

Exploiting parsed corpora in grammar teaching

SEAN WALLIS, *Survey of English Usage, University College London*

IAN CUSHING, *Department of Education, Brunel University London*

BAS AARTS, *Survey of English Usage, University College London*

Abstract

The principal barrier to the uptake of technologies in schools is not technological, but social and political. Teachers must be convinced of the pedagogical benefits of a particular curriculum before they will agree to learn the means to teach it. The teaching of formal grammar to first language students in schools is no exception to this rule.

Over the last three decades, most schools in England have been legally required to teach grammatical subject knowledge, i.e. linguistic knowledge of grammar terms and structure, to children age five and upwards as part of the national curriculum in English. A mandatory set of curriculum specifications for England and Wales was published in 2014, and elsewhere similar requirements were imposed.

However, few current English school teachers were taught grammar themselves, and the dominant view has long been in favour of ‘real books’ rather than the teaching of a formal grammar. English grammar teaching thus faces multiple challenges: to convince teachers of the value of grammar in their own teaching, to teach the teachers the knowledge they need, and to develop relevant resources to use in the classroom. Alongside subject knowledge, teachers need pedagogical knowledge – how to teach grammar effectively and how to integrate this teaching into other kinds of language learning.

The paper introduces the *Englicious*¹ web platform for schools, and

¹The *Englicious* project was funded by the UK research councils AHRC

summarises its development and impact since publication. *Englicious* draws data from the fully-parsed *British Component of the International Corpus of English*, ICE-GB. The corpus offers plentiful examples of genuine natural language, speech and writing, with context and potentially audio playback. However, corpus examples may be age-inappropriate or over-complex, and without grammar training, teachers are insufficiently equipped to use them.

In the absence of grammatical knowledge among teachers, it is insufficient simply to give teachers and children access to a corpus. Whereas so-called ‘classroom concordancing’ approaches offer access to tools and encourage bottom-up learning, *Englicious* approaches the question of grammar teaching in a concept-driven, top-down way. It contains a modular series of professional development resources, lessons and exercises focused on each concept in turn, in which corpus examples are used extensively. Teachers must be able to discuss with a class why, for instance, *work* is a noun in a particular sentence, rather than merely report that it is.

The paper describes the development of *Englicious* from secondary to primary, and outlines some of the practical challenges facing the design of this type of teaching resource. A key question, the ‘selection problem’, concerns how tools parameterise the selection of relevant examples for teaching purposes. Finally we discuss curricula for teaching teachers and the evaluation of the effectiveness of the intervention.

1 Introduction

Parsed corpora, or ‘treebanks’, have a wide range of applications, from research into language production and variation, to providing the knowledge base for natural language processing algorithms. In language education, parsed corpora have untapped potential in the development of teaching resources and curricula, at a wide range of levels. This application is the focus of this paper.

For three decades in the UK, the school curriculum has been directed top-down by government. Since the 1990s, L1 (first language) English teaching has had an increasing emphasis on concepts of grammar, especially in the most recent version of the curriculum, which has been taught in schools since 2014 (Department for Education 2013). This shift in emphasis has faced opposition from mainstream teacher opinion, which tends to see grammar teaching as prescriptive, of limited benefit, and contrary to a child’s linguistic creativity. Yet both primary

(AH/H015787/1 and AH/L004550/1) and EPSRC (a UCL BEAMS Enterprise Award 2011), and other sources.

(junior) and secondary (high) schools have faced increasing demands placed upon them to teach English grammar.² This paper examines how parsed corpora can be used in classroom teaching to address the needs of the new curriculum while offering teachers a descriptive approach to grammar that can be empowering for students.

Corpora have been used in education for a number of years. We first discuss the perspective of ‘data-driven learning’ (DDL, sometimes referred to as ‘classroom concordancing’, Johns and King 1991), which hands students the same research tools as academic researchers, sets them tasks and encourages them to explore the data.

This approach has proved effective in intermediate-to-advanced translation studies for idiomatic expressions. In this application, the translator student has an understanding of the context of the utterance in both languages and can interpret the data they are presented with. What Vygotsky famously referred to as the ‘zone of proximal development’ (Chaiklin 2003) is the novel target phrase for translation, ‘scaffolded’ by the already-understood surrounding context.

But is DDL effective for teaching first-language grammar in schools, as Johns and King advocate? Various linguists, including Kaltenböck and Mehlmauer-Larcher (2005) have questioned the effectiveness of DDL in this context. Without a knowledgeable teacher to direct the class, students learn only that language use is varied, without grasping the reasons for that variation. They can over-generalise from a few examples. Peering into a corpus, we can see *that* something is the case but we cannot see *why* it is. Below we consider how natural language samples, whether drawn from a corpus or not, can be used effectively for L1 English grammar teaching.

1.1 Corpora in the classroom

The idea of using a corpus to assist language learners is intuitive and attractive, beginning with the rise of cheap home computing in the 1980s. Johns (1986) proposed using a corpus of English text in second language learning with *Micro-concord*, a simple ‘Key Word in Context’ (KWIC) concordancing tool running on the Sinclair Spectrum microcomputer. Suddenly, teachers had access to new technology. Concordancing, an idea borrowed from Bible scholars, became the basis for a new methodology of language teaching. Concordancing was simple, intuitive and available. Tools had been developed for research. Could they be used in a classroom context?

²In this paper we will use the UK educational nomenclature of ‘primary’ and ‘secondary’ schools in preference to the US ‘junior’ and ‘high’ school equivalents.

KWIC concordancers produce ‘concordance lines’, the results of a search across the entire corpus, where each matching instance is aligned around the target word or words. They reframe how the reader engages with a text: instead of reading the text from start to end and possibly discovering a target word, they view citations in context, which can cause the reader to identify potential regularities.

According to Johns, the corpus and tool plays the role of ‘informant’, replacing the need for a fluent speaker of the language. Whereas a teacher directs the lesson and the class, responding to feedback, the corpus passively responds to queries the student poses. Johns comments that in his DDL perspective, ‘the language learner is also... a researcher-worker whose learning needs to be driven by access to linguistic data’ (Johns 1991).

The best way to demonstrate this approach is with an example. In Johns’ 1991 paper he describes teaching a language-learning class the appropriate use of *persuade* and *convince*. Do native English speakers say *convince to* or *convince that* in the same way that they might say *persuade to* or *persuade that*? In Figure 1 we reproduce his exercise with a different corpus, the *British Component of the International Corpus of English* (ICE-GB, Nelson *et al.* 2002, see Section 1.5), applying John’s wild card searches (*persuad**, *convinc**).

In Johns’ corpus, all cases of *convince* are followed by a subordinated *that*-clause (two cases are zero-subordinate), i.e. *convince* {someone} *that*, with no instances of *convince to*. On the other hand, *persuade* tended to be found in the pattern *persuade* {someone} *to* (14 out of 18 cases).

Johns rightly notes that the non-existence of *convince to* in his corpus does not make it impossible. Indeed, as the first line in ICE-GB reveals, example (1) below is entirely possible. Nonetheless, in ICE-GB, only one in 15 cases of *to*-clauses with either verb contains *convince*, compared to 26 out of 39 *that*-clauses.

(1) *My Dad was trying to convince me to meet her* [S1A-007 #310]

The corpus does not contain explicit linguistic knowledge of why this different distribution occurs. It cannot reflect or comment on why a particular pattern is found. Students must guess rules inductively. Thus Johns describes how students inferred a rule not suggested by the teacher, namely that *persuade* was connected to actions, whereas *convince* concerned ideas. The teacher originally proposed a conventional explanation concerning differences in grammatical structure. However, given that word choice is largely semantic, the students’ explanation seems more powerful.

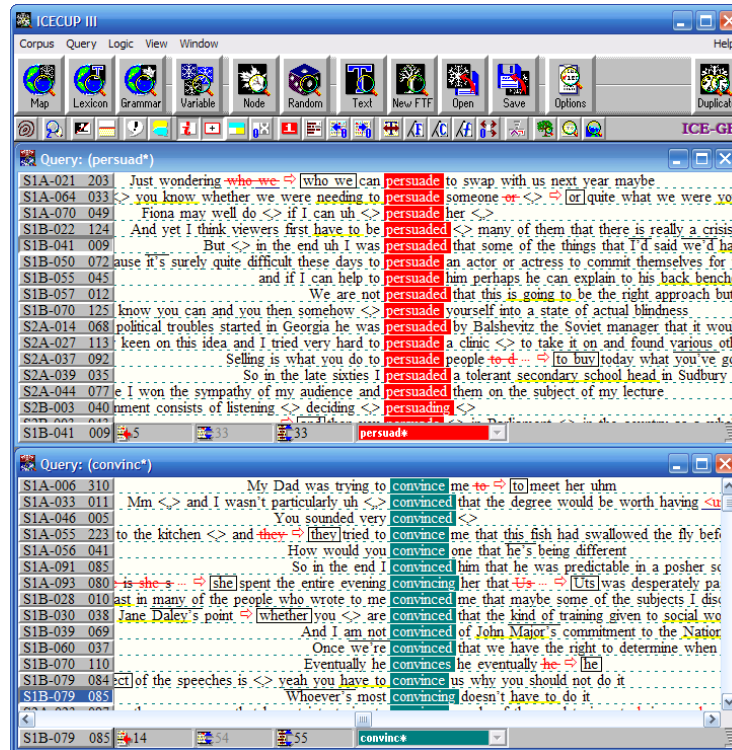


FIGURE 1: Concordance lines of *persuade* vs. *convince* in the ICE-GB corpus, replicating the method of Johns (1991)

Kaltenböck and Mehlmauer-Larcher (2005) argue that corpus data are best thought of as decontextualised (albeit genuine) evidence of performance. In order to comprehend the text, each language learner is obliged to recontextualise the text by imagining themselves in the situation of the speaker or writer. This perspective places an emphasis on offering the learner more context, from full sentences to whole passages. It implies further that Johns and King's pedagogical strategy can only work if students are competent performers or cogent observers of the text.

