CHAPTER 52

ANALYTICITY AND SYNTHETICITY IN THE HISTORY OF ENGLISH

BENEDIKT SZMRECSANYI

1. Introduction

No one interested in typological change in the history of English will manage to avoid the terms "analytic" and "synthetic", terminology that goes back to August Wilhelm von Schlegel (Schlegel 1818). The textbook view on this issue is epitomized in the following seminal quote:

> En Europe les langues dérivées du latin, et *l'anglais*, ont une grammaire tout analytique... synthétiques dans leur origine... elles penchent fortement vers les formes analytiques. (Schlegel 1846: 161)[1]

Thus English is supposed (cf. e.g. Baugh and Cable 1993: 60) to have changed from a rather synthetic language (i.e. one that relies heavily on inflections to code grammatical information) in Old English times into a rather analytic language that draws on word order and function words to convey grammatical information. The wholesale loss of nominal and verbal inflections that started toward the end of the Old English period, so the textbook story goes, has set in motion a long-term drift (on the notion of drift, see also Hawkins, this volume) toward analyticity

[1] 'In Europe, the languages derived from Latin, as well as English, have strongly analytic grammars... synthetic in origin... they tend strongly toward analytic forms' (translation mine).

that is still in operation today (but see e.g. Danchev 1992: 36, for a more nuanced account).

The trouble is that the terms "analytic" and "synthetic" are more often than not seen as referring to notional concepts not amenable to rigorous quantification, and there is also a lot of terminological variance. Moreover, most previous scholarship has looked into the matter by studying some isolated features, in an attempt to generalize from these. In this chapter, I endeavor to rethink the analyticity–syntheticity dichotomy in a way that makes possible precise, systematic, and holistic measurements, drawing on quantitative, frequency-based measures originally developed in quantitative morphological typology (Greenberg 1960). Specifically, I utilize a quantitative, language-internal measure of "overt grammatical analyticity", defined as the text frequency of free grammatical markers, and a measure of "overt grammatical syntheticity", defined as the text frequency of bound grammatical markers (see Haselow, this volume, for a similar approach in the domain of derivation).

Using such measures, Kortmann and Szmrecsanyi (2009) and Szmrecsanyi and Kortmann (2009) (see also Kortmann, this volume) have demonstrated that there is a good deal of variability in contemporary geographic varieties of English. Szmrecsanyi (2009) additionally diagnoses substantial text-type variability in Present-Day English and presents evidence that written standard English, both British and American, has recently become significantly more synthetic. The task before us now is to extend this line of analysis to long-term diachronic change in the history of English. Applying the method, which unlike much extant scholarship is concerned with token frequencies (not with e.g. analyzing lists of linguistic variants), to historical corpus material covering the period between late Old English and Present-Day English, I shall argue that the textbook view à la Schlegel—which, to reiterate, implies a fairly steady drift in the history of English from synthetic to analytic—is in need of some rethinking.

2. Methodical preliminaries

Empirically, I tap the Penn Parsed Corpora of Historical English series, which consists of the following corpora: The Penn-Helsinki Parsed Corpus of Middle English, second edition (PPCME2); the Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME); and the Penn Parsed Corpus of Modern British English (PPCMBE). The three corpora yield a database of 605 texts, which span slightly less than 4 million words of running text. Each of the texts in this database can be assigned not only to well-established periods in the history of English (Middle English, Early Modern English, Late Modern English) but also to specific centuries, starting with the early twelfth century and ending with the early twentieth century. Notice that the texts represent a variety of text types, such as letters, sermons, handbook prose, and history writing.

The texts in the Penn Parsed Corpora of Historical English series are all part-of-speech annotated and syntax-parsed using roughly the same tagsets and annotation schemes. The bonus of drawing on a richly annotated data source like this is that unlike in work based on corpus material which is not part-of-speech annotated *a priori* (e.g. Greenberg 1960; Szmrecsanyi and Kortmann 2009), the numerical findings reported in this chapter derive not from manually annotated random samples but from an exhaustive analysis of the *entire* material (about 4 million words) contained in the Penn Parsed Corpora of Historical English series.

