[published in: *After Half a Century of Slavonic Natural Language Processing* (Festschrift for Karel Pala) Brno, Czech Republic : Masaryk University. pp. 63-80]

The Linguistic Double Helix: Norms and Exploitations

Patrick Hanks

Institute of Formal and Applied Linguistics, Charles University in Prague

1 Linguistic rules and linguistic data

In this paper I propose an approach to analysing the lexicon of a language that is driven by new kinds of evidence that have become available in the past two decades, primarily corpus data. In the course of developing this approach, much of which was undertaken in 2005-8 at the Faculty of Informatics, Masaryk University, Brno, it has been necessary to develop a new, lexically driven theory of language. This was inevitable because received linguistic theories proved inadequate: they were not up to the job of explaining observable facts about the way words are used to create meanings. In particular, patterns of linguistic behavior are observable in corpus data that cannot be directly accounted for by standard linguistic rules, of the kind that govern compositionality.

There has been much confusion, misunderstanding, and even outright hostility about the relation between data and theory in linguistics, so it is necessary to be precise here. Corpus linguists object to invented examples; theoretical linguists question whether corpus data can reveal facts about language as system. I shall not delve into the theoretical linguists' objections, as these have been dealt with more than adequately elsewhere. A balanced summary can be found in Fillmore (1992). Instead, I will comment on the corpus linguists' objection. Except among a few extremists, this objection is not to the use of intuitions to interpret data, for how else could data be interpreted, other than by consulting intuitions? The objection is to the invention of data. As long ago as 1984, in a paper delivered at a conference on Meaning and Translation in Lodz, Poland (a paper that was eventually published as Hanks 1990), I argued that intuitions are a very poor source of evidence, because introspection tends to focus on less common uses, while the really common uses (e.g. the use of ***take*** with expressions of time, as in *It won't take long* and *It took three years to build*) are buried too deep in the subconscious of native speakers to be readily available for recall.

The argument to be developed here is that the so-called "mainstream" in linguistic theory in the late 20th century was out of focus, due to three factors:

> 1) the goal of explaining all possible well-formed utterances within a single monolithic rule system;
>
> 2) the speculative invention of evidence;

3) neglect of the lexicon and the ways in which people actually use words to make meanings.

Put together, these factors resulted in half a century of concentration on syntactic well-formedness, supported by intuitive judgements about the acceptability of invented sentences. Inventing evidence is a hard habit to break. It is insidious. It starts reasonably enough. If a language teacher wants to explain the importance of word order and prepositional phrases in English, what could be more innocent than making up an ordinary, everyday sentence such as *John asked Mary for a pen*? But when applied to speculation about the boundaries of possible usage, the practice of making up evidence has led to the invention of implausible and even misleading examples such as *The box was in the pen; The horse raced past the barn fell*; and *The gardener watered the flowers flat* (all invented by linguists speculating about possibilities rather than analysing data, and all demonstrably implausible in one way or another).

There are, of course, exceptions to this rather sweeping indictment of late 20th-century linguistic theory, notably the hard-nosed insistence of corpus linguists such as John Sinclair on the importance of collecting texts into corpora for use as evidence and the use of computational tools to analyse collocations and other phenomena statistically. This insistence has led inevitably to a demand for a bottom-up system of lexical rules that is both powerful and flexible.

An old-fashioned view of rules is that a rule is not a rule if it is flexible. But the observable facts of everyday language in texts, in corpora, and on the Internet compel us to the uncomfortable conclusion that linguistic rules are both immensely powerful and immensely flexible. Much of both the power and the flexibility of natural language is derived from the interaction between two systems of rules for using words: a primary system that governs normal, conventional usage and a secondary system that governs the exploitation of normal usage. Both these systems of rules are primarily lexical—i.e. rules for using words, rather than rules for constructing sentences. Of course syntactic rules have a role to play: there is interaction between lexis and syntax, but syntax must take second place. Why?

One reason for putting syntax in second place is presented by Wray (2002), who argues that much ordinary communication consists of familiar, conventional phrases and sentences and that, when uttering and understanding these, speakers and writers do not normally analyse them syntactically. Instead, they utter and understand the phrases (and even sentences) as a whole—ready-made, as it were. Wray shows that such ready-made phrases are far more common and widespread than was previously believed, and that, although they are formulaic, they are not "fixed". Operating on a "slot-and-filler" basis (replacing one word or phrase in a formula with another), language users can and do vary their stock of formulas to meet different requirements without necessarily building up utterances from first principles. People resort to syntactic analysis occasionally, but, according to Wray, on the basis of "needs-only analysis". When we want to say something entirely new, or when we hear or read a puzzling or complex sentence, we have the ability to analyse it syntactically. But the fact that we can do this does not entail that we do do it on all occasions. Life is too short and conversation is too quick and (mostly) too trivial to merit or need fundamental syntactic analysis, except in unusual circumstances.

There are other reasons, too, for putting syntax in second place in the analysis of meaningful language. Let us start with a simplified account of syntactic rules. The power of syntactic systems is undeniable, whether they depend on word order (as in English and French) or on inflection (as in Latin and Czech), or on some combination of the two (as in German). Simple rules can often be more powerful than complicated ones. An extremely simple classification of most of the words of many if not all languages is that some of them are nouns and others are verbs. One hugely powerful rule for making meanings is the SVO rule (subject – verb – object): you can take any two nouns, join them to a verb, and (if you have a lot of nouns and several verbs) make a very large number of sentences in which the verb expresses a relationship between the two nouns. This is a very simple, very powerful rule. The conventional order may be SOV in some languages, or VSO or VOS, and under suitable conditions the order can be varied for rhetorical and other effects. In highly inflected languages such as Czech and Latin, inflections rather than word order determine clause roles of nouns in relation to a verb. That does not matter, as long as a language has some way or other of distinguishing subject from object. The crucial point here is the classification of words into parts of speech—verbs and nouns. Several of the possible sentences that result from the simple rule just mentioned – SVO – are actualized and used for some communicative purpose or other. Others remain no more than theoretical possibilities.

Complications begin to set in for this primitive account of linguistic rules when we observe that, for a sentence to be well formed, function words and/or inflections are needed. In English, in particular, nouns in most circumstances need a determiner, so the clause roles S and O are realized, not just by a noun but by a noun phrase (consisting basically of a noun plus a determiner, possibly with some other words such as adjectives, or even an embedded dependent clause, thrown in). Another complication is that a well-formed sentence does not always need two noun phrases—one noun phrase is sufficient with intransitive verbs. The complexities become exponentially greater as other parts of speech—in particular prepositions and adjectives—are added, each bringing with it its own set of rules. Nevertheless, all of these complexities can be expressed in a single monolithic rule system, even though such a system necessarily consists of a rather large number of rules. Most traditional grammars (including transformational and generative grammars) are monolithic rule systems of this kind.