Johns and King's perspective of 'student as researcher', and the premise that it was more effective to get students to look at corpus data, rather than learn formal rules, mirrors a particular methodological perspective in corpus linguistics. This 'corpus-driven' perspective is most associated with Sinclair (1992), also based in Birmingham University during the same period.

Birmingham's rich tradition of applied corpus linguistics in English is ongoing. Their CLiC project (Mahlberg *et al.* 2016), which uses corpus linguistics methods to explore the stylistics of Charles Dickens' prose, may be applied in secondary school English teaching.

Sinclair's starting point is that explicit knowledge handed down from previous generations may be wrong or irrelevant.³ His more general argument (expounded throughout the COBUILD corpus project he led) was that researchers' preconceptions about grammar rules may be wrong, and therefore the appropriate starting point for linguistic study is volumes of plain text, from which researchers must derive rules inductively. As we have seen, Johns' student-as-researcher model reflects a similar perspective in language learning.

Sinclair's approach to corpus linguistics has been challenged by numerous linguists. The majority of corpus linguists today describe themselves as 'corpus-based' rather than 'corpus driven' (exclusively working bottom-up from corpus to theory). See Tognini-Bonelli (2001: 84) for a discussion. Wallis (forthcoming) suggests that the greatest achievement of bottom-up, corpus driven linguistics has been automatic wordclass tagging (where each word is classified by its part of speech and other features). But success creates a paradox. A tagged corpus provides obvious benefits to linguistic research that Sinclair himself could not ignore. A tagged corpus distinguishes, for example, between verb and adjective forms of *convinced*. This success caused Sinclair to modify his position in practice, accepting tagged corpora as an improvement over plain text.⁴

For Johns and King's approach, algorithmic enhancement of text is recognised to be a benefit. Although influenced by Sinclair, their discussion is more pluralistic, recognising that students engage with corpora in an interactive manner supported by a teacher. Explicit taught rules are to be engaged with using the corpus. Corpus results are discussed in class and new hypotheses formed. With advanced tools and larger corpora it becomes possible to test hypotheses in a more rigorous way. Indeed the advent of parsed 'treebank' corpora (see Section 1.5) permits us to construct grammatical queries to exhaustively identify examples.

Whereas Johns and King used a plain-text corpus, Kaltenböck and

³Thus, for language learners, recognising the subtle semantic differences between near-synonyms is part of their acquisition and refinement of vocabulary. The grammatical rule concerned the use of *that-* and *to-* clauses, but this is actually a secondary question after the semantic choice.

⁴Indeed, working with Fred Karlsson, he accepted morphological and syntactic tagging and automatic parsing using a constraint grammar scheme, ENCG (Järvinen 1994).

Mehlmauer-Larcher (2005) consider the use of the parsed ICE-GB corpus to teach second language learners a number of practical grammatical and semantic distinctions, such as to determine the optimal preposition in cases like the following.

- (2) *The building is adjacent the train station.*
- (3) *It is usually a good idea to abide the law.*
- (4) *You should give clear indication your intentions.*
- (5) *He was aghast the violence he witnessed..*

To solve this problem, the student is encouraged to search for the preceding word, e.g. ‘*adjacent* <PREP>’ on the basis that it is the semantics of the preceding word that governs the choice of preposition. This strategy is mostly successful. However low-frequency words such as *aghast* only obtain a single instance, whereas the more frequent *indication* obtains *indication in*, *indication as to*, and *indication from* in addition to *indication of*. How does the student decide between these different outcomes? Can a single instance be relied upon?

How might data-driven learning be applied to the problem of teaching English grammar (rather than semantic word choice, or an author’s style) to first language learners? Can this approach be applied to primary school classes, where children are below the age of 12?

We have already seen that for DDL to be successful, the role played by the teacher in guiding learning and structuring lessons is central. Direction is invariably left to the teacher, both to set up the task (drawing attention to the preceding word) and to help students draw conclusions. What do teachers need to know?

1.2 The knowledge gap

There is a substantial problem in UK English grammar teaching. A key missing element in applying corpora to teaching lies at the interface between students and corpus: *the knowledge of schoolteachers*.

There are two main areas of knowledge about language that teachers need to teach grammar (Aarts *et al.* 2019, Cushing and Aarts 2019). These are

- *subject knowledge*, i.e. knowledge of the conceptual framework (‘meta-language’) and rules of grammar, and
- *pedagogical knowledge*, knowledge of how to teach grammar effectively, based on the first (Giovanelli and Clayton 2016).

British schoolteachers face a particular historical problem (Crystal 2017). Unfortunately, the level of grammar subject knowledge among UK English teachers is not high (e.g. Borg 2006), and in its absence,

pedagogical knowledge is unobtainable. Most English teachers have received limited linguistic training at teacher training college. Hudson and Walmsley (2005: 616) write:

Most younger teachers know very little grammar and are suspicious of explicit grammar teaching. Not surprisingly, therefore, new recruits entering teacher-training courses typically either know very little grammar... or have no confidence in their knowledge, presumably because they have picked it up in an un-systematic way... This situation raises obvious problems for the implementation of the official programme.

Our own experience in working with teachers has reinforced that view. Schoolteachers in the UK have tended not to acquire grammatical knowledge. Giovanelli (2015) identifies several potential explanatory factors. Teacher-training courses in English tend to have few applicants from a language or linguistics background, and rarely is there opportunity on courses to acquire linguistic analysis skills. The tendency of 'English' to be primarily associated with literature has been long-term and is self-sustaining. Indeed, many teachers are actively antagonistic to grammar teaching, for reasons discussed in the next section (see also Paterson 2010, Aarts 2018). Ultimately a lack of knowledge of grammar and the belief that it is unnecessary tend to be mutually reinforcing. Opposition to explicit grammar teaching is not wholly negative. The same teachers express a positive perception of teaching whole books, and of inspiring students to read and write in order to express themselves. It is crucial to recognise the underlying positive motivation towards encouraging child self-expression that lies behind much teacher opposition to grammar in the curriculum. Only by doing this might it be possible to persuade them that explicit grammar knowledge could be beneficial towards this aim.⁵

However, first we must contend with the knowledge gap. In the UK, a generation of schoolteachers are increasingly required to teach English grammar without ever being formally taught the subject. There is, consequently, a substantial demand for high-quality resources.

However, without a single point of reference, teachers say they often rely on internet searches for grammatical information. This inevitably

⁵Giovanelli (2015) interviewed UK 'A' level English Language teachers who had received linguistics training after making the transition from literature to language. The teachers reported that making the transition was demanding. But as a result they became more confident in their ability to teach grammar and re-evaluated their own role as English teachers. Strikingly, they commented that language work had a positive impact on their English teaching methods more generally.

finds disagreement between grammatical frameworks, and these differences inevitably add to their sense of confusion.⁶ Some teachers relied on 1960s grammar books or *English as a Foreign Language* (EFL) text books.

1.3 Prescriptivism in grammar teaching

The problem of limited subject knowledge among a generation of teachers is compounded by a second factor. Grammar teaching in the UK is widely perceived by teachers and parents as being concerned with ‘good’ English and prescriptive approaches to grammar (Cameron 1995).

There is a long tradition of prescriptivism in English language teaching (Paterson 2010). Whereas prescriptivism in second language learning may be relatively frictionless, the same is not true for first language students. They are required to unlearn ‘bad habits’ at best. At worst they are presented with negative value judgements about their own language, making them feel inferior and criticised.

In a prescriptivist framework, ‘grammar’ is frequently conflated with style and register. It tends to be narrowly focused on what not to do (don’t split the infinitive, don’t use the agentless passive, etc.). In addition, this approach tends towards a selective coverage of ‘grammar’ itself, focusing on ‘common errors’ rather than comprehensive understanding (Myhill 2005).⁷

Although grammar is not officially presented in a prescriptive framework, official documents present grammar alongside ‘vocabulary’ as an aspect of language competence:

The quality and variety of language that pupils hear and speak are vital for developing their vocabulary and grammar and their understanding for reading and writing... Good comprehension draws from linguistic knowledge (in particular of vocabulary and grammar) and on knowledge of the world. (Department for Education 2013)

A ‘competence’ perspective is ambiguous. It might be read as justifying empowering children with skills, or labelling students as ‘incompetent’. In practice, an emphasis on testing and standards tends toward the second interpretation.

An unresolved issue concerns *which* linguistic knowledge is empow-

⁶See Aarts’s blog <https://grammariansm.wordpress.com/2015/05/01/grammar-terminology-confusion>

⁷Even today, English grammar struggles to overcome its association with prescriptivism. As internet advertising encourages, “Correct All Grammar Errors And Enhance Your Writing. Get Grammarly Now!”

ering, and *which* knowledge might be largely irrelevant. Thus Andrews (2009) argues, with some justification, that there is little evidence that merely teaching explicit grammatical meta-language improves student writing ability. Wyse and Torgerson (2017) reach a similar conclusion. But Andrews also notes there is some evidence that teaching grammar in a task-oriented manner is beneficial. However, much more research is needed on this topic. We return to this question in the conclusion.

