Using machine learning for semi-automatic expansion of the *Historical Thesaurus of the Oxford English Dictionary*

James McCracken

Oxford University Press
E-mail: james.mccracken@oup.com

Abstract

The *Historical Thesaurus of the Oxford English Dictionary* (HTOED) provides a highly granular taxonomic classification of the contents of the OED. However, HTOED was based largely on the first edition of the OED (plus supplements), and has not been updated to include content added more recently, or changed content emerging from third-edition revisions. This means that 32% of lexical items in the current OED data set are unclassified.

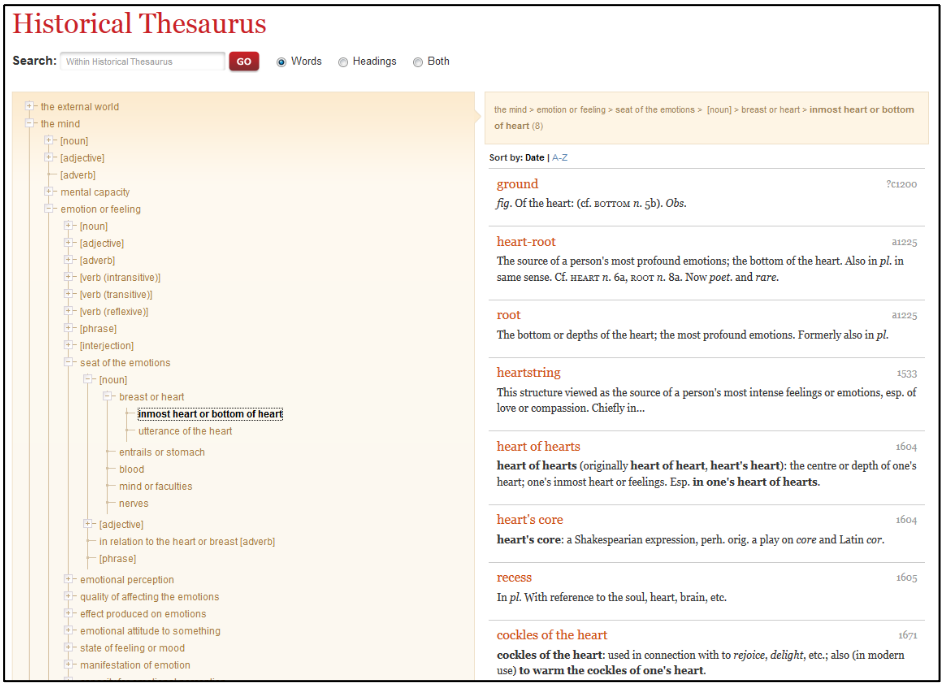
We use the existing HTOED classifications as training data to classify this 'missing' content. The classification system works as a two-stage process. Firstly, for a given input sense, a Bayesian classifier identifies the general topic (high-level thesaurus branch) to which the sense belongs; secondly, a battery of similarity measures identifies possible target nodes within this branch. The system looks for consensus or proximity among the outputs of these methods, in order to pinpoint the optimal node(s) to which the sense should be assigned.

The system is currently able to classify 25% of input senses to the correct node, and a further 40% of input senses to the right neighbourhood (a parent, child, or sibling of the correct node). A web-based UI facilitates the manual checking, approval, and adjustment of proposed classifications.

Keywords: Oxford English Dictionary; Historical Thesaurus; machine learning; lexical ontology; feature extraction

1. Introduction

The *Historical Thesaurus of the Oxford English Dictionary* (HTOED) is a taxonomic classification of the content of the *Oxford English Dictionary* (OED), compiled at the English Language department of the University of Glasgow between 1965 and 2008. The HTOED data were integrated with the OED data in 2010, and now form a core part of OED Online (www.oed.com/thesaurus). The HTOED is also available as a standalone resource at http://historicalthesaurus.arts.gla.ac.uk/, and is published as a two-volume book (Kay et al., 2009).



the mind > emotion or feeling > seat of the emotions > [noun] > breast or heart > **inmost heart or bottom of heart** (8)

Sort by: **Date** | A-Z

ground ?c1200

fig. Of the heart: (cf. BOTTOM *n*. 5b). *Obs*.

heart-root a1225

The source of a person's most profound emotions; the bottom of the heart. Also in *pl*. in same sense. Cf. HEART *n*. 6a, ROOT *n*. 8a. Now *poet*. and *rare*.

root a1225

The bottom or depths of the heart; the most profound emotions. Formerly also in *pl*.

heartstring 1533

This structure viewed as the source of a person's most intense feelings or emotions, esp. of love or compassion. Chiefly in...

heart of hearts 1604

heart of hearts (originally **heart of heart**, **heart's heart**): the centre or depth of one's heart; one's inmost heart or feelings. Esp. **in one's heart of hearts**.

heart's core 1604

heart's core: a Shakespearian expression, perh. orig. a play on *core* and Latin *cor*.

recess 1605

In *pl*. With reference to the soul, heart, brain, etc.

cockles of the heart 1671

cockles of the heart: used in connection with to *rejoice*, *delight*, etc.; also (in modern use) **to warm the cockles of one's heart**.

Figure 1: HTOED integrated with OED Online. The taxonomy is shown on the left; the senses in a selected node are shown on the right.

HTOED was based largely on the first edition of the OED (1888–1928) and its supplementary volumes (1933; 1972–1986), and latterly extended to include new material from the OED *Additions* volumes (1993–97).[1] This is now incomplete relative to the current state of the OED, in two main respects:

1. Certain categories of OED material, such as undefined compound lemmas, were systematically omitted from HTOED;

2. HTOED has not been updated to cover new material added to OED since 1997, or new and changed sense distinctions emerging from the Third-edition revision programme which began in 1993.

Consequently, a third of all senses in the current OED data set (264,000 out of 821,000) are not covered in HTOED.[2]

A project within the OED programme is currently attempting to ‘complete’ the HTOED by assigning an HTOED classification to as many of these 264,000 ‘missing’

[1] It also contains material from several Old English dictionaries, also published separately as *A Thesaurus of Old English* (Roberts & Kay, 1995). Much of this material falls outside the scope of the OED.

[2] Throughout this paper, I use ‘sense’ to mean any semantically distinct unit of an OED entry, including both senses of main headwords and sublemmas.

senses as possible. This is being done semi-automatically: a supervised machine-learning process uses the existing classifications as training data to classify the input senses (in some cases generating two or three 'candidate' classifications); these classifications are then accepted, rejected, or adjusted by human reviewers.

