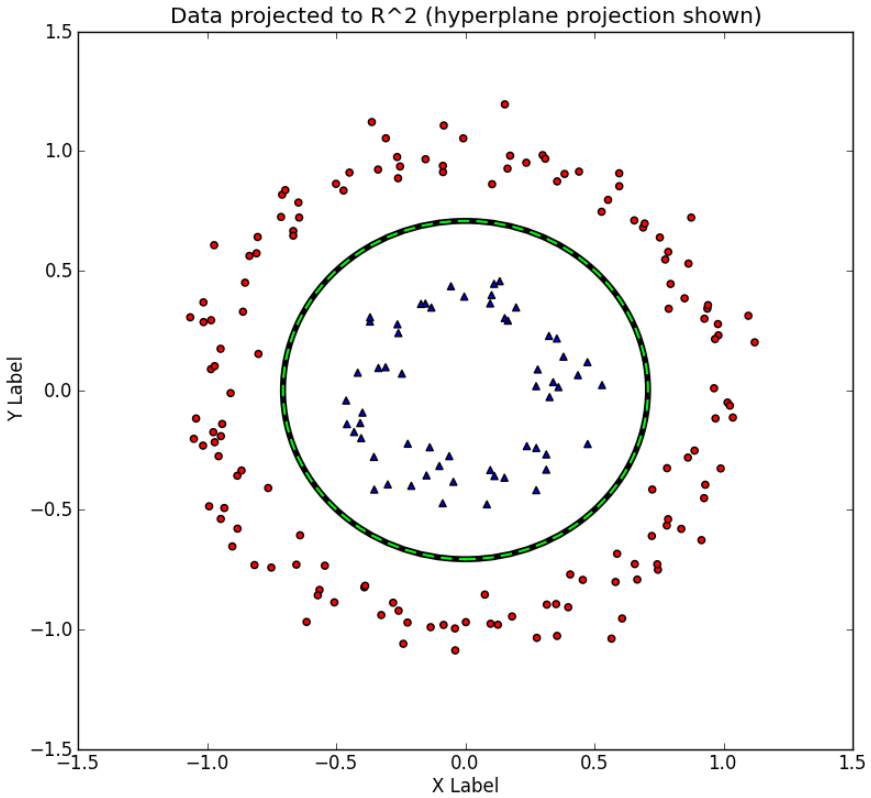


# Selecting Model Parameters

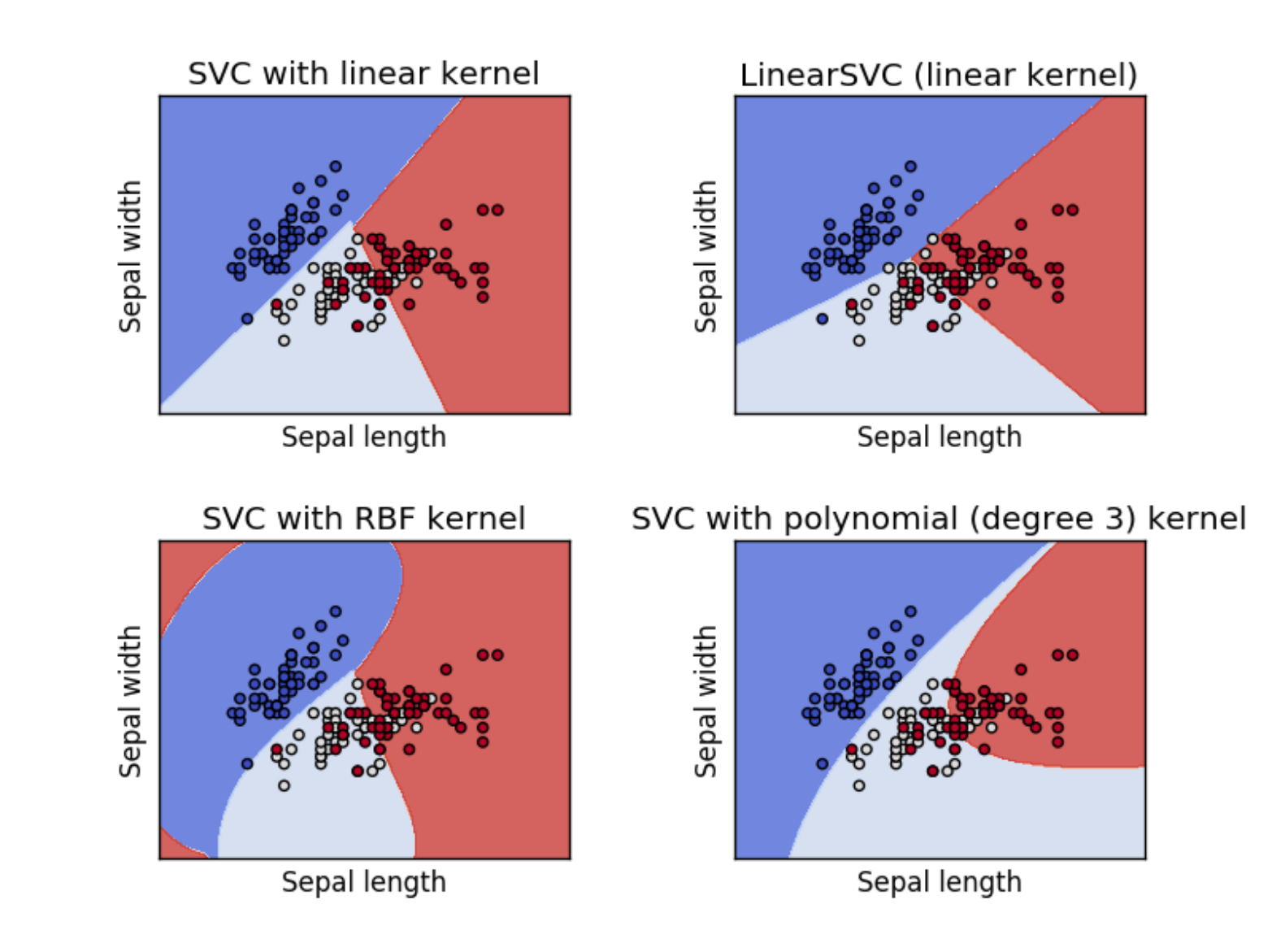


# Selecting Model Parameters



# Selecting Model Parameters



# Results with default RBF kernel

| Featureset | Accuracy |
|---|---|
| BoW | 44% |
| BoW + characters | 46% |
| BoW + characters + centuries | 47% |