Formal grammar teaching in English was withdrawn from the UK school curriculum in the mid-1960s. Teachers and parents developed an opposition to prescriptivism in favour of child-centred learning, linguistic diversity and expression. Part of that critique concerned a perception of bias by social class (selective public-sector schools were even called ‘grammar schools’) and race. The result was that to a substantial extent, grammar simply disappeared from the English language curriculum. Hudson and Walmsley (2005) also observe a link between the lack of grammar teaching in schools and the absence of grammar teaching research in universities.

Following the Education Reform Act of 1988, the UK government brought back explicit grammar teaching in schools as part of their new National Curriculum. Grammar was conceived of as a ‘writing skill’, like spelling and vocabulary. Grammar teaching was introduced on a piecemeal basis, but grammar knowledge was not tested directly. Grammar first became a formal requirement in 2000 under the *National Literacy Strategy* (Department for Education and Employment 2000),⁸ and it was another fifteen years before it was codified further.

In 2011, the then UK Education Minister Michael Gove announced that GCSE examinations in all subjects would be penalised by as many as ten percentage points for poor spelling, punctuation and grammar.⁹ ‘Grammar’ was explicitly prescriptive, and errors might cost a student their grades. Then in 2013, he announced further reforms ‘to compete with rigorous overseas education’. One result was the introduction of a test in grammar, punctuation and spelling (the ‘GPS’ test, see Section 2.5) as one of a series of ‘standard assessment tests’ in the final year of primary education. The GPS test asks students to recognise grammatical concepts in sentences, suggest missing words to complete sentences or correct them by substituting a word or punctuation. A new obliga-

⁸The first innovation of the National Literacy Strategy in primary English teaching was the widespread adoption of synthetic phonics. This clearly demonstrates the potential for teachers to adopt new methods if they can be shown to improve levels of literacy.

⁹See <https://www.telegraph.co.uk/education/educationnews/8721387/GCSEs-sloppy-grammar-will-cost-pupils-one-in-10-marks.html>

tory National Curriculum in English was published in 2014, alongside a ‘non-statutory’ (optional) grammar glossary. This set out the types of grammar knowledge to be attained by a particular age. See below.

Is there scope within this new framework for a descriptive and empowering approach to grammar teaching? How might we equip students to express themselves more ably and better appreciate the writing of others?

Once English grammar is motivated as an enabler of language performance, rather than as a set of negative prescriptions, there are three consequences. First, this perspective need not be in opposition to the sensible pedagogical perspective of teachers who focus on ‘child-centred learning’ (if not the child, then who else?), self-expression and whole books. Get the child to read first, and analyse and reflect later.

Second, the ability to use real language texts to explore grammatical ideas becomes central. We need corpora, and ideally pre-analysed (parsed) corpora. A focus on corpus data offers a chance for teachers to redress this balance, initially by drawing on real examples and ultimately by project work. Rather than being told what should or should not be written or said, the student can be directed towards the data and encouraged to appreciate distinctions and consistency across registers. Which ‘rules’ are broken, and which are not?

Finally, for grammar to be empowering it must be presented as useful knowledge rather than knowledge for its own sake, what we might term a ‘toolkit’ approach, or what Myhill *et al.* (2012) call ‘a repertoire of possibilities’. Just as one would not teach mathematical operators (plus, minus, divide, multiply, etc.) by only learning their definitions, grammatical concepts should be taught as operators to be applied to phrases and sentences.

1.4 The UK National Curriculum

In 2014, the UK Government published a new set of specifications for the National Curriculum in English. This was the first time since the 1970s that a single standard official grammatical scheme for all schools and levels was published.

This was an important event. Students in secondary education (age 12 upwards) are required to obtain a pass in the General Certificate in Secondary Education examination (or ‘GCSE’). As a result, the UK curriculum in English is internationally recognised. Reaching this standard is a minimum requirement for students to progress in education, such as to attend a UK university, or to commence study for professional qualifications.

The previous absence of standardisation created significant problems

in evaluating pupils' subject knowledge. With no standard grammar, students might be taught contradictory concepts. Questions occasionally employed outdated terms, in particular the overused term 'connective'.

Sometimes, the lack of understanding of grammar displayed by the question-setter was exposed by the question itself, an infamous example of which involved a request to name an 'appropriate adverb' in the 2013 GPS test.¹⁰

One important change introduced in this curriculum was the retirement of the old-fashioned extended use of the term 'connective'. This had been defined as covering co-ordinating and subordinating conjunctions (*and* and *although*, respectively), and sentence-initial linking adverbs (*however*, *nonetheless*...). Co-ordination, subordination and sentence-linking are distinct grammatical and pragmatic operations, so placing them conceptually under the same label is patently unhelpful.¹¹

Primary school level children are required to learn subject knowledge, which tends towards a prescriptive interpretation. By contrast, secondary school children are merely 'expected' to learn more advanced grammar, but what this 'advanced grammar' consists of is not specified in comparable detail. Evaluation of this grammatical knowledge is to be performed through marking writing for clarity and accuracy of communication, and the student's ability to use linguistic terminology in text analysis.

At Key Stage 3, the National Curriculum has this to say about grammar (Department for Education 2013):

Pupils should be taught to consolidate and build on their knowledge of grammar and vocabulary through:

- extending and applying the grammatical knowledge set out in English Appendix 2 to the Key Stage 1 and 2 programmes of study to analyse more challenging texts;
- studying the effectiveness and impact of the grammati-

¹⁰Students were asked to complete this test sentence: *The sun shone in the sky*. The marking scheme used by examiners stipulated the following: 'Accept any appropriate adverb, e.g. *brightly*, *beautifully*'. The implication was that *bright* was wrong, despite the fact that *bright* in this context is an adverb (and one with a long history). Poetic answers such as *dutifully* were also deemed 'inappropriate', despite the fact that this is a semantic, rather than a grammatical distinction. See <http://david-crystal.blogspot.co.uk/2013/09/on-not-very-bright-grammar-test.html>

¹¹The Key Stage 2 standard assessment test for grammar (the GPS test) persisted with this loose definition of "connective" until 2015 at least. See <https://grammariansm.wordpress.com/2015/04/28/connectives>

- cal features of the texts they read;
- drawing on new vocabulary and grammatical constructions from their reading and listening, and using these consciously in their writing and speech to achieve particular effects;
- knowing and understanding the differences between spoken and written language, including differences associated with formal and informal registers, and between Standard English and other varieties of English;
- using Standard English confidently in their own writing and speech;
- discussing reading, writing and spoken language with precise and confident use of linguistic and literary terminology.

Students are supposed to ‘extend and apply’ their primary school (‘Key Stage 1 and 2’) knowledge, but there is no detail in the National Curriculum specifications as to what such an extension might involve.

How are students to ‘draw on’ grammatical constructions they observe and ‘use them consciously’, unless they have the means to recognise these constructions and decide how they should be properly applied? It appears that the responsibility for extending the incomplete National Curriculum grammar specifications for primary into secondary education has fallen not just to teachers, but to pupils!

Secondary school students and teachers need an extended glossary and resources aimed at supporting teachers, many of which we had developed in the first project (see Section 2.1). Indeed Cushing (2018) argues that Key Stage 3 secondary school teachers need particular help to support children making the transition from primary to secondary English.

In the absence of such resources, the risk is that grammar teaching simply disappears from the English curriculum and the rating of ‘clarity of exposition’ is left to prescriptive, subjective and informal criteria. Text analysis may use out-dated terminology that the students themselves were not taught at primary school.

The 2014 curriculum was implemented on a rolling basis. In September 2014, Year 5 (age 10) children were introduced to it, in preparation for the GPS test at the end of Year 6. The 2016 GPS test was the first to use the new grammar terms. These pupils then graduated to secondary school in the summer of 2016. This timeframe shaped how we implemented changes to our teaching platform, a process we discuss in Section 2.5 below.

1.5 ICE-GB and ICECUP

In 1998, the Survey of English Usage at UCL published the million-word *British Component of the International Corpus of English* (ICE-GB, Nelson *et al.* 2002). This parsed corpus was published with a specially-designed corpus exploration research platform called the *International Corpus of English Corpus Utility Program* (ICECUP).

The full corpus consists of some 83,000 sentences of written and spoken English produced by educated British adults. Participants had completed secondary education and none were under 18 years old. Speech data was transcribed word-for-word and then analysed in exactly the same way as written text.

‘Sentences’¹² in the corpus are all fully parsed, i.e. each is given a grammatical analysis in the form of a phrase structure tree. Whereas some treebanks have employed a simplified grammatical scheme, the ICE-GB grammar was an extension of an extremely well-known and respected framework (Quirk *et al.* 1985). Familiar concepts of grammar such as ‘noun phrase’ and ‘subject’ are explicitly labelled, and the ICECUP software has an intuitive grammatical search engine. In principle, given some training, a teacher or student could find a complete list of subject clauses in the corpus (say) with a few keystrokes.

A 10,000-word sample corpus plus software, sufficient for many teaching tasks, was freely available from 1998 onwards. However, only a small number of enthusiast teachers, usually those with linguistics training, reported to the Survey that they were able to use it, usually with the most able students.

One aspect of this problem is likely to be information overload. ICECUP simply presents texts or series of sentences by default, hiding grammatical annotation (see Figure 1). But one keystroke opens a complete tree analysis (see Figure 3 below), containing terms that are unlikely to be familiar. ICECUP has a very open-ended, exploratory interface, so users risk being distracted by the options available at any point.

