The growth and evolution of the meaning space spanned by the English language over the last thousand years

Fletcher Forrest Hazlehurst,[1,*] Christopher M. Danforth,[1,†] and Peter Sheridan Dodds[1,‡]

[1] *Department of Mathematics & Statistics, Vermont Complex Systems Center, Computational Story Lab, & the Vermont Advanced Computing Core, The University of Vermont, Burlington, VT 05401, USA*

(Dated: March 1, 2017)

Language is an essential part of being human and with the advent of computers and large text corpora, a new avenue of research is available to advance the science of linguistics. Here, we present a computational framework for empirically investigating how the shared orthographic representations of words, whose meanings are organized by a hierarchy of semantic categories, provide insight into the current and past structure of the English language. We demonstrate how changes in the English language arising from many different linguistic processes reflect large-scale societal trends and events that have occurred. For example, we find the vocabulary relating to *Leisure* to have dramatically risen in the last 200 years, resulting in *Leisure* becoming the current dominant semantic domain. We also use our framework to provide inspiration for future work on examining the structure and dynamics of words and their semantic domains.

I. INTRODUCTION

Language as a communicative device is a fundamental aspect of being human, but language may provide more than a means to communicate. Studying human language may also provide insight into one of the final frontiers of science, human cognition. Lakoff and Johnson [1] demonstrated that conceptual metaphor, a linguistic process consisting of using language organized by one conceptual domain to describe and think about another, is pervasive in the English language. Examples from their book include: *I saved a lot of time* as an example of the conceptual metaphor TIME IS MONEY and *I attacked his weak points* as an example of ARGUMENT IS WAR. The book argued that the widespread use of conceptual metaphors indicates how this linguistic process may reflect a deeper cognitive process dedicated to our comprehension of many different domains, especially abstract ones. If this is true, then there is much to be gained by analyzing the different types of metaphor that we use to better understand language and it's role in human culture and cognition. Specifically, if we can systematically map out how words shift from one semantic domain to another over time, we can learn more about how conceptual metaphor and other similar linguistic processes develop, and how humans make use of these processes to assist in and or maybe even structure their cognition.

In this spirit, the Mapping Metaphor Project [2] has attempted to manually identify all metaphors in the history of the English language. This was accomplished using a categorical hierarchy of all words in the English language compiled in The Historical Thesaurus [3]. Semantic categories such as *Anger*, *Poverty*, and *Punishment* are compared to others such as *Chemistry*, *Vision*, and *Weather* to find metaphors in specific words that link the categories. An example metaphor between two categories could be the use of vocabulary describing the weather (*storm* or *tornado*) to also describe anger. The Mapping Metaphor Project found over 10,000 metaphors between 411 major semantic domains of English [2].

In this paper, we use the inspiration behind the Mapping Metaphor Project to create a computational framework for a general analysis of the major semantic domains of English, and visualize connections among them changing over time. This framework includes an interactive website [4] with different plots and tables to explore the semantic domains of English, the connections between them, and how these domains and connections have changed over the last 1000 years. Here, we describe our methodology to create the framework and then present five stories relating linguistic change to large-scale societal trends and events.

This paper is organized as follows: in Sec. II, we introduce the Historical Thesaurus and the categorical hierarchy it contains; in Sec. III, we review the work done by the Mapping Metaphor Project; Sec. IV presents our methodology and explains each part of the framework; in Sec. V, we introduce five sets of figures from the framework and explain their significance; finally, Sec. VI offers some commentary on the project and summarizes some possible directions for future work.

II. THE HISTORICAL THESAURUS

The Historical Thesaurus [3] is a categorical hierarchy of the English language from Old English (~1000 AD) to the year 2000 created using the *Oxford English Dictionary* and *A Thesaurus of Old English* [5]. Almost 800,000 words are contained in this electronic thesaurus,

* fletcher.haz@gmail.com
† chris.danforth@uvm.edu
‡ peter.dodds@uvm.edu

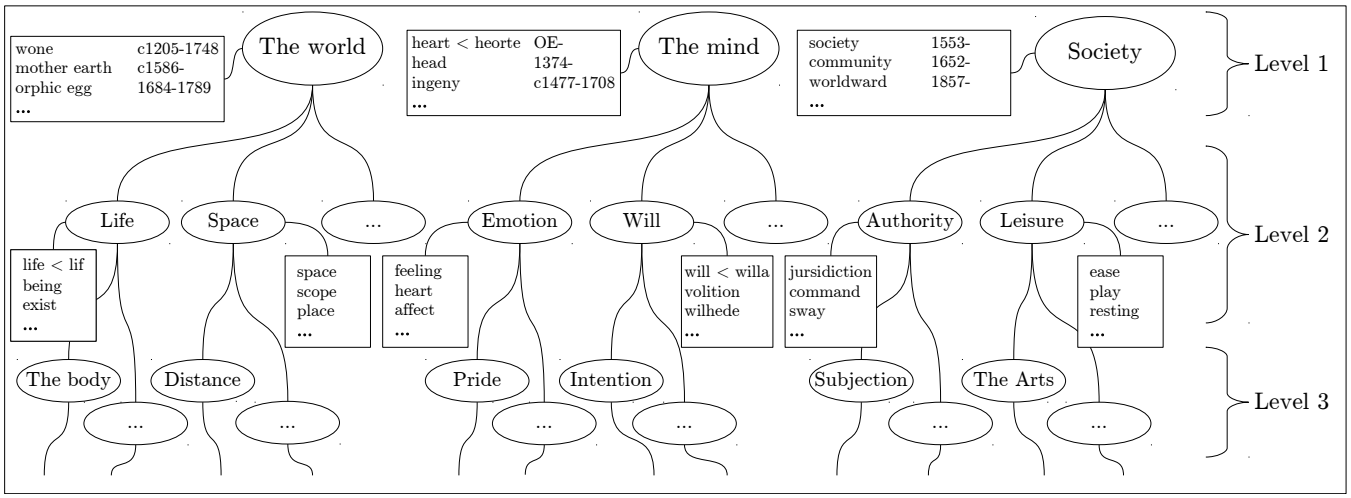

FIG. 1. Top of the category hierarchy in the Historical Thesaurus. Shown are three categories at level one: **The world**[1], **the mind**[1], and **Society**[1] and several of the descendants of those categories appear in lower levels. All categories are in ellipses and word lists of synonyms for the category title appear in rectangular boxes.