2. Viability of machine learning

On the face of it, this is a very attractive machine-learning task: 557,000 senses manually classified by a team of well-trained researchers within an academic department should make for a very rich and reliable set of training data.

But there are some complicating factors:

1. The HTOED taxonomy is highly granular: for any given input sense, there are over 200,000 candidate labels (i.e. taxonomy nodes), so the amount of training data decreases sharply as you go down a taxonomic branch. The number of training-data senses usually drops to single figures by the fifth or sixth level down.

2. The input senses are not altogether similar to the training-data senses. That is to say, the input senses are not a random subset of the population as a whole. For example, input senses tend (on average) to be more recent, more technical, or more minor than the training-data senses.

3. Although the training data as a whole is very rich, each individual document (dictionary sense) tends to be short and feature-poor. So for a given input sense, it may be difficult to extract a set of features good enough to support comparison with the training-data models.

4. The OED's unrestricted defining vocabulary means that individual feature values (e.g. a specific word or phrase in a definition) may be very sparse.[3]

The HTOED taxonomy also presents some problems:

1. The HTOED taxonomy was developed 'bottom-up', largely determined by the material that happened to be in the first and second editions of OED (Kay et al., 2009: p.xix). At the very fine-grained level (leaf or near-leaf nodes), the HTOED taxonomy expresses a variety of relations, e.g. meronymy and other associative relations, as well as hypernymy. This fine-grained structure tends to be determined by the specific definitions of the original member senses of each node; hence at this level it becomes harder and harder to determine that an input sense belongs to a given node, and probabilistic approaches break down.

[3] This problem is most acute when dealing with superordinates; see section 9.

2. Moreover, a new input sense may represent a concept not currently accounted for by HTOED; so there may be no correct classification in terms of the existing taxonomy.

2.1 Two-stage system

For these reasons, we found that no single machine-learning model was adequate for the task. Instead, we developed a two-stage system:

1. For a given input sense, a naïve Bayes classifier (the Topic_classifier module) is used first to identify the probable topic(s), i.e. a relatively high-level branch of the taxonomy;
2. A range of more targeted methods (often with their own Bayesian models) are then applied to determine a specific node within that branch.

These results are collated by a top-level module (the Central_classifier) to determine the final classification assigned to the input sense.

Although particular methods may require some parsing and analysis (e.g. to identify superordinates within a definition; see section 9), in general terms this approach is statistical rather than rule-based. That is to say, an input sense is classified by comparing its features to the training data, rather than by any direct attempt to decode its definition. This allows for models that are adaptable to the very variable nature of OED senses.

2.2 Summary of classification methods

The Central_classifier first uses the Topic_classifier to restrict the search-space within the taxonomy to a particular branch or branches. The following set of methods is then applied to try to find a more specific node or region within that branch:

- **Cross-reference**: If an input sense cross-refers to another sense that has already been classified, this may indicate how the input sense should be classified. See section 4.
- **Taxonomic binomial/genus term**: An animal- or plant-name definition often includes a binomial name, or at least a genus name; this can be used to find an exact classification. See section 5.
- **Synonyms**: If an input sense definition includes one or more synonyms, the classification of the synonym words may indicate how the input sense should be classified. See section 6.

- **Morphology**: A derived form can usually be assumed to be semantically close to its root, or to sibling terms derived from the same root. See section 7.
- **Compound form**: The elements of a compound lemma may be indicative of sense. See section 8.
- **Superordinate**: If the superordinate term can be extracted from the definition of an input sense, this can be compared with other senses with the same or similar superordinate term. See section 9.

For each input sense, all methods are attempted, effectively in parallel.[4] If a number of different classifications are returned, the following procedure is applied:

1. The Central_classifier polls the results to look for cases where two or more methods have returned the same (or very similar) classifications;
2. If multiple classifications still remain, the classification chosen by the most reliable method is preferred; the remainder are treated as runners-up;
3. If no classifications remain (or if no classifications were returned in the first place), the Central_classifier defaults to the Topic_classifier's best-guess branch.[5]

As step 2 indicates, the system is often dependent on *a priori* rankings of different methods (e.g. for a typical sense, classification by cross-reference is ranked as more reliable than classification by synonyms). This system, therefore, does not always take individual circumstances into account (e.g. there may be occasional senses where synonyms are a better bet than a cross-reference).

2.3 'Runner-up' classifications

If the Central_classifier retrieves multiple candidate classifications, one of these will be selected as the 'winner' and treated as the primary classification. If any others remain, the top one or two are selected as 'runners-up'. Runners-up usually indicate different lines of attack that were considered by the Central_classifier.

In some cases, a supposed runner-up may turn out to be a better classification than the winning classification. The editorial interface provides a means to promote a runner-up ahead of the primary classifications; see section 11.

[4] Not all methods succeed in all cases, of course; for example, the cross-reference method will fail if the input sense has no cross-references. In such cases, a null result is returned, and is discarded.

[5] This will almost always be too high up in the taxonomy to be correct as it stands, but often provides a good starting point for human checking to identify the correct node further down.

3. Topic_classifier module

The Topic_classifier module is responsible for generating a ranked list of the three or four most likely topics (high-level branches of the HTOED taxonomy) for each input sense. This is used to restrict the search-space available to the more targeted classification methods employed by the Central_classifier. It may also be used to assist some of those methods more directly, e.g. to help pick likely senses of a synonym.

3.1 Flattened categories

The set of labels (i.e. the categories to which the Topic_classifier can assign a sense) is the set of thesaurus branches which contain 2000+ senses. This adds up to about 200 branches in total, some of which are sub-branches of others. The Topic_classifier treats these 200 branches as a flat list of disjoint labels.

This ‘flattened’ method may seem counter-intuitive. I spent some time experimenting with ‘taxonomy-aware’ classifiers, e.g. using decision trees (classifying first at level 1, then at level 2, level 3, etc.), but these approaches proved less successful. In practice, so long as each branch is reasonably well-populated, the Topic_classifier does not really need to know about the taxonomy. For a given sense, probabilities are calculated for each label in turn, and the label with the highest score wins. This may turn out to be a branch at any of the upper levels of the taxonomy.