However, a rule system such as the one just outlined, complex though it may be, does not have the slightest chance of coming anywhere close to descriptive adequacy, i.e. of describing the realities of actual human linguistic behaviour. Great theorists of the past have attempted to deal with this mismatch by idealizing the language system (*langue*, competence) and distinguishing it from the everyday reality (*parol*e, performance). But these idealizations simply won't do. The exceptions to the rules are so numerous, and so obviously well motivated, that they cannot possibly be dismissed as mere 'performance errors'. Something else is going on.

J. R. Firth rejected Saussure's parole/langue distinction (as he would no doubt have dismissed the competence/performance distinction had he lived long enough to encounter it). Instead, he insisted on the close observation of actual linguistic behaviour. It is ironic, therefore, that close inspection of the textual traces of actual linguistic behaviour, looking at words in context within a neo-Firthian framework, compels us to the conclusion that the only satisfactory way of accounting for the observed facts is once again to postulate a duality. In this case, the duality is not between an idealized system and everyday reality, but rather between two interactive systems of rules governing linguistic behavior: rules for norms and rules for exploitations. Without such a theory, perfectly well-formed, meaningful sentences such as 'I hazarded various Stuartesque destinations', 'Her eyelids yawn', 'Always vacuum your moose from the snout up', and 'Never invite two China trips to the same dinner party' – all attested in real data – would either be inexplicable or would require selectional restrictions set so wide that no meaningful study of collocations would be possible and therefore the investigation of meaning in language would be unable to proceed. The only reasonable conclusion is that "selectional restrictions" are not really restrictions at all, but rather preferences, and that preferences are rule-governed, but governed by a different set of rules from the rules that govern normal utterances. These rules yield probabilities, not determinations.

I am not the only person in recent years to have observed that collocations are preferences rather than restrictions. From a bottom-up perspective, it seems obvious. Only top-down theorists think in terms of restrictions. Equally obvious is the fact that people exploit normal usage for rhetorical and other effects. This fact has been observed in a great variety of realizations and discussed by various writers on language and meaning over the past two thousand years, dating back at least to the Roman teacher of rhetoric Quintilian (1st century AD), if not to Aristotle. What is new here is the theoretical status given to exploitations, releasing the theory and rules of well-formed normal usage from the need to account in the same breath for equally well-formed but abnormal usage. Also new—though obvious when you come to think about it—is the finding that the distinction between norm and exploitation is a matter of degree (some utterances are more normal than others). The methodology for determining the extent to which any given utterance is normal depends on statistical measurement of corpus or textual evidence. Two very common secondary rules—ellipsis and semantically anomalous arguments—may be counted as exploitations, although they are not found among the tropes and figures of speech discussed by classical rhetoricians: they are too mundane to count as rhetorical devices.

2. Genesis and summary of the theory

The theory of norms and exploitations (TNE) had its genesis in a marriage between lexicography and corpus linguistics. It is a bottom-up theory, created in response to the general question, how can we account for the ways in which people use words to make meanings? How can we account for the intuition shared by most empirical language analysts—strongly reinforced by observation of corpus data—that there are patterns and regularities lurking just below the surface of everyday usage? How does language really work, at the lexical, semantic, and pragmatic level? What are the general principles that govern word use, and what generalizations can be made about the relationship between word use and word meanings?

TNE proposes that, in natural languages, a set of rules governing the normal, conventional use of words is intertwined with a second-order set of rules governing the ways in which those norms are exploited. As its name suggest, TNE is a theory with two main components, as but unlike many other theories, its two components are not sharply distinguished. Rather, they are poles at opposite ends of a cline. Some norms are more normal than others; some exploitations are more outrageous than others. And in the middle are alternations: lexical alternations, where one word can be substituted for another without change of meaning; syntactic alternations, of the kind described by Levin (1993); and semantic-type alternations, which are a device for selecting a different focus when describing what is basically the same event type (you can talk about *calming someone* or alternatively, with a slightly different focus, about *calming someone's anxiety;* you can talk about *repairing a car* or you can focus on the presupposition and talk about *repairing the damage*).

First and foremost, TNE is a theory of prototypes and preferences, based on extensive analysis of actual traces of linguistic behavior – what people say and what they may be supposed to mean – as recorded in large corpora. Analyzing corpus data is an exercise in syntagmatics. The lexical analyst looks at large quantities of text data in various ways, using a variety of corpus-analytic tools such as a KWIC index (a concordance) and the statistical analyses of the Sketch Engine, and immediately perceives that there are patterns in the way the words are used. More thorough analysis reveals further patterns, hidden below the surface. The whole language is riddled with interconnecting patterns. But as analysis of corpus data proceeds, something very alarming happens: the patterns in a concordance which seemed so obvious and which caught the eye at first glance begin to seem more and more difficult to formalize semantically, as more and more unusual cases are noticed. More and more exceptions show up as the data accumulates.

Different patterns are found at different levels of delicacy: discovering that what people *hazard* is usually *a guess* is a very coarse-grained discovery, easily made given a handful of corpus lines from a general corpus of English. At the other extreme, the discovery that *firing a smile at someone* is also part of a pattern (a conventional metaphor that extends also to clauses such as *She fired a shy glance at him*) is a more fine-grained, delicate discovery. At this level of delicacy, it is hard to know where to draw a line between normal and abnormal usage. More thorough examination of data leads to the conclusion that, usually, there is no line to be drawn—only a broader or narrower grey area. Likewise, it is sometimes hard to know where to draw a dividing line between any two patterns of normal usage. Nevertheless, it is usually easy to identify a few prototypical examples, around which other uses may be grouped. It seems from this that a new, prototype-based approach to linguistic formalisms needs to be developed.

When a word is associated with more than one pattern of normal use, it is usually but not always the case that different patterns activate different meanings. *Hazarding a guess* (= stating a proposition without confidence that it is true) activates a different meaning from *hazarding one's life* or *hazarding one's money at the* *roulette table* (= putting one's money or life at risk in the hope of some good outcome).

On the other hand, *firing a gun* and *firing at a target* have different patterns (syntactic structures) but activate the same basic meaning. The relationship between patterns and meanings is strong, but not straightforward. It takes many forms.