What makes a tool powerful for linguistics research can be a disadvantage in a pedagogical setting. ICECUP demands a high level of teacher knowledge, and teachers needed help navigating a class through using the tool. ‘Complexity’ is not merely a problem of tool design but ultimately it lies in the detail of the grammatical analysis. Ultimately we are faced with the problem of teacher knowledge discussed in Section

¹²In the case of speech, we might more accurately call these ‘putative sentences’ or ‘text units’ as sentences are not punctuated and sentence units had to be determined during the parsing process.

1.2. Simply making the corpus available could not bridge the grammar gap.

The parent project, the *International Corpus of English*, is a long-running international project collecting comparable corpora of speech and writing in over twenty countries¹³ where English is the first or official second language. Research teams in each country collected a one million-word corpus of their own national or regional variety of English. Each team followed a common corpus design and a common annotation scheme, in order to ensure maximum comparability between the components (Nelson 1996). Data from countries in the first phase, including Britain, was collected between 1990 and 1992.

This common corpus design ensured that data was collected across a broad range of text types, including multiple types of speech and writing, in similar proportions. Spoken data included face-to-face conversations in classrooms, broadcast interviews, commentaries and telephone calls. Written data ranged from handwritten student essays to official printed documents, newspaper articles and novels. Although each corpus had only one million words, the idea was to capture English in as many different contexts as possible. Since this was the 1990s, there was no internet material, and no email or Twitter data.

The next stage of the development of ICE-GB involved annotation. Spoken data was transcribed word-for-word, and then annotated structurally. This structural annotation includes identifying slips of tongue and self-correction, noting ambiguous or hard-to-transcribe words, identifying turn-taking and indicating speaker overlap. Written data was annotated in a similar structural manner.

The following is a simple example spoken ‘sentence’ in ICE-GB.

(6) ~~We’ve als~~ we’ve also got tapes of things like <,> [S1A-012 #11]

The first part is corrected by the speaker, so is here represented by the strike-through. The final symbol, ‘<,>’, represents a pause.

The first stage of grammatical annotation is to allocate a part of speech for every word, a process called ‘tagging’. Every word in ICE-GB was given a meaningful word class tag, such as “N(prop, sing)” for a proper singular noun or ‘PRON(refl)’ for a reflexive pronoun. So (6) becomes the following (with a gloss to the right):

¹³Currently there are 26 research teams collecting corpora. See <https://www.ice-corpora.uzh.ch/en.html>

(6')	We	PRON(pres,plu)	personal, plural pronoun
	've	AUX(perf,pres,encl)	perfect, present tense, enclitic auxiliary verb
	als	UNTAG	untagged element
	we	PRON(pers,plu)	
	've	AUX(perf,pres,encl)	
	also	ADV(add)	additive adverb
	got	V(montr,edp)	monotransitive, <i>-ed</i> participle verb
	tapes	N(com,plu)	common, plural noun
	of	PREP(ge)	general preposition
	things	N(com,plu)	
	like	CONNEC(appos)	appositive connective
	<, >	PAUSE(short)	short pause

This level of analysis is very common, because automatic tagging programs can obtain an accurate classification in excess of 95% accuracy. Most corpora available today are automatically tagged for word-class (sometimes termed ‘part of speech’ corpora), but are not manually checked.

From a pedagogical perspective, an unparsed corpus has three obvious disadvantages.

Some simple patterns may be identifiable in such corpora using regular expressions (e.g. ‘N of N’ examples like *tapes of things*). But they cannot be exhaustively identified if intervening constructions are allowable (*tapes of very old things*). However, a key benefit of corpus examples is that by showing permutations they draw students’ attention to the options for modifying phrases in their own compositions. In a tagged corpus, with regular expressions these elaborated examples are not reliably revealed.¹⁴

Second, a central topic in the grammar curriculum is verb complementation. Reliably identifying subjects, objects and complements, and deciding on the transitivity of the verb, are parse-level annotation decisions. An ‘Identify the Subject’ task is not possible unless subjects are properly annotated.

A third reason for working with a parsed corpus does not concern the focus of teaching, but example context. We frequently wish to screen out over-complex examples in teaching. See Section 2.4 below. We may wish to avoid examples with particular structures or elements that are

¹⁴Some tools such as collocation analysis (Stefanowitsch and Gries 2003) or *SketchEngine* (Kilgarriff *et al.* 2014) might be capable of performing a ‘proto-parsing’ process to find expanded examples during search. But since algorithmic results are imperfect, this step displaces the problem of parsing accuracy to the student.

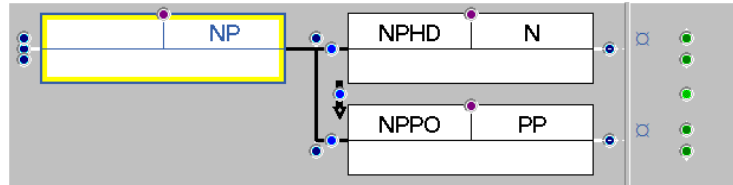


FIGURE 2: An FTF for a noun phrase ('NP') that contains three boxes ('nodes') in succession: the top level NP, a head noun ('NPHD, N') followed immediately by a noun phrase postmodifier in the form of a prepositional phrase ('NPPO, PP'). The tree is drawn from left to right for space reasons.

not found in the curriculum (like clausal subjects), or are not taught to students at a particular level. Or we might wish to exclude sentences with more than one embedded subordinating clause, sentences with more than three clauses, etc.

In ICE-GB this annotation was used as the basis for a full parse analysis. The phrase structure tree for (6) is shown below in Figure 3.

As we have noted, this corpus was fully parsed using a grammatical analysis scheme based on Quirk *et al.* (1985). Parsing a corpus is a much more complex and labour intensive process than tagging. Parsers tend to be more inaccurate than taggers, especially on natural language. Parsing spoken data also involves a degree of interpretation of the stream-of-speech into sentence segments, and the removal of self-correction. Thus for (6), the parser would be given the task of parsing *we've also got tapes of things like*. (Note the ambiguity of the final *like*, especially in the absence of audio.)

Each sentence in a text would be parsed separately, and once parsed, previously-removed elements would be reinstated. The parsed text was then reviewed by linguists and every tree in it corrected manually (Wallis and Nelson 1997). After the corpus was assembled, a second wave of manual review could begin. This process, termed 'transverse' or 'cross-sectional correction', reviewed the corrected corpus construction-by-construction.¹⁵ Whereas a sentence-by-sentence review could be carried out with simple tree editor working through a series of files, performing cross-sectional correction necessitated significant software support in the form of a database platform on which search tools and editors can operate.

ICECUP 3 has proven to be a robust database platform with a set

¹⁵Needless to say, the parsing of ICE-GB was a lengthy process involving an estimated fifteen person-years of effort, on top of the ten person-years spent collecting and transcribing data.

of effective search tools. The same software was used for editing the corpus (in a ‘search and edit’ mode) as for performing research. Thus, as the corpus was corrected it was possible to begin to perform research using it, and simultaneously identify errors in the annotation. Overview tools such as the lexicon can be used to spot coding errors.

ICECUP’s main search engine is a fast and efficient artificial intelligence pattern-matching algorithm for grammatical tree searches and a sophisticated corpus indexing system. This supports a visual grammatical query framework called *Fuzzy Tree Fragments* (FTFs, Nelson *et al.* 2002). An FTF is like a ‘grammatical wild card’. A simple example is given in Figure 2.

In the ICE framework, each node contains three parts. The upper left section contains a grammatical function label (e.g. ‘subject’, ‘direct object’); the upper right section contains the category or word class (‘noun’, ‘adjective’), and the lower half of the node contains various kinds of features that further specify the node.

In an FTF, specifying each section is optional. The user can include or omit information from every part of this tree. For example, the user might say that they would like to require that the noun phrase in Figure 2 be a direct object (‘OD’ in the Quirk notation).

Alternatively, they might alter a link. Swapping the black ‘immediate’ arrow between the two ‘child’ nodes for a white ‘eventual’ one relaxes the requirement that the two nodes follow in order without any intermediate element. Every spot (‘radio button’) in Figure 2 represents a switch that can be selected to change the geometry of the FTF in some way.

Nodes and words can be added or removed. In Figure 2, no words are specified, as indicated by the ‘ \square ’ symbols on the right hand side of the figure. These ‘word slots’ can be replaced with single words or wild cards, or a set of words and wild cards. An option of marking words to be excluded allows us to specify an ‘exclusion set’, such as ‘*ing thing’ meaning “every word ending in *-ing* that is not *thing*”.

A node in an FTF looks like a node in a tree, with the same set of functions, categories and features available to be included or left out. However, more sophisticated options are available, for example to require that a node has the function of either direct object or indirect object (a member of a set, ‘OD, OI’), or that the node itself is specified by a logical expression, e.g. neither an *-ed* participle adjective nor a verb (‘ $\neg(\text{ADJ}(\text{edp}) \vee \text{V})$ ’).