# Results with linear kernel

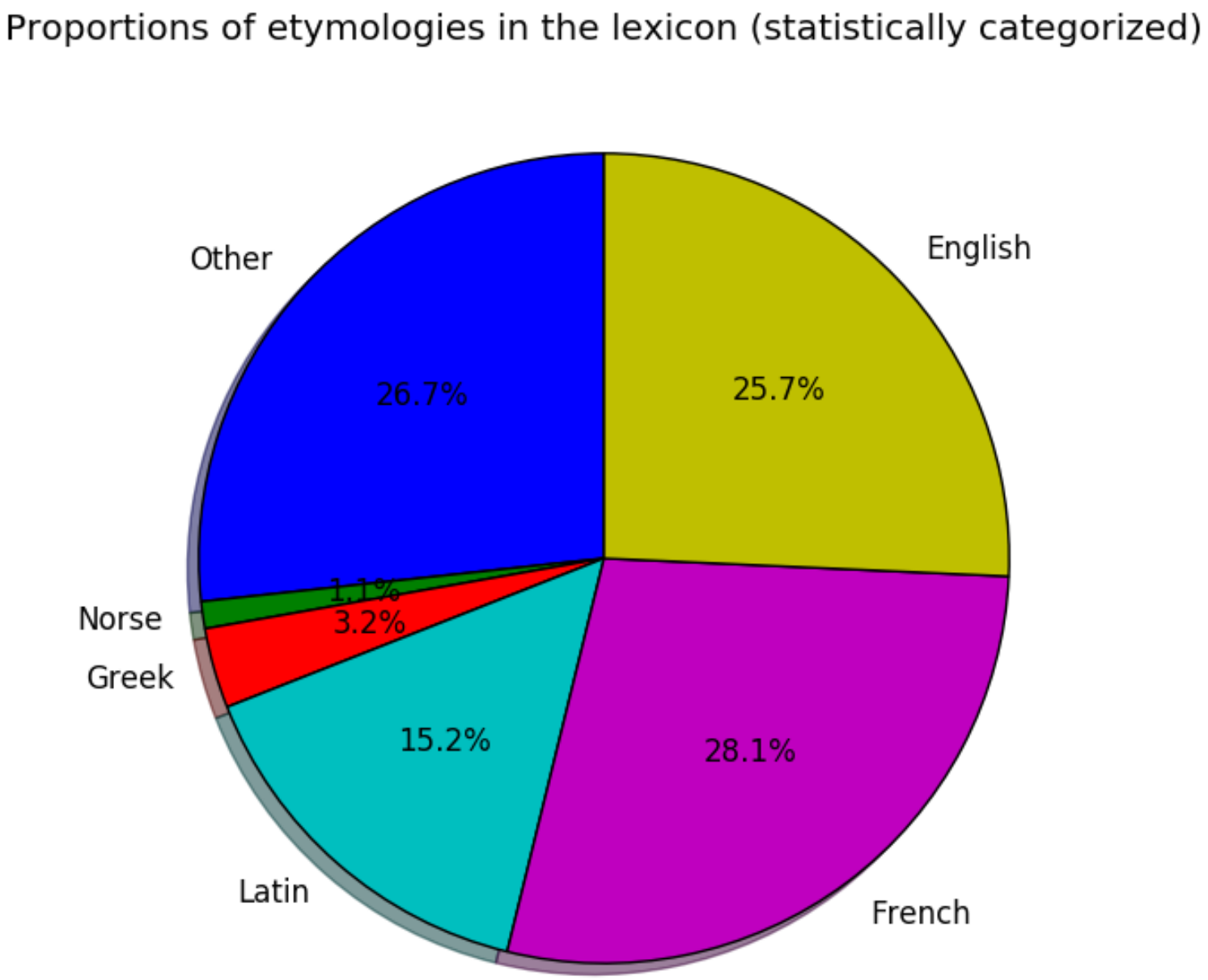
| Dataset | F-score |
|---|---|
| BoW | 88.96% |
| BoW + characters | 89.08% |
| BoW + characters + centuries | 89.15% |

# Results with linear kernel

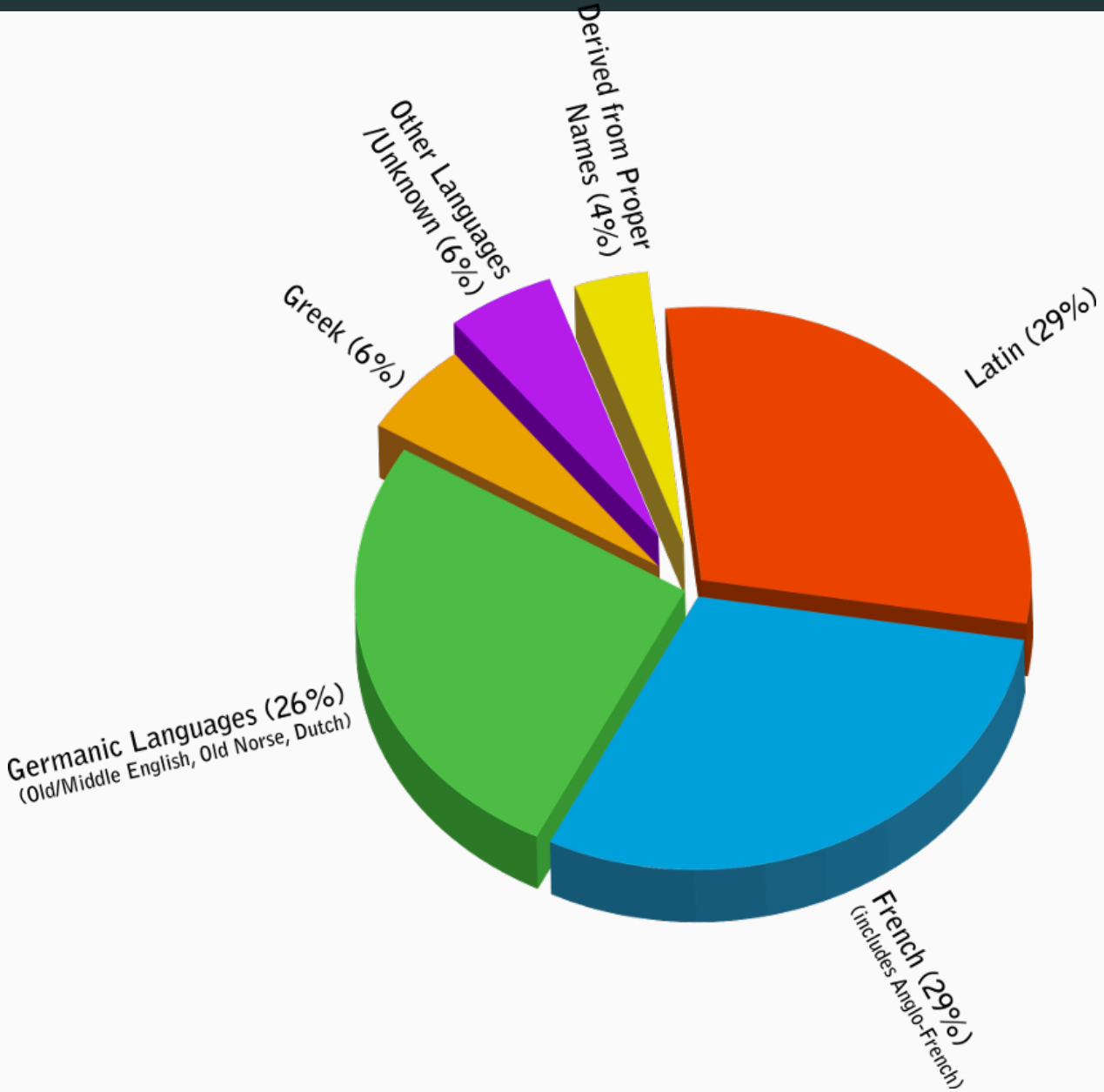
| Features | Accuracy |
|---|---|
| Syllables | 0.48 |
| BagofWords | 0.71 |
| Characters | 0.72 |
| Syllables + BagofWords | 0.73 |
| Syllables + Characters | 0.75 |
| BagofWords + Characters | 0.81 |
| BagofWords + Characters + Syllables | 0.82 |
| Syllables + Characters + Century | 0.75 |
| BagofWords + Characters + Century | 0.81 |
| All Four Features | 0.82 |

