Software for an Inflectional Network
Project to be funded by ESRC 2001-2
Principal Investigator: Richard Hudson, UCL

1. One-paragraph synopsis

The project will produce a software package which will allow a user to interact with a database of information about English inflectional morphology. The user will be able to add new facts about morphological patterns, inflectional categories, lexemes and lexical relations, and to test the database's ability to analyse or produce new forms. The theoretical framework will be Word Grammar, in which the theory of inflectional morphology is already sufficiently well formalised for this project. The software will be designed in such a way that it can also be applied (in future work) to other areas of language, including syntax, and will include a simple model of spreading activation which can be expanded in future work.

2. Aims and objectives

1. To produce a computer system which will allow testing of the existing Word Grammar theory of inflectional morphology. At present the current theory exists only on paper, so it is important to check whether the assumed version of default logic does in fact produce the conclusions that have been claimed for it.

2. To produce a user-friendly interface for grammar development which will:
· display the relevant portion of the grammar (as a network),
· allow the grammar to answer questions about word forms as a means of testing it,
· allow the user to input new data to the grammar.

3. To produce a database of facts for English verbal inflections which cover both regular inflections and a representative sample of irregularities, including some sub-regularities (such as the sing-sang, ring-rang set).

4. To study the feasibility of including spreading activation in such a network, in preparation for a further grant application.

3. Non-technical description

Inflectional morphology is the area of language which is responsible for variations in word forms which reflect tense, number and so on. In English these variations are rather trivial compared with highly inflected languages such as Latin and its descendants, because a regular English verb has just four alternative forms (including its basic form): e.g. walk - walks - walked - walking. However even in English there are non-trivial problems:

· How to accommodate irregular verbs such as take (past: took, not taked) in the same analysis as the regular ones? For example, does took contain both or either of take and the ed suffix?

· How to relate the morphological structure - e.g. the presence of ed - to the more abstract inflectional categories such as 'past tense' and 'past participle'? Are there two distinct ed suffixes, each mapped onto a different inflectional category, or just one ambiguous one?

· How to accommodate sub-regularities such as the tendency for monosyllabic verbs which end in *t* to have identical base and past forms: cut, hit, let, shut?

Any linguistic theory which aims to cover inflectional morphology has to answer these questions, and answers have already been suggested in Word Grammar, as explained below.

The proposal is for a grant to build a computer system for inflectional morphology in Word Grammar. This will allow us to test the Word Grammar answers for consistency, and if results are positive it can be used freely in future work for analysing inflectional morphology in other languages where the facts are much richer and (in some ways) more challenging. This kind of work really requires a computer model as a tool because of the amount of detail and the interactions between different kinds of data. It is one thing to present the facts in the form of a paradigm of inflected forms, as in traditional grammars, but it is much more difficult to present the general patterns in a formal analysis.

This area of language has recently attracted a great deal of attention from psychologists, as popularised in Pinker's Words and Rules (Weidenfeld and Nicolson 1999). The question is whether inflected forms are all generated by connectionist networks, or (as Pinker thinks) some are handled in this way but others are handled by a completely different mechanism, rules. The debate has involved both experimentation and computer modelling by both sides, so the present proposal can be seen as a small contribution to this research. It will be small, because it will not attempt to duplicate the experimental findings by modelling spreading activation. (This may be possible in future research using the software developed in this project.) However it will be sufficiently different from both the competing alternatives to suggest the possibility of a middle way.

What, then, are the Word Grammar answers to the various research questions that have been raised? The main claim of the theory is that language is a symbolic network - a network of nodes and connections each of which is 'symbolic' in the sense that it can be related individually to some aspect of reality such as a word, a relationship or a concept. This steers a middle way between the two approaches that Pinker contrasts because (like rules) it is symbolic, but like connectionist networks it assumes a network structure. Linguists generally favour symbolic analyses (in this sense), but few linguistic theories other than Word Grammar assume network structures.