¹⁶Gloss: PU = parsing unit, CL = clause, main = main clause, montr = mono-transitive, NPHD = noun phrase head, PRON = pronoun, pers = personal, plu =

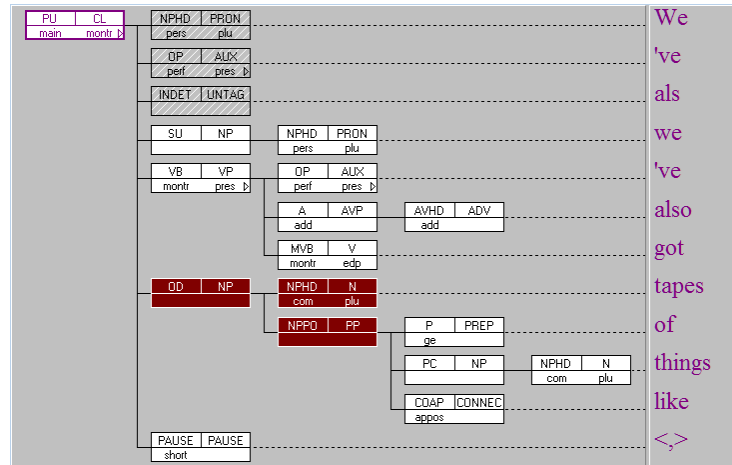


FIGURE 3: Applying the FTF in Figure 2 to the corpus finds sentences like (6), parsed.¹⁶

The essential idea is extremely simple. The FTF looks like a tree, and it can be applied to the corpus to find other trees.

Given the complexity of any complete grammatical system, how can linguists ever hope to learn the grammar and use the tool? ICECUP is designed from the ground up to address the ‘exploration problem’. Users, whether new or experienced, cannot be said to truly know the grammatical scheme until they have seen how it is instantiated in the corpus, but they need to know the scheme in order to search the corpus.

ICECUP addresses this problem by supporting a ‘search-browse-refine’ cycle with an open-ended interface with multiple entry points. A user could build an FTF like Figure 2 ‘top down’ by editing a blank FTF. But that requires them to anticipate that the corpus contains structures like this.

Alternatively, a user could perform a simpler search (e.g. a lexical search or just browse a text), browse the sentences and trees within it, and then select nodes in the tree to construct a new FTF that includes them. The user can find a construction they are interested in and ask

plural, OP = operator, AUX = auxiliary verb, perf = perfect, pres = present tense, INDET = indeterminate, UNTAG = untagged, SU = subject, NP = noun phrase, VB = verbal, VP = verb phrase, A = adverbial, AVP = adverb phrase, add = additive, AVHD = adverb phrase head, ADV = adverb, MVB = main verb, V = verb, edp = *-ed* participle, OD = direct object, NPPO = noun phrase postmodifier, PP = prepositional phrase, P = prepositional, PREP = preposition, ge = general, PC = prepositional complement, COAP = appositive connective, CONNEC = connective, appos = appositive, PAUSE = pause, short = short pause.

ICECUP to obtain a complete list of similar constructions.

For example, the lexical search ‘<N> of’ (any noun followed by the word *of*) will find over 20,000 examples in ICE-GB to pick from, including the tree in Figure 3.

ICECUP contains an ‘FTF creation ‘wizard’ tool that builds new FTFs from corpus trees. Figure 2 can be obtained by selecting the highlighted nodes in Figure 3 and removing information to make it more general (e.g. clearing the function of the NP node).

This means that if a user does not know the grammar they can still immerse themselves in it in the corpus. They can see how a particular sentence was analysed and follow their curiosity to find other trees with similar analyses. They can answer the ‘how’ question: *how is this sentence analysed?* But the user still needs a basic understanding of English grammar to make sense of the analysis found – the ‘why’ question – *why is this sentence analysed like this?*

Teachers need additional help.

2 Teaching English Grammar in Schools

In 2009, the Survey of English Usage began a project aimed at developing grammar teaching resources for UK secondary (high) school English teachers. The starting point of this project, called *Teaching English Grammar in Schools* (Aarts and Smith-Dennis 2018), was the observation that simply giving teachers and students access to a corpus (a ‘DDL’ approach) is insufficient, even with a corpus and tool as sophisticated as ICE-GB and ICECUP. As we have seen, most teachers’ existing knowledge is limited, whether this be knowledge of the subject or of pedagogical strategies for teaching it.

ICECUP contains a complete help file with a glossary of grammatical terms, and the help system can even trigger searches for examples in ICECUP. But the learning curve for teachers with no training in grammar is simply too steep, and their time too limited, for this strategy to be sufficient.

2.1 A platform for secondary school grammar teaching

While we were developing corpus resources for research purposes, government educational policy was moving in the direction of instructing English L1 teachers to teach grammar. The initial impetus for the project was a review of Key Stage 3 English Grammar Teaching (the early years of secondary school) carried out in 2007. This concluded that teaching should make use of formal and informal English in different settings, and that grammar teaching must be driven by real examples. In our view, as far as possible, therefore, resources would draw examples

from ICE-GB.

However, it was agreed from the outset that a requirement to take examples from our corpus would not be a limiting factor. Our principal goal was to develop engaging lesson plans and assessments that teachers actively wished to use. As a ‘knowledge transfer’ project, success of the project was ultimately a question of how many teachers used the platform.

We developed a website platform called *Englicious* (<http://www.englicious.org>) to house a large collection of starter tasks, lesson plans, exercises, assessments, student project outlines, ‘continuous professional development’ (CPD, training materials for teachers) and reference materials. An experienced former schoolteacher on our team with a linguistics background developed resources and evaluated them in his sixth-form (senior) classroom.

The original plan was to access corpus examples directly in real time. In other words, when a teacher displays an exercise or a lesson script on a computer or whiteboard, examples are drawn from the corpus on the spot. The lesson script contains an FTF rather than a hard-wired example. When the page is viewed, the webserver applies the query to the ICECUP database on the server, and merges the output with the webpage, reformatting as necessary. Unlike the DDL approach, where search results are presented without pedagogical mediation, the technology integrates corpus examples with the lesson script.

Such an approach is appealing due to its flexibility. It separates the two layers – the ‘pedagogical layer’ of the lesson script and the ‘content layer’ of corpus examples – for maximum flexibility and re-use. Multiple scripts can access the same examples. The teacher can restrict examples to a particular text type or genre (such as limiting examples to formal letters, novels, conversations, etc.); or the student can press a button to access the surrounding context of an example.

However, the principal problem with an unfiltered direct-access corpus approach is simply that it is very difficult to control which examples are presented in this way. Examples may be inappropriate for a variety of reasons, including age-inappropriateness of vocabulary. Problems of excessive complexity can afflict the target structure (a noun phrase, say) or the complexity of the whole sentence. We refer to the problem of selecting examples for teaching purposes as ‘the selection problem’, and discuss it in Section 2.4 below.

There was a second problem related to drawing examples direct from ICE-GB. This is the fact that the grammar taught in schools, while not dissimilar to the ICE-GB grammar, is different in some important

respects.¹⁷

Leaving aside variations in terminology, the school grammar has limited coverage, whereas the corpus grammar had to be applicable for every sentence. As a result, a randomly-selected example from ICE-GB, *like* example (6), may include grammatical concepts that the teacher is ill-equipped to discuss. Why, for example, is *like* in example (6) not a verb but an ‘appositive connector’ (see Figure 3)? Alternatively, why should *I mean* and *you know* in (7) both be considered ‘formulaic discourse markers’, rather than pronouns and verbs?

(7) *I mean I you know I shared a flat with him for over a year*
[S1A-093 #104]

If example (7) were presented in a ‘spot the verb’ dynamic exercise, the student would almost certainly select the semantically-bleached *mean* and *know*. In the absence of a concept of ‘discourse markers’, a literal interpretation would be valid.

The most immediate practical solution to the selection problem was to address many of the factors by pre-screening examples from the corpus manually. We drew examples from the corpus, considered their appropriateness and created smaller pools of twenty to fifty pre-filtered examples for each assessment task. These pools were then used in dynamic exercises such as the one in Figure 4.

The assessed exercise tasks on the *Englicious* platform fall into a number of different types. These include

- Selecting simple alternatives by clicking on a ‘radio button’, including ‘cloze’ exercises with a different set of alternatives for each example;
- ‘Independent selection’ exercises – any lexical item (a word or part-word) can be individually selected (e.g. to select all nouns in a sentence);
- ‘Sequence selection’ exercises – any sequence of words may be individually selected (e.g. to identify the subject of a clause);
- ‘Keyboard completion’ exercises – a word is completed by typing (e.g. to give the correct spelling of a word);
- ‘Card adjoining’ exercises – virtual cards are dragged together on the screen (e.g. to form a new word or phrase).

¹⁷This grammar was later standardised in the 2014 National Curriculum Specifications (see Section 2.3).

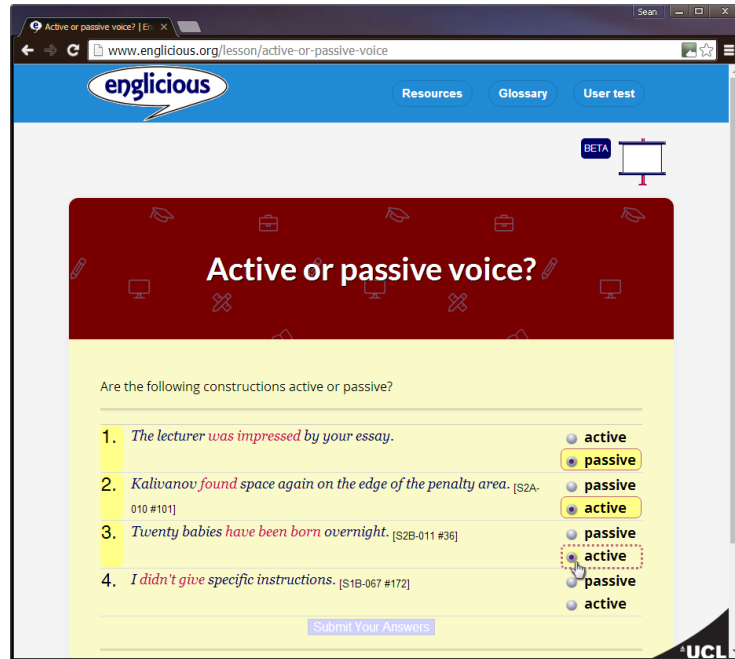


FIGURE 4: A dynamic preview of an *Englicious* assessment task for secondary school pupils. Examples are drawn from ICE-GB. Completing the exercise reveals a score and an explanation of the distinction between active and passive. At the top right is the image of a small projection screen. If this is clicked, the exercise is displayed in a full-screen mode on, e.g. an interactive whiteboard in the classroom. The Resources button pulls up a complete set of searchable resources, and Glossary shows the extended glossary.