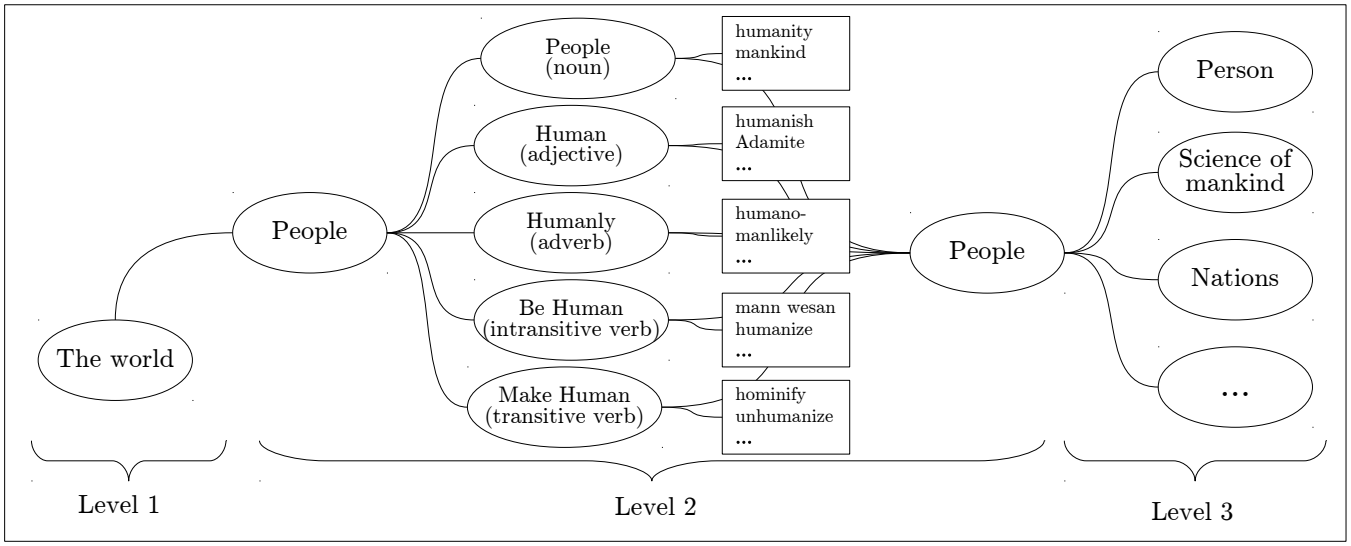

FIG. 2. Expansion of a single category to display the different parts of speech. The configuration shown here is how the Historical Thesaurus accounts for parts of speech. For our work, all parts of speech (middle five ellipses) and their word lists (boxes) are combined into one category usually named after the noun, **People**[2] in this instance.

and the authors of the thesaurus created a hierarchical semantic structure to organize the words. Each word was placed into a category based upon the word's definition, along with it's synonyms. Each category was then positioned in the hierarchy, of which the highest levels are shown in Fig. 1.

At the highest level of the thesaurus are three categories: **The world**[1], **The mind**[1], and **Society**[1]. We refer to this level of the thesaurus as level 1 with subordinate levels increasing in number. In the rest of this paper, we will denote categories of the thesaurus as follows: **Category Name**^{CategoryLevel}. Note that in Fig. 1 there are words, in rectangular boxes, attached to each category. These are the synonyms of the category name. For example, *mother earth* and *wone* are just two of the synonyms for *The world*.

Each level 1 category also has a number of level 2 categories underneath it. As shown in Fig. 1, **Life**[2] and **Space**[2] are both underneath **The world**[1], while **Emotion**[2] and **Will**[2] are underneath **The mind**[1]. **Life**[2] and **Space**[2] also have synonym lists associated with them, shown in attached rectangular boxes in Fig. 1. Similarly, **Life**[2] and **Space**[2] both have more categories underneath, at level 3. Finally, **The body**[3] and **Distance**[3] are under **Life**[2] and **Space**[2] respectively.

There are 3 level-1 categories, 37 level-2 categories, 377 level-3 categories, and 1,927 level-4 categories. The hierarchy tree is not a complete tree, for example many

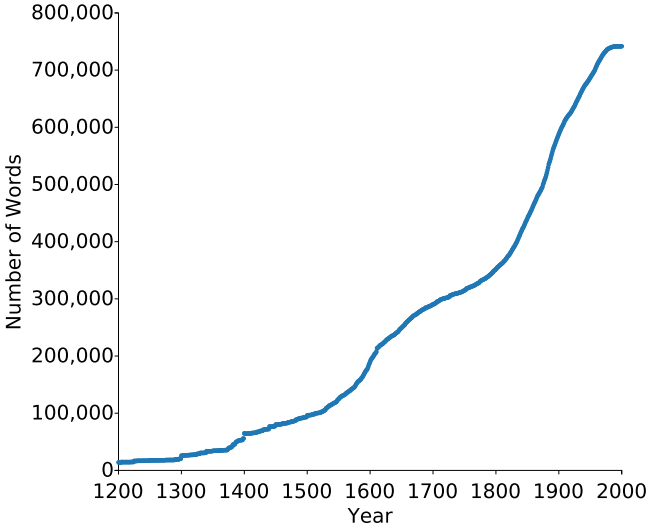

FIG. 3. Number of words (not word forms) in the Historical Thesaurus over time. Words are added and lost at every year. Old English words are included only if they exist after 1200. Resolution is on one year intervals

paths stop at a depth of three or four. The maximum depth of a category in the tree is 7 and there are a total of 8,561 distinct categories in the hierarchy.

However, the above description of the hierarchy is complicated by a feature of the thesaurus not shown in Fig. 1 and this feature is how the thesaurus separates linguistic representations of the category by their part of speech. **The world**[1] is actually three different categories: **The world**[1] (nouns), **Pertaining to the world**[1] (adjectives), and **Earthly**[1] (adverbs). Each of these three categories has a separate list of synonyms associated with it, but all three categories share the same children and parent. This is shown in Fig. 2, with **People**[2] as the category in focus. Here we can see that **The world**[1] is the parent category and no distinction is made for part of speech when displaying the parent in the hierarchy of the category in focus. **People**[2] is split into five different parts of speech: noun, adjective, adverb, transitive verb, and intransitive verb. Each of these parts of speech has a list of synonyms (shown in rectangular boxes). Finally, all the parts of speech of **People**[2] share the same children because, as before, part of speech is not represented except for the category in focus when displaying the hierarchy.

Also shown in the word lists for level 1 categories in Fig. 1 are the dates that each word was in use. Dates, which can range from *OE* for Old English (circa AD 1000) to 2000, are listed for all words in the thesaurus. Importantly, we must consider the accuracy of a word's starting and ending when drawing conclusions from the historical thesaurus data. Words are dated by of their first known use, which can be notoriously difficult to pinpoint accurately. Surprisingly, the curve representing the number of words currently in use, shown in Fig. 3, is smooth even when the resolution is brought down to one year intervals. We do not see sudden jumps or lulls, which we would expect from rounding or other estimation errors made by the lexicographers. Instead, word additions and losses over time appear as smooth curves with slowly changing derivatives. The one point where there is an abrupt change is the transition from Old English to Middle English (not shown in Fig. 3). Due to the extensive changes in the language during this time, our analyses in this paper begin after the transition from Old English.

The most important note about the Historical Thesaurus is that words with multiple distinct definitions can be placed in multiple different categories. For example, *hand* has a definition referring to the human body part at the end of the arm that dates back to Old English. This definition is categorized under **Hand**[7] → **Extremities**[6] → **Limb**[5] → **External parts of body**[4] → **The body**[3] → **Life**[2] → **The world**[1]. *Hand* also has a definition for the arrow on a clock face, which dates to 1575. This definition is categorized under **Hand**[6] → **Parts of a Clock**[5] → **Clock**[4] → **Instruments for measuring time**[3] → **Time**[2] → **The world**[1]. Yet another definition meaning *to give* comes from the United States in 1901 and is categorized under **Give**[3] → **Possession**[2] → **The mind**[1].

A single orthographic representation, a written word, can have many different meanings or definitions that appear in each of the three level 1 categories and each definition can have a different beginning and ending date associated with it. Here, each definition will be referred to as a *word*, while the orthographic representation referring to all possible definitions will be called the *word form*. Note that not every definition of a word form corresponds to a distinct word in the thesaurus. Some are left out based upon the methodology used by the thesaurus creators.

III. THE MAPPING METAPHOR PROJECT

The Mapping Metaphor project [2] expanded upon the Historical Thesaurus with the key insight of using the shared orthographic representations of words to identify relationships between categories of the Historical Thesaurus. Specifically, the task was to identify vocabulary from one category, which was also used as vocabulary for another category in a metaphorical nature.