The other major component of TNE arises out of the observation that some uses of words are highly abnormal or unusual and do not fit into a pattern very well at all, and yet there is no reason to believe that they are mistakes. In fact, rather the reverse. Unusual expressions like *vacuuming a moose (from the snout up)* and *urging one's car through a forest* are communicatively effective and memorable precisely because they are unusual and stretch the boundaries of normal, patterned usage.

The principle governing pattern analysis in TNE is **collocation**: grouping collocates together. Different groups of word (lexical sets) have a preference for the company of certain other lexical sets, large or small. The lexical sets so grouped can in turn be mapped, as colligations, onto syntactic structures. Indeed, they must be so mapped in order to enable speakers to utter meaningful sentence at all—though not through any conscious effort on the part of the speaker. The groupings are integral to the system that each speaker has internalized since birth (see Hoey 2005). Thus, meaning is dependent on lexical sets grouped as colligations, both according to their normal contexts and permitting exploitations of normal contexts.

The patterns associated with each word (strictly speaking, each content word) are complex because they do not merely relate to one another syntagmatically and paradigmatically; they also serve as representations of non-lexical cognitive entities, for example of beliefs about the world, of a speaker's subjective emotions, of stored recollections of reactions to past events, sensations, hopes, fears, expectations, and so on. At the same time, this complex mass of private attitudes and beliefs in an individual speaker's brain has to interact somehow with similar but not identical complex masses of private attitudes and beliefs in the brains of other users of the same language, for the whole purpose of language is communication—interaction with others—not merely the expression of private beliefs and sensations. Each content word in a language is like a huge railroad station, with trains departing to and arriving from other words, other cognitive elements, and other speakers. We humans are not merely cognitive beings but also social creatures, and language is the instrument of our sociability. For this reason, the conventional patterns and uses of each content word in a language constitute a more or less complex linguistic gestalt. The gestalt for normal uses of the English verb *sentence* is very simple and straightforward, boiling down to one single pattern – *a judge sentences a convicted criminal to a punishment*. The gestalt for a verb such as scratch or throw is extremely complex, with a wide variety of syntagmatics, meanings, and pragmatic implicatures, which would take many pages to explain. The astonishing fact is that, somehow or other, all native speakers (including people with otherwise limited educational attainment) manage to internalize at least a substantial part of this gestalt for almost all common, everyday words, as well as many less common ones, depending on their particular interests and life circumstances.

The whole picture is further complicated by the necessary introduction of a diachronic perspective. Whether we know it or not, the language we use today is dependent on and shaped by the language of past generations. Most exploitations of norms are lost as soon as uttered, but every now and again one of them catches on and becomes established as a new secondary norm in its own right.

3. Theory and application

What applications can be envisioned for TNE? It is for others to judge how useful the theory is and how or whether they want to make use of it. Here I shall mention just three areas in which I believe that it has some relevance: natural language processing by computer, language teaching, and linguistic theory.

To take the first two of these, we may note that there are, broadly speaking, two main aspects to the practical application of any linguistic theory: productive applications and receptive applications. Productive applications use a theory to understand the creation of linguistic events, and even to create them: for example, to help language learners or computers to generate well-formed and relevant utterances. Receptive applications are designed to facilitate understanding: the computer or the human must understand what is being said in order to respond appropriately. (Appropriate responses include learning—the assimilation or rejection of new information and the formation of new beliefs.) In both cases (production and reception), there is an underlying assumption that a linguistic theory serves as a basis for creating an inventory of linguistic items. The particular inventory predicted by TNE is an inventory of patterns associated with each content word in the language.

TNE can be seen as a tool for creating tools, for lexicographic resources themselves are tools for use in applications such as language learning by people, language understanding by people and machines, and language processing by computer. But of course the theory is a theory of language, not of tool building, so if it has any value, that value must be applicable directly in activities such as natural language processing by computer, language teaching, and literary studies (what Jakobson called ‘poetics’). In all of these fields, it seems likely that applying a theory that focuses on normal language use, that has a special role for creativity, that refuses to be distracted by speculation about remote possibilities, and that insists on close empirical analysis of data has potential applications that will yield rich dividends. In addition to the applications just mentioned, the theory probably has something to contribute to cognitive science and our understanding of the way the human mind works, but that is not its main focus.

There are many areas in which the relevance of TNE could be discussed—for example grammar and grammatical theory, literary stylistics, cognitive linguistics, translation studies, and many others. In this section, I shall confine myself to sketching out a few comments in three major areas of potential application: computational linguistics, language teaching, and lexicography.

3.1 The Semantic Web, NLP, and AI

One more motive for exploring new approaches to lexical analysis and develop a lexically based theory of language with a focus on normal usage is the current buzz of excitement surrounding the infinite possibilities of the so-called Semantic Web. The dream of the Semantic Web (see Berners-Lee et al. 2001) is to “enable computers to manipulate data meaningfully”. Up till now (2009) work on realizing the dream has done little more than propose the construction of ontologies (see Chapter 1 above, section 1.8) and the addition of tags to documents and elements of documents, to structure them and improve their machine-tractability, without engaging with their semantic contents. It is a fair prediction that, sooner or later, if it is going to fulfil the dream of enabling computers to “manipulate data meaningfully”, the Semantic Web will have to engage with natural language in all its messy imprecision. The stated aim of manipulating data meaningfully could, of course, be taken in any of a number of ways, depending on what counts as data. Current assumptions in the SW industry are that “data” means tagged data, and “manipulating data meaningfully” means little more than matching patterns and processing tags. However, Berners-Lee et al. (2001) also said:

> Web technology must not discriminate between the scribbled draft and the polished performance.

This would seem to be a clear indication that the original vision, though vague, included being able to process the meaning and implicatures of free text. But how is this to be done?