Each exercise has a pool of test examples, each one including the right answer and information to generate some feedback hints to the student. Examples are drawn from this pool using a randomised sequencing algorithm. This orders examples in a sequence and then remembers that sequence between page visits. A student returning to the same exercise or pressing a ‘Try again’ button will be sure to see new examples until the sequence is exhausted.

Additional controls are then used to balance examples (see Section 2.4). In the case of ‘radio button’ exercises like Figure 4, the algorithm also attempts to serve up examples to have an approximately equal number of each type, so in Figure 4 two are active and two passive.

With the the exercise types mentioned above it is possible to keep

scores. However, not all *Englicious* activities have a ‘right’ answer. We also developed a wide range of interactive activities to help teachers explore ideas in the classroom.

2.2 Extending the platform to primary schools

The secondary school *Englicious* project was completed in 2012, and was followed by a second project in which we extended the initial project to primary schools (Key Stages 1 and 2, up to age 11). We carried out work in partnership with a North London primary school, and teachers experimented with material in their classes as we developed it.

We saw in Section 1.3 how UK primary schoolteachers have been required to teach English grammar for some years and that the government instituted the ‘Grammar, Punctuation and Spelling test’ (more colloquially termed the ‘SPaG’ test) to be administered at the end of primary education (Aarts forthcoming). The National Curriculum standards (see Section 1.4 and below) were in place from 2014. Unlike secondary school teachers, who are specialised by subject, primary school teachers are almost always generalist teachers, teaching everything from mathematics to poetry.

As we saw in Section 1.4, government curriculum requirements for grammar on primary school teachers are more explicit than the requirements at secondary level. Primary teachers are expected to teach grammar to children from an early age, and this knowledge is evaluated by the grammar test at the end of the child’s final year in the school. The results of these tests can impact on a school’s government league table ranking. Primary school students tend to be less challenging of the information presented than older children, if only because they are younger. However, both primary and secondary teachers express anxiety about questions they cannot answer.

The obvious advantage of extending *Englicious* from secondary to primary level is that it can provide an integrated pathway for English grammar with consistent terminology and resources. It provides the opportunity for secondary school teachers and students to refresh their knowledge by drawing on resources that span the primary/secondary divide.

The development of a teaching platform for primary children required us to develop a software interface that was expressly geared towards young children’s activity-based learning and indeed play. Corpus examples from ICE-GB, whether randomly-selected sentences, clauses or words, would likely not be appropriate simply due to the fact that it contains adult English language unfamiliarity of some of the language

to young children.

Developing for very young children meant designing specific classroom activities, and simplifying other activities to make them usable for younger children. The slideshow mode hides distracting extraneous material. Words are presented in a large font where possible. Tasks were aligned by age and curriculum goals. Interactive activities aimed specifically at younger children included card sorting, where the child placed ‘cards’ containing words in containers, and a novel tool we term the ‘slot machine’ interface.

The strength of the DDL approach for teaching purpose lies in the breadth of linguistic examples used and the illustration of the effect of grammatical rules by induction. Selecting and aligning similar sentences in a concordance draws attention to regularities on either side of the target concept. However, using a concordance for teaching very young children was unlikely to be successful, especially with the adult language ICE-GB corpus.

Instead, we turned the DDL idea of exposing the regularities of sentences by grammatical alignment into a classroom activity. We developed the ‘slot machine’ interface. A noun phrase version of this tool is in Figure 5. Like a corpus concordance, the activity emphasises grammatical regularity by employing a repeating pattern. Unlike a concordance, elements within each slot are randomised and mobile, generating random noun phrases. These are clearly not attested sentences from ‘real language’, but for this activity this is not necessary.

The activity generates classroom discussion, whether the phrases make sense or not. Teachers reported that this exercise in particular created extensive classroom discussion, in two ways.

Firstly, it caused children to see how word order was important in their own writing and thus to see the relevance of grammar rules. In a concordance of attested sentences, grammar rules will rarely be broken, and the student can only induce a rule on the basis of positive examples. They don’t see examples of rule breach. The slot machine allows rules to be broken. Designed as a classroom activity, it generates classroom discussion.

The teacher does not have to be confident in the meta-language in the column titles to show, for instance, that switching the order of columns in Figure 5 generates incoherent phrases. Very young children may not immediately spot where determiner-noun agreement is broken, but older ones have learned this rule and comment on it.

At the same time, the exercise encouraged children to reflect creatively on their own language use, especially when teachers asked the class to think of semantically odd – yet grammatically permissible –

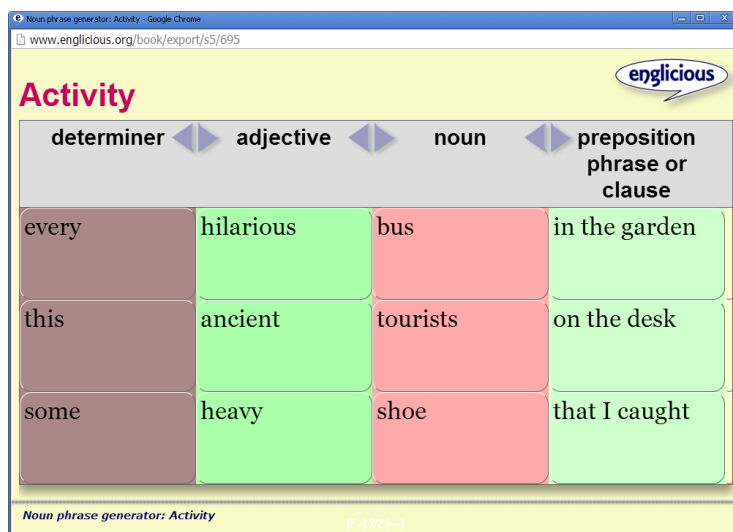


FIGURE 5: A primary school classroom activity, the noun phrase generator ‘slot machine’, in slideshow mode. Columns are independently vertically scrolled or dragged, causing different words and phrases to align horizontally and generate new noun phrases. The tool generates engaging results and draws attention to grammar rules, such as demonstrating determiner-noun agreement, by breaking them. The blue double-triangles indicate that the columns may be moved sideways. In this case (unlike e.g. adverbials) the teacher can show that changing column order is grammatically prevented.

sentences involving ‘mouldy tourists’ or ‘hilarious buses’. Teachers at our North London primary school partner commented in particular how their pupils’ subsequent writing seemed to benefit from this activity.¹⁸

2.3 Implementing the 2014 UK National Curriculum

The publication of the 2014 National Curriculum in English (see Section 1.5) prompted a thorough revision of *Englicious* to align the grammatical framework used throughout the site to the new ‘non statutory’ glossary of terms. This was not just a question of primary resources. We needed to align resources for secondary schools with the framework. Children taught these terms at primary school would soon progress to secondary.

We reproduced the statutory glossary in *Englicious*, and extended

¹⁸See, for example, <http://www.englicious.org/lesson/getting-started/englicious-classroom>

it in several ways. First, existing entries were reproduced and supplemented with longer explanations where these were considered helpful. Second, new entries for terms that teachers and students were likely to come across were added, clarifying that these were not part of the official glossary. In particular, entries were added to make the framework more systematic for secondary school teachers and students. Whereas we did not attempt to create a similar ‘secondary school glossary’ for grammar, we could identify some of the missing concepts that students are likely to come across or need.

This process was not just a question of adjusting the glossary and updating a few pages. It requires a thorough review. For example, the treatment of clauses generated a particular issue for secondary school students. In the National Curriculum, a clause is considered a kind of phrase whose head is a verb. Sentences are subdivided into single clause sentences and multi clause sentences (i.e. combining ‘compound’ and ‘complex’ sentences). Resources for secondary school students that show how students can build more complex sentences by combining clauses had to be rewritten.

As we saw in Section 1.3, secondary school teachers tend to be specialised English literature teachers who are more resistant to a government-imposed requirement to teach English grammar subject knowledge. Resources aimed at secondary teachers need first and foremost to be practically-oriented, to help students to apply grammar learned at primary school to writing assignments. In fact, we needed to support two transitions: the annual transition of cohorts of children from primary to secondary, and the year-by-year implementation of the new curriculum as children progressed (it was only in 2017 that pupils exposed to the 2014 curriculum moved to secondary school). The language syllabus also focuses on spoken language, register, and variation. In particular, resources need to help teachers develop their linguistic subject knowledge.

The idea of consolidating and applying existing knowledge for students brings us back to the use of real natural language, whether from set texts or from corpora. On the one hand this risks identifying gaps in the curriculum and teacher knowledge, on the other it offers an opportunity for project-based work with corpora like ICE-GB.