Another peculiarity of Word Grammar is that the networks are not simply associative networks which show that node A is associated in some way with node B. Instead the links between nodes are all classified in terms of a general system of link-types which is itself clearly defined. It is common in AI to present knowledge structures in this way, though less so in linguistics; but what may be unique to Word Grammar is the assumption that link-types are themselves organised in an 'isa' (i.e. classification) hierarchy. This seems psychologically plausible, as it solves the question of where the link-types 'come from': more general types can be learned inductively by generalisations across specific sub-types.

When applied to inflectional morphology, this theory assumes a network which translates quite directly into a traditional description of morphology. Categories such as 'past tense' are word classes, alongside lexically based classes such as 'verb' and lexical items such as WALK; so the past tense of WALK is defined as the intersection of the classes 'past tense' and 'WALK'. This is represented in a Word Grammar analysis by the label WALK:past, but in the network it is shown as a node with 'isa' connections to 'past tense' and to WALK. This word's 'stem' is the part contributed by the lexical item, i.e. walk, and its 'suffix' is the part contributed by its inflection; so its completely inflected form (called its 'whole', for lack of an established term) consists of walk followed by ed. For an irregular form such as took we say that its 'stem vowel' is oo, which overrides the usual stem vowel of take. And if there is a sub-regularity such as the one involving cut, hit, let and shut, this can be expressed by postulating a sub-class of verb-stems whose 'prototype' is defined by a combination of monosyllabicity and a final t, and whose past tense is always the same as its present.

In general, then, this theory builds heavily on traditional informal descriptions. However, there is one important respect in which it breaks with tradition. The analysis assumes the logic of default inheritance, whereby default properties are inherited unless overridden by more specific ones. In this kind of analysis we take basic and 'unmarked' characteristics as defaults, which has the effect of removing the traditional distinction between a super-category and its unmarked sub-category - for example, the distinction between Noun and Singular noun. This gives a two-category analysis which contrasts Noun (= Singular noun) and Plural noun, instead of the more traditional Noun - Singular noun - Plural noun. When applied to verbs this approach makes the analysis even simpler than a more traditional one.

Apart from the use of default inheritance, the main advance on traditional analyses is the possibility of strict formalisation. There are two fundamental components to the formalisation:

· link-types, which accommodate relations such as 'stem of' and 'suffix of';

· the logic of multiple default inheritance, which allows the inflected word to inherit characteristics from multiple models (namely, the lexical item and the inflectional category), and to inherit only 'by default', so that exceptions can override the default patterns.

As is to be expected, there are other theories which show partial similarities, but none which combine characteristics in quite this way. Moreover the theory of inflectional morphology is integrated into a much larger theory (Word Grammar) which applies the same basic ideas to other areas of language including syntax and semantics. No competing theory of morphology has such a close relationship to an over-arching theory of language.

However it is important to stress that the proposed project has a very limited goal, commensurate with the size of the grant requested. The theory described is already available, and it will be easy to build a small database for a sample of English verbs. All that is needed is software which will allow the following elementary operations:

· displaying a selected part of the database as a network on screen;

· adding to the database and editing it;

· creating new temporary facts by multiple default inheritance;

· using default inheritance to define queries in either direction (i.e. querying the form of a specified word classification or querying the classification of a specified form).

Once it is ready the software will be made generally available on the internet so that it can be used by other researchers and by students to allow easier development of theories and descriptions, which are impossibly complex without computer testing.

Another benefit of this system will be the possibility of incorporating a model of spreading activation into it. There will not be time in this project to take this idea far, but we shall at least aim to test the feasibility of developing a more sophisticated system in later work.

This project fits into a plan for the longer term which includes various extensions which will be able to build on the same computer system with minimal changes.

· As just mentioned, the system will model spreading activation, so that the answer to a query will be the most active node that has the specified characteristics.

· The system will apply to derivational morphology (e.g. the link between WALK and WALKER), lexical semantics (e.g. the link between Walking and other semantic nodes such as Go, Leg and Run), syntax and compositional semantics (e.g. the result of combining John with walks).

These extensions will require a series of separate projects, which we hope ESRC will be able to fund.