The protagonists of the Semantic Web drama, who are nothing if not canny, have avoided getting embroiled in the messy imprecision that underlies the ordinary—and sometimes precise—use of words in ordinary language. The Semantic Web’s RDF (Resource Description Framework) confines itself to using and processing HTML tags and strictly defined technical terms. Insofar as ordinary words are assigned strict definitions for computational processing, scientific research, rules of games, and other purposes, they acquire the status of technical terms and are no longer part of ordinary language. Technical terms are essential for many logical, technological, and computational applications, but they cannot be used to say new and unusual things or to grapple with phenomena that have previously lain outside the scope of the imagination of the definer, which is one of the most important things that can be done with ordinary language. The notion that the words of human language could all be rigorously defined was a dream that tantalized great thinkers of the European Enlightenment, in particular Wilkins and Leibniz. Their disgust with the fuzziness of word meaning was shared by philosophers up to Russell, and was indeed a factor in the latter’s breach with Wittgenstein, who invited us to “look and see” what is actually going on when people use words to make meanings. But the vagueness and indeterminacy that Wilkins, Leibniz, and Russell (among others) considered to be faults in natural language may now be seen as essential design features. Sooner or later, the Semantic Web must engage with this design feature, the imprecision of natural language, if it is to fulfil its own dream. The theoretical approach to the lexicon outlined in this book lays part of the foundation for such an engagement. This dream cannot be fulfilled without an inventory of the content words of a language, describing their normal patterns of usage and implicatures of each pattern, together with sets of rules that govern exploitations and alternations and procedures for matching usage in free text preferentially onto the patterns. This is a remote dream at the time of writing, but it does not seem unachievable in principle.

If this dream is to be fulfilled, it seems important to proceed methodically, step by step (in the right direction, of course) and to abandon—ort at least suspend for the time being—the yearning for instant solution by a magic bullet that is so typical of computational linguists.

Semantic Web research is not the only computational application that stands to benefit from a long, hard look at how the lexicon actually works. In recent years, ‘knowledge-poor’ statistical methods in computational linguistics have achieved remarkably—some would say astonishingly—good results, at a coarse-grained level, in applications such as machine translation, message understanding, information retrieval, and idiomatic text generation. At the same time, refined methods based on syntactic and valency theory have yielded largely disappointing results. The same is true of methods based on using machine-readable versions of dictionaries that were designed for human beings. However, statistical methods, in principle, have a ceiling, while deterministic methods point to the need for a reappraisal of the relationship between lexis and syntax. TNE points a possible way forward, toward an integration of statistical and deterministic methods. Some procedures in computational linguistics and artificial intelligence—‘knowledge-rich’ approaches—still lean heavily on linguistic theories that are not empirically well-founded and lexical resources that are based more on speculation and intuition than analysis. Whether it acknowledges it or not, the computational linguistics community will continue to encounter fundamental difficulties, at least insofar as the serious analysis of meaning is concerned, until it starts to build and use lexical resources that are based on empirical analysis of actual use of language. Any strategy other than bypassing meaning entirely (which is what statistical methods do) will need a theoretical approach of the kind outlined by TNE.

A	B	C
In-	-script-	-ion
Pre-	-vent-	-ive
De-	-vict-	-ible
Con-	-duc-	
Pro-		

Figure 1: predictability of meaning based on Latinate morphemes.

Consider predictions of meaning in English and French words based on Latin morphology, as in Figure 1.   Most Romance languages contain sets of words that consist of one item taken from column A and one from column B and one from column C. However, the system is not totally productive or predictable. There can be gaps here and there, e.g. there is no word **deviction*, although, if someone chose to invent such a word, its meaning should be to some extent predictable on the basis of this table. Moreover, the meaning of lexical items can be non-compositional. There is nothing in Latin morphology to explain why ***prescription*** has something to do with doctors and drugs in English, rather than writing something in advance. And this non-compositional meaning of collocated morphemes may or may not carry over to some other Romance language.

Much the same applies to rules and phraseology. A rather trivial but telling anecdote seems relevant. Some years ago I was involved with a software company doing, among other things, information retrieval. When asked to retrieve information about "nursing mothers", the search engine retrieved vast quantities of information about care homes for the elderly.[1] This was, of course, wrong. Dictionaries do not say so, but in English the specific meaning of the collocation ***nursing mother*** is non-compositional. It means a mother who has recently had a baby and is in the phase of feeding it with milk from her breasts (rather than from a bottle). This expression yields 19 hits in BNC. Technically, syntactic analysis suggests that it could mean something else. In actuality, it does not.

A lexical resource built on the principles of TNE would show the specific normal meaning, 'mother who is breast-feeding', of this collocation and would treat it as a single lexical item, contrasting it with the normal, compositional meanings of the verb ***nurse***. This verb normally means 'tend (a sick or injured person)'. It also means 'harbour (bad feelings): *nurse a grievance, nurse a grudge*'. The sense 'feed (a baby) at the breast' is in almost all current dictionaries, but in actual usage this is not really compositional, for it is vanishingly rare (only 3 hits in BNC). What's more, although *child* is a close synonym of *baby*, the expression *nursing a child* is not used as a synonym for breast-feeding. If a mother is nursing her child, the child is sick or injured. This is not a matter of certainty based on syntactic analysis (which gets it wrong); it is a matter of statistical probability based on collocational analysis.

Now multiply this anecdote by some number in the hundreds of thousands, and you will have some idea of the number of semantic traps that lurk in waiting for ostrich-like computational linguists. Clever algorithms solve many problems, but in matters of the relationship between word use and word meaning, clever algorithms create more problems than they solve.

3.2 Language learning, language teaching, and the lexicon

In broad brushstrokes, the next most important area of applied linguistics in terms of money spent and number of people affected, after applications in computational linguistics and artificial intelligence, is language learning. Literally hundreds of millions of people at any given time are currently learning one or more foreign languages. Language teaching is big business, world wide. Some learners are very proficient and seem to be able to pick up other languages with apparent ease, regardless of the teaching methods are used. Others struggle mightily. But even the

[1] It should be pointed out that this particular search engine application was aiming to "break the tyranny of text matching" (in the words of Greg Notess).

proficient ones welcome well-organized help, while badly organized help can add to the struggles of the less proficient. Moreover, it seems that different learning strategies suit certain individual learners better than others. A few gifted individuals respond well to an emphasis on formal grammar, which used to be fashionable; others respond better to an emphasis on analogy and ‘communicative competence’. Even apparently irrelevant factors such as the student’s personal goals (short-term and long-term), the personality of the teacher, the commercial strength of different language communities, the vibrancy of different cultures, and even the beauty of the countryside, can play a part in motivating learners. TNE cannot, of course, help with any of the motivating factors just mentioned, but it does have a contribution to make in helping teachers, syllabus designers, course-book writers, lexicographers, and learners themselves to get the lexicon in perspective, make an organized selection, and to give a high priority in their teaching and learning to the most normal patterns of usage associated with particular words. In other words, it can help with a focus on lexical relevance.

This idea is not new, of course. It is what A. S. Hornby and his colleagues (Gatenby and Wakefield) tried to do in their *Idiomatic and Syntactic Dictionary* (ISED; 1942)—a remarkable work, subsequently re-published as the *Oxford Advanced Learners’ Dictionary* and greatly inflated in its second and subsequent editions. For a fuller discussion, see Hanks (2008).

