The big data of the Old English lexicon

Javier Martín Arista
Nerthus Project-Universidad de La Rioja

1. Looking at the lexicon from two angles

There are two main ways to the Old English lexicon, through the texts and through a dictionary or a lexical database. The textual and lexicographical sources offer different but complementary perspectives. The former gives access to the raw data, whereas the latter offers the structured data gathered and organised by the lexicographer or the database compiler. In the presentation that follows *The Dictionary of Old English Corpus* has been selected for textual analysis, while the lexicographical source has been the *Online Lexical Database of Old English Nerthus* (consulted on January 2016), which is based on the dictionary by Hall (although the dictionaries by Bosworth-Toller and Sweet were also used for its compilation). These two perspectives, along with their corresponding sources, are combined in this presentation into an analysis of big data for insights that lead to a better understanding of the lexicon of Old English. The visual tools have been provided by Watson Analytics (https://watson.analytics.ibmcloud.com).

2. The Corpus: token analysis

The written records of the language as presented by *The Dictionary of Old English Corpus* total about three million words, including around 187000 different attestations. Figure 1 shows the most frequent words in the Corpus (with 1000 or more occurrences), whose figure rises to 291. They belong to the grammatical classes of the article-determiner (*se*-sēo-*þæt*), the personal pronoun (*ic-þū-hē-hit-hēo*), the genitive of the personal pronoun functioning as a possessive adjective or pronoun (*mīn-þīn-his-hire*), the negative word (*ne*), the interrogative pronoun (*hwā-hwæt*), the conjunction (*and, gif, ac, butan, þēah, þe*), the relative pronoun (*þe),* the indefinite pronoun (*sum, man*), the numeral (*ān, twēone*), the quantifier (*micle, mā, fela*) and the preposition (*tō, of, mid, in, þonne, in, ofer, æt, fram, æfter, þurh, wið, ǣr, ūp, under*).

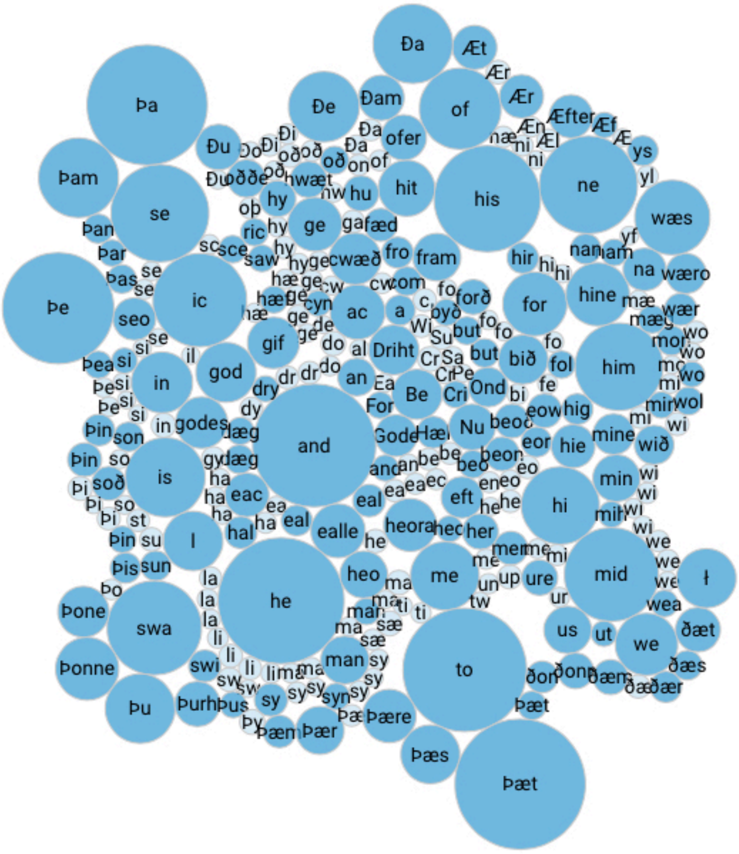

Figure 1: Over 1000 occurrences in *The Dictionary of Old English Corpus*.

Turning to the lexical classes, figure 1 presents remarkably high textual frequences of the adverbs *þā* ‘then’, *þǣr* ‘there’, *swīðe* ‘very’, *nū* ‘now’, *eft* ‘again, then’, *wel* well, *nǣfre* ‘never’, *sōðlīce* ‘truly’, *ǣfre* ‘ever’, *witodlīce* ‘truly’; the irregular verbs *bēon* ‘to be’, *dōn* ‘to do’, *gān* ‘to go’, *willan* ‘to wish’; the preterite-present verbs *magan* ‘to be able’, *sculan* ‘to owe’; the lexical verbs *cweðan* ‘to say’, *faran* ‘to go, travel’, *cuman* ‘to come’, *habban* ‘to have’, *weorðan* ‘to become’, *secgan* ‘to say’, *hātan* ‘to order’, *sendan* ‘to send’, *sellan* ‘to give’; the nouns *mann* ‘man’, *fæder* ‘father’, *God* ‘God’, *nama* ‘name’, *driht* ‘lord’, *rīce* ‘kingdom’, *dæg* ‘day’, *eorðe* ‘earth’, *sunu* ‘son’, *hælend* ‘saviour’, *folc* ‘people’, *sāwol* ‘soul’, *miht* ‘power’, *heorte* ‘heart’, *cyning* ‘king’, *word* ‘word’, *þing* ‘thing’, *lif* ‘life’, *gāst* ‘spirit’, *stōw* ‘place’, *woruld* ‘world’, *sǣ* ‘sea’, *ece* ‘pain’, *land* ‘land’, *bearn* ‘heir’, *hand* ‘hand’, *biscop* ‘bishop’, *lichama* ‘body, corpse’, *tīde* ‘time’, *lār* ‘teaching’, *weg* ‘way’, *wita* ‘wise man, inhabitant’, *synn* ‘sin, injury, wrong’, *yfel* ‘evil’, *mōd* ‘mind’, *here* ‘army’, *niht* ‘night’, *ǣ* ‘law’, *deað* ‘death’; the adjectives *eall* ‘all’, *sōð* ‘true’, *hālga* ‘holy’, *gelēaf* ‘believing’, *georn* ‘eager’; and the names *Crist* and *Petrus*. It must be borne in mind that textual occurrences have not been lemmatised. This does not affect invariable classes but has the important consequence of throwing lower textual frequences of variable classes, with which the nouns, adjectives and verbs listed above would present even higher frequences if all the inflectional forms were attributed to the same lemma.