# Lexicon

# Lexicon


Proportions of etymologies in the lexicon (statistically categorized)
Other
26.7%
English
25.7%
Norse
1.1%
Greek
3.2%
15.2%
Latin
28.1%
French

# Lexicon


French (29%)
(Includes Anglo-French)
Germanic Languages (26%)
(Old/Middle English, Old Norse, Dutch)
Greek (6%)
Other Languages
/Unknown (6%)
Derived from Proper
Names (4%)
Latin (29%)

# Other? (12209)

['discography (n.)', 'governmental (adj.)', 'shinny (n.)', 'red herring (n.)', 'play-dough (n.)', 'disorientation (n.)', 'prophetess (n.)', 'industrialisation (n.)', 'quirky (adj.)', 'uncirculated (adj.)', 'transmittal (n.)', 'Guinevere', 'acridity (n.)', 'lit (n.2)', 'Lee-Enfield', 'prat (n.)', 'duffel', 'karyotype (n.)', 'Clarisse', 'movies (n.)', 'sporran (n.)', 'piss-pot (n.)', 'merman (n.)', 'animalism (n.)', 'glow (n.)', 'consortia (n.)', 'Levantine (adj.)', 'linguistic (adj.)', 'mutagen (n.)', 'masochistic (adj.)', 'profusive (adj.)', 'blur (v.)', 'quizzical (adj.)', 'flap (n.)', 'unheeded (adj.)', 'realistic (adj.)', 'proselytization (n.)', 'po-face (adj.)', 'maturate (v.)', 'time-out (n.)', 'unrestricted (adj.)', 'Sheol (n.)', 'usb', 'impermanent (adj.)', 'deterministic (adj.)', 'misappropriation (n.)', 'Gabriel', 'clabber (n.)', 'inauthentic (adj.)', 'Nathan', 'Chinaman (n.)', 'extremism (n.)', 'metrosexual (adj.)', 'wham (n.)', 'chad (n.3)', 'demarcate (v.)', 'shivaree (n.)', 'knack (n.)', 'spherical (adj.)', "Hell's Kitchen", 'dogwood (n.)', 'Spackle (n.)', 'festschrift (n.)', 'homogenization (n.)', 'visualize (v.)', 'Sacramento', 'troika (n.)', 'ceriph (n.)', 'latency (n.)', 'Yahweh', 'antagonise (v.)', 'disambiguation (n.)', 'squander (v.)', 'litre (n.)', 'red

# Other? (12209)

**linguistic (adj.)** "of or pertaining to the study of language," 1824, from German linguistisch (1807); see linguist + -ic.

**Nathan** masc. proper name, biblical prophet, from Hebrew Nathan, literally "he has given," from verb nathan, related to mattan "gift."

**consortia (n.)** plural of consortium

**quirky (adj.)** 1806, "shifty," from quirk + -y (2). Sense of "idiosyncratic" first recorded 1960. Related: Quirkily; quirkiness.

**mutagen (n.)** 1946, from mutation + -gen "thing that produces." Related: Mutagenic; mutagenesis; mutagenize.

# Testing

# Testing dataset

# Testing dataset

**The Brown Corpus**

# Testing dataset

**The Brown Corpus**

~1 million words

# Testing dataset

**The Brown Corpus**

~1 million words

Categorized

# Testing dataset

| | Broad text category | Text category letter and description ("genre") | | Number of texts | | | |
|---|---|---|---|---|---|---|---|
| | | | | Brown | Frown | LOB | FLOB |
| Informative | Press | A | Press: Reportage | 44 | same as Brown | | |
| | | B | Press: Editorial | 27 | " | " | " |
| | | C | Press: Reviews | 17 | " | " | " |
| | General Prose | D | Religion | 17 | " | " | " |
| | | E | Skills, Trades and Hobbies | 36 | | 38 | |
| | | F | Popular Lore | 48 | | 44 | |
| | | G | Belles Lettres, Biographies, Essays | 75 | | 77 | |
| | Learned writing | H | Miscellaneous: Government documents, industrial reports etc. | 30 | same as Brown | | |
| | | J | Science | 80 | " | " | " |
| Imaginative | Fiction | K | General Fiction | 29 | " | " | " |
| | | L | Mystery and Detective Fiction | 24 | " | " | " |
| | | M | Science fiction | 6 | " | " | " |
| | | N | Adventure and Western | 29 | " | " | " |
| | | P | Romance and Love story | 29 | " | " | " |
| | | R | Humour | 9 | " | " | " |

# Testing dataset

| Tag | Meaning | Tag | Meaning |
|---|---|---|---|
| . | (. ; ? *) | NN | singular or mass no |
| BER | are, art | NNS | plural noun |
| BEZ | is | NP | proper noun |
| CC | conjunction | PN | nominal pronoun |
| CD | cardinal numeral | PPL | (myself) |
| CS | subordinator | PPLS | (ourselves) |
| DO | do | PPS | 3(he, she, it, one) |
| HVN | had | RN | nominal adverb |
| IN | preposition | UH | interjection, exclam |
| JJ | adjective | VB | verb, base form |
| JJR | comparative adjective | VBD | verb, past tense |
| JJS | superlative adjective | VBN | verb, past participle |
| MD | modal auxiliary | WQL | wh- qualifier |