For example, the adjective *boiling* is defined as "Bubbling up under the influence of heat; at boiling temperature" [6]. This definition dates to c1320 [6]. In the Historical Thesaurus, the first definition, about temperature, is classified under **Boiling**[6] → **Of/Pertaining to heat**[5] → **Having temperature of a specific kind**[4] → **Properties of materials**[3] → **Matter**[2] → **The world**[1]. In 1579, *boiling* appeared with a new definition, "Inflamed, in a state of passionate agitation, bursting with passion, etc." [6]. This second definition, regarding emotion, is classified under **Inflamed with passion**[5] →

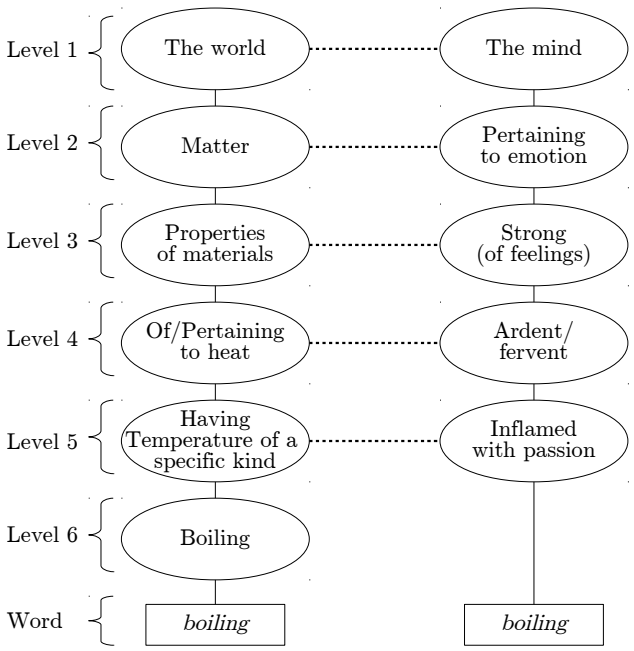

FIG. 4. Connections created between two paths of the hierarchy sharing a common descendent word, *boiling*. Each level of one path is connected to the same level on the other path. Connections are never made between categories on different levels (i.e. there is no connection between **The mind**[1] and **Matter**[2] even though they share the word *boiling*).

Ardent/fervent[4] → **Strong (of feelings)**[3] → **Pertaining to emotions**[2] → **The mind**[1]. This shared word form creates a connection between all the categories *boiling* is under, as shown by dashed lines in Fig. 4. This connection is clearly a conceptual metaphor as we are using a concrete physical process to describe an abstract mental emotion.

The Mapping Metaphor project is an attempt to systematically identify all of these metaphors using the Historical Thesaurus categories as a knowledge base for identifying the semantic domains of a word definition. The project chose to focus on level 3 categories from the thesaurus due to the appropriate specificity of these categories for creating conceptual metaphors. They started with 411 categories, roughly approximating the categories from level 3 of the Historical Thesaurus, although some changes were made (for unspecified reasons). The project then created word lists for each of these categories using words from the thesaurus. By comparing these word lists for matching word forms, connections similar to the ones shown in Fig. 4 were made, when possible, between all pairs of categories.

Each connection was then manually inspected for "systematic evidence of metaphor" [2] and classified as either showing strong, weak, or no metaphoricity. The Mapping Metaphor Project found about 10,000 metaphors labeled as strong or weak. One example is the connection between *Poverty* and *Bodily shape/physique*, which has shared word forms including "pinched, starved, withered, poorness, [and] feeble" [2] and is classified as a strong metaphorical connection. Manual inspection was necessary because the relationship between the orthographic representations could be due to metaphor, metonymy, noise or some other linguistic process [2].

However, manual inspection is expensive and time consuming and ignoring other linguistic processes may mean discarding useful information. Thus, we approach the problem computationally and examine all shared word forms dynamically in an attempt to learn about the history of these connections and how the vocabulary of the English language has expressed linkages of major semantic domains over time.

IV. MODEL

We focused upon level 2 and 3 of the historical thesaurus for reasons similar to decisions made by the Mapping Metaphor project. These categories appear, qualitatively, to be of the best specificity for the information we wish to gather. Furthermore, as categories become more granular, they become increasingly difficult to qualitatively analyze. Level 4 contains 1,927 categories, a number too large for the visualizations made to interpret the data.

In contrast to the Mapping Metaphor project, we exactly followed the level 2 and 3 categories of the thesaurus. When creating the list of level 3 categories, a choice must be made to either collapse the different parts of speech into one category or keep them separate. Separating the categories dramatically increases their number (from 377 to 1,914 for level 3) and therefore the complexity of the analysis. We performed our analysis using collapsed parts of speech for this reason and because of our interest in broad relationships between general semantic areas. However, our interactive website provides the ability to examine the results using either methodology.

Each category had a word list consisting of all words at or below that category. For example, **The mind**[1] has a word list of consisting of all words for the mind (attached rectangular box in Fig. 1), and all words for the categories underneath it: **Emotion**[2], **Will**[2], **Pride**[3], **Intention**[3], etc. In this work we combine the word lists of each part of speech to generate the word list for each category. In Fig. 2, all words under the five parts of speech of **People**[2] are considered part of the **People**[2] category. The name of the category, **People**[2] in this case, is given by the first part of speech encountered in the list of categories in the thesaurus. This is generally the noun part of speech, but is always the most important part speech for that category.

Since every category has a word list associated with it, we create connections among categories by comparing the word lists between every pair of categories at a given level. A word form will be in the shared word form list if it is used in both categories. A connection is creat-

ed between categories if there is at least one word form shared at some point in time between the two categories. The connection's weight is the number of word forms the two categories share.

Furthermore, because every word has a beginning and ending date associated with it, the connection has differing weights at every year. Connections between categories can increase in strength as words are added and lose strength when words are lost. Many connections do not exist in Old English but do in modern English. These connections between nodes create a network that we can then analyze.

We produced stacked area charts for category size, node degree, edge weight, and normalized edge weight. Category size is simply the raw size of the word list for each category at the given level. An example is shown in Fig. 5. The positive series represent the current number of words in the category and are ranked from bottom to top by size in the year 2000. The ability to rank by any year is available on the website. The negatives series represent the total number of lost words for the given category up to that year. Thus, the negative series are all monotonically increasing but the positive series are not.

Node degree is the degree (number of connections) each node (category) has at each year. The positive series are the node degree at the given year (not monotonic) while the negative series show the total number of lost connections up to the given year (monotonically increasing). Again, these plots are ranked from bottom to top based upon a specified year. We do not use any of these plots for this paper, but note that they are available on the accompanying website [4].

Edge weight is the connection strength or the number of shared word forms between two categories. If we consider each category to consist of a set of the word forms then the edge is the intersection of two sets and the edge weight is the cardinality of this intersections. We can also normalize by the cardinality of the union of the two sets to achieve a normalized edge weight. These plots inform about which connections have the greatest absolute strength and which connections are the strongest given the number of possible words to share. Plots of raw edge weights have positive series representing the current value and negative series representing the sum of lost values (examples are in Fig. 12), while plots of normalized edge weights only have positive series of the current value.

Finally, each of the four plots described above have two other variants that display the changes occurring in each year. One type, not used in this paper, displays the number of added words, degrees, or edge weight as positive series and lost values as negative series in a stacked area chart. Instead of totals, every 10 year increment is independent of the others. The other type, shown in Fig. 6, represents net change (added - lost) for independent 10 year increments, although the curve is smoothed to show general trends using a cardinal spline and tension of 0.7.