3.2 Feature set

The feature set used includes the following:

- lemma (or lemma elements, in the case of MWEs);
- subject labels;
- register and usage labels;
- tokens from definition text;
- tokens from modern quotation text;
- tokens from quotation titles;
- author names;
- presence/absence of taxonomic binomials;
- first date (binned by 50-year periods);

- part of speech.

Tokens are all case-stripped, Porter-stemmed, and truncated to a maximum of eight characters. For example, *historically* and *historicism* are both normalized to *historic.*

3.3 Confidence score

A confidence score between 0 and 10 is associated with the ranked list that the Topic_classifier computes for each input sense. (A zero score indicates that the Topic_classifier has failed altogether, usually because the input sense provides insufficient features.)

The confidence score is a measure of the number of features provided by the input sense, and the margin by which the top two or three labels outscored the rest. If the confidence score is low, the Topic_classifier may be partly or wholly disregarded by other classification methods (i.e. the search-space is not restricted), and the Topic_classifier will not be used as a fallback if the other classification methods fail.

3.4 Sanity check

The Topic_classifier module acts as a kind of sanity check on some of the more deterministic methods described below. It tends to preclude or at least deprecate some of the more egregious errors that can arise from mistakes in a particular classification method: misinterpreting a word in a definition, misidentifying a superordinate, failing to correctly separate metalanguage from gloss, etc.

At the same time, the use of confidence measures prevents the Topic_classifier from being too aggressive in pruning away candidates.

4. Cross-references

Cross-references are a valuable way to contextualize a given input sense. Uniquely among the classification techniques discussed here, cross-references can be used deterministically rather than probabilistically, meaning that classifications made in this way tend to be both more accurate and more reliable.

4.1 'Equals'-type cross-references

An equals-type cross-reference provides the easiest win for the Central_classifier: if the target sense is classified, the input sense can simply adopt the classification of the target sense.

For example, *emulsin* is defined as:

> A neutral substance contained in almonds; = SYNAPTASE n.

Here the Central_classifier can safely ignore the definition and any other features of the input sense, and just copy the existing classification of the target sense of *synaptase.*

There are various formulae that can be treated in this way: not only a leading equals sign as in the *emulsin* example, but also formulae like 'another name for…', 'short for…', 'variant of…', etc.

About 15,000 senses are classified this way (7% of classified senses).

4.2 'Cf.'-type cross-references

'Cf.'-type cross-references do not provide such a direct and positive means of classification; but they nevertheless provide a good indicator of the right branch, at a fairly granular level.

For example, *generically* 2 is defined as:

> *Biol.* In a generic manner; with reference to genus. Cf. GENUS n. 2a.

So we can be fairly confident that *generically* 2 belongs in the adverb branch parallel to the noun branch in which *genus* n. 2a is found.

About 8,000 senses are classified this way (4% of classified senses).

4.3 Other cross-references

Other cross-references are useful not as classification methods in their own right, but as ways to improve the performance of other methods.

In particular, parenthetical cross-references within a definition often serve to disambiguate keywords, especially to make clear that a word is not being used in its primary modern sense.

4.4 Problems with cross-references

Cross-references can be susceptible to the kind of problem described in section 6.3 in relation to synonyms; namely that focussing on a cross-reference to the exclusion of the rest of the definition risks ending up with a classification that only captures one aspect of the sense, not its primary meaning.

For example, *general servant* is defined as:

> A servant whose duties are general rather than limited to a particular sphere; *spec.* = *maid-of-all-work*.

Because the Central_classifier focusses on the cross-reference, it ends up with the specific classification of ‘housemaid’ rather than the more general classification suggested by the main gloss.

5. Taxonomic binomial and genus names

Definitions for animal and plant names often include an explicitly tagged taxonomic binomial or genus names. By indexing all such names in the training data, we build a model mapping binomials to HTOED classifications. This can then be used to classify any input sense containing a taxonomic term in its definition.

For example, *Java lemon* is defined as:

> A small lime tree, *Citrus aurantifolia* (formerly *C. javanica*), originating in South-East Asia...

This sense can therefore be classified by checking the classification of training senses which also include *Citrus aurantifolia*. Failing that, the right branch can be found by checking the classification of training senses which include some other *Citrus ...* binomial.

About 4,500 senses are classified this way (2.3% of classified senses).

6. Synonyms

Although OED senses do not identify synonyms explicitly, OED definitions are very rich in synonym-like terms. These provide a useful aid to classification. If an input sense includes a synonym that can be reliably identified and disambiguated, then the classification of that synonym will be a good indicator of how the input sense should be classified.

It’s unusual for an OED definition to depend wholly on synonyms, but it’s quite common for definitions to include synonyms in some form as an adjunct or support to the main definitional gloss. This can be particularly valuable when dealing with adjectives; somewhat less valuable when dealing with verbs and adverbs; and least useful when dealing with nouns.

About 12,000 senses are classified using synonyms (6% of all classified senses).

6.1 Patterns

The prototypical pattern for a synonym-rich definition is something like this:

> Main gloss here; foo, bar, or baz.

where *foo*, *bar* and *baz* are the synonyms.

For example, *abhorred* is defined as:

> Regarded with disgust or hatred; detested, loathed, abominated.

where *detested*, *loathed*, and *abominated* serve as synonyms.

Beyond this prototypical pattern, there are nine or 10 other patterns which can also be used to identify synonyms within a definition. Slightly different patterns apply to different wordclasses.

6.2 Disambiguating synonyms

Having identified a synonym or synonyms for a given definition, the system then looks up the synonym's own OED entry, finds the appropriate sense, and examines how that sense has been classified.

'Finding the appropriate sense' is the difficult bit. It is tempting to assume that synonyms will usually be used in their main modern sense; but in practice this turns out not to be the case. Since a definition usually consists of a gloss followed by one or more synonyms (as with *abhorred* above), the gloss serves to prime a particular sense of the synonym word – which may or may not be the main sense.

For example, *generous* 4b is defined as:

> Of an action, a gift, etc.: readily done or given; more than is strictly necessary or expected; large, ample, bounteous.

where *large*, *ample*, *bounteous* can be identified as synonyms. *Large* here does not have its usual modern sense, but rather has the (now somewhat unusual) sense of 'liberal', primed by the preceding gloss.

Similarly, in a list of two or more synonyms, the meaning of each synonym may be primed by the others in the list. For example, *gleg* 1b is defined as:

> Of the eye: quick, sharp.

where *quick* and *sharp* are primed by each other so that we understand them in their 'shrewd' sense rather than in their more prototypical 'speedy' or 'keen-edged' senses.