Apart from the very frequent words, the number of occurrences of the words in the Corpus deserves attention.

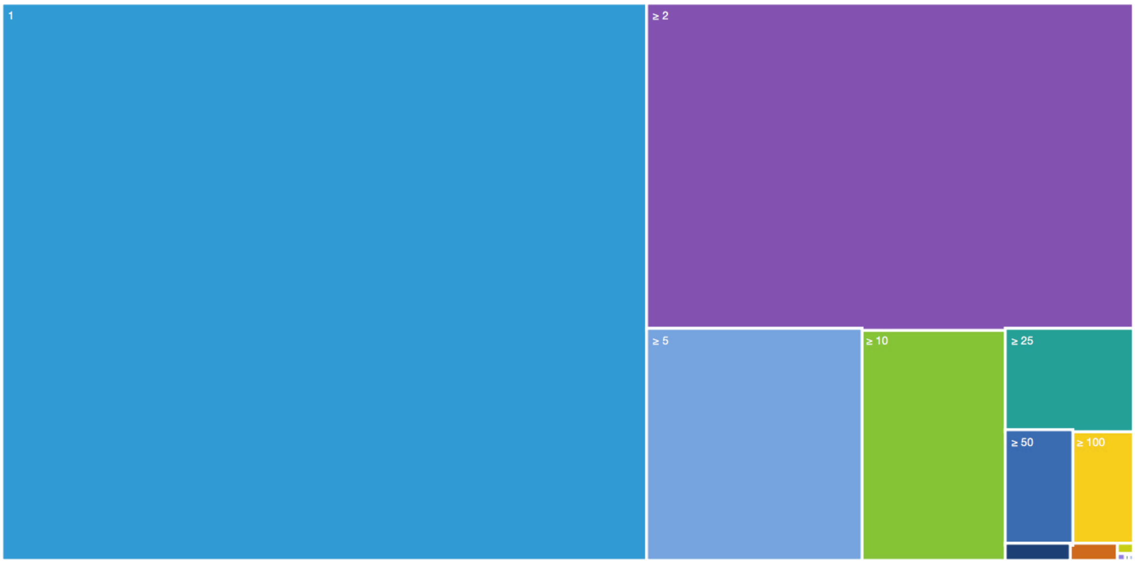

Figure 2: Hapax legomena and dis legomena in *The Dictionary of Old English Corpus.*

As can be seen in figure 2, approximately one half of the total of attested forms appear once only, thus constituting unique formations or *hapax legomena* in token analysis. Moreover, about one fourth of the attestations are *dis legomena* or forms with two occurrences in the Corpus. Hapax legomena and dis legomena together represent over two thirds of the occurrences in the Corpus.

Although not all of them can be analysed as derivatives to which the prefix ge- has been attached (see figure 11), about one ninth of the words in *The Dictionary of Old English Corpus* begin with the sequence ge-. Although the same reasoning is applicable to word endings, it can be assumed that most of the words which end with the sequences included in the next figure display actual inflectional endings (-e is probably the less reliable in this respect).

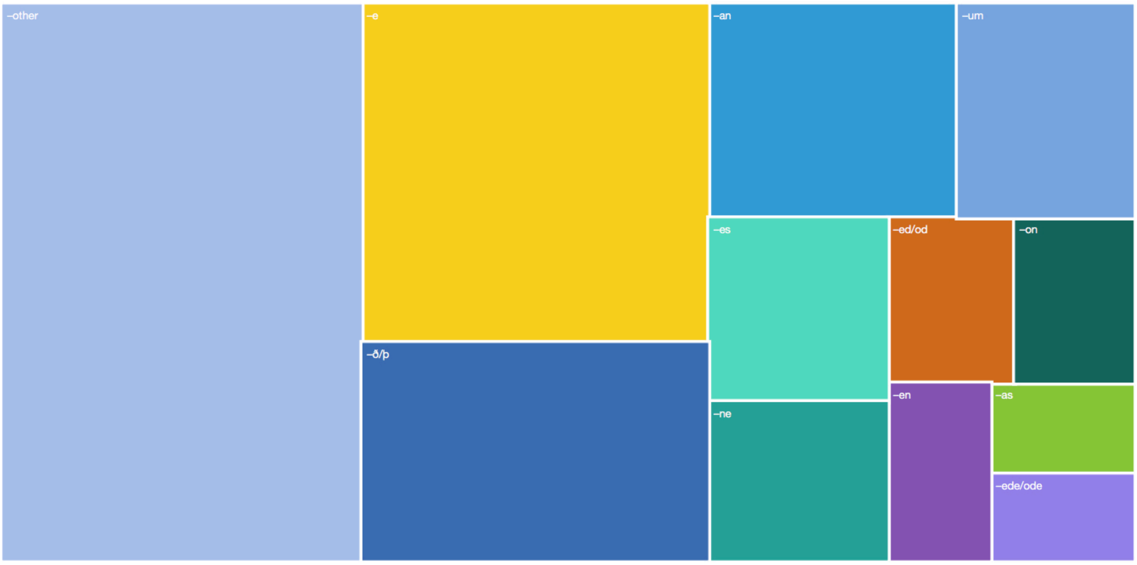

Figure 3: Inflectional endings in *The Dictionary of Old English Corpus*.