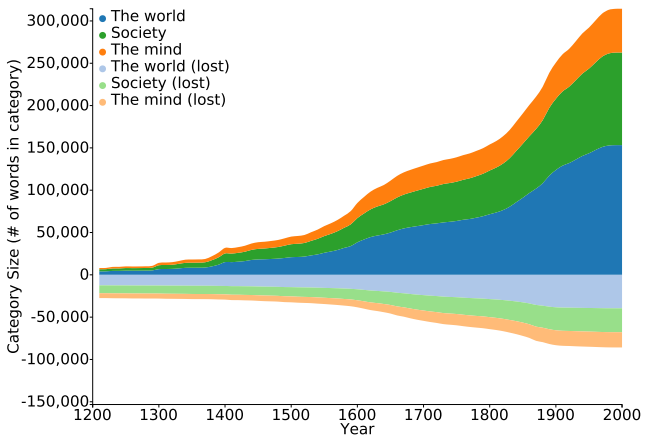

FIG. 5. Current number of words (positive) and total number of lost words (negative) for level 1 categories.

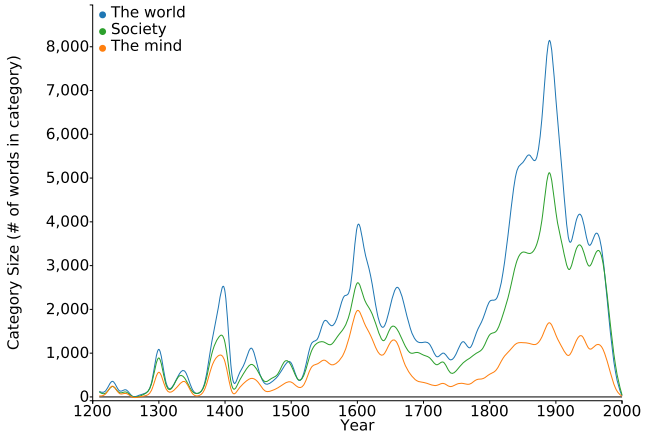

FIG. 6. Net change (added word - lost words) for level 1 category size for 10 year intervals. Curve is smoothed using a cardinal spline with a tension of 0.7.

Note that for all plots we use a resolution of 10 years in order to smooth the curves and alleviate some of the rounding or estimation errors possible in the dating of word usage. For example, all changes from 1991 through 2000 are considered to have happened at the year 2000. Furthermore, since the transition between Old English and Middle English is abrupt and disruptive to interpreting many of the plots, we have elected to remove those years from the work in this paper and instead focus on the years 1210-2000.

V. RESULTS

A. The Physical World is Dominant

When examining the stacked area charts for level 1 there is a large disparity between **The world**[1] and the

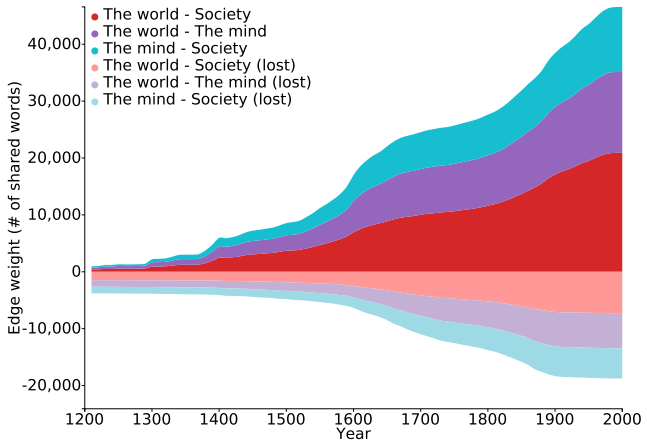

FIG. 7. Weight of connection (number of shared words) between level 1 categories.

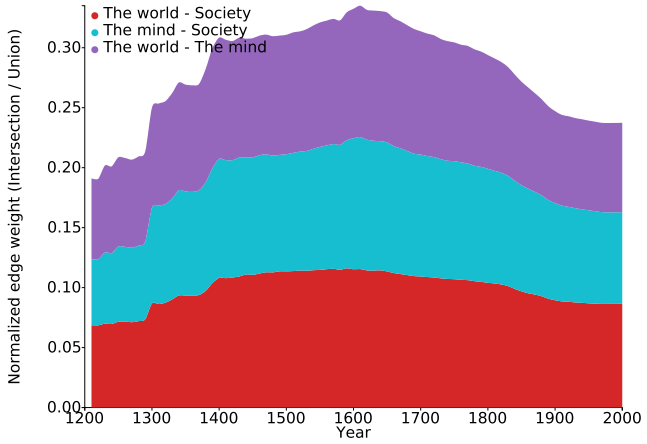

FIG. 8. Weight of connection (number of shared words) between level 1 categories normalized by possible number of words to share.

other two categories. The category size (number of words underneath in the hierarchy) over time for level 1 categories is shown in Fig. 5. **The world**[1] is clearly dominant throughout time and **Society**[1] is a clear second place. In large part, this is due to the many words used for classification of plants and animals as can be seen by examining the different parts of speech, which shows that noun categories are clearly where the majority of words reside. However, a coherent story of human language and cognition would predict many more words to describe the physical world than the mental and social worlds and this prediction is certainly borne out by Fig. 5.

Looking at the net change of category size for level 1 in Fig. 6, we can see only one period (1460-1490) through which **Society**[1] has a greater net change than **The world**[1]. Not only is **The world**[1] significantly greater than the other two, but it also has always been adding words at a faster rate and this reinforces the theory above about the primacy of representing the physical world in human communication. It is fascinating, however, the similarity in the dynamics of all three categories. The three series exhibit the same shape almost perfectly, even though the individual magnitudes of the series differ. Thus, it appears our language evolution has been spread relatively evenly over the three categories, although when making this conclusion one must not forget the highly contrived nature of the Historical Thesaurus.

A similar story is found when examining the edge weights at level 1 (Fig. 7). The two connections involving **The world**[1] are the largest two connections. However, normalized edge weights (Fig. 8) tell a story of evenly distributed edge weights when the disproportionate size of **The world**[1] category is taken into account. We also note the decrease in all three edge weights after 1600 and go into more detail later when investigating the same plots for level 3.

Overall, the data appears to show proportional growth in category size and uniform normalized connection weights at the highest level of the Historical Thesaurus. But when the absolute strength of the disproportionate category size of **The world**[1] is allowed in, it dominates the connection weights by a significant margin.

B. How Category Size Has Changed

The top ten categories of level 2 when ranked by category size at 100 year intervals are displayed in Tab. I. Category size represents an estimate of the vocabulary available when discussing the given category. Given that only 16 categories occupy top ten ranks at all times, several fascinating patterns emerge when examining this table.

Leisure[2] enters the top ten in 1600 and gains rank steadily through 2000. At the year 2000, **Leisure**[2] is 125% greater than the next biggest category, a significant margin. Much of this vocabulary relates to different types and styles of **Music**[3], **Literature**[3], and **Visual arts**[3] while **Sports**[3], **Performance arts**[3], and **Pastimes**[3] are the next three largest categories under **Leisure**[2]. It appears we have dramatically increased the variety of leisurely activities we engage in.

However, there are only 26,000 words in the English vocabulary in the category **Leisure**[2], which is about 6% of all words currently in use today. While **Leisure**[2] is by far the biggest category at level 2 it is not dominating the language. Even so, we can state that in the last 400 years, leisure has become an increasingly important part of the English language and the societies that speak English.