In line with the methodology outlined in more detail in Szmrecsanyi (2009), I set the scene by defining four broad categories, identifiable via part-of-speech annotation, into which individual word tokens in the database may be grouped. These categories precisely define the present chapter's take on "analyticity" and "syntheticity" (all examples given below are from Present-Day English).

1. *Analytic word tokens* (i.e. synsemantic words or function words), which I in turn define as being members of synchronically closed word classes: complementizers (as in *he thinks* ***that*** *he will go*), coordinating conjunctions (*he sleeps* ***and*** *she reads*), determiners (***the*** *house*), infinitive markers (*he needs* ***to*** *go*), modals (*he* ***can*** *go*), negators (*she will* ***not*** *leave*), existential THERE (***there*** *are many examples*), pronouns (*I, you, me,...*), prepositions (*of, in, at,...*), comparative and superlative quantifiers (*more* and *most*), and auxiliary BE (*I* ***was*** *called*), DO (***do*** *you read me?*), and HAVE (*I* ***have*** *eaten lunch*). Note that this definition of analyticity and of what should count as a function word is a fairly uncontroversial one in accordance with standard reference works (e.g. Bussmann 1996: 22, 471). I stress right at the outset that what is not captured in this definition of analyticity is, of course, syntax and word order.
2. *Synthetic word tokens*, which carry bound grammatical markers (including clitics) such as verbal (*I walk > he walk-s*), nominal (*one dog > two dog-s*), and adjectival inflectional affixes (*small > small-er/small-est*). I also include allomorphies such as ablaut phenomena (*sing > sang*), *i*-mutation (*goose > geese*), and other non-regular yet clearly bound grammatical markers. Observe therefore that the underlying model of morphological analysis is essentially an "item-and-process" model (Hockett 1954). In more general terms, I thus categorize into this class inflected comparative and superlative adjectives, the possessive marker or clitic (the "Saxon Genitive"), plural inflected nouns, inflected verb forms (i.e. present participles, past forms, perfect participles, passive participles, and the inflected present tense forms *-st*, *-th*, *-s*, and *-en*). What is not included in this notion of syntheticity is the (in any event questionable) construct of "zero" inflections; see Szmrecsanyi (2009) for a more detailed discussion of why zero inflections do not mesh well with frequency-based metrics.

3. *Simultaneously analytic and synthetic word tokens*, for example, inflected auxiliary verbs (BE, DO, and HAVE), as in *he* ***has*** *eaten lunch*.
4. *Purely lexical word tokens*, such as singular nouns. This is a wastebasket category that is uninteresting for present purposes.

Subsequently, we establish the text frequencies of the relevant tokens in each of the 605 texts subject to analysis. At this stage, the Penn Parsed Corpora of Historical English series' uniform part-of-speech tagset (which makes automatic retrieval of most of the above tokens easy) takes center stage. Once we have these frequencies, it is a trivial task to generate a table that details two frequency indices, by corpus text or century:

- An "analyticity index" (henceforth AI), which is calculated as the ratio of the number of free grammatical markers (i.e. function words) in a text to the total number of words in the text, normalized to a sample size of 1,000 tokens.
- The "syntheticity index" (henceforth SI), which is in parallel calculated as the ratio of the number of words in a text that bear a bound grammatical marker to the total number of words in the sample text, normalized to a sample size of 1,000 tokens.

This exercise in index calculation is essentially the method proposed by Greenberg (1960), who in turn was inspired by Edward Sapir. Hence, the present study's interest in token frequencies, rather than inventory size, is fully in keeping with Greenberg's approach. The rationale is that the present chapter seeks to avoid "intuitive estimates based on over-all impressions" (Greenberg 1960: 185), and instead much prefers the objectivity and "sufficient rigor" (Greenberg 1960: 185) that can be attained by measuring text frequencies. Notice in particular that the reliance on "words" (instead of e.g. grams) as the basic unit of analysis is mandated by the method outlined in Greenberg (1960).