4. Full research Proposal

Relations to the previous proposal

The present application is based on the much more ambitious proposal R000239358 called "Parsing by spreading activation in Word Grammar". This was described as follows by the Board Assessor:

"This is a project that has considerable promise in that it may well provide a more realistic model of human syntactic processing than the mainstream approach that has been provided in theoretical linguistics. However, the project as it stands is very ambitious, very expensive particularly given the number of potential problems that developing it may well imply. It is thought that a more piecemeal approach, which inevitably will mean a more modest proposal, may be a better way forward for this very innovative and interesting approach to modelling."

The present proposal is the first step in the piecemeal approach recommended. It corresponds to the preparatory phase of the larger project, but with a more specific focus on inflectional morphology. Neither of these parts of the earlier proposal attracted any adverse comments from referees. However the proposed project can stand on its own, regardless of the success of future grant applications. This is guaranteed by the focus on inflectional morphology of English verbs, which is a worthwhile research topic in its own right.

Inflectional morphology in Word Grammar

The analysis of an inflected word such as *walked* involves the following elements:

1. a lexeme: WALK
2. an inflectional category: Past
3. a word class: Verb
4. two categories: Word, Morpheme
5. a word-type defined by the intersection of WALK and Past: Walk:past
6. two morphemes: *walk*, *ed*
7. five variables which by default are instantiated: 1, 3, 4, 5, 6
8. one variable which by default is not instantiated: -2
9. the 'isa' relationship from
    10. WALK and Past to Verb,
    11. Verb to Word
    12. *walk*, *ed*, 1 and -2 to Morpheme
    13. 3 to *walk* and 4 to *ed*
    14. 5 to 3 and 6 to 4
15. the 'stem' relationship from
    16. WALK to 3
    17. WALK:past to 5
    18. Word to 1
19. the 'suffix' relationship from
    20. Past and WALK:past to *ed*
    21. Word to -2
22. the 'whole' relationship from
    23. WALK:past to the sequence {5, 6}
    24. Word to the sequence {1, -2}
25. the 'predecessor' relationship from
    26. *-2* to 1
    27. 6 to 5

These elements define the network shown in Figure 1.

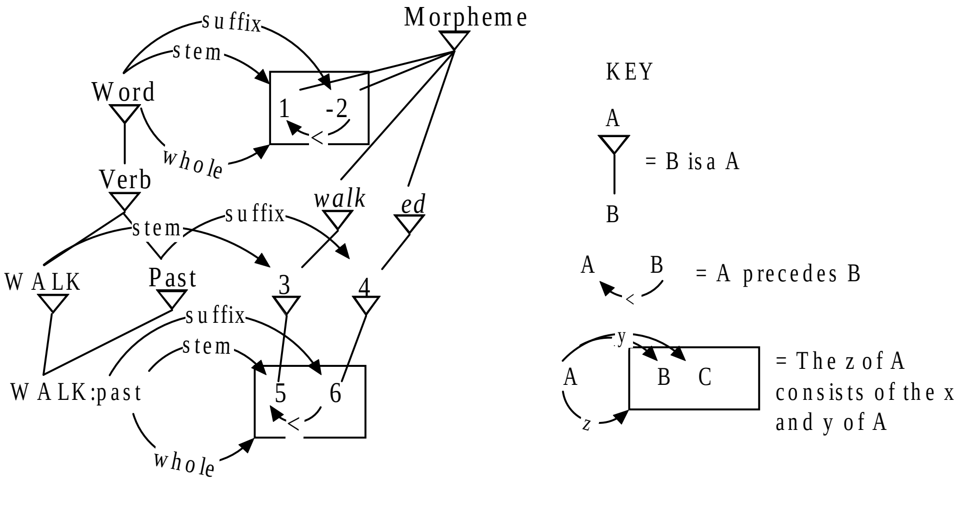

Figure 1

All of the elements in this network are part of the stored grammar except for those in the bottom right-hand corner - the variables 5 and 6, and all the links to them. These stand for the form *walked*, composed out of examples of the morphemes *walk* and *ed* in a pattern that can be inferred by multiple default inheritance from the stored elements. For example, because WALK:past isa WALK, its stem isa 3, which in turn isa *walk*; and because WALK:past isa Word (inherited via WALK and Word), this token of *walk* precedes the token of *ed*. The term 'whole' means the word's fully inflected form; all the other terms are traditional and self-explanatory.