# Testing dataset

Problem: Etymonline uses the base form of every word, and a much simpler tag system

# Testing dataset

Problem: Etymonline uses the base form of every word, and a much simpler tag system

Fix: Big tag decomposer

1. Is the word in the dictionary alone?
2. Is the word (lower case) in the dictionary alone?
3. Is the word in the dictionary followed by " ("
4. Reduce Brown POS tag (skip punctuation, B,H,V -¿ v, N -¿ n J -¿ adj)
5. Is the word with the tag in the dictionary?
6. Is the word with the tag and a number in the dictionary?
7. Change tags to wordnet format (adj -¿ a adv -¿ v)
8. Lemmatize and recurse (stopping at one recursion level)

# Example Runs

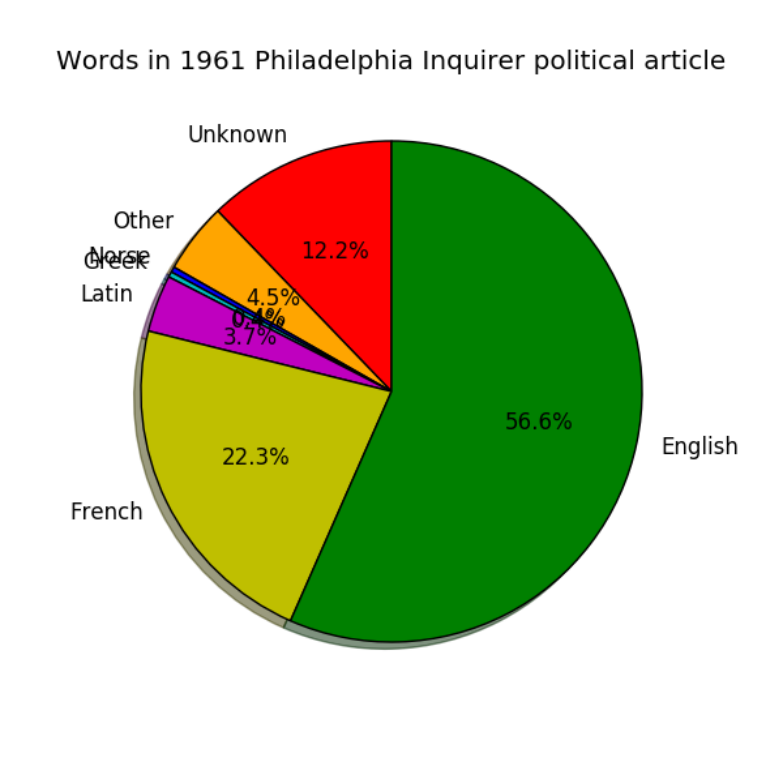
# Reminder of Hypotheses from October

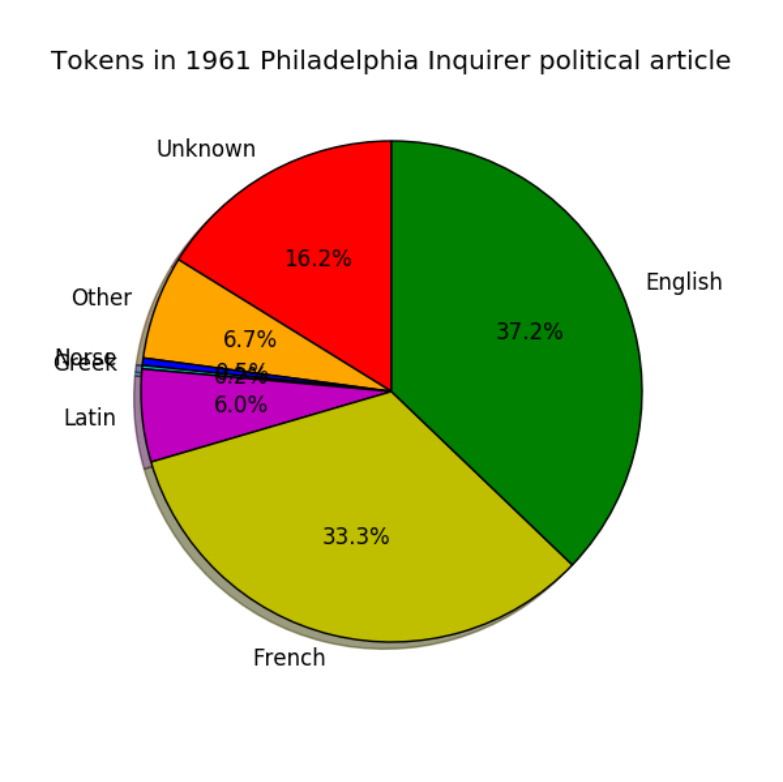
**Abstract**

I expect to see English literature as mostly Germanic, upwards of 60%. In scientific texts, I expect to see Latin and Greek push out a lot of that English vocabulary, and French in legal texts