Language teachers have a problem with the lexicon. There is simply too much of it. Learning phonology and syntax in a classroom environment can yield valuable generalizations for learners comparatively rapidly, applicable to vast swathes of language. But as far as the lexicon is concerned, what can be taught? It is harder to make a case for “getting the words in” (Bolinger 1971) than for teaching syntax. Isn’t the lexicon just a vast list of “basic irregularities”, with no predictability, “an appendix of the grammar”, as Bloomfield (1933) famously remarked? If so, getting the words in can, indeed must, be left to happenstance.

Even if we agree with Bolinger that the words must be ‘got in’, it remains to be decided precisely what should be got in. Any language learner is faced with the daunting task of learning how to use many thousands of words in a language in order to be able to make meanings and even more if they are to understand what is said. As this book has shown, many words constitute daunting challenges of complexity in themselves.

Against this is the obvious necessity to learn at least some of the meaning potential and usage patterns of at least some words in order to be able to use a language at all. Slightly less obvious is the impossibility, even at a theoretical level, of learning *all* the words of a language. Selectivity is essential. Even less obvious, at first glance, is the impossibility of learning all possible uses of a given word. To those who use them, words seem so simple, so obvious. Surely they must be constrained by clear-cut finite boundaries of meaning and usage? But the awful fact is that such boundaries are not clearly defined. They are fuzzy and complex, and the full power of a word as a linguistic gestalt is sometimes of awesome complexity, as was demonstrated in Chapter 12. These, in reality, are among the problems facing the unfortunate learners of a language. They not only have the immense problem of productive usage—generating idiomatically well-formed utterances in a language whose conventions are different—often subtly different, full of traps—from those of their native language; they also have to prepare themselves for receptive usage. And on the receptive side, learners never know quite what will be thrown at them. Who knows what a native speaker is going to say next? The argument of TNE is that this latter point is true, but nevertheless it is possible to predict probabilities, set up defaults, and focus attention on interpreting normal phraseology.

It has become fashionable in some places to provide learners with corpus-access tools such as WordSmith or the Sketch Engine and let them loose on a large corpus without further ado. The motivation for this practice is highly commendable: bringing learners face to face with the realities of actual usage and engaging them collaboratively in the process of learning how words are used and in solving their language problems. In every classroom I have ever visited where this is done (even—or perhaps, especially—if done chaotically), the excitement of engaging with real data and trying to solve real problems is palpable. However, without good principles of selection and organization of data, it can lead to disappointment and even confusion. Guidance is needed on such matters as what to expect from raw corpus data, how to select and sort corpus data, and how to deal with the unexpected. TNE offers a theoretical basis for developing this kind of guidance. In short, learners looking at raw data need not only to be encouraged to inspect the data thoroughly and look for patterns, but also to be informed about principles for interpreting data, to be prepared for the complexities of ordinary usage, to expect exceptions to the patterns, and to be shown how to make effective hypotheses about what the various patterns mean.

Recently, there has been a revival of interest in the lexical approach to language teaching. Pioneers of the "lexical approach" to language teaching were Sinclair, Willis, and Lewis. Following Sinclair (1988), Willis (1990) proposed a 'lexical syllabus' for language learning. Sinclair and Willis were writing in the very earliest days of corpus linguistics, when a corpus 18 million tokens was regard as large and before the full enormity of the challenge posed by corpus data to established theories had even begun to be recognized. Willis's proposal was implemented as the Cobuild English Course. It must be acknowledge that, despite its innovative approach, the Cobuild English Course was not a huge success. Why was this? No doubt part of the reason was bad marketing and off-putting presentation by the publisher: a web search reveals comments on such things as "cognitive overload" (http://www.usingenglish.com/forum/) and a general sense that the pages are unpleasantly cluttered. Such problems could be easily fixed. More germane reasons may have included the non-existence of a systematic body of research into what counts as a pattern and the absence of reliable information about the relative frequency of different patterns and senses. Willis's approach to a lexical syllabus was also hampered by the absence of a thoroughly worked-out theoretical distinction between patterns in which a word can participate and patterns in which it normally participates.

In the same stream, Lewis (1993) argued that "language consists of grammaticalized lexis, not lexicalized grammar" and that the interests of language learners are seriously impaired by excessive concentration on teaching grammar rather than lexis. Lewis's lexical approach concentrates on developing learners' proficiency with lexis—i.e. words and syntagmatics—and 'chunks' of formulaic language of the kind that was subsequently to be discussed more fully in Wray (2002).

It seems a matter of obvious common sense that lexical research should contribute to syllabus design. Words and patterns of word use are far too important to be left to happenstance or the whims of individual teachers. For almost all groups of learners, it is absurd to give a high priority to teaching such terms such as ***umbrella, overcoat, hat,*** and ***cloakroom attendant***. But then, many words that deserve a high priority in a lexical syllabus, e.g. ***need, search, hope, look, find,*** are semantically complex: not only the words themselves but also the most normal uses of such words needs to be prioritized. When we examine them, the issues regarding the lexical contribution to syllabus design turn out to be rather complex. At least the following points must be taken into account:

- Integration with other approaches to syllabus design

- The distinction between function words and content words: function words should, perhaps, be considered as part of a grammatical component of a syllabus, while only content words are organized into the lexical component

- The role of pro-forms: learners have a higher-than-normal need for effective use of semantic pro-forms such as ***thing, something, anything,*** and ***do***, to help them fill lexical gaps and achieve fluency

- The relative frequency of different phraseological patterns and senses of polysemous words: selectivity is just as important at this microstructural level as at the macrostructural level of the lexical component

- Pragmatic functions of lexical items such as ***broadly, you know, I think, it seems***.

A lexical syllabus is not a magic bullet. The different interests, goals, and abilities of different individuals and different groups of learners are relevant and need to be taken into account by individual teachers working within a general framework of lexical selection. At a more general level, the best pedagogical approaches to a lexical syllabus have must necessarily understate the rich complexity of the *possible* uses of each word, while in language teaching more generally, prioritization has all too often been left to common sense and happenstance, or to the intuitions of the teacher, which are typically skewed towards boundary cases, for reasons which have been discussed throughout this book. Part of the argument here is that each word in a language has a core set of one or more prototypical uses, which can be discovered only by painstaking lexical analysis. Some prototypical uses are general; others tend to be domain-specific. Each prototypical use is associated with a prototypical meaning. Prototypical uses can be exploited in regular ways. All of these facts can and should be part of the foundations for prioritization of a lexical syllabus, based on relevant corpus evidence. A lexical syllabus goes hand in hand with a grammatical syllabus, and both need to be empirically well founded.