Faith[2] begins as rank 4 in Old English and then slowly loses rank until 1600 when it is no longer in the top ten. **Life**[2] follows a similar pattern to **Faith**[2] and disappears from the top ten by 1600 as well. Both categories are gaining words at every interval, excepting the switch from Old English to Middle English, but other categories are gaining words at a faster rate. This is probably more

Rank	Old English	1200	1300	1400	1500
1	Action/operation (3,543)	Action/operation (986)	Action/operation (2,011)	Action/operation (4,288)	Action/operation (5,539)
2	Emotion (2,917)	Emotion (735)	Emotion (1,432)	Emotion (2,840)	Space (3,540)
3	Space (2,407)	Space (658)	Space (1,224)	Space (2,680)	Emotion (3,501)
4	Faith (1,875)	Mental capacity (545)	Movement (987)	Mental capacity (2,192)	Mental capacity (2,868)
5	Movement (1,859)	Faith (499)	Mental capacity (925)	Movement (2,025)	Food and drink (2,853)
6	Life (1,859)	Life (497)	Faith (844)	Curious, inquisitive (1,903)	Authority (2,564)
7	Mental capacity (1,836)	Movement (486)	Life (842)	Authority (1,794)	Curious, inquisitive (2,559)
8	Authority (1,770)	Food and drink (449)	Authority (822)	Faith (1,727)	Movement (2,433)
9	Food and drink (1,707)	Authority (432)	Curious, inquisitive (742)	Life (1,697)	Faith (2,317)
10	Curious, inquisitive (1,684)	The earth (381)	Food and drink (709)	Food and drink (1,681)	Life (2,093)

Rank	1600	1700	1800	1900	2000
1	Action/operation (9,819)	Action/operation (13,470)	Action/operation (14,525)	Leisure (18,904)	Leisure (26,131)
2	Space (6,348)	Mental capacity (9,923)	Mental capacity (10,937)	Action/operation (18,191)	Action/operation (20,903)
3	Emotion (6,192)	Space (9,322)	Space (10,505)	Animals (17,578)	Animals (19,596)
4	Mental capacity (6,065)	Emotion (8,538)	Emotion (9,539)	Mental capacity (14,978)	Mental capacity (18,153)
5	Curious, inquisitive (5,318)	Curious, inquisitive (7,628)	Leisure (9,314)	Food and drink (14,690)	Health and disease (18,011)
6	Food and drink (4,731)	Leisure (6,907)	Food and drink (8,660)	Space (13,477)	Food and drink (17,692)
7	Authority (4,520)	Food and drink (6,839)	Curious, inquisitive (8,412)	Plants (13,322)	Life (16,048)
8	Leisure (4,197)	Authority (6,508)	Animals (7,988)	Health and disease (13,276)	Occupation and work (15,665)
9	Movement (4,041)	Health and disease (6,327)	Plants (7,820)	Emotion (12,352)	Plants (14,476)
10	Plants (3,989)	Animals (5,629)	Health and disease (7,349)	Occupation and work (11,855)	The earth (14,449)

TABLE I. The top ten categories ranked by category size (number of words) at each century and in Old English. Numbers in parentheses reflect the category sizes.

due to the static nature of those topics than the decline of our interest in those topics. Only 8 words have been added to **Faith**[2] since 1980, a very low number compared to other level 2 categories. Furthermore, an examination of the words added to **Faith**[2] since 1900 shows that many of these are the names of newly encountered forms of faith recently exposed to English speakers (e.g *Tantricism* (1959), *Rastafarianism* (1968), *panentheist* (1974), etc.), meaning that fewer words describing new aspects of the faith concept have been added recently.

Life[2] has had a major resurgence in recent years and an examination of the change in word lists shows this is due to a large vocabulary from scientific research in biology. However, much of the early vocabulary describing our basic ideas surrounding **Life**[2] and contained in the category **Source/principle of life**[3] and **Death**[3] still exists.

Faith[2] and **Life**[2] have many old words and add new words slowly causing them to fall out of the top ten at level 2 before **Life**[2] has a comeback with biology. Of interest for future work would be devising quantitative measures of oldness and stability to compare different categories at both level 2 and level 3. This inspiration for quantitative future work is exactly the purpose of our framework for general analysis of the Historical Thesaurus categories and the connections between them.

Curious, inquisitive[2] has a downward opening parabola shape with the peak around 1600-1700, a period that corresponds with the scientific revolution and age of enlightenment. By 1900, **Curious, inquisitive**[2] had dropped off the top ten but **Occupation and work**[2] appears in the top ten as most of the words from the industrial revolution were added in the 19th century. The correlation with the societal movements of the time is truly fantastic.

The categories of **Plants**[2] and **Animals**[2] both increase dramatically to make the top ten from 1800-2000 and an examination of the words in these categories shows that this is due to the large naming and taxonomy efforts made for plants and animals since 1800. Finally, **Health and disease**[2] appears in 1700 and slowly works up to 5th place by 2000 reflecting our society's increased focus upon health and sanitation.

Many of the changes in ranking of the categories by size can be attributed to large-scale societal factors driving the topics that are important to English speakers. Seeing these changes in the vocabulary mirror society reflects the importance of language as a proxy for our understanding

about ourselves and the world we live in and navigate everyday.

C. Vocabulary Size in Different Categories Reflects Large-Scale Societal Changes

We have discussed how the top categories change over time and how these changes in language reflect large societal changes. But how exactly those changes are occurring can be further investigated. Fig. 10 shows the net change in category size for some of the level 2 categories. Fig. 10A-H are the top eight categories when ranked by most positive net change at any point. **Animals**[2] has the most positive net change in 1890 with a net increase of 1,713 words. Fig. 10I-P are the top eight categories when ranked by most negative net change at any point. **Action/operation**[2] has the most negative net change in 1410 with a net decrease of 83 words.

First, we can see that many of these categories correspond to the categories seen in Tab. I, which is understandable because the categories rank highly must add the most words. We also note that very few categories have negative net change. The largest decrease in category size at level 2 is 83 words while the largest gain is orders of magnitude higher. This emphasizes the point that the addition of words to the language far outstrips the loss of words. Furthermore, this difference may increase in the future as new forms of media improve our capacity to remember older words. We also note that in the top eight positive changes (Fig. 10A-H), only two high peaks occur before 1800 in **Action/operation**[2] and **Leisure**[2]. Here, we reinforce the conclusion that the most major English language changes have occurred in the last 200 years.

By examining the top eight negative change time series (Fig. 10A-H) we find series with much less parity. These categories are all older categories, with much smaller changes since 1800. Interestingly, these categories all have their maximal negative net change in 1410. An examination of the word lists for the year 1410 shows an large amount of Old English words being lost in that year. Around 1400 there appears to be a critical transition from Old English and Middle English to Modern English because of the dramatic changes taking place in the graphs and word lists. However, the largest negative net change (-83, for level 2) is quite small compared to the sizes of the positive changes. Due to the transition and the difficulty of estimating the date when a word ceased to be used, we do not draw any conclusions from these plots except to note that these older categories have significantly less change in the last 200 years, possibly indicating a stability not found in more recent categories.

By combining all these individual graphs, we arrive at Fig. 9A, which displays time series of the net change (added words minus lost words) for the top ten level 2 categories when ranked by maximum positive change. Note that the curves in Fig. 9 are drawn using a cardinal spline with a tension of 0.7 to smooth the transitions. In this plot, we can see a general trend of many additions to the language after 1800. However, there are also high peaks in 1600 and 1400 that correspond to changes in the slope of the line for number of words in the language (Fig. 3). This is a general trend that can be seen in all the individual plots in Fig. 10 and when they are put together it becomes more apparent. This trend also holds when examining the changes in level 3 categories in Fig. 9B. Although the peaks are less pronounced because the raw category sizes are smaller, there are clearly three peaks at 1400, 1600, and 1900, indicating that these three periods of time were critical in the history of the English language.

In order to examine the changes in level 3 categories more closely, we constructed individual plots for the top eight categories with positive and negative change in the same way as we did for level 2 categories. These are shown in Fig. 11A-P. On the same scale as level 2, the peaks are clearly less pronounced, as they should be, but many of the categories correspond to the categories shown in Fig. 10. Specifically, seven out of eight of the top categories from level 3 (Fig. 11A-H) are child categories of one of the top eight from level 2: **The Arts**[3] is the largest child category of **Leisure**[2], **Invertebrates**[3] is the largest child of **Animals**[3], **Ill-health**[f] or **Health and illness**, **Biology**[3] and **The body**[3] for **Life**[2], **Physics**[3] for **Matter**[2], and **Equipment**[3] for **Occupation and work**[2]. This tells us that there are a few key categories which have large amounts of vocabulary added to make the changes we see at level 2 and these few categories are shown in Fig. 11A-H.