The occurrences of the most frequent inflectional endings rise to over two thirds of the attestations, as can be seen in figure 3. In the breakdown of the endings, -e totals nearly one fifth of the attestations. This inflection corresponds, among others, to the dative singular of the strong noun and adjective, the nominative of the feminine and neuter weak noun, the feminine and neuter nominative of the weak declension of the adjective, the first person singular of the present indicative of strong and weak verbs, the singular of the present subjunctive of strong and weak verbs. The inflectional ending -an, which marks the accusative, genitive and dative singular as well as the nominative and accusative plural of the weak declension of the noun and the adjective, and the uninflected infinitive of strong and weak verbs, is present in nearly one tenth of the attestations. The ending for the third person singular and all the plural of the present indicative of strong and weak verbs, -ð/-þ, throws a total of approximately one eighth of the attestations. Other verbal inflections throw lower figures: the ending -ed/-od for the past participle of weak verbs, the ending, -on for the preterite plural of strong and weak verbs, -en for the past participle of strong verbs and -ede/-ode for the first person singular of the preterite indicative and the singular of the preterite subjunctive. By nominal cases, the -um ending characteristic of the dative plural of the strong and weak noun declension and the dative singular and plural of the strong and weak declension of the adjective stands out with respect to the -es ending for the genitive singular ending of the strong declension of the noun and the adjective and the -ne ending for the accusative of the strong declension of the adjective and the pronouns.

3. The lexical database: type analysis

To shift to the lexicographical perspective, the average Old English dictionary contains between 30000 and 35000 headword entries. *Nerthus* comprises around 31000 files, which can be broken down by lexical category as shown in figure 4. Nouns represent about one half of the entries to a dictionary and verbs about one fourth, the remaining fourth comprising adjectives, adverbs and the grammatical classes.

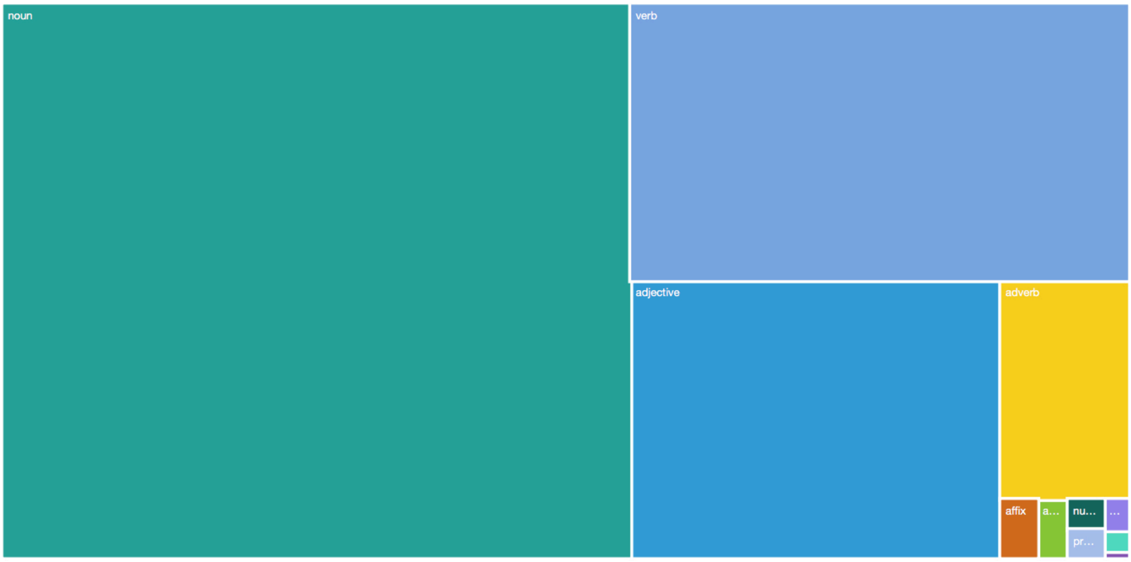

Figure 4: Dictionary entries by lexical class.

If the focus is put on nouns, there are more masculine than feminine nouns, these two genders comprising about four fifths of the noun total and the remaining fifth corresponding to the neuter gender. This is shown in figure 5.

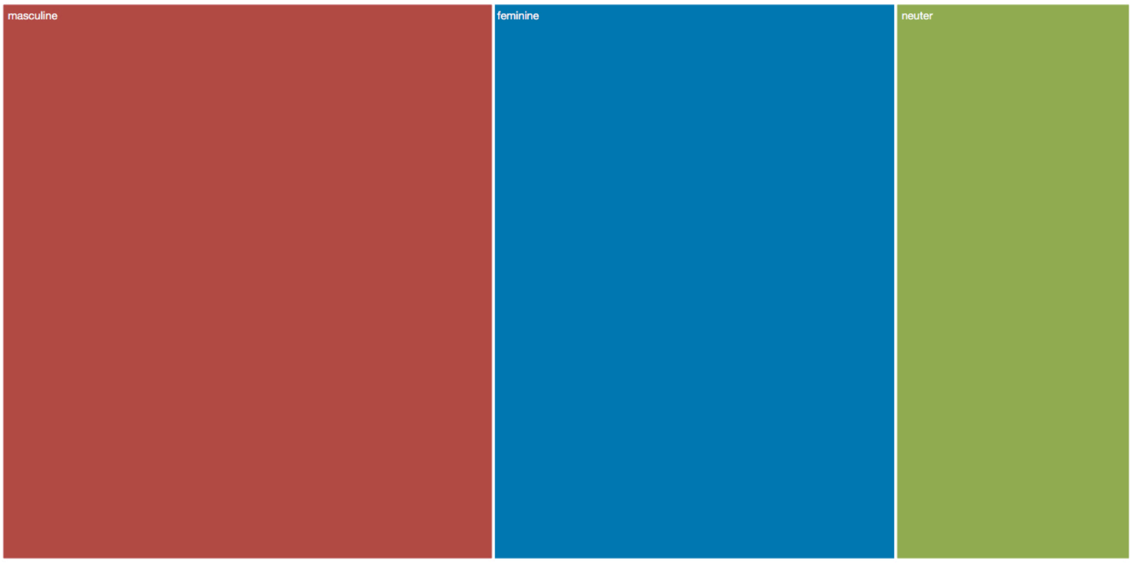

Figure 5: Nominal gender.