3.3 Electronic lexicography

TNE was sired by lexicography upon corpus linguistics, and it would not be a runner at all if, as its owner and trainer, I did not believe that it could be entered in the Language Theory Stakes as a potential winner[2]. It has, I believe, the potential to inspire new directions in electronic lexicography. The preceding two main sections have both mentioned ‘resources’. Among the new resources that need to be developed for all such applications and no doubt many others, are corpus-driven pattern dictionaries.

3.3.1 Pattern dictionaries

TNE is an essential foundation for a new kind of dictionary which, on the basis of corpus analysis, will report the patterns of usage most associated with each word (strictly speaking, each content word) in a language. The great advantage of such a dictionary is that, for activities such as natural language processing by computer, it enables meanings to be attached to patterns, rather than to the word in isolation. This facilitates pattern matching. Thus, if a word has more than one sense, a pattern will have already identified the conditions under which each sense is activated before any attempt is made to state the meaning and do anything with it. The sense is ‘anchored’ to the normal phraseology with which each sense of each word is associated.

The first such dictionary is already in progress at the time of writing. It is the *Pattern Dictionary of English Verbs* (PDEV: http://nlp.fi.muni.cz/projects/cpa/), an on-line resource, in which each entry consists of four components:

A. The verb lemma together with a list of the phraseological patterns with which it is associated, expressed in terms of argument structure and subargumental cues.

B. The primary implicatures associated with each pattern (roughly equivalent to a dictionary definition, but ‘anchored’ to the arguments in the pattern, rather than floating freely, as dictionary definitions tend to do).

C. A training set of actual uses of each verb illustrating its use in each pattern, taken from the British National Corpus

D. A shallow hierarchical ontology of the semantic types of nouns (see Pustejovsky et al. 2004), populated with a lexical set of nouns to which each argument is related. As far as the data permits, nouns are related to argument patterns of verbs according to their semantic type.

Compiling a pattern dictionary—indeed, compiling any dictionary—is a long, slow process. At the time of writing (June 2009), approximately 10% of PDEV is complete, after four years work. At the current rate of progress, if there is not a substantial injection of funds to build up a professional lexicographic staff, the project will not be completed until 2040, when the author will be 100 years old.

[2] To the uninitiated, it should be explained that this sentence is an extended and somewhat contrived horse-racing metaphor of a peculiarly British kind.

However, one of the great benefits of online publishing and internet access is that such work can be published as work in progress.

PDEV is an exploration of one possibility for practical implementations of TNE. A project with some similarities to PDEV is FrameNet. Both PDEV and FrameNet are pointers to new future roles for lexical analysis. However, there are important differences. Some of them are as follows.

- FrameNet expresses the deep semantics of a number of prototypical situations (frames) associated with different lexical items. PDEV investigates syntagmatic criteria for distinguishing different meanings of polysemous words, in a ‘semantically shallow’ way, using semantic types as a grouping mechanism.
- FrameNet proceeds frame by frame and analyses situations in terms of frame elements. PDEV proceeds word by word and analyses patterns of use of individual words (verbs).
- FrameNet studies differences and similarities of meaning between different words in a frame, with no systematic attention to polysemy. PDEV studies differences of the relationship between usage and meaning for each polysemous verb.
- FrameNet does not analyse corpus data systematically, but goes fishing in corpora for examples in support of hypotheses. PDEV is driven by a systematic analysis of corpus data and provides statistically valid data for the comparative frequency of a verb’s different meanings.
- There is considerable overlap between closely related frames in FrameNet; it does not seem to have clear criteria for distinguishing frames, and does not seem to be aiming at an inventory of all possible (or all normal) frames. The number of frames seems to be open-ended. This perhaps relates to the impossibility of postulating that the world is organized into neat and finite hierarchies of semantic frames.
- PDEV attempts to group observed lexical sets in a hierarchical ontology, according to a) their shared co-occurrence, and b) their shared semantic types. This raises interesting theoretical and practical issues, which cannot be discussed here; they could be the subject of a whole separate book.
- FrameNet does not seem to have any criterion for completeness. Exploration of FrameNet’s frames reveals that many of the lexical items that ought to be members of a particular frame have been missed. Of course, once this is noticed, case by case, it is easy to bring additional lexical items into a frame, but as things stand (June 2009), there is no way of telling whether either a frame or a lexical item has been fully analysed.

3.3.2 Historical pattern dictionaries

Literary scholars need to ask, not only in what respects does the phraseology used by a great writer of the past differ from the normal, conventional phraseology of present-day English, but also in what respects it differs from the normal, conventional phraseology of the language of his or her own time. Unfortunately, historical records of spoken language are non-existent before the 20th century invention of recording devices. However, for many periods in English, among other languages, written records of ordinary language survive in sufficient quantities to make a pattern dictionary of the language of that period a theoretical possibility. Moreover, the writings of great writers themselves are not ruled out as evidence for patterns, insofar as their usage overlaps with the usage of other writers of the same period, for proof of pattern depends on extrapolation from many sources, including sources which may include (elsewhere in them) idiosyncrasies of usage. Translating the theoretical possibility of a historical pattern dictionary into a practical reality would, of course, depend on the usual necessary combination of scholarly interest and funding.

The masses of evidence collected and analyzed over the past century and a half by the great historical dictionaries such as OED would be a valuable resource for a historical pattern dictionary of this kind, but it needs a stronger theoretical foundation. In principle, the OED evidence, which is substantial, could be reanalyzed to give an account of the normal, shared phraseological patterns in use at any given period in the history of the language. In practice, it would be desirable to supplement the OED citation evidence for word use in each period with a corpus of whole texts of the same period, especially a corpus containing informal, non-literary texts (bulletins, broadsheets, journals, private letters, and suchlike). This is desirable because, inevitably and quite properly, OED has a literary bias and because it is based largely on citations collected by human citation readers, not on corpus analysis. As Murray (1878) noted, human citation readers have a natural tendency to select citations for rare words and unusual uses and to overlook common, everyday, familiar words and uses. It would be an odd citation reader who copied out all the uses of the lemmas ***give*** or ***take*** in a text being read for citations. But it is precisely the common, everyday uses of words such as ***give*** and ***take*** that a pattern dictionary concerns itself with and for which corpus technology can provide evidence 'at the press of a button'. A further issue concerns the comparative frequency of patterns. A citation can prove the existence of a word, phrase, or sense at a given period, but only a corpus (ideally, a balanced corpus of 'representative' texts) of the same period can give an approximate idea of the relative frequency of different patterns of use of the same word.