3. Historical analyticity–syntheticity variability by numbers

Figure 1 plots mean analyticity indices, by century, against mean syntheticity indices. How do we read this plot? Consider the data point for the thirteenth century: its coordinates indicate that in 1,000 words of thirteenth-century running text, we find on average 434 word tokens that are analytic function words and 151 word tokens that carry a bound grammatical marker (i.e. an inflection). In short, the scatterplot calls into question the notion that analyticity-syntheticity variability after Old English can be described in terms of a steady trend. Observe that the dataset as a whole yields a mean AI of 471 free grammatical per thousand words of running text (minimum:

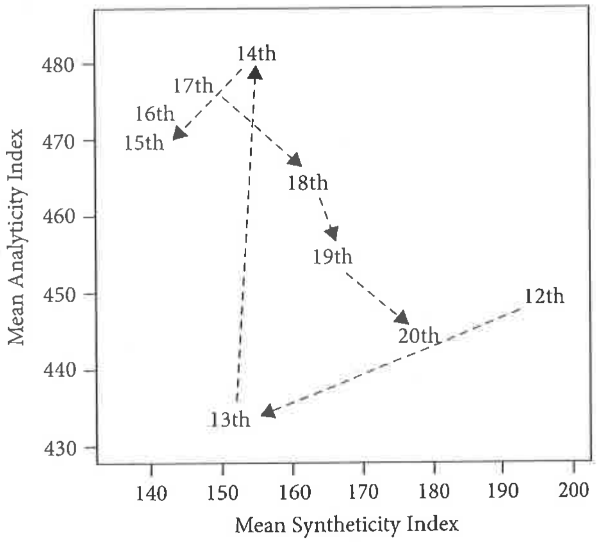

Figure 1. Mean analyticity indices against mean syntheticity indices; aggregation to century

326, maximum: 566, standard deviation (sd): 32.8 index points), and a mean SI of 148 bound grammatical markers per thousand words of running text (minimum: 67, maximum: 220, sd: 30.2 index points).[2] Now, as for variability in real time we start out with a data point for the twelfth century that is the most synthetic one in Figure 1 (196 SI points) but only averagely analytic (449 AI points). By the thirteenth century, mean syntheticity has decreased significantly to 151 SI points; the difference compared to the twelfth century is significant at p = .013, according to an independent samples t-test. Analyticity has also actually decreased somewhat to 434 AI points, although the differential is not significant (p = .64). From the thirteenth to the fourteenth century, we witness a huge and significant (p = .002) surge in analyticity to 481 AI points. Syntheticity levels stay roughly the same (151 versus 155 SI points, p = .70). Nothing much happens between the fourteenth and the seventeenth century—both analyticity and syntheticity index levels remain roughly the same (one-way ANOVAs, with century as between-group variable, yield insignificant p values of .42 for the analyticity index and .09 for the syntheticity index). Between the seventeenth and the twentieth centuries, however, we observe a steady and monotonous drift toward more syntheticity and less analyticity. The data point for the twentieth century is both significantly more synthetic (p = .011) and less analytic (p = .012) than the data point for the seventeenth century. In any case, with an AI score of 444 points and an SI score of 178 points, the twentieth century has come almost full circle back to where we started in the twelfth century.

[2] The measures of central dispersion, as well as t-tests and ANOVAs (ANalysis Of VAriance) here and throughout the article are calculated on the basis of index values for individual texts in the Penn Parsed Corpora of Historical English corpus series (total N = 605).