Hence there are two main ways to disambiguate a synonym:

1. Use the Topic_classifier to determine the broad subject area of the sense, then

look for a sense of the synonym that falls within this subject area;

2. If the synonym is one of a list of synonyms, look for senses of the synonyms which cluster on a particular branch of the HTOED taxonomy (e.g. the ‘shrewd’ senses of *quick* and *sharp* are clustered on the ‘sharpness, shrewdness, insight’ branch of the HTOED taxonomy).

In practice there can be problems with both methods:

- Method #1 can fail because apparent synonyms are not always direct semantic equivalents, for the reasons discussed in section 6.3 below;
- Method #2 can fail because a list of synonyms may not be synonyms of each other, and so may not lie on the same taxonomic branch: the purpose of a list of synonyms is often to stake out the wider semantic territory, rather than to indicate a specific single meaning.

Because disambiguation can be problematic, it is often easier to focus on unambiguous synonyms words. For example, *graith* 2c is defined as:

> Of a stroke: clean, unimpeded.

where *clean* and *unimpeded* can be identified as synonyms. But because *clean* is polysemous, it is easier to focus instead on the less ambiguous *unimpeded.* However, this can exacerbate the problems discussed in section 6.3 below: the more unambiguous synonyms are often the more partial.

6.3 Are these really synonyms?

The patterns mentioned in section 6.1 above identify words that occupy a synonym-like slot in the definition; but this does not guarantee that they are actually synonyms in the strict sense. In fact, in the prototypical ‘gloss + synonyms’ pattern, the supposed synonyms may really be extensions, generalizations, or weakenings of the main gloss, rather than restatements of it.

A consequence of this is that a classification based on a synonym may capture certain aspects of the sense, but miss the core meaning.

For example, *musing* 2 is defined as:

> Given to or characterized by meditation; contemplative, thoughtful, dreamy.

where *contemplative*, *thoughtful*, *dreamy* are identified as synonyms. But *dreamy* here is rather different from the main gloss *given to or characterized by meditation.* If the Central_classifier focusses on *dreamy*, the sense will end up with a classification that reflects a minor extension of the sense rather than its core meaning.

So although synonyms are in principle a very direct and widely-available aid to classification, in practice the issues of disambiguation mean that these are not always usable. Moreover, some apparent synonyms may really be distractions from the core sense. It is often better to treat synonyms as a supplement to other methods, rather than as a classification method in their own right.

7. Morphology

A derivative form can usually be assumed to be on the same branch of the HTOED taxonomy as its parent or root word. If the derivative is in a different wordclass from its root (e.g. an *–ize* verb derived from an adjective), it can be assumed to be in a branch of the HTOED taxonomy parallel to that of its root.

If the root word has more than one sense, a run-on derivative lemma can usually be assumed to be related to the main sense of the root. However, this becomes more problematic if the root word has many possible senses; classification by morphology is not usually attempted in such cases.

This approach can be adapted for 'sibling' derivatives, i.e. two derivative subentries derived from the same root word. For example, the likely classification of *causationism* can be inferred from the existing classification of its sibling *causationist.*

About 16,000 senses are classified this way (8% of classified senses).

8. Compounds

About 34% of all input senses are compound subentries. There are also many main-entry senses which have a compound form. A special module (the Compound_classifier) is dedicated to determining candidate HTOED classifications based on the compound form itself.[6]

8.1 Initial assumptions

Our initial approach to handling compounds was to assume by default that the meaning of a compound lemma (and therefore its HTOED classification) is encoded in the lemma form, i.e. that the compound is endocentric. This assumption was strongest in the case of undefined subentries (21% of all input senses).

Most compounds (especially nominals) were taken to be head-final, i.e. the last element is a hypernym, and the first element is a qualifier.

[6] This draws on an extensive body of research into compounding and semantics in English; see Bauer (2009), Booij, (2007) and Lieber (2004).

The appropriate HTOED branch (if not the specific HTOED class) was therefore assumed to be related to one of the senses of the last element (and usually one of its main senses). Thus *furniture-van* is a hyponym of one of the main senses of *van*; *wheat-maggot* is a hyponym of one of the main senses of *maggot*.

But early testing found that these assumptions produced poor results. In particular, the assumption that the meaning of a compound can be deduced from the main sense of its last element turned out to be flawed in many cases. For example:

- *ship-jumper* is not a hyponym of any listed sense of *jumper*;
- *character assassin* is not a hyponym of any listed sense of *assassin*.

8.2 Probabilistic model of compounding

This led to a different strategy: rather than assuming compounds to be endocentric and head-final, we built a Bayesian model of compound semantics within OED, using both the first and last elements of the lemma. For each training-data sense with a compound lemma, the first and last elements are indexed against the HTOED classification of the sense. For a given input sense with a compound lemma, the most likely branch(es) of the HTOED taxonomy can then be predicted from these models.

Having identified a branch, the Compound_classifier can then revert to the more naïve assumption: the specific class within this branch is identified by focusing on the last element; either by looking for other compounds within the selected branch that have the same last element, or by looking for a sense of the last element that falls within the branch.

For example, the undefined compound *matrimonial broker* is classified as follows:

1. The Compound_classifier evaluates the two elements *matrimonial* and *broker* against the Bayesian model. This finds that initial *matrimonial* is strongly correlated with the *community » kinship or relationship* branch, whereas final *broker* is most strongly correlated with *occupation » trade and commerce*, and more weakly correlated with *community » kinship or relationship*.

2. The net result is that *community » kinship or relationship* is selected as the most likely branch.

3. The Compound_classifier then tries to find the specific class within the *community » kinship or relationship* branch. It tries two approaches in parallel: (a) it checks for senses of *broker* that fall within this branch; (b) it checks for other compounds with *broker* as the last element which falls within this branch. Approach (a) fails in this instance, but approach (b) finds a cluster of -*broker* compounds in the *community » kinship or relationship » marriage or wedlock »*

match-making » match-maker class (*match-broker, flesh-broker, wife-broker*, etc.). This is therefore selected as the class to which *matrimonial broker* will be assigned.

The process works very neatly with the example of *matrimonial broker*, but many examples are not so clear-cut. Often the Compound_classifier will draw on the Topic_classifier to help arbitrate between competing possibilities.