In this model the role of inflectional morphology is to handle any differences there may be between a word's stem and its whole. In some cases this may be done directly by a rule such as "The whole of a Past verb consists of its stem followed by *ed*", as in the above example. However there are in fact reasons for handling this particular pattern (and others) in two steps rather than one. One such reason is the need to explain systematic syncretism; for example, the fact that a verb's past tense is usually the same as its past participle, and (more strikingly) that the past participle is always the same (in form) as the passive participle, however irregular the verb may be in other respects. In order to explain this syncretism we can assume an 'intermediate' morphological function called 'ed-form' which stands between the stem and the whole. This allows a more explanatory analysis than would otherwise be possible:

- A verb's ed-form consists of its stem followed by *ed*.
- A past tense verb's whole is its ed-form.
- A past-participle verb's whole is its ed-form.
- A passive-participle verb's whole is its ed-form.

Morphological irregularity is then located in the definition of the ed-form (which may be further subdivided if the past tense is different from the participles). For simplicity, none of this is shown in Figure 1.

The network in Figure 1 deals with the morpho-syntax of *walked*, but it does not make contact with the sounds of which these words are composed (which is important for reasons explained below). The morpho-phonology for *walked* requires at least the following additional elements:

- The categories Phoneme, Vowel and Consonant.
- The phonemes /ɔ, w, k, t/.
- Four obligatory variables (1, 2, 3, 4) standing for the four segments.
- The 'isa' relationship from
  - Vowel and Consonant to Phoneme
  - /ɔ/ to Vowel
  - /w, k, t/ to Consonant
  - 1 to /w/, 2 to /ɔ/, 3 to /k/, 4 to /t/.
- The 'shape' relationship from
  - *walk* to the sequence {1, 2, 3}
  - *ed* to 4
- The 'stem vowel' relationship from *walk* to 2.
- The 'predecessor' relations from 2 to 1 and from 3 to 2.

This analysis involves a minimum of assumptions about phonological structure because WG has not been applied seriously to this area of language; presumably the individual phonemes are further classified by multiple default inheritance in terms of some set of phonetic categories which we shall not attempt to show. Nor shall we try to show the allomorphy of *ed* (whose default shape may in fact be /d/ rather than /t/). The network for these facts is shown in Figure 2. (For simplicity this figure omits the 'predecessor' relations among the phonemes inside *walk*.) A different network would be needed for the graphological structure (defined in terms of letters).

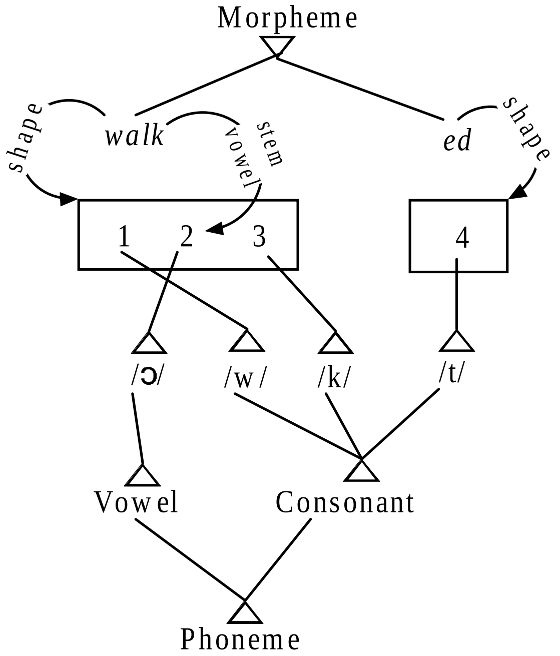

Figure 2

Why is it necessary to deal with morpho-phonology as well as with morpho-syntax? There are three reasons.