Note that **Particular plants**[3] in Fig. 11D does not have a corresponding level 2 category in the top eight, likely due to it's more even spread of word addition over time. Its also useful to note that **Food and drink**[2] and **Action/operation**[2] do not have corresponding level 3 categories in the top eight. For **Action/operation**[2] this is probably due to the oldness of the category, which means a slower rate of change in word addition. But for **Food and drink**[2] it appears the lack of a top eight category in level 3 is due to the rates of word changes within the child categories. **Food**[3], **Drink**[3], **Farming**[3], and **Hunting**[3] are all adding words at medium rates leading **Food and drink**[2] to be on the top eight but none of the children to be in the top eight for level 3.

In summary, we repeat the observation that there have been three key moments of great change in the English language (1400, 1600, and 1900) corresponding to large-scale societal changes (early renaissance, age of enlightenment, and industrial revolution). Because this phenomenon is spread across all categories we know that the general trend is not due to a single innovation or expansion in a key area, but a combination of innovation and expansion in many categories. By examining which categories increased during each of these three time periods, the exact nature of each of these three language changes

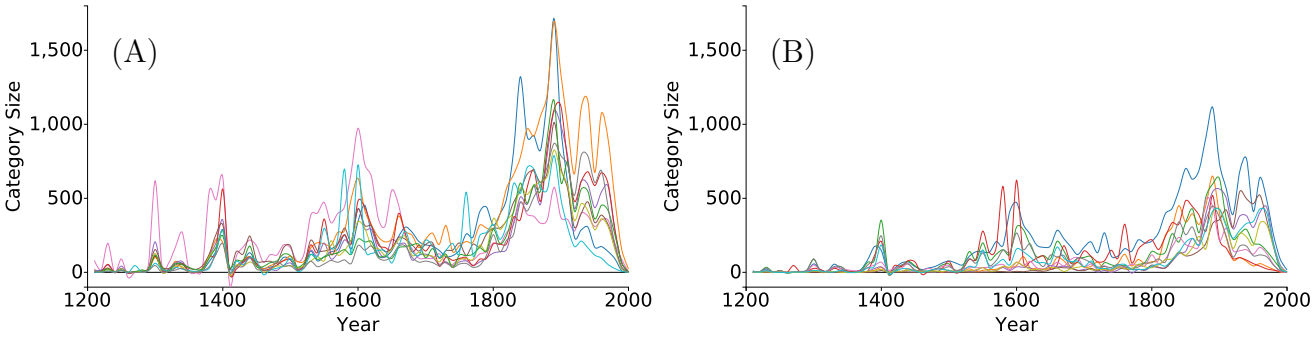

FIG. 9. The net change in category size (added words - lost words) binned in 10 year intervals from 1200-2000 for level 2 (A) and level 3 (B) categories. Curves are smoothed using a cardinal spline with tension 0.7.

is revealed, and can be shown to intuitively correspond to the societal changes mentioned.

D. Early Connections Are Fairly Static

In Fig. 12A and Fig. 12B we show a stacked area chart of the top 20% of connections between level 2 categories when ranked by connection size in the year 1300 and 2000 respectively. Fig. 12C and Fig. 12D show the top 20% of normalized connections between level 2 categories and are ranked in 1300 and 2000 respectively as well. The height of the area represents the strength of the connection (the size of intersection of each category's word lists). Fig. 12C and Fig. 12D entail the same connections as the plots above them but are normalized by the size of the union of the category's word lists. Fig. 12A and Fig. 12B also show the total number of lost words for connections in faded colors. Those connections with lost words have also been ranked by the year (1300 or 2000) and only the top 20% are shown.

Fig. 12C shows the connections with the greatest normalized weight in 1300 steeply losing normalized weight from the year 1600 onward. Fig. 12D displays a similar trend, but with a slower descent after 1600. Although it could not be shown in these two plots, the transition from ranking the series in the year 1300 to the year 2000 follows a smooth progression in diminishing the gradient of the produced plot after 1600. The descent after 1600 shown in these plots means that connections among categories are made more slowly than words are added after the year 1600. This effect is more pronounced in connections highly ranked in the year 1300 (Fig. 12C) than in the year 2000 (Fig. 12D) meaning that the categories that make up those connections in the former case add many words that are not shared among the categories. In other words, the words added to categories highly ranked in the earlier era (1300) are bound to their semantic category and less likely to have connections to other categories.

Furthermore, since the series in Fig. 12C start off above and climb slower than the series in Fig. 12D we conclude that those early strong connections do not add words as rapidly as the connections ranked highly in 2000. Thus, connections made in 1300 do not add words rapidly prior to or after 1600. The change in average slope of the series from positive to negative is an effect of rapid word expansion with slower increases in connection strength. These conclusions indicate high stability in early connections from 1300 and lower stability in connections highly ranked in 2000. Further quantitative investigation remains to be done to reinforce this conclusion.

Fig. 12A and Fig. 12B provide some support for the above conclusion. The slope of the edge weight series from 1600 to 2000 is much greater in Fig. 12B demonstrating faster growth, which corresponds to a small slope after 1600 in corresponding Fig. 12D. Connections displayed in Fig. 12A show significantly less growth confirming the conclusion that early connections are more stable.

Interestingly, the total loss in connections (shown in the negative series of Fig. 12A-B) appears to be similar for both plots. We do not have an explanation for this particular trend but merely note it as one of many fascinating phenomena that appear when investigating these connections between concepts as expressed in the structure of the English language through time.

E. What About The Words

In order to demonstrate exactly how the shared word lists of connections demonstrate related meaning between two categories, thus returning to questions about metaphor, we examine two specific connections in detail here.

The first connection is between **Mammals**[3] and **Badness/evil**[3] and the list of all words shared between the two categories is in Tab. II. Here we can see *pig* referring to the animal in **Mammals**[3] and *pig* being used as a word for an offensive thing, person, or place in **Badness/evil**[3]. *Honeycomb* appears in **Mammals**[3] as a word describing the sides of the stomach of a goat,