# Political Article

# Political Article

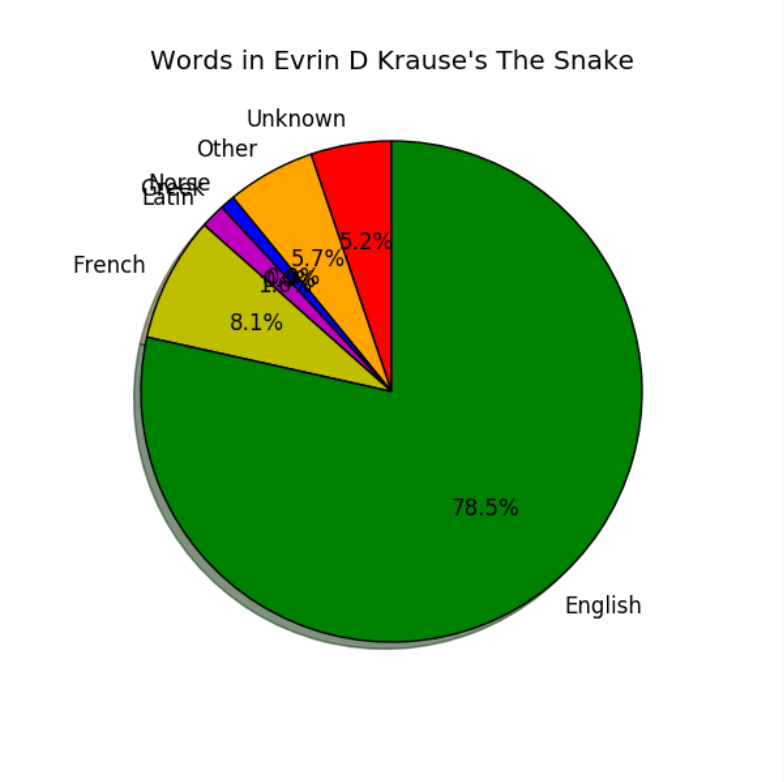


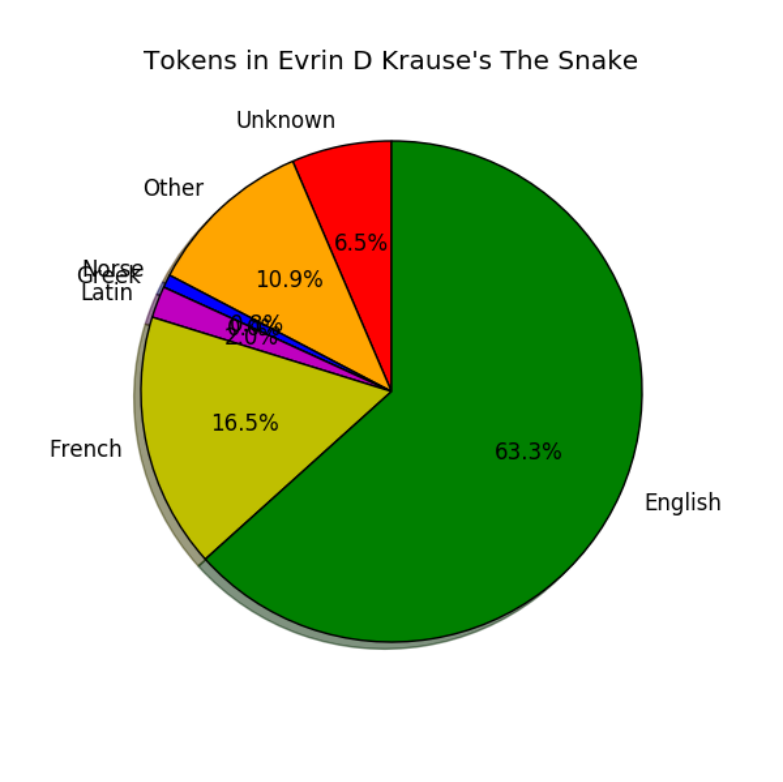



# Unknown?

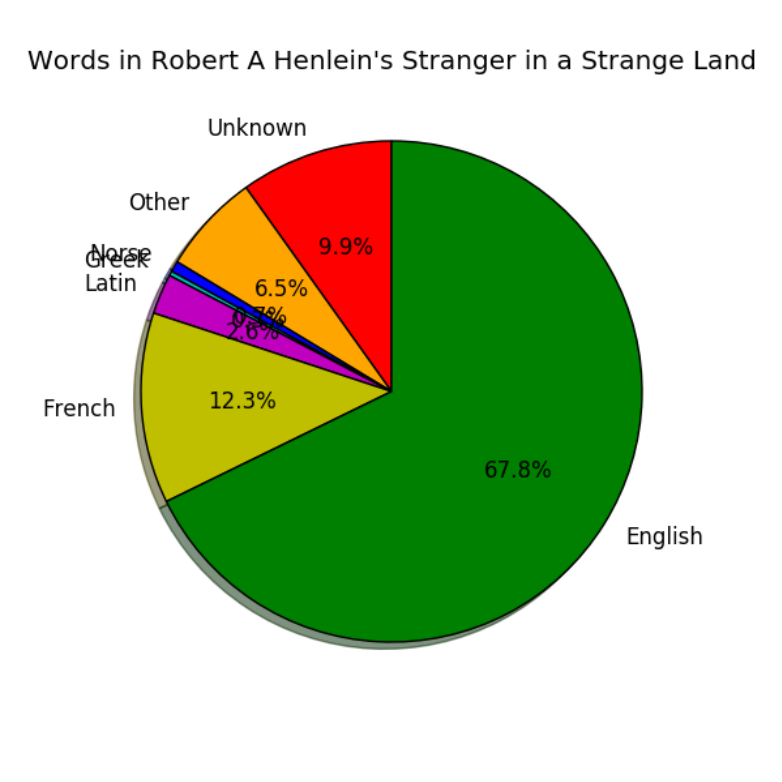
('Hemphill', 'NP'), ('Frankford', 'NP-TL'), ("city's", 'NN$'), ('$344,000', 'NNS'), ('$200,000', 'NNS'), ('Hemphill', 'NP'), ('Hughes', 'NP-TL'), ('impossibly', 'QL'), ('Hughes', 'NP'), ('Hemphill', 'NP'), ("Controller's", 'NN$-TL'), ('investigation', 'NN'), ('$172,400', 'NNS'), ('investigating', 'VBG'), ('its', 'PP$'), ('V.', 'NP'), ('Varani', 'NP'), ('vouchers', 'NNS'), ('Varani', 'NP'), ('C.', 'NP'), ('Wagner', 'NP'), ('Hemphill', 'NP'), ('by', 'IN'), ('Berger', 'NP'), ('Wagner', 'NP'), ('PTC', 'NN'), ('its', 'PP$'), ('Hughes', 'NP'), ('N.', 'JJ-TL'), ('2d', 'OD-TL'), ('St.', 'NN-TL'), ('Hughes', 'NP'), ('its', 'PP$'), ('investigation', 'NN'), ('by', 'IN'), ('Hemphill', 'NP'), ('C.', 'NP'), ('Crumlish', 'NP'), ('Jr.', 'NP'), ('Hughes', 'NP'), ("Berger's", 'NP$'), ('Hughes', 'NP'), ('U.', 'NP-TL'), ('S.', 'NP-TL'), ('Berger', 'NP'), ('mostly', 'RB'), ('involve', 'VB'), ('overhauling', 'NN'), ('Hemphill', 'NP'), ('Hemphill', 'NP'), ('Hughes', 'NP'), ('$500', 'NNS'), ('A.', 'NP'), ('Belanger', 'NP'), ('Mass.', 'NP'), ('Hughes', 'NP'), ('$600', 'NNS'), ('Hemphill', 'NP'), ('$2400', 'NNS'), ('$3100', 'NNS'), ("Berger's", 'NP$'), ('by', 'IN'), ('Wagner', 'NP'), ('Wagner', 'NP'), ('$37,500', 'NNS'), ('Hughes', 'NP'), ('know', 'VB'),