Implementing the 2014 National Curriculum was not merely a matter of updating the glossary. It required us to undergo a thorough review of *Engliscious*, to align lesson plans, exercises and other content with this new syllabus. The grammatical terminology changed, what’s more, the age level at which material was targeted was more ambitious than before.

2.4 The selection problem

The selection problem (Mehl *et al.* 2016) is a key challenge for deploying corpus resources in a teaching context. It can be summarised as follows.

Can we guarantee that a set of examples drawn from a corpus are appropriate for the teaching task?

The best way to discuss this is through an example. The exercise in Figure 4 has four example sentences drawn from ICE-GB which are used to evaluate a secondary-school child's ability to discriminate between active and passive voice.

One way we could create such an exercise would be to randomly draw examples of active and passive clauses directly from a corpus like ICE-GB. But were we to do this we might find that some clauses were too difficult or contained distracting constructions. We need to develop principles to govern how examples are selected. We have already seen that ICE-GB uses some grammatical concepts that are not documented in the UK National Curriculum.

The selection problem can be considered along two dimensions. The 'single selection problem' refers to constraints on selecting examples that should be applied to each example separately. These can simply be used to exclude or edit examples from selection. The second aspect is the 'group selection problem', where constraints apply to the group of examples collectively: if example 1 is selected, example 2 should be excluded.

There are two main single selection constraints. The first is the *age-appropriateness* of the language used, including its general readability, and relevance to the age, interests and tasks that children are set in school. A random example drawn from a corpus might include adult topics that are of no interest to younger children, or are in a register, such as legal or parliamentary language, far removed from their experience. If the justification for using a corpus is 'natural' language, it is worth reminding ourselves that what is natural for a twelve-year-old may not be the same as for a lawyer or politician.

The second constraint of this kind is the *grammatical complexity* of the entire sentence, including whether or not it contains grammatical features that a young child would find unfamiliar, such as, for example, sentences containing clauses that function as subject. These are not covered by the National Curriculum. We have already seen example sentences from ICE-GB (6, 7) that contain constructions that will be unfamiliar to students and likely to be distracting. Finally, the same sentence might have more than one instance of a particular target concept in it, for example in an exercise that asks students to 'identify the

subject' there may be more than one subject.

In most cases, examples are not presented as single examples to the student or class, but as in Figure 4, they are presented in a group. This creates a 'group selection problem'. How do we ensure that the group of examples is well-chosen when they are considered together? First, we would want to ensure that the examples are *varied* and exemplify different aspects of the test concept. We already discussed how 'radio button' examples may be selected to maximise their variation and avoid repetition from a limited set.

Alongside variation is another principle – *comparability*. It is frequently beneficial to juxtapose related examples. Consider the following pair from an *Engliscious* assessment task that asks students to decide whether a word is an 'adverb or adjective'. In each example students need to decide whether the highlighted word is an adjective or adverb.

- (8) *Computers work best if you kick them.*
 (9) *You've just ruined my best shirt.*

A key lesson of grammar is that the same word can belong to a different word class in different contexts. In these cases *best* is an adverb in (8), but an adjective in (9). Juxtaposing the two examples reminds the student of this fact.

Finally, examples must be *independent*, that is, no part of one example presented at the same time as another should help the student answer another question. For example, in a spelling test where obscured test words are presented in context sentences, no test word should appear in any form in another sentence.

A practical initial solution to the selection problem is simply to avoid directly drawing examples from a corpus (allowing, perhaps, for an exception in the case of the most advanced students). If we pre-select examples from a corpus to populate an example pool for each exercise, we can apply these principles manually. Age-inappropriate examples and complex utterances can thus be edited or excluded.

Ideally, we would wish to draw examples from a corpus directly, without relying on developers or teachers to manually select examples. General principles would need to be implemented programmatically, in the form of screening heuristics and algorithms. Reliable algorithms to rate 'age appropriate' language, readability or complexity are non-trivial to develop, but approximate heuristics are feasible. For example, we can pre-calculate a 'SMOG' readability score for every sentence and use this to filter search results.¹⁹

¹⁹The 'simple measure of gobbledegook' (McLaughlin 1969).

Ideally we would wish to extend our corpus data with more recent spoken and written texts, and more child language data. Including new data widens the pool of available data, but it does not address the selection problem. Indeed, child language data, like all data, needs ‘scaffolding’ (see introduction), through structured teaching materials and careful selection and direction, to make it effective.²⁰

2.5 Teaching the teachers

The same thirty-year period that saw the gradual reintroduction of grammar into the English curriculum in the UK also saw changes in government policy which increased competition between schools and created more pressure to focus teaching around achieving high test scores (termed ‘teaching to the test’). Today, teachers in UK primary and secondary schools find themselves in a situation where their school performance (and potentially their own) is judged by test results.

Test results are thus not only a matter for each child, but can lead to significant consequences for schools. For secondary schools, examination passes at ‘GCSE’ (compulsory education completion level at age 16) and the higher ‘A’ level (university entry level) have become a proxy for educational quality. In primary education, standard assessment test scores are used by government inspectors to rate schools.

These ratings are published and openly discussed in the press, especially local newspapers. High average scores can lead to increased government funding of a school and increased applications for places from parents.

Prior to the publication of the 2014 National Curriculum, primary teachers may have relied on improvised and outdated sources. However once the primary school GPS test drew on a standard meta-language, it gained predictability. It would be possible to argue that there was a documented ‘right answer’ to questions, and the assessment process became more credible. As a result, pressure has increased on primary teachers to learn grammar meta-language in a more consistent and structured manner than before. At the same time they are justifiably uncertain as to whether this kind of teaching actually benefits their charges’ linguistic ability.

As we have seen, schoolteachers need guidance and direction on how

²⁰One of our reviewers suggested using unparsed corpora such as COCA or web corpora as a source. Leaving aside the fact that such corpora, being unparsed, are not reliable sources of clauses, subjects, etc. (see Section 1.5), the text types available are not obviously appropriate. Teaching grammatical *concepts*, as distinct from spotting patterns and discussing them (Section 1.1), does not require vast quantities of data, but it does require reference-quality reliable grammatical analysis, especially when teachers themselves are unsure of definitions.

to teach English grammar in addition to content knowledge of what to teach. It is not sufficient to create lesson plans, classroom assignments and assessment tests for students, just it is inadequate to give teachers access to a corpus and expect them to ‘get on with it.’

In the first place teachers need to acquire explicit linguistic subject knowledge. A major component of *Englicious*, and the component that has grown the most, are teacher-training resources (termed Professional Development resources). These outline the basic definitions of each grammatical term and importantly explain the rationale for these definitions. Resources are then backed up by the extended glossary of grammar terms. Lesson plans for teachers to use in class are subdivided into two parts: an explanation for teachers and an activity to be performed in the classroom.

We made *Englicious* freely available for teachers worldwide with a login, although due to its focus on the UK curriculum, this was primarily of benefit to UK teachers. Alongside the provision of *Englicious* to school teachers, it became clear that the project team needed to engage in a range of supplementary activities.

In part, these activities had to generate income. We needed to sustain *Englicious* as a ‘social enterprise’, to pay for maintenance and upgrade work. The project commenced in 2010, and has been funded through a series of research funding grants, but it will ultimately survive only if stable income to maintain it is earned. But the principal impetus for these activities was the identification of a real need: simply providing resources was not enough.

There remains a clear need for course provision for teachers, alongside the teacher-training ‘professional development’ aspect of the *Englicious* website. At the time of writing, the Survey of English Usage runs two one-day courses for teachers – a subject knowledge course called *English Grammar for Teachers*, and a pedagogical knowledge course *Teaching English Grammar in Context*. Courses are run in the university and are also offered as in-school training. A new collaboration with professional teacher-trainers will also allow them to offer these courses on a franchise basis.

For an enterprise based on the web, a return to print media may appear a retrograde step. Nonetheless, we have produced print materials for classroom and home use: a ‘grammar knowledge organiser’ (i.e. a portable guide to key terms), a set of ‘flashcards’ for testing student knowledge and a set of posters for display.²¹

We also produced a series of short YouTube videos for children and

²¹ See <https://www.ucl.ac.uk/english-usage/projects/grammar-teaching/print>

parents, and a mobile phone app for practising the GPS test.²² Finally, our internet Grammar of English app (Mehl *et al.* 2016) was updated to explain how the grammar in that app related to the UK National Curriculum grammar framework.

3 Conclusions

Englicious is both a proof-of-concept and a strategic pedagogical project. It could not be successful if ‘knowledge transfer’ was conceived of as being one-way, from academics to teachers. By contrast, resources have been developed by working with teachers, in developing resources at both secondary and primary school level and in evaluating them in practice, developing guidance and teacher-training materials alongside lesson plans.

It is a highly ‘political’ project in two senses. First, a generation of schoolteachers need to be persuaded that grammar is relevant to the English curriculum – a reflection of the history of English grammar teaching in schools and the highly politicised rejection of prescriptive grammar in the past. Second, it cannot help but be political in the wider macro-politics of education policy. Unfortunately the wider issue of primary school children being made to sit a test on which their future (and the future of their school) risks overwhelming a proper pedagogical debate about the best way to empower children to express themselves by understanding the rules of sentence formation.

The most convincing justification for teaching grammar in schools must ultimately be that grammar knowledge is useful for children to acquire. The next stage of the Teaching English Grammar in Schools project must concern evaluation. The following question is crucial:

Does teaching explicit grammar knowledge improve children’s literacy skills?