The figures represented by figure 5 have been calculated on the basis of the main gender of nouns, some of which are attested with two and even the three genders. For instance, the following nouns can be found in the masculine, the feminine and the neuter: *ǣspryng* ‘pallor; gloom’, *ǣspryng* ‘spring, fountain; departure’, *ǣt* ‘once, formerly’, *bismer* ‘filthiness; disgrace; shame; reproach, mockery, insult, blasphemy’ , *brēost* ‘breast; stomach, womb; heart, mind, thought; disposition’, *ceder* ‘cedar’, *cwild* ‘death; destruction; plague’, *ðrēa* ‘reproof; threat, menace, abuse; oppression, attack; punishment’, *ðrūh* ‘pipe; trough; sarcophagus’, *ealdorwisa* ‘chief’, *ēarliprica* ‘flap of the ear’, *fulwiht* ‘baptism’, *græft* ‘carved or graven image, sculpture’, *gūðfana* ‘standard, ensign, banner’, *līget* ‘lightning, flash of lightning’, *lyft* ‘air, breeze; clouds, sky, atmosphere, heavens’, *onweald* ‘authority, power; territory, jurisdiction’, *pæð* ‘path, track; valley’, *slōh* ‘miry place’, *swelgend* ‘glutton; drunkard’, *unseht* ‘disagreement, discord, quarrel’, *unwæstm* ‘weed; barrenness’, *ūplyft* ‘sky’, *wæstm* ‘growth, increase; plant, fruit, offspring, product; result; benefit, interest, usury’, *weorðmynd* ‘dignity, honour, nobleness, glory’, *wēsten* ‘desert, wilderness’, *wōl* ‘disease, pestilence; mortality’.

Within the lexical class of the verb, there are three times as many weak verbs as strong verbs, the number of irregular verbs being practically negligible (*dōn* ‘to do’, *gān* ‘to go’, *wesan* ‘to be’, *willan* ‘to wish’ along with their compounds and derivatives). This is represented in figure 6.

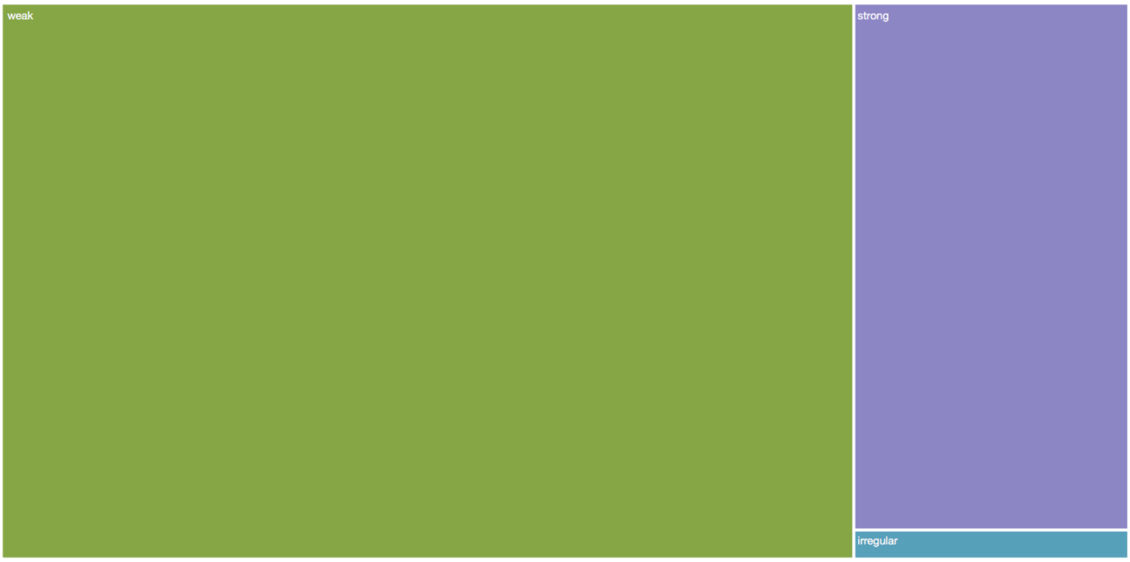

Figure 6: Weak, strong and irregular verbs.

As shown in figure 7, the first and second classes constitute the vast majority of weak verbs. The third weak class is restricted to four verbs as well as their compounds and derivatives: *habban* ‘to have, own, possess; to keep’ (compounds and derivatives: *æthabban, forhabban, ofhabban, wiðerhabban, wiðhabban, ymbhabban*); *(ge)hycgan* ‘to think, consider, concieve, study; to understand; to determine; to remember’ (compounds and derivatives: *āhycgan, behycgan, forhycgan, oferhycgean, onhycgan, wiðhycgan*); *libban* ‘to live; to exist’ (compounds and derivatives: *ālibban, belibban, eftlibban, mislibban, oferlibban*); *secgan* ‘to say, explain, tell of, speak; signify’ (compounds and derivatives: *āsecgan, besecgan, foregesecgan, foresecgan, forsecgan, forðsecgan, gesecgan, ofsecgan, onsecgan, sōðsecgan, tōsecgan, wiðsecgan*).

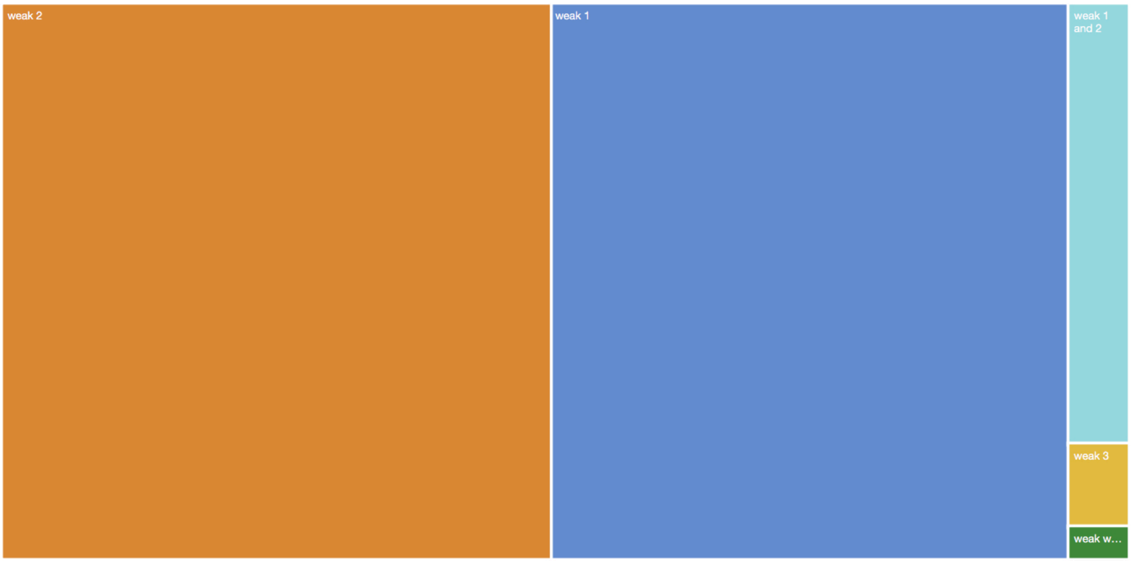

Figure 7: Weak verbs.