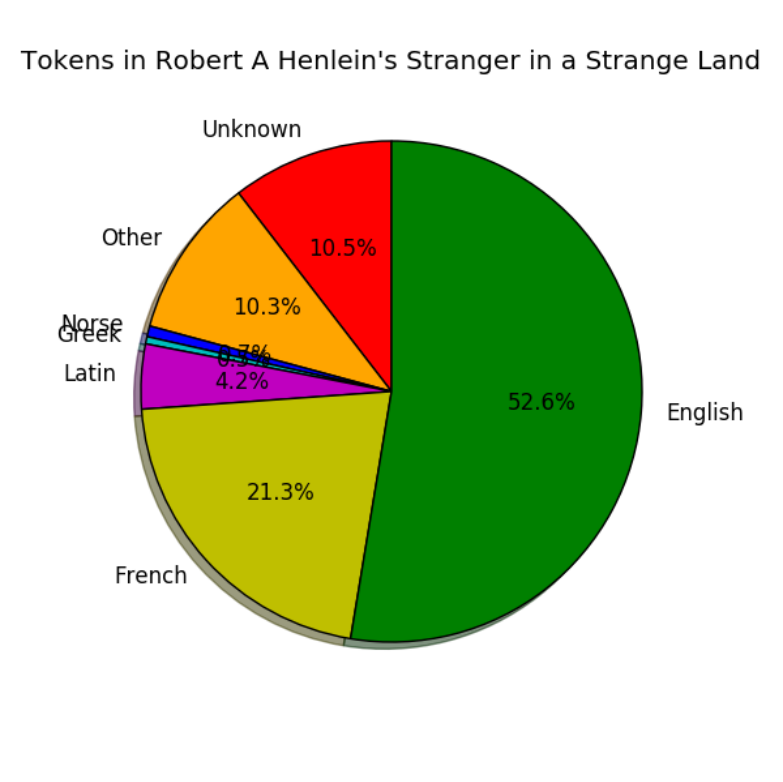
# Romance Novel



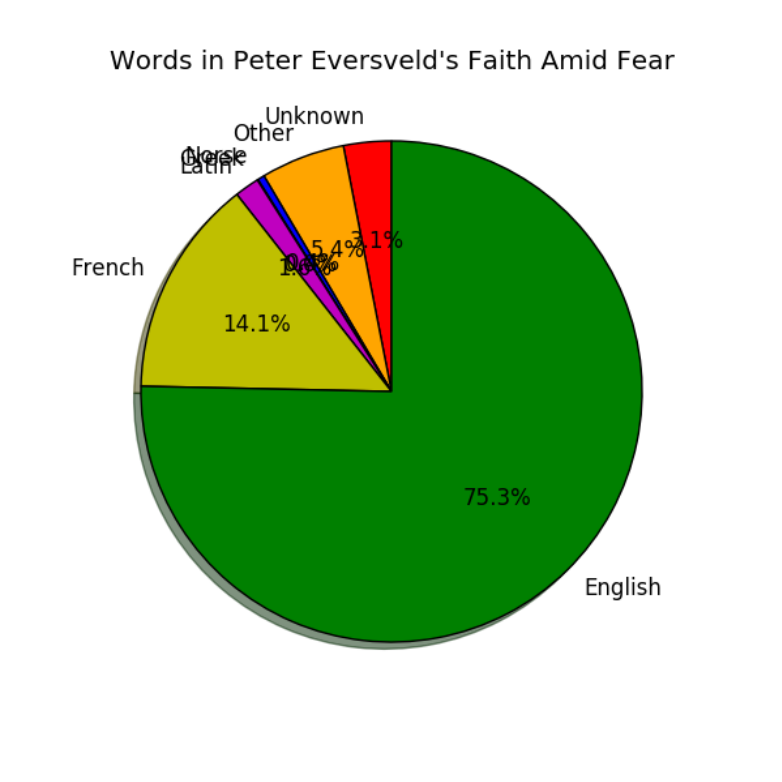



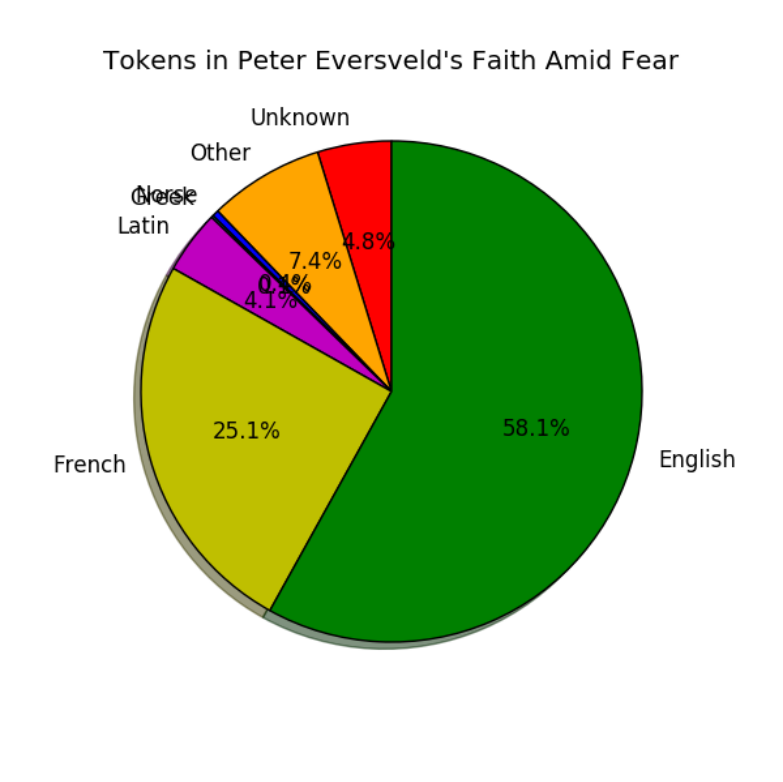
# Sci-Fi


Words in Robert A Henlein's Stranger in a Strange Land
Unknown
Other
Norse
Greek
Latin
French
English
9.9%
6.5%
12.3%
67.8%