4. The broader picture

TNE is closer to Tomasello's account of human cognition (1999, 2003) than to Leibnizian primitives or Chomsky's predicate logic. In a series of studies, Tomasello and his colleagues compared the developmental behavior of human children with that of other primates (chimpanzees, etc.). On this basis, he argues that what distinguishes humans from apes is the ability to recognize other members of the species (conspecifics) as intention-governed individuals. This makes possible shared purposeful actions (cooperative behavior) and prediction of the likely actions and reactions of others. In other words, his conclusion that what distinguishes humans is the ability to put oneself in other's shoes. Language plays a crucial role in this ability. It is, therefore, a biological and cultural phenomenon rather than a mathematical one. He says:

> The understanding of conspecifics as intentional beings like the self is a uniquely human cognitive competency that accounts, either directly on its own or indirectly through cultural processes, for many of the unique features of human cognition. —Tomasello (1999: p. 56).

Tomasello argues that there simply has not been enough time, in evolutionary terms, for these unique features to have developed by genetic evolution. There must be another explanation – and there is, namely cultural transmission.

When an intelligent ape or other mammal makes an important discovery—for example, how to use a stick as a tool—the discovery is useful to that individual, it may be remembered and repeated by that individual. It may even be imitated by other members of the species in the same clan, pack, or social group. However, indivudals of nonhuman species have no means of sharing, recording, and transmitting their discoveries, so that sooner or later each dscovery is lost.[3] When a human makes a discovery, on the other hand, it is (or can be) disseminated throughout the community, not lost, because humans have a mechanism for sharing and storing the knowledge gained. This mechanism is language. It operates what Tomasello calls "the ratchet effect': faithful dissemination and storage of knowledge acts as a rachet, preventing backward slippage that would cause knowledge to be lost. This has enabled *Homo sapiens* to evolve at an astonishing speed compared with the genetically bound evolution of other species.

Thus, human linguistic behavior is cooperative social behavior. It involves, among other things, the sharing of knowledge. The relevance of all this to TNE lies in the Gricean mechanism described in section 4.2 of Chapter 4 above. In order to communicate, a human relies on the ability of other members of her species (her conspecific interlocutors) recognizing her intention to communicate, together with an underlying body of shared communicative conventions to encode the message. These shared conventions are words and phrases and their meanings. TNE shows how these conventions work and provides a theoretical framework for compiling an inventory of the conventions in any given culture on which successful communication depends.

Thus, TNE provides a basis for explaining, within Tomasello's Darwinian model and Grice's theory of conversational cooperation, what the shared conventions of linguistic behavior in any given community are and how they are flexible enough to encompass and develop novel ideas and novel situations s well as repetition of the norm.

Tomasello goes on to argue that the diversity of human language is too great to be accounted for by the Innate Universal Grammar hypothesis: there just are not enough linguistic universals to explain anything of any great interest about the rich, culture-specific complexities of human linguistic behavior. Again, the explanation lies in cultural transmission.

[3] Such discoveries may, of course, be rediscovered independently by other members of the species at other times and other places.

5. The need for interdisciplinarity

The world has changed dramatically since I started writing this book over sixteen years ago, and so have attitudes to and understanding of the lexicon. New developments have come so thick and fast that at times it seemed impossible to keep up with them all and I began to fear that the book would never be finished. As it is, I am acutely aware that the treatment of other research has been superficial in many places, while I fear that it seems inevitable that I must have overlooked much important work by others. To them, my apologies. My excuse is that, in this book, I have focused mainly on the empirical analysis of data, rather than on trying to provide a full account of the theories, analyses, and speculations of others.

In the intervening period since work started, I have been changing, too (as we all do)—and learning. I have had the privilege and pleasure of teaching courses in lexicology in several universities in different countries of Europe, as well as at Brandeis in Massachusetts, working in departments of computer science, informatics, linguistics, and English studies, with concomitant feedback from an unusually wide range of students and colleagues with widely different interests and different approaches to language theory and applications. This has been both beneficial and challenging. I have published two 6-volume collections of papers—one on lexicology and the other (edited with Rachel Giora [in press]) on figurative language—tracing our understanding of how lexical meaning works, from Aristotle through Wittgenstein to modern empirical work by philosophers of language, anthropologists, cognitive psychologists, linguists, and others.

Despite many exciting new developments, however, the overall picture of research into and understanding of the lexicon is still sadly disjointed and uncoordinated. Cognitive linguists, lexicographers, and corpus linguists talk to each other perhaps less often than they should. I can testify from personal experience that, when they do, the effect is occasionally total mutual incomprehension—lack of enough common ground to even begin a useful conversation—but more often, especially among people analyzing empirical data of one sort another, the effects can be most beneficial. Addressing a shared problem or topic of mutual interest, the work of other people, who may come from vastly different backgrounds, with different training and different kinds of expertise, has a habit of unexpectedly ringing bells.

The crucial phrase here is ‘analyzing empirical data’. Interaction with speculative linguists does not ring bells in the same way. In the past, speculative linguists, including computational linguists, who professed an interest in “natural” language, have tended to shy away from the complexity and volatility of real word use and meaning, preferring to deal in logical abstractions, artificially constructed ontologies, and logical necessities and sufficiencies that are relatively easy to compute, rather than to engage with the vast mass of subtle variations, probabilities, and analogies that characterizes meaning in ordinary language. Fortunately, they are a dwindling band. Nevertheless, perhaps their legacy is one reason why Semantic Web research currently focuses more on creating strictly defined ontologies and processing tagged and pre-processed documents than on engaging with the semantics of natural language in the raw.

Study of the lexicon demonstrates more clearly than any other branch of linguistics that a natural language is a puzzling and complex mixture of both logical and analogical processes and structures. Meanwhile, historical lexicographers continue to make advances in understanding the mechanisms of meaning change and continue to remind us that word meaning is in reality very unstable. But they, too, have until recently had comparatively little engagement with other researchers with different interests in the lexicon.

This sense of disjointedness is partly due, therefore, to the many different approaches to the lexicon by researchers with very different interests. Each word (or at any rate each content word) is like a vast railway junction on several different levels. On the cognitive level, there are strong and weak (fast and slow, frequent and rare) connections within the idiolect of each speaker, which is unique to that speaker and yet must necessarily closely resemble the set of connections in the brain of each other speaker with whom he or she communicates, otherwise effective communication could not take place. These are connections not only to other words but also to a whole system of beliefs, to non-verbal memory, stored and sorted perceptions, and telic motivations.