In other words, is the teaching of grammar worthwhile terms of literacy outcomes? As Myhill (2005) observed, with regard to writing, an answer to this apparently simple question has been elusive in the past. In part this was because the teaching of grammar was dominated by prescriptivism. If simply teaching meta-language alone has not improved pupil literacy, either the meta-language is irrelevant (as some teachers argue) or more effort must be made to show teachers and children how to apply it.

There is growing evidence for the hypothesis that teaching clause structure in context can improve secondary students’ writing (Andrews

²²See <https://www.englicious.org/content-type/videos> and https://www.ucl.ac.uk/english-us_age/apps/gpks2

2009). In one of the largest randomised control trials yet undertaken, Myhill *et al.* (2012) found that the contextualised teaching of grammar as a ‘design tool’ offering ‘a repertoire of possibilities’ to the writer had a significant and positive effect on their writing development. This study concerned an intervention that most benefited the more able secondary school students, and questions remain regarding strategies directed at less able children. Perhaps unsurprisingly, this study also highlighted the importance of the linguistic subject knowledge of the school teacher in directing the class.

By contrast, despite the statutory emphasis on primary school education, the evidence that explicit grammar teaching improves primary literacy remains mixed and under-researched. One question we are concerned with is whether some of the innovations in primary play-oriented learning that *Englicious* contains (such as the ‘slot machine’) can improve children’s ability to apply grammar rules in their own work, as well as perhaps motivate the learning of grammatical meta-language.

What is the most appropriate role of real language and corpora? As we developed *Englicious*, ICE-GB has been the principal source of examples for both exemplification and evaluation, but the readability level and relevance of many sentences in the adult British English corpus has meant that without filtering they are only suitable for later years in secondary school. Valuable resources at this school level (Key Stage 4 and ‘A’ level include whole texts of speech and writing, and project work exploring spoken language and register variation. On the other hand as we have seen the development team had to manually select or edit examples for younger children, or simply create their own.

The fact that the corpus is parsed means that it is a source of a vast number of potential example constructions, but the limited definitions of the National Curriculum mean that not all constructions we are likely to find are explained. Were the non-statutory glossary to be developed further at secondary level, some of these issues would likely be resolved.

- To what extent is an intervention based on *Englicious* effective in improving young children’s writing?
- What are the main implications for teacher practice after implementing such an intervention, and, more generally, for the evidence-informed teaching of writing?
- In what ways do the outcomes of the research have implications for the teaching of writing in the National Curriculum for primary schools in England?

Englicious developed out of a series of project proposals, practi-

cal development challenges and engagement with school teachers. The main priority for the project was attempting to address the ‘knowledge gap’, by developing practical resources, driving uptake, and convincing teachers of the benefits of integrating grammar into their English teaching.

A key technical problem for automating and scaling the platform remains the ‘selection problem’. The challenge is to select relevant data from a corpus for teaching purposes. At the greatest remove, the problem is analogous to the task that corpus linguistics researchers face in formulating queries for research purposes, termed the ‘abstraction problem’ (Wallis forthcoming), but it has particular characteristics because pedagogical and research goals are very different. Currently the solution requires a degree of intelligent hand-crafting and practical proof-of-concept work, although we anticipate that more effective heuristic solutions will emerge as we better understand teachers’ and students’ needs.

References

- Aarts, Bas. 2018. Long read: do teachers really hate teaching grammar? <https://www.tes.com/news/school-news/breaking-views/long-read-do-teachers-really-hate-teaching-grammar>. Times Educational Supplement.
- Aarts, Bas. forthcoming. English in the National Curriculum. *Languages, Society and Policy*.
- Aarts, Bas, Ian Cushing, and Richard Hudson. 2019. *How to teach grammar*. Oxford: Oxford University Press.
- Aarts, Bas and Ellen Smith-Dennis. 2018. Using corpora for English language teaching and learning. In D. McIntyre and H. Price, eds., *Applying Linguistics: Language and the Impact agenda*, 163–175. London: Routledge.
- Andrews, Richard. 2009. Teaching sentence-level grammar for writing: the evidence so far. In T. Locke, ed., *Beyond the Grammar Wars*. London: Routledge.
- Borg, Simon. 2006. *Teacher Cognition and Language Education: Research and Practise*. London: Continuum.
- Cameron, Deborah. 1995. *Verbal Hygiene*. London: Routledge.
- Chaiklin, Seth. 2003. The zone of proximal development in Vygotsky’s analysis of learning and instruction. In A. Kozulin, B. Gindis, V. S. Ageyev, and S. M. Miller, eds., *Vygotsky’s Educational Theory in Cultural Context*, 39–64. Cambridge: Cambridge University Press.
- Crystal, David. 2017. English grammar in the UK: a political history. <http://www.davidcrystal.com/?fileid=-5222>. Supplementary material to *Making Sense: the Glamorous Story of English Grammar*.

- Cushing, Ian. 2018. 'Suddenly, I am part of the poem': texts as worlds, reader-response and grammar in teaching poetry. *English in Education* 521:7–19.
- Cushing, Ian and Bas Aarts. 2019. Making grammar meaningful: grammatical subject knowledge and pedagogical principles for grammar teaching. *Teaching English* 19:52–54.
- Department for Education. 2013. National Curriculum in England: English Programmes of Study.
<https://www.gov.uk/government/publications/national-curriculum-in-england-english-programmes-of-study/national-curriculum-in-england-english-programmes-of-study>.
- Department for Education and Employment. 2000. Grammar for Writing (*Key Stage 2*).
<https://webarchive.nationalarchives.gov.uk/20130103071913/>,
<https://www.education.gov.uk/schools/toolsandinitiatives/nationalstrategies>.
- Giovanelli, Marcello. 2015. Becoming an English language teacher: linguistic knowledge, anxieties and the shifting sense of identity. *Language and Education* 29(5):416–429.
- Giovanelli, Marcello and Dan Clayton, eds. 2016. *Knowing About Language: Linguistics and the Secondary English Classroom*. London: Routledge.
- Hudson, Richard and John Walmsley. 2005. The English Patient: English grammar and teaching in the Twentieth Century. *Journal of Linguistics* 41(3):593–622.
- Järvinen, Timo. 1994. Annotating 200 Million Words: The Bank of English Project. <http://www2.lingsoft.fi/doc/engcg/Bank-of-English.html>.
- Johns, Tim. 1986. Micro-concord: a language-learner's research tool. *System* 14(2).
- Johns, Tim. 1991. Should you be persuaded. Two samples of data-driven learning materials. *Classroom Concordancing: ELR Journal* 4:1–16.
- Johns, Tim and Philip King. 1991. Classroom Concordancing. *English Language Research Journal (New Series)* 4.
- Kaltenböck, Gunther and Barbara Mehlmauer-Larcher. 2005. Computer corpora and the language classroom: on the potential and limitations of computer corpora in language teaching. *ReCALL* 17(1):65–84.
- Kilgariff, Adam, Vit Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vit Suchomel. 2014. *The Sketch Engine: Ten Years On*. https://www.sketchengine.co.uk/wp-content/uploads/The_Sketch_Engine_2014.pdf.
- Mahlberg, Michela, Peter Stockwell, Johan de Joode, Catherine Smith, and Matthew Brook O'Donnell. 2016. CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics. *Corpora* 11(3):433–463.
- McLaughlin, Harry. 1969. SMOG Grading - a New Readability Formula. *Journal of Reading* 12(8):639–646.

- Mehl, Seth, Sean Wallis, and Bas Aarts. 2016. Language learning at your fingertips: deploying corpora in mobile teaching apps. In K. Corrigan and A. Mearns, eds., *Creating and Digitizing Language Corpora. Volume 3: Databases for Public Engagement*, 211–239. Basingstoke: Palgrave.
- Myhill, Debra. 2005. Ways of Knowing: Writing with Grammar in Mind. *English Teaching: Practice and Critique* 4(3):77–96.
- Myhill, Debra, Susan Jones, Helen Lines, and Annabel Watson. 2012. Re-thinking grammar: the impact of embedded grammar teaching on students' writing and students' metalinguistic understanding. *Research Papers in Education* 27:139–166.
- Nelson, Gerald. 1996. The design of the corpus. In S. Greenbaum, ed., *Comparing English Worldwide: The International Corpus of English*, 27–35. Oxford: Clarendon.
- Nelson, Gerald, Sean Wallis, and Bas Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: Benjamins.
- Paterson, Laura Louise. 2010. Grammar and the English National Curriculum. *Language and Education* 24(6):473–484.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Sinclair, John. 1992. The automatic analysis of corpora. In J. Svartvik, ed., *Directions in Corpus Linguistics*, 379–397. Berlin: Mouton de Gruyter.
- Stefanowitsch, Anatol and Stefan Th. Gries. 2003. Collocations: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2):209–243.
- Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.
- Wallis, Sean. forthcoming. Grammar and corpus methodology. In B. Aarts, G. Popova, and J. Bowie, eds., *Oxford Handbook of English Grammar*, 59–83. Oxford: Oxford University Press.
- Wallis, Sean and Gerald Nelson. 1997. Syntactic parsing as a knowledge acquisition problem. In *Proceedings of 10th European Knowledge Acquisition Workshop*, 285–300. Berlin: Springer Verlag.
- Wyse, Dominic and Carole Torgerson. 2017. Experimental trials and 'what works?' in education: the case of grammar for writing. *British Educational Research Journal* 43(6):1019–1047.