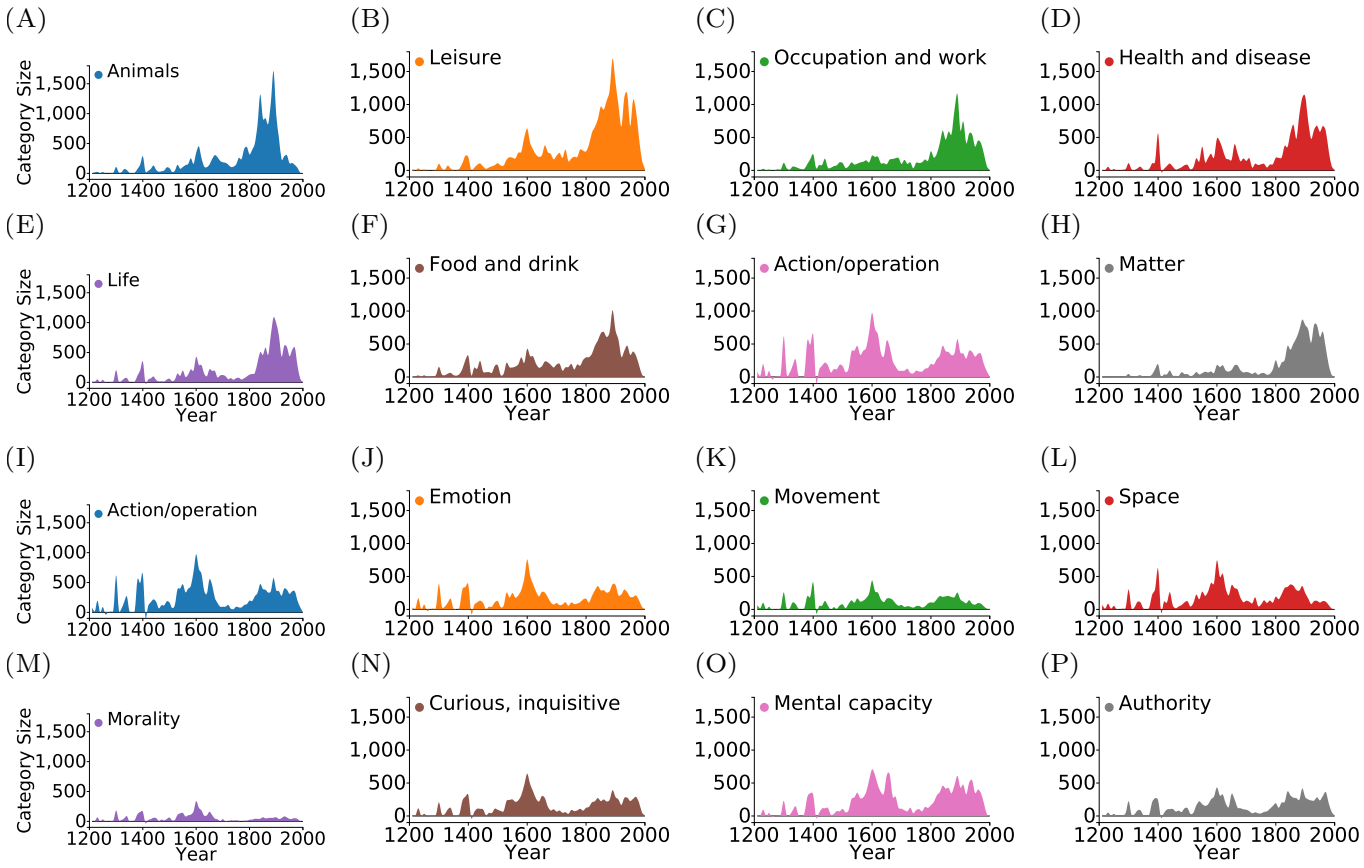

FIG. 10. The net change in category size (added words - lost words) binned in 10 year intervals from 1200-2000 for level 2 categories. A-H show the top ten categories when ranked by most positive net change for any decade. I-P show the top ten categories when ranked by most negative net change for any decade.

Mammals[3] to **Badness/evil**[3]	**Rate of motion**[3] to **Manner of action**[3]
pig	quick
sable	slow
foul	haste
fell	smartly
decidence	tarry
vicious	speedy
honeycomb	rash
wrack	sharp
slide	sluggish
skunkdom	rackly
set	...

TABLE II. Shared word lists for two connections at level 3. These words were taken from the shared word lists in the year 2000.

cow, or sheep while it appears in **Badness/evil**[3] because it has a definition as a verb meaning "to render hollow, rotten, or weak; to undermine" [6]. *Skunkdom* means to be of "skunkish character" or "skunks collectively" [6]. *Vicious* describes the behavior of mammals and is deemed to be a bad or evil behavior. *Foul* and *decidence* also have meanings in both categories that show similar patterns. Thus, overall we see a strong connection between words describing certain types of mammal behaviors or other forms of description to the words expressing **Badness/evil**[3].

However, some shared word forms (*sable*, *fell*, *wrack*, *slide*, *set*) do not demonstrate a meaningful relationship. Instead, these word forms appear in both categories for different reasons, usually just noise due to the orthographic version of English borrowing from many other languages, and are confounding our analysis. It is a difficult problem to automatically differentiate between the meaningful and non-meaningful shared word forms because humans make the decision based upon the semantic meaning of the words, something computers do not understand, yet. For this reason the Mapping Metaphor project manually inspected all of these shared word lists to find metaphors. The inability to automatically determine exactly which words demonstrate a linguistic process and not a chance overlap of orthographic word form must be taken into account. That said, our en masse analysis to find general patterns is broad enough to still find interesting patterns of metaphor or other meaningful linguistic relationships.

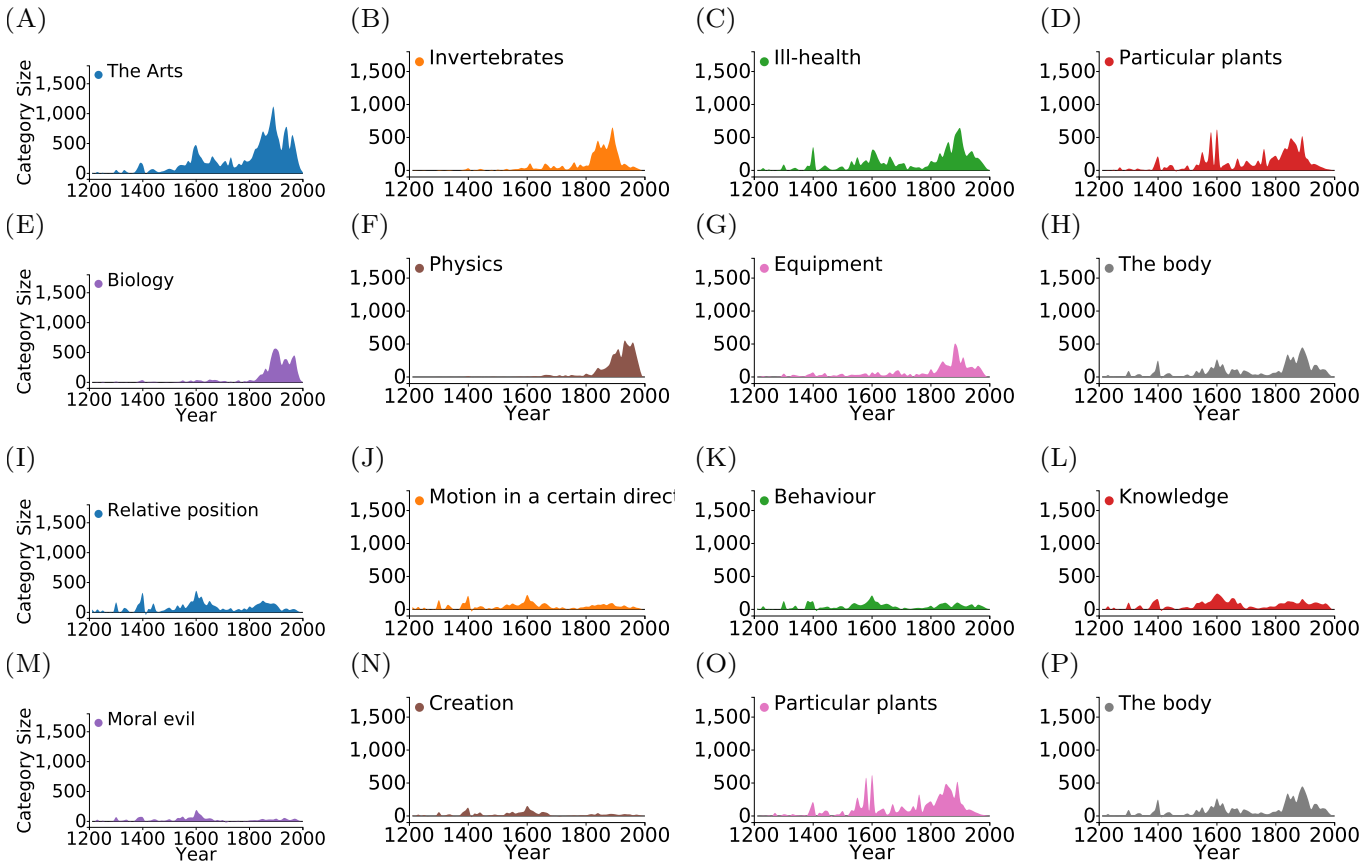

FIG. 11. The net change in category size (added words - lost words) binned in 10 year intervals from 1200-2000 for level 3 categories. A-H show the top ten categories when ranked by most positive net change for any decade. I-P show the top ten categories when ranked by most negative net change for any decade.