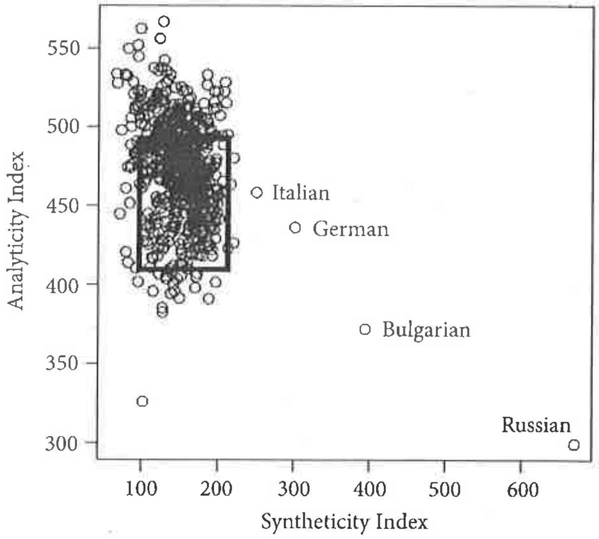

Figure 2. Analyticity indices against syntheticity indices, selected European languages versus Penn Parsed Corpora of Historical English texts. Box depicts variation space in Figure 1.

Figure 2 puts historical analyticity-syntheticity variability in English into a wider crosslinguistic context. On the basis of an application of the coding scheme outlined in section 2 to contemporary newspaper prose in other languages, Figure 2 plots analyticity and syntheticity index levels for four contemporary European languages: Italian, German, Bulgarian, and Russian.[3] The coordinates of these data points are, as a matter of fact, nothing to write home about—we knew before that Russian is a particularly synthetic language, and that, say, Italian is less synthetic and more analytic than Russian. However, Figure 2 also depicts an "English diachrony cloud" in which every data point corresponds to one individual text in the Penn Parsed Corpora of Historical English corpus series; the box indicates the (century-based and thus aggregated) variation space delimited in Figure 1. Figure 2 makes amply clear two things. First, against the backdrop of crosslinguistic variation, historical analyticity variability in English is considerable but syntheticity variability is not. Second, even eight centuries ago English was less synthetic than, say, Italian or its contemporary cousin, German.

Let us now have a closer look at the historical development of each index in turn. Figure 3 reports two box plots that depict variance in index levels (on the *y*-axis) by century (on the *x*-axis). Analyticity (cf. the left plot in Figure 3) increased between the twelfth and the fourteenth century, plateaued until the seventeenth century, and decreased subsequently. The right-hand diagram in Figure 3, which box-plots the development of the SI in real time, is a fairly felicitous mirror image to that. Syntheticity decreased rather robustly between the twelfth and the thirteenth century, stayed in a valley until the seventeenth century, and bounced back

[3] So, for example, in 1,000 words of running Italian text, 458 word tokens are analytic function words while 250 word tokens carry a bound grammatical marker (i.e. an inflection).

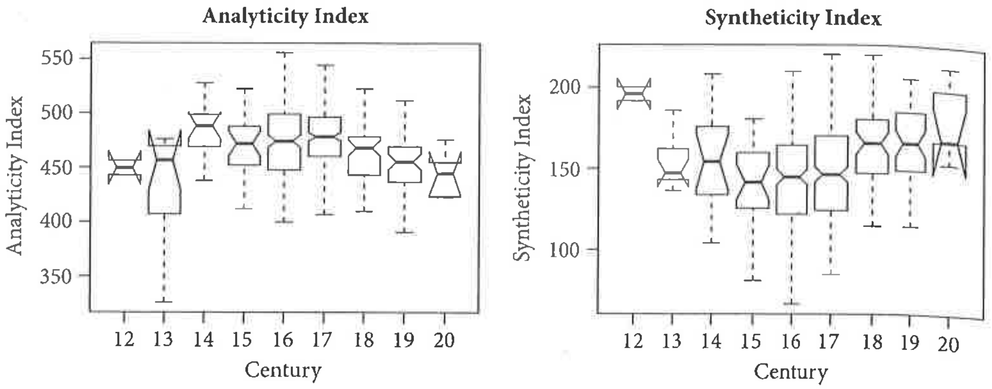

Figure 3. Box plots. Left: analyticity indices. Right: syntheticity indices. Boxes depict the interquartile index range comprising the middle 50 percent of index level observations for the relevant texts, with the thick line in the boxes indicating the median. The whiskers above and below the boxes extend to data points that score no more than 1.5 times the interquartile range.