In the strong classes, the verbs from the first, second, third and seventh strong classes together represent about two thirds of the total. The fourth class is the smallest, comprising only the following verbs as well as their compounds and derivatives: *becuman* ‘to become’, *beran* ‘to bear’, *brecan* ‘to break’, *cuman* ‘to go, come’, *cwelan* ‘to die’, *delan* ‘to fall, sink’, *dwel*an ‘to err’, *ðwer*an ‘to stir’, *felan* ‘to stick’, *gelan* ‘to pour’, *helan* ‘to cover, hide, conceal’, *hlecan* to unite’, *hwelan* ‘to roar, rage’, *niman* ‘to take’, *scieran* ‘to cut, hew, cleave’, *swelan* ‘to burn’, *teran* to tear’. The visualization of these data is given in figure 8.

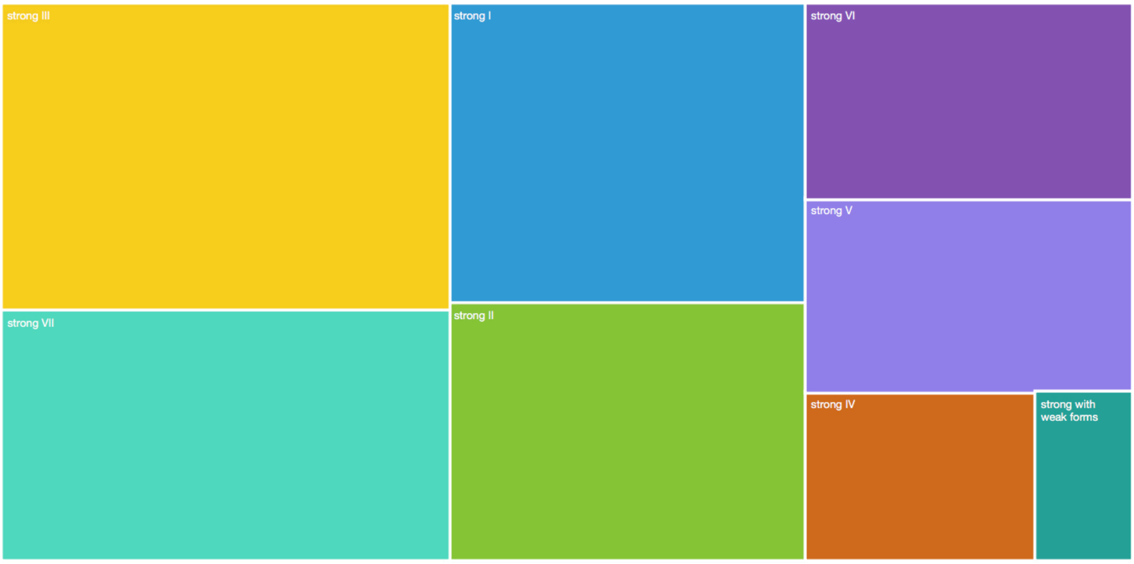

Figure 8: Strong verbs.

As regards word-formation, most words are morphologically related to other words by derivational processes. For example, the derivational paradigm of the prime *beald* ‘bold, brave, confident, strong; presumptuous, impudent’ comprises: *bealde* ‘boldly, courageously, confidently; immediately’, *bealdian* ‘to be bold’, *bealdlīce* ‘boldly’, *bealdnes* ‘boldness’, *bealdor* ‘lord, master, hero’, *bealdwyrde* ‘bold of speech’, *beldan* ‘to encourage, excite, impel, exhort, confirm’, *bieldo* ‘boldness, courage, arrogance, confidence’, *cynebeald* ‘royally bold, very brave’, *forðbylding* ‘emboldening, encouragement’, *fulbealdlīce* ‘full boldly, very boldly’, *(ge)bieldan* ‘to encourage, excite, impel, exhort, confirm’, *gebæeldan* ‘to encourage, excite, impel, exhort, confirm’, *gebild* ‘bold, brave, confident, corageous’, *hēafodbald* ‘impudent’, unbeald ‘cowardly, timid, weak, irresolute, distrustful’, *unbieldo* ‘want of boldness, diffidence, timidity’.

As can be seen in figure 9, derivation (prefixation, suffixation and *zero derivation* or derivation without explicit derivational morphemes) and compounding together give rise to about five sixths of the total lexical stock, while primitives rise to approximately one tenth. This means that the great majority of lexical items are morphologically related to other lexical items, as, for instance in the derivation *flōwan* ‘to flow’ > *oferflōwan* ‘to flow over’ > *oferflōwend* ‘superflous’ > *oferflōwendnes* ‘excess’. As regards morphologically unrelated words, most of them are nouns, thus *anclēow* ‘ankle’, *clyne* ‘lump of metal’, *fæs* ‘fringe, border’, *fenester* ‘window’, *gellet* ‘bowl’, *hēcen* ‘kid’, *īre* ‘a monetary unit’, *max* ‘net’, *oll* ‘contumely, contempt, scorn, insult’, *slīm* ‘slime’, *ȳr* ‘back of axe’, etc.