The complexity of lexical items and the variety of their interconnections is such that it is likely to yield to nothing less than a concerted effort by researchers with a variety of different expertise, working in interdisciplinary harmony on an empirical basis of evidence.

At the social level, communication depends on words and the relevant, shared, conventional beliefs about word meaning that must be held by all participants in a conversation or discourse, including the discourse between writers and their readers, if communication is to be effective. This book does not pretend by any means to cover all aspects of words, word usage, and word meaning. Still less is it an exhaustive study of the interface between lexical and other kinds of linguistic theory.[4] It does, however, attempt to make a central contribution to a crucial area, namely the ways in which words activate meanings. If the argument of this book had to be put in a single sentence, it would be something like this: understanding the phenomenon of natural language depends crucially on distinguishing the normal, typical conventions of word use and of shared beliefs about word meanings, on which effective communication depends, from the creative exploitation of those conventions. There is no sharp dividing line between conventions and exploitations, and to make matters more complicated, exploitations can be recursive, i.e., as we saw in the chapters on metaphor and similes, there can be conventional exploitations of a literal meaning as well as innovative, creative ones.

In the 1990s syntactocentric generative grammatical theory, with its comparatively impoverished treatment of the lexicon, reigned supreme in many university linguistics departments (though not among lexicographers). Corpus analysis, whether for syntactic, semantic, prosodic, or other purposes, was an untested new technology. Now, courses on collocations and corpus analysis are the norm rather than the exception, while it is impossible to publish a major new

[4] A thought-provoking approach to such an interface is outlined in Jackendoff (2002).

dictionary anywhere (except, perhaps, in America—and even the Americans are slowly catching on, except, let it be said, in the most conservative backwaters) without a strong basis in the empirical evidence provided by computational analysis of the behavior of words in large corpora. More and more studies of syntax, discourse, and other aspects of language are corpus-based.

6. Conclusion

This chapter has proposed a contribution to the empirical foundations for a lexical theory of language, pointing towards new ways of exploring meaning in text and conversation, the development of new research methodologies, and new insights into relationships between lexis and grammar, lexis and cognition, and lexis and the world. It will encourage a reappraisal of all pre-corpus theories of language and abandonment of the sloppy habit of inventing evidence to support research.

At the same time it points to the need for a fresh approach to studying the relationship between logic and analogy. A human language is a curious mixture of logical and analogical processes. Tidy-minded thinkers such as Wilkins (1668), Leibniz, and Russell tended to regard the analogical aspect as a fault, but more careful observers such as Wittgenstein and Rosch have laid foundations for an approach that regards the analogical aspect as an essential design feature. Corpus-driven lexicology can build on these foundations.

The title of this chapter is intended to mark a new departure. Corpus linguistics—approaching language through the analysis of patterns of lexis observable in corpus and textual data—is in its infancy, for the simple reason that corpora large enough for this purpose did not exist until about 20 years ago. The first astonishing finding of corpus linguistics has been an apparent contradiction: the regularities are much more regular than most pre-corpus linguists expected, while the irregularities are much more irregular. The theory of norms and exploitations outlined here shows how this apparent contradiction can be reconciled.

Systematic corpus analysis of the whole lexicon of all languages is called for, leading to new lexicons and new grammars. Some aspects of existing linguistic theories will receive confirmation from such an exercise; others will have to be jettisoned.

The Theory of Norms and Exploitations is a response to the challenges posed by corpus data, offering a contribution to the study of meaning in language. It can be seen, if you like, as a step towards making explicit the nature of the conventions on which interlocutors rely in expecting to be understood and to understand—i.e. the Gricean mechanism of conversational cooperation. It is to be hoped that much future linguistic research will be bottom-up, driven by empirical analysis of lexis, and will focus on exploring the nature both of conventions and of alternating probabilities. In this way, new light can be expected to be shed on the nature of human behaviour and human linguistic creativity.

References

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. ‘The Semantic Web’ In: *The Scientific American*, May 2001.

Bloomfield, Leonard. 1933. *Language.* Holt, Rinehart, Winston.

Bolinger, Dwight. 1970. ‘Getting the words in’. In *American Speech*, 45. Reprinted in Raven I. McDavid and Audrey R. Duckert (eds., 1973), *Lexicography in English,* New York Academy of Sciences.

Fillmore, Charles J. 1992. ‘“Corpus linguistics” or “Computer-aided linguistics”?’ In: J. Svartvik (ed.): *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82.* Stockholm, August 1991.

Hanks, Patrick. 1984 [1990]. ‘Evidence and intuition in lexicography’. In Jerzy Tomaszczyk and Barbara Lewandowska-Tomaszczyk (eds.), *Meaning and Lexicography*. Benjamins.

Hanks, Patrick. 2008. ‘Lexical Patterns: from Hornby to Hunston and beyond’ (the Hornby Lecture). In: E. Bernal and J. de Cesaris (eds.) P*roceedings of the XIII Euralex International Congress.* 9 Sèrie Activitats 20. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada.

Hoey, Michael. 2005. *Lexical Priming: a New Theory of Words and Language*. Routledge.

Jackendoff, Ray. 2002. *Foundations of Language.* Oxford University Press.

Levin, Beth. 1993. *English Verb Classes and Alternations.* University of Chicago Press.

Lewis, Michael. 1993. *The Lexical approach*. Hove: Language Teaching Publications.

Murray, James. 1878. Presidential Address to the Philological Society.

Pustejovsky, James, Anna Rumshisky, and Patrick Hanks. 2004. ‘Automated induction of sense in context’. In *COLING 2004 Proceedings*. Geneva

Sampson, Geoffrey. 2001. *Empirical Linguistics.* Continuum.

Sinclair, John. 1988. ‘A lexical syllabus for language learning’. In M. J. McCarthy and R. A. Carter (eds.) *Vocabulary in Language Teaching*. Longman.

Tomasello, Michael. 1999. *The Cultural Origins of Human Cognition*. Harvard University Press.

Tomasello, Michael. 2008. *Origins of Human Communication.* MIT Press.

Wilkins, John. 1668. *Essay towards a Real Character, and a Philosophical Language*. The Royal Society, London. Excepts reprinted in Hanks (ed., 2008a)

Willis, Dave. 1990. *The Lexical Syllabus*. HarperCollins.

Wray, Alison. 2002. *Formulaic Language and the Lexicon.* Cambridge University Press.