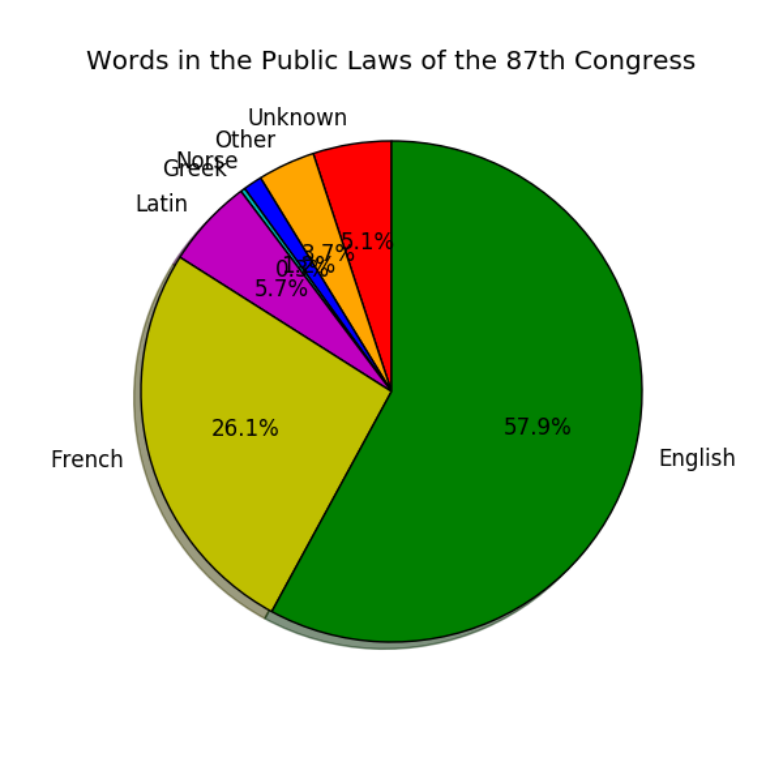

Tokens in Robert A Henlein's Stranger in a Strange Land
Unknown
Other
Norse
Greek
Latin
French
English
10.5%
10.3%
4.2%
52.6%
21.3%

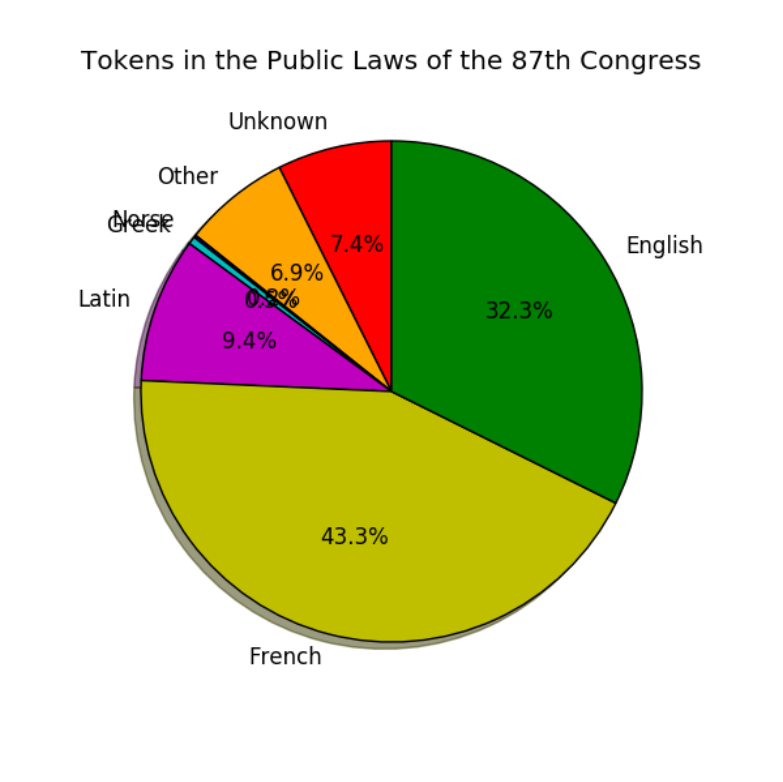
# Religion


Words in Peter Eversveld's Faith Amid Fear
Unknown
Other
Latin
French
14.1%
5.4%
3.1%
75.3%
English

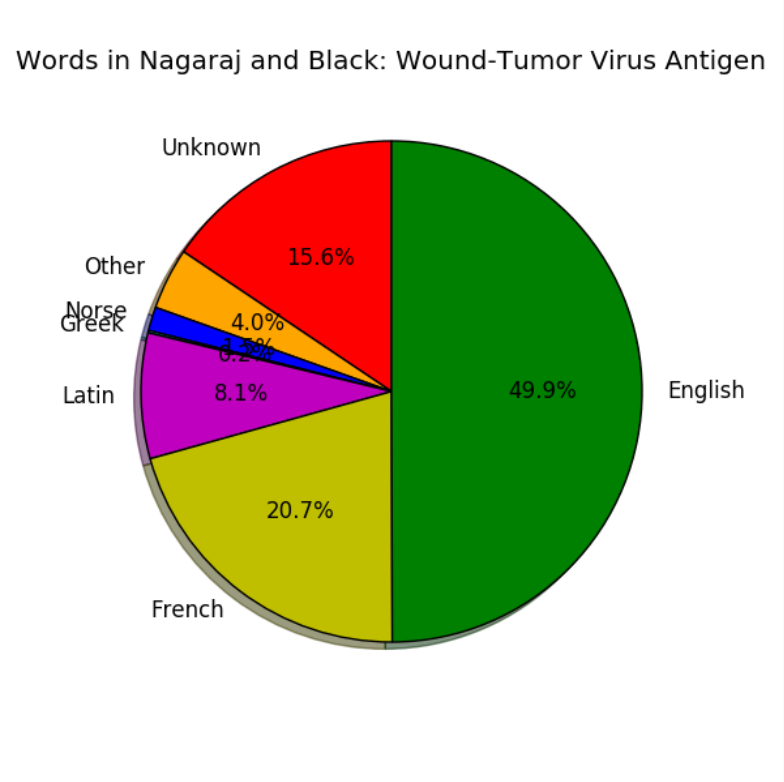

Tokens in Peter Eversveld's Faith Amid Fear
Unknown
Other
Latin
7.4%
4.8%
4.1%
25.1%
58.1%
French
English