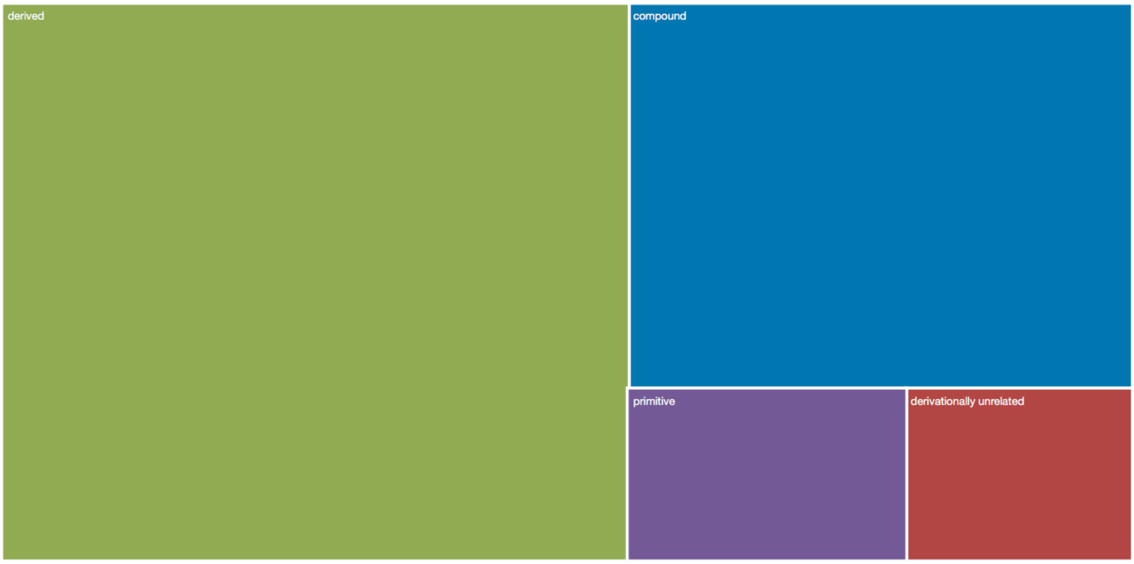

Figure 9: Derivational relatedness.

As is displayed in figure 10, three fourths of derivational primes are nouns, the remaining fourth corresponding to verbs and adjectives. The sum of all the derivational paradigms based on nominal primes represents more than one half of the lexicon. This is clearly a consequence of the existence of large series of compounds with nominal bases such as *bīspellbōc* ‘book of parables’, *blētsingbōc* ‘benedictional’, *canonbōc* ‘book of canons’, *ciricbōc* ‘church-book’, *cnēorisbōc* ‘book of genealogy’, *cræftbōc* ‘commentary’, *Cristesbōc* ‘the Gospel’, *cwidbōc* ‘Book of Proverbs’, *ðēnungbōc* ‘mass-book’, *dōmbōc* ‘doom-book’, *færelbōc* ‘itinerary’, *fīfbēc* ‘Pentateuch’, *fōrbōc* ‘itinerary’, *frēolsbōc* ‘charter of freedom’, *Gecyndbōc* ‘Genesis’, *frōforbōc* ‘book of consolation’, *hālgungbōc* ‘benedictional’, *gerīmbōc* ‘calendar’, *geanbōc* ‘duplicate charter’, *godspellbōc* ‘book containing the four gospels’, *handbōc* ‘hand-book’, *healsbōc* ‘amulet’, *hierdebōc* ‘pastoral book’, *yrfebēc* ‘will’, *lǣcebōc* ‘book on medicine, book of recipes’, *lǣdenbōc* ‘Latin book’, *landbōc* ‘charter in which land is granted’, *lārbōc* ‘book containing instruction’, *mæssebōc* ‘mass-book, missal’, *mynsterbōc* ‘book belonging to a monastery’, *nambōc* ‘register of names’, *healsungbōc* ‘book of exorcism’, *pistolbōc* ‘book containing the Epistles’, *rǣdingbōc* ‘lectionary’, *sangbōc* ‘singing book; service-book’, *scriftbōc* ‘penitential’, *sealmbōc* ‘psalter’, *sīðbōc* ‘itinerary’, sinoðbōc ‘book containing the decrees of a synod’, *spellbōc* ‘book of sermons’, *sumorbōc* ‘summer lectionary’, *trahtbōc* ‘treatise’, wīsbōc ‘book of wisdom’, *wītegungbōc* ‘book of prophecy’, *ymenbōc* ‘book of hymns.

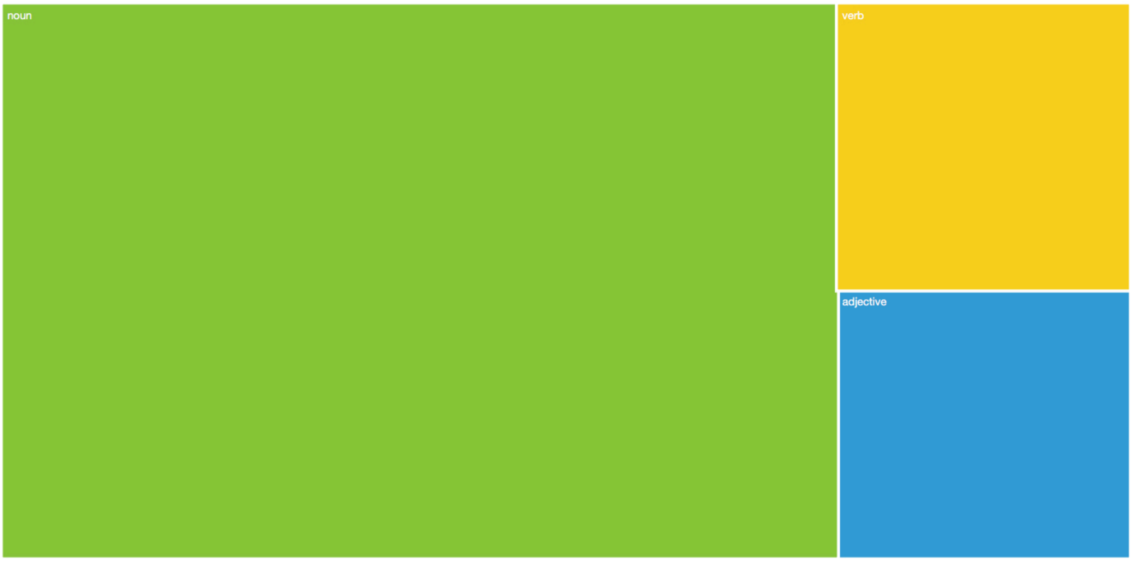

Figure 10: Derivational primes by lexical class.