8.3 Successes

The following are examples of compounds which were incorrectly classified by methods based on the earlier endocentric, head-final assumptions, but which are correctly classified by the more probabilistic approach of the Compound_classifier:

- *truth-speaking*: classified as *mental capacity » faculty of knowing » conformity with what is known, truth » sincerity, freedom from deceit » sincere*
- *mimosa scrub*: classified as *the earth » land » landscape » fertile land or place » land with vegetation » wooded land*
- *vision-monger*: classified as *mental capacity » expectation, looking forward » foresight, foreknowledge » prediction, foretelling*
- *quiet-footed*: classified as *sensation » hearing » inaudibility » inaudible » silent » of footsteps*
- *vine-clad*: classified as *the earth » land » landscape » fertile land or place » land with vegetation » covered with vegetation » wooded*

8.4 Casualties

Not all compound-handling is improved by the Compound_classifier; some compounds were better served by the earlier approach.

For example, *junction piece* (which seems to be something to do with plumbing) gets classified as *travel » travel by railway » railway system or organization*, due to the fact that *junction* is strongly correlated with railways.

Still, when the Compound_classifier gets things wrong, it at least tends to do so with a certain wit, as when it misclassifies *butt mark* (an archery term) as *...animal husbandry » animal keeping practices general » branding or marking.*

8.5 Compounds with unusual elements

There are cases where the Compound_classifier draws a blank for a given input sense, because either the first or last element of the lemma is unusual and so does not appear in the predictive model.

In such cases (for undefined compounds, at least), the Central_classifier will disregard the Compound_classifier and fall back to a more naïve approach, usually reverting to the assumption that the lemma is a hyponym of the main sense of its last element. Failing that, it may just leave the sense unclassified.

8.6 Figurative, poetic, and metaphoric compounds

Many of the OED's undefined compounds are figurative or metaphoric to some extent. The intended meaning is often vague or unclear (often there is only a single quotation).

For example, *strife-race*, which has the single quotation:

> The strife-race, for we must run, and fight as we run, strive also to outstrip our fellow-racers,

gets classified as *leisure » sport and outdoor games » types of sport or game » racing or race » racing on foot* – which is not bad, except that it completely misses the fact that *-race* here is used metaphorically.

Some of the more poetic compounds involve deliberate repurposing of the first or second elements. For example, *panther-peopled* ('Amid the panther-peopled forests...') means not 'peopled' at all, but rather 'occupied by panthers'. The Compound_classifier does not really get to grips with such compounds at all.

It is debatable whether it is even worth attempting to include such compounds in the classification exercise. But that is a moot point, given that currently there is no sure way to distinguish between literal and figurative compounds.[7]

9. Superordinates

For noun senses in particular, identifying the superordinate within the definition is often a critical part of the classification process.

The classification process is based primarily on the training data: having identified the superordinate of a given input sense, the Central_classifier checks for training senses

[7] On the relationship between HTOED and metaphor, see Alexander & Bramwell, 2012.

that have an identical or similar superordinate, and examines how these are classified. Contextual information, notably the Topic_classifier's evaluation, may be used to arbitrate in the case of several competing possibilities.

About 33,000 senses are classified using superordinates (17% of all classified senses).

9.1 Process

This process can be broken down into a series of subtasks:

1. Separate the definition proper (the core gloss) from any metalanguage or secondary clauses;

2. Tokenize and p.o.s.-tag the gloss;

3. Chunk into noun phrases; the first noun phrase is presumed to be the superordinate in raw form;

4. Normalize the superordinate to allow fuzzy matching;

5. Retrieve training senses with (fuzzily) matching superordinates;

6. Cluster matching training senses into candidate HTOED branches;

7. Select the best HTOED branch, if there is more than one candidate (using the Topic_classifier or other secondary indicators).

9.2 Difficulties

There are potential difficulties with each of these steps, but the critical problems lie in steps 1 and 4. The general problem with step 1 (extracting the core gloss from metalanguage) is discussed in section 12.3. With respect to superordinates, this issue means that a metalanguage phrase may be erroneously identified as the superordinate.

Step 4 (normalization of the superordinate noun phrase) is required because, taken literally, many superordinates are unique or near-unique noun phrases. For example, *lagre* is defined as:

> In sheet-glass making: A sheet of perfectly smooth glass, placed between the flattening stone and the cylinder to be flattened.

The noun phrase containing the superordinate here is identified as *a sheet of perfectly smooth glass*. Since no other sense is defined in exactly the same way, this would draw a blank with the training data. However, if this is normalized to *glass sheet* (rearranging the syntax, and omitting possibly extraneous words), this now has more chance of matching training-data senses (given that the training data is also normalized in the same way).

Normalization of this kind is difficult: it is difficult to figure out what can be omitted, and sometimes difficult to reorganize into an optimal form. It is also tricky to figure out how far to normalize. For example, in some cases it may be beneficial to normalize synonyms towards their prototypes (so that e.g. *tracts of arable land* and *tilled field* would both be normalized to *field*); but in other cases this would over-generalize.

9.3 Uninformative superordinates

The most common superordinate is *person* (and its variant *one*, as in 'One who…'), closely followed by *man*. These provide no real help with classification, since *person/man* senses are distributed pretty evenly across the HTOED taxonomy.

A *person/man* superordinate can be made more specific by extending the 'scope' of the superordinate to include the following clause (normalized as outlined above). Some of this has already been attempted, but more work is needed.

9.4 Ontological bias

The weight given to the superordinate within a definition tends to give the classifier an ontological rather than functional bias. That is to say, it tends to classify according to what a thing actually *is*, rather than what a thing does or is used for.

For example, *alum curd* is defined as:

> Milk or egg white curdled with alum, used chiefly as a poultice.

This ends up being classified as *the external world » the living world » food and drink » food » dairy produce » milk » curds*. From an ontological point of view, this is perfect (that is exactly what alum curd is). But it overlooks the medical function, which is arguably the more salient aspect here. The HTOED taxonomy tends to be organized from a functional and human-oriented point of view, rather than from a strictly ontological point of view.

9.5 Adjectives

Strictly, the superordinate-based method described above only really applies to noun senses. However, the principle can be extended to certain kinds of adjective sense. In particular, adjectives defined in terms of a noun phrase (introduced with phrases like 'of or relating to', 'designating', etc.) are susceptible to superordinate-like classification.

For example, *all-in* adj. 2 is defined as:

> Designating a form of wrestling with few or no restrictions on the tactics that may be employed; of, relating to, or involved in this kind of wrestling.