Another example of a productive connection is between **Rate of motion**[3] and **Manner of action**[3], which is ranked 12th by normalized edge weight out of all level 3 connections. A selection of the 96 words connecting these two categories in the year 2000 is shown in Tab. II. We can see from words such as: *quick*, *slow*, *smartly*, etc., that the two connections are highly interrelated. Due to the larger size of the connection, we must rely upon the normalized ranking to demonstrate exactly how strong the connection is.

This type of connection is another great example of the metaphor the Mapping Metaphor Project was looking for. Concrete language describing how object dynamics in the physical word is used to communicate about the abstract concept of performing an action. A qualitative examination of the word dates shows that many of these words were first introduced into the concrete category (**Rate of motion**[3]) and then later into the abstract category (**Manner of action**[3]). These are just two examples, but there are tens of thousands of these connections at level 3 that demonstrate how language evolves and changes in the case of English, and potentially in all human languages. In future work, we would like to devise quantitative measures to analyze these vocabulary shifts based upon the qualitative investigations we have performed on many of these word lists.

VI. CONCLUDING REMARKS

In this work, we presented a computational framework for analyzing the semantic categorical hierarchy of The Historical Thesaurus and how it has undergone change over the last 1000 years through the introduction and loss of English language words. We discovered many trends, particularly in category size, but also in our non-hierarchical connections, often corresponding to large-scale societal change. Especially important was the identification of three major periods of great language change and the identification of several key semantic domains (categories), which contribute to those changes.

While this work has revealed much about the nature of changes in English vocabulary over time, it has asked many more questions than it has answered. Several different avenues of future research are suggested and an interactive website has been created to pursue those ideas. Devising quantitative methods to verify the conclusions made in this paper about vocabulary size as

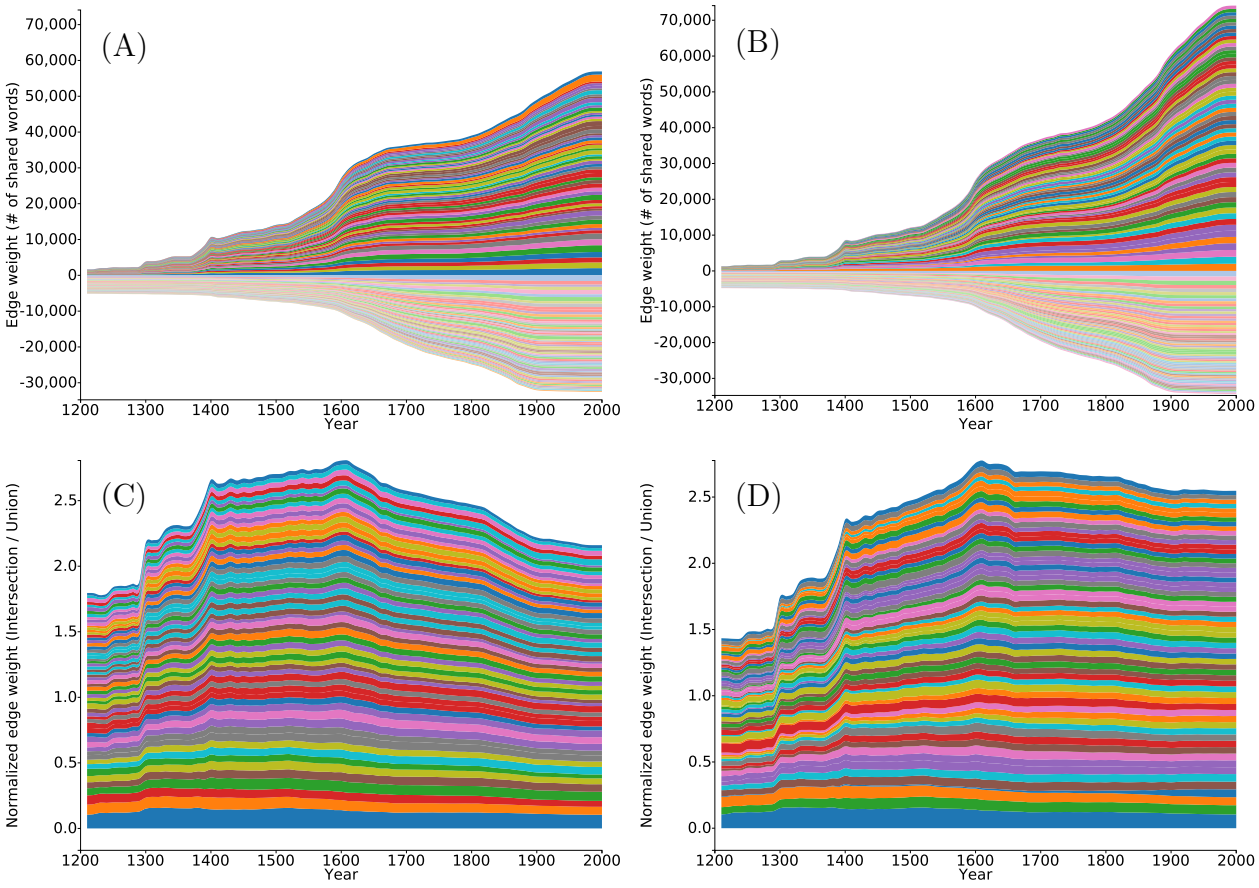

FIG. 12. (A) and (B) display connection weights between categories for level 2 ranked by the year 1300 and 2000 respectively. (C) and (D) display normalized connection weights between categories for level 2 ranked by the year 1300 and 2000 respectively. Connection weights in (C) and (D) are normalized by the union of the two words lists corresponding to the categories being connected. In all four graphs, only the top 20% of connections are shown.

a representation of the large-scale societal changes and the static nature of older categories are two key research areas.

Furthermore, the original inspiration behind this work was to attempt to automatically replicate the Mapping Metaphor Project's results. While that task is difficult given the nature of metaphor, we would like to attack the classification problem using the litany of options available in the machine learning literature.

Finally, we are interested in more closely examining how words shift from category to category by comparing the dates given in the Historical Thesaurus and the Oxford English Dictionary. Not only would this provide us with valuable knowledge about how language changes on a broad scale, but it could potentially provide some insight into exactly how and when words from one category are "co-opted" by another. This could provide us with knowledge about how metaphors have been created in the past, which may in turn allow us to understand why and when metaphors are used. Given that metaphor is thought to be very closely linked to cognition and comprehension, understanding these phenomena has the potential to greatly advance research in these two important areas.

ACKNOWLEDGMENTS

We are grateful for ...

[1] G. Lakoff and M. Johnson, *Metaphors we live by* (University of Chicago press, 2008).
[2] M. Alexander and E. Bramwell, Studies in the Digital Humanities (2014).
[3] C. Kay, J. Roberts, M. Samuels, and I. Wotherspoon, (2009).
[4] Http://fletcherhaz.github.io/sorel/.

[5] J. Roberts, C. Kay, and L. Grundy, *A Thesaurus of Old English: Index*, Vol. 2 (Rodopi, 2000).
[6] O. E. Dictionary, "Oxford: Oxford university press," (1989).