In prefixation, the negative prefix un- and the Germanic verbal prefixes ge-, ā-, be-, on-, for-, and tō- qualify as the most frequent affixes. Formations with these prefixes include, among many others, *unāwendendlic* ‘unchangeable’, *getredan* ‘to tread’, *ābacan* ‘to bake’, *begylpan* ‘to boast’, *onblōtan* ‘to sacrifice’, *forcwolstan* ‘to swallow’, *tōsceacan* ‘to shake off’. The visualization is presented in figure 11.

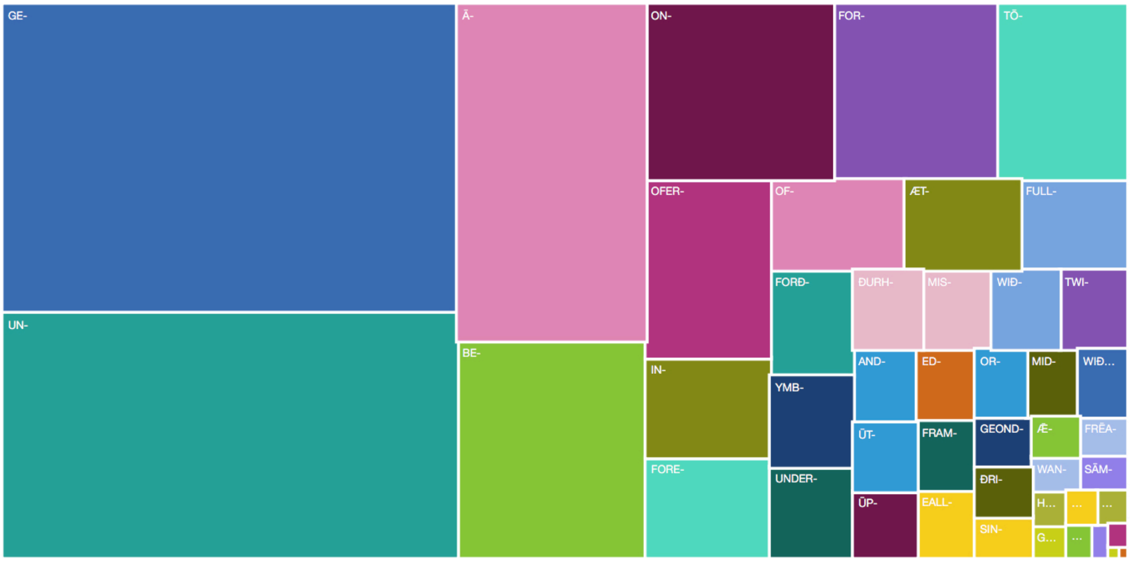

Figure 11: Prefixes.

The frequency of the suffixes -nes and -lic (as in *gefēgnes* ‘association’ and *wundorlic* ‘wonderful’) clearly sticks out, although -ung, -e, -ing, -end, -ere, -līce and -ig are also quite frequent (thus, for instance, *gecīgung* ‘summons’, *wrāðe* ‘angrily’, *bōcrǣding* ‘reading of books’, *æfterfylgend* ‘follower’, *folgere* ‘follower’, *egesig* ‘terrible’, *selflīce ‘egotistic’)*. Figure 12 represents this point.

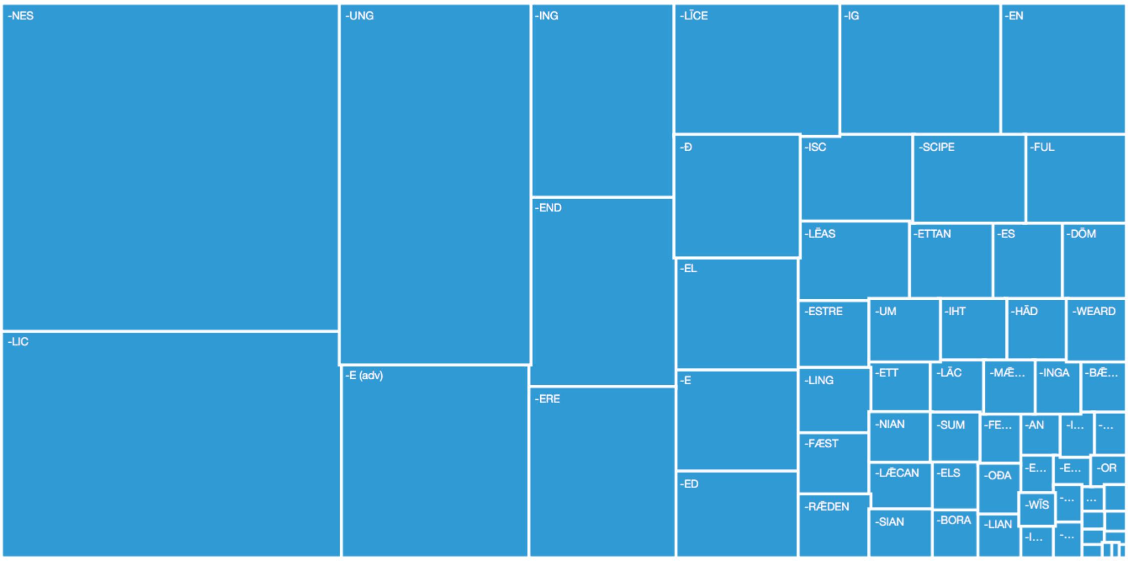

Figure 12: Suffixes.

To finish the analysis of the big data of the lexicon of Old English, the question of meaning is addressed in terms of the semantic primes of the Natural Semantic Metalanguage Research Programme (Goddard and Wierzbicka 2014).