Here we can say that *a form of wrestling* is a kind of superordinate, not of the adjective sense itself, but of its nominal equivalent. So we can 'pretend' that this is a noun sense with the superordinate *a form of wrestling*, classify it accordingly, and then convert that classification to an equivalent adjective branch.

About 2,500 adjective senses are classified in this way (1.3% of all classified senses).

10. Results and evaluation

Of the 821,000 senses in the OED data set:

- 557,000 (68%) are training senses, i.e. senses that already have at least one HTOED classification;
- 264,000 (32%) are input senses, i.e. senses for which a new HTOED classification is to be computed.

10.1 Output summary

Of the 264,000 input senses processed by the classifier:

- 227,000 (86%) were assigned a classification;
- 25,000 (9.5%) were left unclassified (i.e. the classifier failed to find any classification);
- 12,000 (4.5%) were rejected as intractable.[8]

10.2 Evaluation

The accuracy of the classifier was evaluated by taking a random sample of 1000 senses from the 227,000 senses assigned a classification. For each sense, an evaluator was asked to judge whether the assigned classification was accurate, i.e. represented a valid categorization of the definition.

[8] These include senses in wordclasses not covered by HTOED (chiefly prepositions, conjunctions, and pronouns); and senses whose definition indicates that they are semantically too vague to be meaningfully classified (e.g. proverb senses, senses with a long list of lemmas, senses defined as 'miscellaneous').

Note that this rubric is designed to check for *a* valid categorization, not all possible valid categorizations: see the discussion of multi-part definitions at section 12.2.

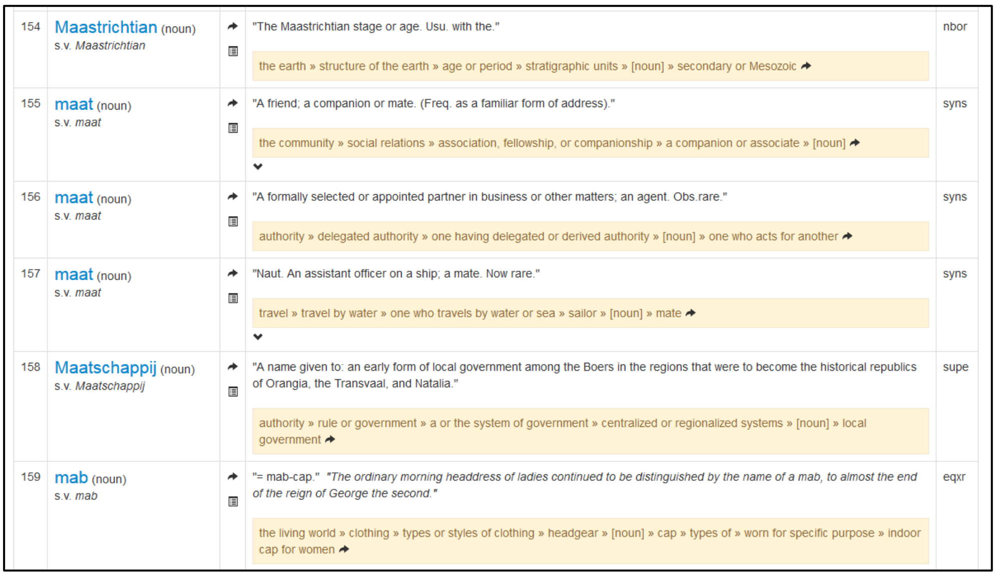
Overall, we found that:

- 25% of classifications were accurate, i.e. the correct node of the HTOED taxonomy had been identified;
- 22% of classifications were immediate neighbours of the correct node, i.e. a parent, child, or direct sibling node;
- 18% of classifications were second- or third-generation ancestors of the correct node, i.e. on the correct branch but not specific enough;
- 33% were either straightforwardly incorrect (i.e. on the wrong branch) or were not specific enough to be of any use (i.e. on the right branch, but too high up from the correct node).
- A small residue (<2%) were cases were the evaluator was uncertain of the correct classification (chiefly technical definitions, and obscure undefined compounds).

Only primary classifications were considered; runner-up classifications (see section 2.3) were disregarded.

11. Editing interface

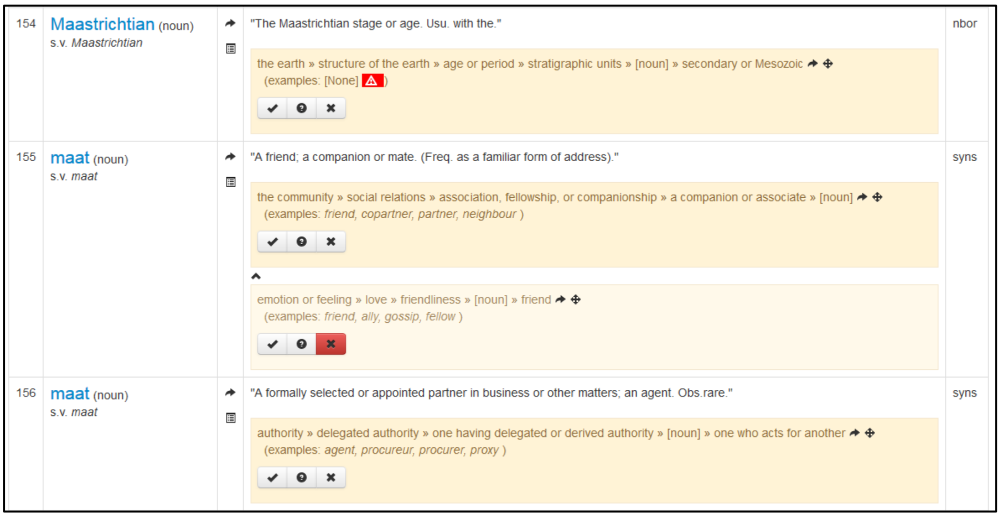
A web-based interface allows results to be reviewed and analysed by a number of different features, including wordclass, sense type (main sense or subentry, defined or undefined), HTOED branch, and principal method of classification:



154	Maastrichtian (noun) s.v. *Maastrichtian*	"The Maastrichtian stage or age. Usu. with the." the earth » structure of the earth » age or period » stratigraphic units » [noun] » secondary or Mesozoic	nbor
155	maat (noun) s.v. *maat*	"A friend; a companion or mate. (Freq. as a familiar form of address)." the community » social relations » association, fellowship, or companionship » a companion or associate » [noun]	syns
156	maat (noun) s.v. *maat*	"A formally selected or appointed partner in business or other matters; an agent. Obs.rare." authority » delegated authority » one having delegated or derived authority » [noun] » one who acts for another	syns
157	maat (noun) s.v. *maat*	"Naut. An assistant officer on a ship; a mate. Now rare." travel » travel by water » one who travels by water or sea » sailor » [noun] » mate	syns
158	Maatschappij (noun) s.v. *Maatschappij*	"A name given to: an early form of local government among the Boers in the regions that were to become the historical republics of Orangia, the Transvaal, and Natalia." authority » rule or government » a or the system of government » centralized or regionalized systems » [noun] » local government	supe
159	mab (noun) s.v. *mab*	"= mab-cap." *"The ordinary morning headdress of ladies continued to be distinguished by the name of a mab, to almost the end of the reign of George the second."* the living world » clothing » types or styles of clothing » headgear » [noun] » cap » types of » worn for specific purpose » indoor cap for women	eqxr

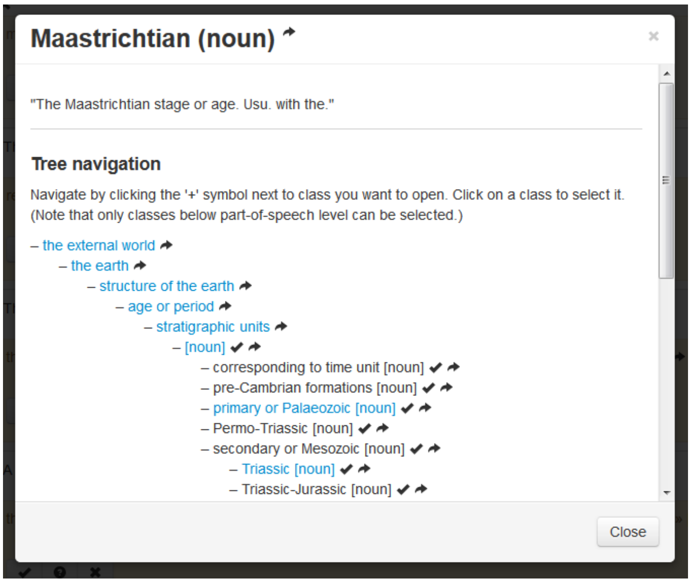
Figure 2: Editorial interface in review mode

The interface also has an ‘edit mode’ which provides controls for a user to approve, reject, or adjust a classification:



154 Maastrichtian (noun) s.v. *Maastrichtian*	"The Maastrichtian stage or age. Usu. with the." the earth » structure of the earth » age or period » stratigraphic units » [noun] » secondary or Mesozoic (examples: [None] ⚠) ✓ ❓ ✗	nbor
155 maat (noun) s.v. *maat*	"A friend; a companion or mate. (Freq. as a familiar form of address)." the community » social relations » association, fellowship, or companionship » a companion or associate » [noun] (examples: *friend, copartner, partner, neighbour*) ✓ ❓ ✗ emotion or feeling » love » friendliness » [noun] » friend (examples: *friend, ally, gossip, fellow*) ✓ ❓ ✗	syns
156 maat (noun) s.v. *maat*	"A formally selected or appointed partner in business or other matters; an agent. Obs.rare." authority » delegated authority » one having delegated or derived authority » [noun] » one who acts for another (examples: *agent, procureur, procurer, proxy*) ✓ ❓ ✗	syns

Figure 3: Editorial interface in edit mode



Maastrichtian (noun)

"The Maastrichtian stage or age. Usu. with the."

Tree navigation

Navigate by clicking the '+' symbol next to class you want to open. Click on a class to select it. (Note that only classes below part-of-speech level can be selected.)

- the external world
  - the earth
    - structure of the earth
      - age or period
        - stratigraphic units
          - [noun] ✓
            - corresponding to time unit [noun] ✓
            - pre-Cambrian formations [noun] ✓
            - primary or Palaeozoic [noun] ✓
            - Permo-Triassic [noun] ✓
            - secondary or Mesozoic [noun] ✓
              - Triassic [noun] ✓
              - Triassic-Jurassic [noun] ✓

Figure 4: Modal dialogue for adjusting an incorrect classification

We currently have a programme under way to systematically check and approve classifications. Approved classifications are fed back to the source database, becoming part of the training data next time round. This allows for an ongoing iterative process.

12. Limitations and further development

12.1 Taxonomy

A key limitation of the project is that it only attempts to classify input senses in terms of the existing HTOED taxonomy; it does not suggest or create new categories. This means that often there is no correct node to which a given input sense could be assigned: the 'bottom-up' construction of the taxonomy means that it is shaped by the existing OED content, with no provision for new senses representing new concepts.

12.2 Multi-part definitions

In general, the classifier treats each input sense as atomic: that is to say, it assumes that a single sense represents a single coherent meaning or usage.

In reality this assumption is flawed, because many individual senses can be decomposed into two or three distinct meanings. Indeed, the original editors of HTOED routinely interpreted OED senses in this way, and so many training-data senses have multiple HTOED classifications.

But the multiple meanings within a single sense can be signalled in more or less explicit ways, and can be hard to distinguish from single-meaning senses. For example, the definition of *scene queen* has two quite different meanings presented as semicolon-separated clauses:

> A woman who is prominent in a particular scene, esp. a particular music scene; (esp. in gay usage) a homosexual man who goes to gay bars, clubs, etc…

The definition of *overpower* v. 3 also has semicolon-separated clauses; but here these are really just restatements or nuances of the same core meaning:

> Of an emotion, fatigue, etc.: to overcome (a person, etc.) by intensity; to be too much or too intense for; to overwhelm.

It is very hard to define formally what differentiates the *scene queen*-type multi-part definition from the *overpower*-type single-sense definition.

We allow the classifier to treat certain input senses as having multiple meanings (and therefore to assign multiple HTOED classifications), where this is unambiguous; but the default approach of treating each input sense as atomic means that the assigned classification often fails to reflect the semantic range indicated by the definition.

12.3 Gloss and metalanguage

OED definitions consist broadly of two kinds of material:

- semantic gloss;
- metalanguage: various forms of grammatical, contextual, and usage information.