afterwards. So far, the story is perfectly in line with our earlier observations based on Figure 1. However, the extension of the whiskers in Figure 3 also demonstrates that century-internal variance—and thus, between-text variability—itself varies in real time. Intertextual variability in terms of both analyticity and syntheticity was especially pronounced during the Early Modern English period and decreased afterwards. The most plausible explanation of this heterogeneity is that it reflects the wide range of genres covered in the Early Modern English period, which is ultimately a corpus compilation issue. I speculate further that the ensuing narrowing of between-text variability may reflect the growing impact of standardization, which seeks to "cut out" variation (Stein 1994).

I emphasize that a more detailed account would have to factor in text type variability as well. I reserve a discussion of this issue for another occasion, though it should be mentioned in passing that in sermons, for instance, syntheticity does not appear to have soared much after the seventeenth century. In stark contrast, syntheticity levels collapsed more dramatically than elsewhere in history writing after the eighteenth century.

4. The linguistic sources for analyticity-syntheticity variability

We have so far not discussed which individual grammatical markers are most robustly involved in historical analyticity-syntheticity variability. A one-way ANOVA which uses a three-partite sub-corpus distinction (PPCME2 versus PPCEME versus PPCMBE) as grouping variable identifies four analytic features (or feature groups)

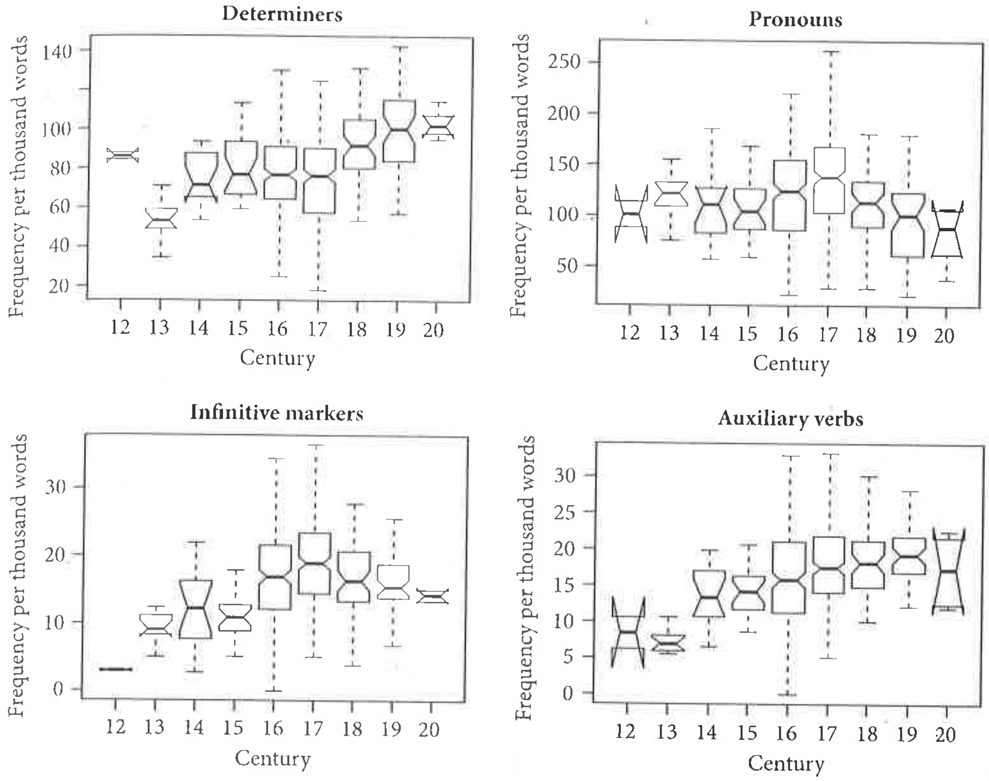

Figure 4. Box plots. Four component features loading on the analyticity index

that exhibit the most extensive real-time variance: determiners, pronouns, infinitive markers, and auxiliary verbs. Figure 4 shows four corresponding box plots that depict these features' diachronic variance, with the frequencies of determiners, as in (1), and auxiliary verbs, as in (2), exhibiting a fairly steady upward trend in real time.