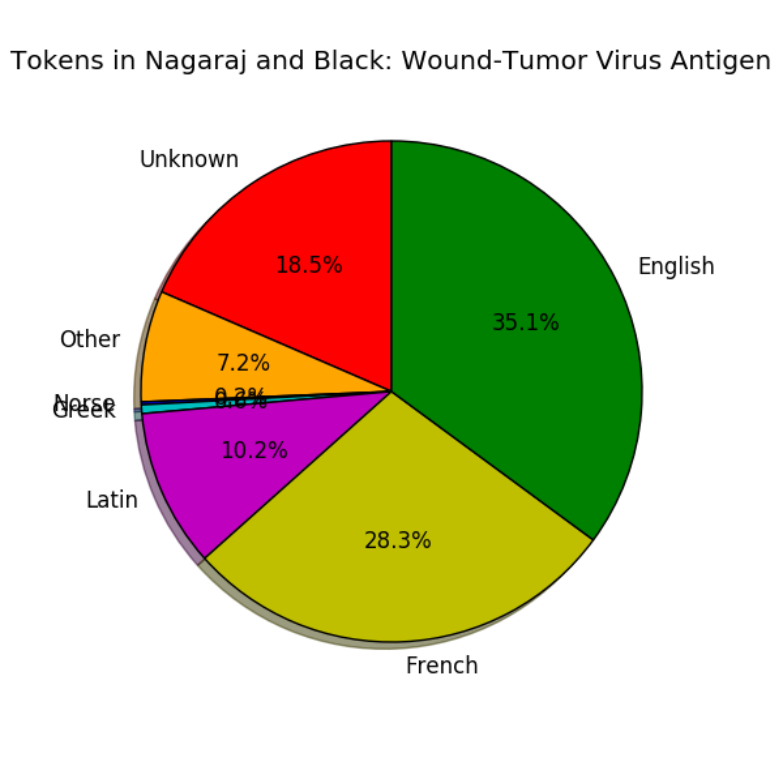
# Law


Words in the Public Laws of the 87th Congress
Unknown
Other
Norse
Greek
Latin
5.1%
3.7%
5.7%
26.1%
57.9%
French
English

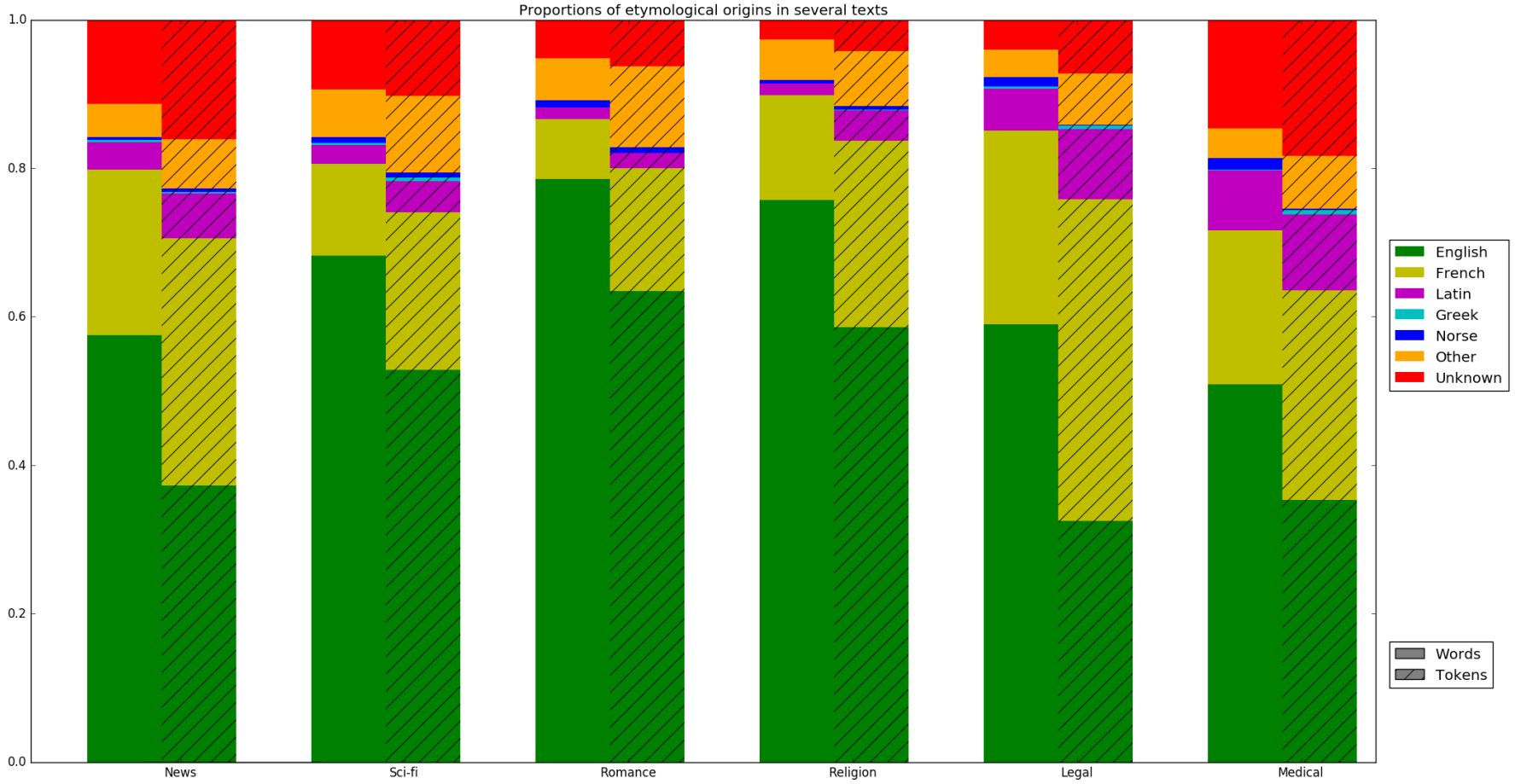

Tokens in the Public Laws of the 87th Congress
Unknown
Other
Norse
Greek
Latin
7.4%
6.9%
9.4%
32.3%
43.3%
English
French

# Medical


Words in Nagaraj and Black: Wound-Tumor Virus Antigen
Unknown
15.6%
Other
4.0%
Norse
Greek
Latin
8.1%
49.9%
English
20.7%
French


Tokens in Nagaraj and Black: Wound-Tumor Virus Antigen
Unknown
18.5%
English
35.1%
Other
7.2%
Norse
Greek
10.2%
Latin
28.3%
French

# All Test Texts



# Next Steps

- Go to ACL

# Next Steps

- Go to ACL
- Word-only classifier

# Next Steps

- Go to ACL
- Word-only classifier
- Tagger

# Next Steps

- Go to ACL
- Word-only classifier
- Tagger
- User-friendly frontend

# Next Steps

- Go to ACL
- Word-only classifier
- Tagger
- User-friendly frontend
- Advertisements

# Next Steps

- Go to ACL
- Word-only classifier
- Tagger
- User-friendly frontend
- Advertisements
- Reddit

# Next Steps

- Go to ACL
- Word-only classifier
- Tagger
- User-friendly frontend
- Advertisements
- Reddit
- **$$$$$$$$$$$$$$**