For the purposes of HTOED classification, the metalanguage is usually redundant, and is best jettisoned so that the classifier can focus on the gloss. This is a necessary first step for many of the analytic strategies described above. If metalanguage is confused for gloss, or vice versa, this can cause some significant problems.

In practice, separating gloss from metalanguage can be difficult, since OED definitions do not explicitly demarcate them.

Certain known patterns can be tested, for example, metalanguage often precedes and/or follows the gloss as separate sentences (sometimes bracketed). For example, in *mash* n. 3b:

> (Without article.) The state of being mashed or reduced to a soft pulp. Chiefly in *to beat (also boil, etc.) to mash.* Also in extended use.

The gloss is *the state of being mashed or reduced to a soft pulp*; the preceding and following sentences are metalanguage which can be discarded.

But gloss and metalanguage are often more fluidly integrated, making automatic separation more difficult. For example, *club-ball* is defined as:

> A term applied by Strutt and subsequent writers to games in which a ball is struck by a club or bat, esp. to the earlier types of these.

where the definition proper is *game[s] in which a ball is struck by a club or bat*, and the rest is metalanguage. But the classifier currently misconstrues *a term applied by...* as the start of the definition proper; this leads to the sense being misclassified.

There is no magic-bullet solution to the general problem of separating gloss from metalanguage. Really, it is just a matter of trying to account for more and more patterns as they are observed; this gradually improves performance, but is unlikely ever to be exhaustive.

12.4 Identifying the main sense of a word

When analysing a sense, a typical task that the classifier needs to perform is to find the meaning of certain keywords within the definition, e.g. a superordinate or synonym term. For example, in the definition *Stocks or shares in a mining company*, we need to be able to determine the sense in which *stocks* and *shares* are being used.

When a word appears in a definition, particularly as a synonym, the default assumption is that the word is being used in its primary modern sense. Although not impossible, it is unusual for an OED definition to use a word in an obscure, historical, figurative, dialect, or slang sense – at least not without some explicit indication.

Hence, to analyse a definition effectively, any system needs to be able to:

1. identify the primary modern sense of a word, as given in OED;

2. determine when the default assumption does not apply, i.e. when there is some indication that the word is being used in a different sense.

The first is an interesting problem in its own right, given that OED lists senses in chronological order, rather than by frequency or prototypicality. There are several promising approaches to this, both internal (evaluating the structure and relative significance of senses within the entry) and external (comparing senses in the OED entry with the corresponding entry in dictionaries which *do* rank senses by prototypicality). But these are not altogether reliable.

The second task – determining when the main-sense assumption does not apply – is handled by looking for explicit markers (e.g. the word in question is followed by a cross-reference pointing to a particular sense of that word); or by testing if the topic of the sense as a whole suggests a more technical sense of a given word within the definition. For example, *prosiphon* is defined as:

> The primitive siphon in an embryonic ammonoid, consisting of a kind of ligament attached to the protoconch.

Here the Topic_classifier establishes that the sense as a whole is zoological; so in analyzing the superordinate *siphon*, the classifier is able to prefer the specifically zoological sense of *siphon* over the more general main sense.

As these examples suggest, there is no attempt to perform full word-sense disambiguation of terms in definitions. Instead, a more primitive default/exception model is employed: by default, a term is assumed to be used in its main sense, unless the contextual evidence suggests that something else may be preferred.

12.5 External methods

All the methods discussed so far are internal methods, to the extent that they only draw on data from within OED and HTOED.

It is also worth considering what other resources could be brought to bear on the problem, especially resources that deal in hypernymy (e.g. Wordnet) or synonymy (e.g. Wiktionary). In general, external resources are of limited value because of the rarefied nature of OED content: most OED lexemes and senses do not appear in other lexical resources, and this is even more true of the input senses considered here. Still, for those OED terms which *do* appear in a resource like Wordnet or Wiktionary, these may provide more direct evidence for a classifier.

13. Acknowledgements

Thanks to Kate Wild, Andrew Ball, Liz Ashdowne and Michael Proffitt (all at OED) for reviewing this project at each stage of development, and for numerous valuable suggestions.

14. References

Alexander, M. & Bramwell, E. (2012). Mapping Metaphors of Wealth and Want: A Digital Approach. In Mills, C., Pidd, M. & Ward, E. (eds.) *Proceedings of the Digital Humanities Congress 2012. Studies in the Digital Humanities.* Sheffield: HRI Online Publications, 2014. Available online at: http://www.hrionline.ac.uk/openbook/chapter/dhc2012-alexander

Bauer, L. (2009). Typology of compounding. In Lieber, R. & Štekauer, P. (eds.) *The Oxford Handbook of Compounding.* Oxford: Oxford University Press, pp. 343-356.

Booij, G. (2007). *The Grammar of Words: An Introduction to Linguistic Morphology.* 2nd edition. Oxford: Oxford University Press.

Crystal, D. (2014). *Words in Time and Place: Exploring Language Through the Historical Thesaurus of the Oxford English Dictionary.* Oxford: Oxford University Press.

Kay, C. & Wotherspoon, I. (2002). Turning the dictionary inside out: some issues in the compilation of a historical thesaurus. In J. E. Diaz Vera (ed.) *A Changing World of Words: Studies in English Historical Semantics and Lexis.* Amsterdam: Rodopi, pp. 109-135.

Kay, C., Roberts, J., Samuels, M. & Wotherspoon, W. (2009). *Historical Thesaurus of the Oxford English Dictionary: With additional material from A Thesaurus of Old English.* Oxford: Oxford University Press.

Levin, B. & Hovav, M. R. (1998). Morphology and lexical semantics. In Spencer, A. & Zwicky, A. (eds.) *Handbook of Morphology.* Oxford: Blackwell, pp. 248-271.

Lieber, R. (2004). *Morphology and Lexical Semantics.* Cambridge: Cambridge University Press.

Mooney, R. J. (2005). Machine Learning. In Mitkov, R. (ed.) *The Oxford Handbook of Computational Linguistics.* Oxford: Oxford University Press, pp. 376-394.

Murphy, M. L. (2003). *Semantic Relations and the Lexicon: Antonymy, Synonymy, and other Paradigms.* Cambridge: Cambridge University Press.

Roberts, J. & Kay, C. (1995). *A Thesaurus of Old English*. London: King's College London Medieval Studies XI.

Taylor, J. (2003). *Linguistic Categorization*. 3rd edition. Oxford: Oxford University Press.