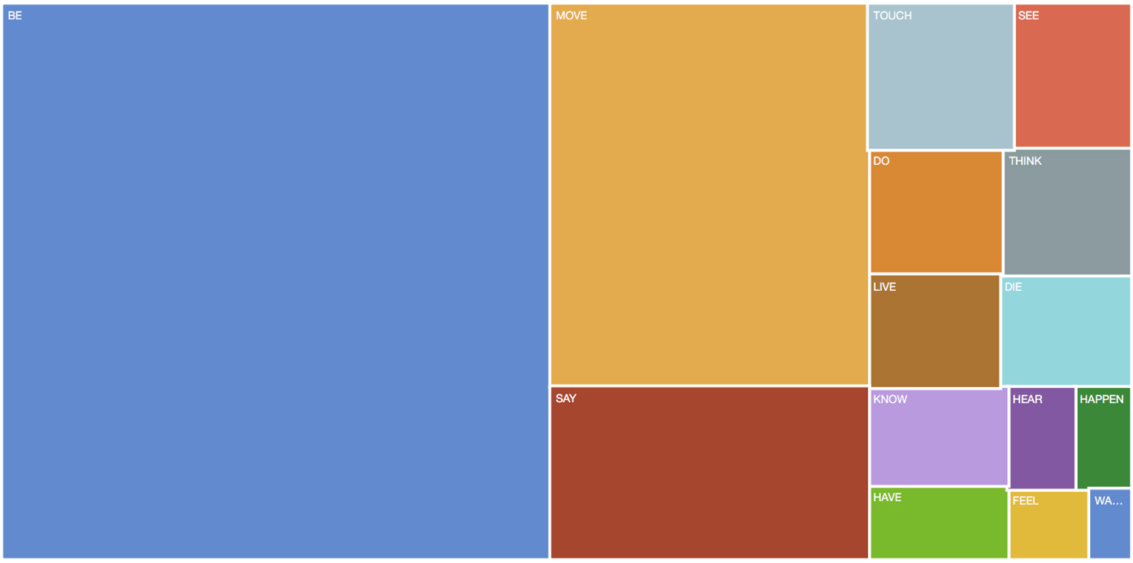

Figure 13: Exponence of semantic universals in verbs.

As can be seen in figure 13, nearly one half of the verbs in the dictionary convey the meaning of the semantic prime BE (including BECOME and the causative TO CAUSE TO BECOME). The meanings MOVE and SAY are expressed by about one fourth of the verbs, while others like TOUCH, DO, SEE, THINK, LIVE, DIE, KNOW and HAVE appear less often.

4. Some conclusions

This presentation has looked at the lexicon of Old English from the textual and the lexicographical angles, for which *The Dictionary of Old English Corpus* and the lexical database *Nerthus* have been selected as data sources. The big data analysis has been determined by the nature of these sources: a corpus allows for token analysis whereas a lexical database based on a dictionary necessarily restricts the scope to type analysis. On the other hand, the raw data from a corpus leads to more general conclusions than the structured data presented by a database. In this respect, the consequences of the lack of lemmatisation have been mentioned. The ambiguity has also been noted of word beginnings and endings with prefixes and inflectional endings, respectively.

In token analysis, hapax legomena and dis legomena together constitute three fourths of the Corpus. As in other corpora, hapax legomena rise to one half of the attested forms, but the fact that dis legomena are as many as one fourth of the Corpus probably deserves more attention in furture research. It must also be noted that not only grammatical words but also some items from the lexical classes stand out for their textual frequency, thus the verbs *bēon* ‘to be’, *dōn* ‘to do’, *gān* ‘to go’, *willan* ‘to wish’, *cweðan* ‘to say’, *faran* ‘to go, travel’, *cuman* ‘to come’, *habban* ‘to have’, *weorðan* ‘to become’ and *secgan* ‘to say’.

Turning to type analysis, and given that Old English shows generalised and relatively transparent derivational morphology, a significant part of this presentation has been devoted to questions related to the units and relations of word-formation. Some results in the area of morphological relatedness are worthy of consideration. Un-, ge-, ā-, be-, on-, for-, and tō- are the most frequent affixes, while the suffixes -nes and -lic clearly outnumber the other suffixes. More importantly, derivation and compounding together represent about five sixths of the lexicon, while primitives are approximately one tenth and the figure of morphologically unrelated words -most of which are nouns- hardly reaches one fifteenth). On the semantic side, the analysis has been focused on verbs and, more specifically, on the exponence of semantic universals. It has turned out that the most frequent meanings in type analysis are related to the primes BE (about one half of the verbal class) as well as MOVE and SAY (which, together, are present in one fourth of the verbal lexicon). It must be pointed out in this line that, for verbs, the exponence of semantic primes in type analysis strongly correlates with the textual frequency displayed in token analysis: for example, *beon* 'to be' is the most frequent verb and the semantic prime BE the most generalised in the class of verbs.

To conclude, lexical analysis based on big data represents a promising line of research not only because it paves the way for future undertakings but also because it allows the researcher to draw conclusions based on well described empirical evidence and an explicit methodology, thus taking a step from structured data towards linked data.

References

Bosworth, J. and T. N. Toller. 1973 (1898). *An Anglo-Saxon Dictionary*. Oxford: Oxford University Press.

Campbell, A. 1972. *An Anglo-Saxon Dictionary: Enlarged Addenda and Corrigenda*. Oxford: Clarendon Press.

Goddard, C. and A. Wierzbicka. 2014. *Words and Meanings. Lexical Semantics Across Domains, Languages and Cultures*. Oxford: Oxford University Press.

Hall, J. R. Clark. 1996 (1896). *A Concise Anglo-Saxon Dictionary*. Toronto: University of Toronto Press.

Healey, A. diPaolo (ed.) with J. Price Wilkin and X. Xiang. 2004. *The Dictionary of Old English Web Corpus*. Toronto: Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto.

Martín Arista, Javier (ed.), Laura García Fernández, Miguel Lacalle Palacios, Ana Elvira Ojanguren López and Esaúl Ruiz Narbona. 2016. *NerthusV3. Online Lexical Database of Old English*. Nerthus Project. Universidad de La Rioja. [www.nerthusproject.com]

Sweet, H. 1976 (1896). *The Student´s Dictionary of Anglo-Saxon*. Cambridge: Cambridge University Press.

Toller, T. N. 1966 (1921). *An Anglo-Saxon Dictionary: Supplement*. Oxford: Clarendon Press.