(1) For **what** hope of this desirable blessing is there in the present state of things? (PPCMBE, burton-1762)

(2) you whom I **have** so wildly loved (PPCMBE, wilde-1895)

In other words, here we have two features where analyticity has been genuinely if locally on the rise in the dataset at hand. The rise of auxiliaries (such as e.g. auxiliary DO) in the history of English is, needless to say, a well-established fact (cf. e.g. Denison 1985: 201–2).

By contrast, frequency-wise both pronouns, as in (3), and infinitive markers, as in (4), had their heyday in the sixteenth and seventeenth centuries.

(3) **I** hertely recommend **me** to **you** (PPCEME, wplumpt-1530)

(4) and a Man knows not how **to** beleive anie of them (PPCEME, hoxinden-1660-e3-h)

Both features thus mirror the overall development of analyticity levels discussed earlier to some extent, although clearly for different reasons—fluctuations in pronoun

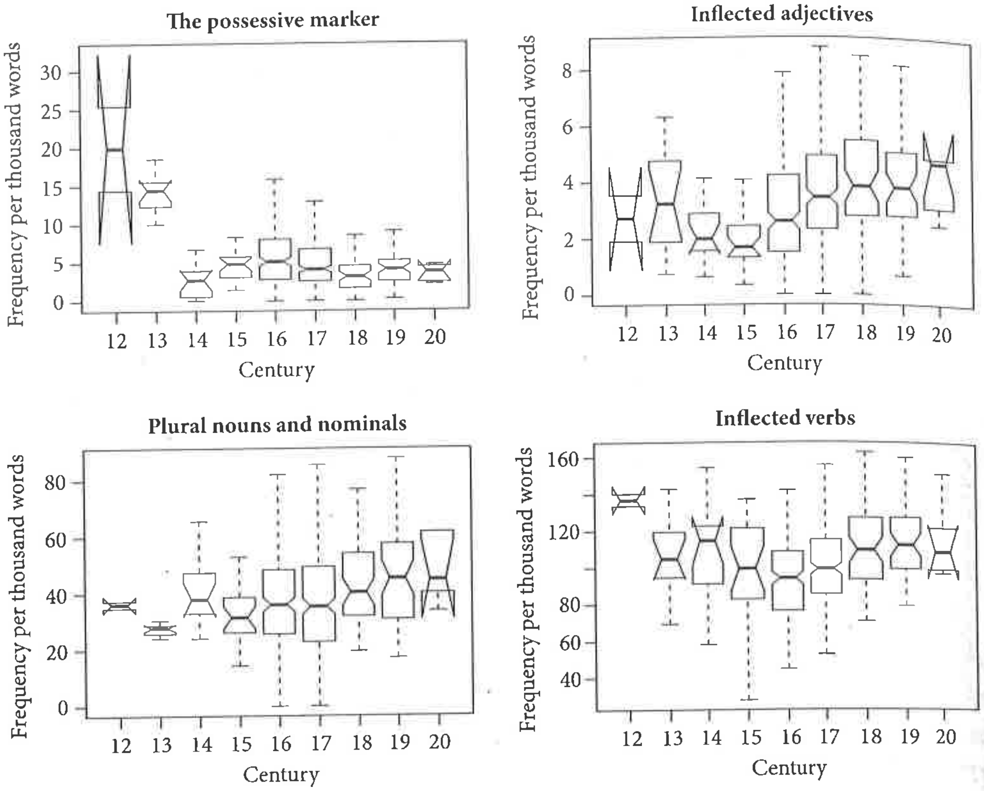

Figure 5. Box plots. The four component features loading on the syntheticity index

frequencies are likely to be partly a stylistic issue, and partly they may be a function of the competition between finite verbs forms (with explicit subjects) and nonfinite verbs forms (with implicit subjects).[4] Having said that, the box plots in Figure 4 show more intertextual variance for infinitive markers, and unlike pronouns infinitive markers also appear to be on the rise overall—a likely outgrowth of an emerging preference for nonfinite verb forms in English (cf. Rohdenburg 1995; Los 2005).

We next turn to the features that load on the syntheticity index: the possessive marker, inflected adjectives, plural nouns and nominals, and inflected verbs (see Figure 5). As for the possessive marker, as in (5), we note a phenomenal collapse of text frequencies after the twelfth century, which is of course fully in accordance with the standard account of the marker's history (Thomas 1931; Mustanoja 1960). Frequencies of the possessive marker after the fourteenth century fluctuate moderately, again in line with the literature (Hinrichs and Szmrecsanyi 2007; Szmrecsanyi, forthcoming).

(5) Eale hwu heh mæden **Godes** moder (PPCME2, cmkentho.m1)

Inflected adjectives, as in (6), exhibit a fairly steady frequency increase, with a minor slump in the fourteenth and fifteenth centuries. In any event, this way of looking at the data does not really suggest that analytic adjective comparison with

[4] I am indebted to Bettelou Los for pointing this out to me.

more and *most* has been taking away market share from synthetic adjective comparison, as is customarily argued (Pound 1901; Leech and Culpeper 1997). What we can see from the box plot, however, is that from the sixteenth century onward we find a lot more intertextual frequency variance than before.

(6) Whethir thou be **hyer** than Petir and **holyer** than Andrew, **worthier** than Iames (PPCME2, cmaelr4.m4)

Inflected verbs, as in (7), follow the U-shaped pattern familiar from our discussion of overall syntheticity levels (cf. Figure 3). Inflected verbs hit bottom, in terms of frequencies, during the sixteenth century. Observe that in Present-Day English, inflected verbs are roughly as frequent as they were during the thirteenth century.

(7) for hit **is** selde **visited** of men (PPCME2, cmwycser.m3)

Lastly, we note that the frequency of plural-marked nominal forms, as in (8), is fairly stable in real time, though since the eighteenth century we do find a tendency for plural forms to become more frequent.

(8) Besides our native hardy **fruits**, **flowers**, and **vegetables**, the horticulturist also has to grow **exotics** from all **parts** of the world (PPCMBE, weathers-1913)

5. Discussion and conclusion

In this chapter, I have endeavored to rethink the story of historical analyticity-syntheticity variability after the Old English period. To this end, I utilized Greenberg-inspired, precise, frequency-based, and holistic measures of grammatical analyticity and syntheticity, and applied these to the Penn Parsed Corpora of Historical English corpus series, which covers the period between the twelfth and the twentieth century. It turns out that the eight centuries covered in the material are clearly not characterized by a steady drift toward more analyticity and less syntheticity. Instead, analyticity was on the rise until the end of the Early Modern English period, but declined subsequently; the reverse is true for syntheticity. That said, the historical variability in English in all the historical periods we investigated was not particularly dramatic. Compared to languages like Italian, German, Bulgarian, and Russian, English scores consistently low on syntheticity in all these periods. An analysis of frequency fluctuation in individual markers further revealed that while in the big picture, twentieth-century English is quantitatively almost back to the analyticity-syntheticity coordinates defining twelfth-century English, modern analyticity and syntheticity seem qualitatively different from their Early English counterparts. For example, determiners have become increasingly important as an analytic category, but pronouns have been on the decline. Conversely, the possessive marker used to be (but is no more) an important synthetic marker, whereas inflected adjectives are on the rise.

In sum, this contribution would seem to have suggested that we can learn a lot from applying terminology, concepts, and ideas developed in quantitative morphological typology to the study of language-internal historical variability. Future research in English historical linguistics will want to extend this line of analysis to also cover Old English material, and outside English linguistics we will eventually need similar analyses in other languages (German, French, Japanese...) to learn more about crosslinguistic regularities in historical analyticity-syntheticity variance.

References

Baugh, Albert C., and Thomas Cable. 1993. *A History of the English Language.* Englewood Cliffs, NJ: Prentice-Hall.

Bussmann, Hadumod. 1996. *Routledge Dictionary of Language and Linguistics*, ed. and trans. Gregory Trauth and Kerstin Kazzazi. London: Routledge.

Danchev, Andrei. 1992. 'The Evidence for Analytic and Synthetic Developments in English'. In *History of Englishes: New Methods and Interpretations in Historical Linguistics*, ed. Matti Rissanen, Ossi Ihalainen, Terttu Nevalainen, and Irma Taavitsainen, 25–41. Berlin: Mouton de Gruyter.

Denison, David. 1985. 'Why Old English Had No Prepositional Passive'. *English Studies* 66: 189–204.

✓ 2 Greenberg, Joseph H. 1960. 'A Quantitative Approach to the Morphological Typology of Language'. *International Journal of American Linguistics* 26: 178–94.

Hinrichs, Lars, and Benedikt Szmrecsanyi. 2007. 'Recent Changes in the Function and Frequency of Standard English Genitive Constructions: A Multivariate Analysis of Tagged Corpora'. *English Language and Linguistics* 11: 437–74.

Hockett, Charles F. 1954. 'Two Models of Grammatical Description'. *Word* 10: 210–31.

Kortmann, Bernd, and Benedikt Szmrecsanyi. 2009. 'World Englishes between Simplification and Complexification'. In *World Englishes—Problems, Properties and Prospects: Selected Papers from the 13th IAWE Conference*, ed. Lucia Siebers and Thomas Hoffmann, 265–85. Amsterdam: Benjamins.

Leech, Geoffrey and Jonathan Culpeper. 1997. 'The Comparison of Adjectives in Recent British English'. In *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*, ed. Terttu Nevalainen and Leena Kahlas-Tarkka, 125–32. Amsterdam: Rodopi.

Los, Bettelou. 2005. *The Rise of the* To-*Infinitive.* Oxford: Oxford University Press.

Mustanoja, Tauno F. 1960. *A Middle English Syntax.* Helsinki: Société Néophilologique.

Pound, Luise. 1901. *The Comparison of Adjectives in English in the XV and the XVI Century.* Anglistische Forschungen, 7. Heidelberg: Carl Winter.

Rohdenburg, Günther. 1995. 'On the Replacement of Finite Complement Clauses by Infinitives in English'. *English Studies* 76: 367–88.

Schlegel, August Wilhelm von. 1818. *Observations sur la langue et la littérature provençales.* Paris: Librairie grecque-latine-allemande.

———. 1846. *Oeuvres de M. Auguste-Guillaume de Schlegel, écrites en français et publiées par Èdouard Böcking.* Leipzig: Weidmann.

Stein, Dieter. 1994. 'Sorting Out the Variants: Standardization and Social Factors in the English Language 1600–1800'. In *Towards a Standard English, 1600–1800*, ed. Dieter Stein and Ingrid Tieken-Boon van Ostade, 1–17. Berlin: Mouton de Gruyter.

Szmrecsanyi, Benedikt. 2009. 'Typological Parameters of Intralingual Variability: Grammatical Analyticity Versus Syntheticity in Varieties of English'. *Language Variation and Change* 21: 319–53.

———. Forthcoming. 'The Great Regression: Genitive Variability in Late Modern English News Texts'. In *Morphosyntactic Categories and the Expression of Possession*, ed. Kersti Börjars, David Denison, and Alan Scott. Amsterdam: Benjamins.

Szmrecsanyi, Benedikt, and Bernd Kortmann. 2009. 'Between Simplification and Complexification: Non-Standard Varieties of English around the World'. In *Language Complexity as a Variable Concept*, ed. Geoffrey Sampson, David Gil, and Peter Trudgill, 64–79. Oxford: Oxford University Press.

Thomas, Russell. 1931. *Syntactical Processes Involved in the Development of the Adnominal Periphrastic Genitive in the English Language.* Ann Arbor: University of Michigan.