- Most irregular verbs involve ablaut - a change of stem vowel between present and past: *sing* - *sang*, *take* - *took*, etc. Ablaut applies directly to phonological structure, by substituting one vowel for another, so we have to push the analysis as far as phonology. Given the analysis outlined above, ablaut is easy to accommodate with the help of the relationship ‘stem vowel’ from the stem to its (stressed) vowel. In this way we can distinguish vowel-change verbs from suppletive ones (*go* - *went*), where the entire stem changes.
- Sub-regularities are defined phonologically, so if we are to represent them we have to be able to define them in phonological terms.
- A phonological analysis will allow us to model psychological processes of spreading activation which are involved in selecting morphemes. This will be important in future work, though less so in the present analysis.

It should be clear why a computer system is important for this kind of work. Even a tiny fragment of the total network for English such as the one illustrated above is already problematic - hard to display and hard to check for completeness and consistency. The problems will increase rapidly as further inflections and lexemes are added, so further serious research is virtually impossible without a computer aid.

Earlier work in Word Grammar

My interest in morphology dates back to my PhD on a highly inflected (Cushitic) language, Beja. I published a formalisation of this morphology in terms of the theory which I then espoused, Systemic Grammar (Hudson 1973), followed by two general articles on morphology (Hudson 1976; Hudson 1977). Each of my monographs on Word Grammar contained a substantial discussion of morphology (Hudson 1984, pages 43-74; Hudson 1990, pages 90-96, 172-80, 225-32), and more recently I have authored or co-authored two articles on morphological theory. The first is a general presentation of Word Grammar morphology with applications to a variety of language types (Creider, Hudson 1999), while the second discusses one particularly interesting detail of English morphology in great detail (Hudson 2000).

In short, I have been thinking about morphological theory for a long time, and know enough about the linguistic and psycholinguistic facts to be able to evaluate my own theory quite realistically. I believe it is as good as any of the alternatives, though some areas are still developing. One reason for requesting this grant is that I feel it may not be possible to make further progress without a computational tool for testing hypotheses.

I also received a grant from the ESRC which allowed me to spend a year in 1987-8 writing a Prolog grammar-tester for Word Grammar. The main thrust of the work was syntax, but the grammar-tester included a morphological analyser which could recognise all the inflections of English regular verbs. In this sense the morphological analyser was successful, but it required a specific module for morphology which contained a very large amount of code. The aim of the present project is to use a very much more general processor in which only a very small part - or better still, no part at all - which is dedicated to morphology.

Relations to other work on inflectional morphology

There are a great many alternative theories of morphology, each of which shares some of the assumptions of Word Grammar. Since the proposed project does not specifically aim to evaluate Word Grammar in relation to other theories it would be irrelevant to list them here. In any case, the theory to be tested is certainly unique in its compatibility with the other tenets of Word Grammar.

However there is a serious question to be answered in relation to DATR, a widely used and tested programming language which was specifically designed for use with lexical analyses which presuppose multiple default inheritance (Evans, Gazdar 1996; Gazdar 1992). Since DATR is freely available, and since it looks at first sight like the tool we need, why do we need to build a new system? This question is especially urgent since DATR has been used as a computer environment for work in a similar theory of morphology, Network Morphology (Brown et al 1996; Corbett, Fraser 1993; Fraser, Corbett 1995). There are various reasons why we cannot use DATR:

- Although DATR allows multiple inheritance, it requires the supercategories to be ranked so that conflicts will always be resolved in favour of one of them. This is not in the spirit of Word Grammar, where it is important for conflicts to remain unresolved. (This is the basis for my explanation of the gap where we expect **I amn't* - Hudson 2000.)
- DATR does not allow the rich vocabulary of elements that we need for relating words, morphemes and phonemes and for distinguishing relationships such as 'stem', 'suffix' and 'whole'.
- DATR uses 'rules of referral' to handle systematic syncretism, for which we use the intermediate functions such as 'ed-form' explained above.
- DATR is designed for use in 'the lexicon', and it is not clear how or even whether it can be extended to other areas of language such as syntax.
- DATR does not include a user-friendly interface for presenting and editing network grammars.

In other respects, however, the proposed software is very much in the spirit of DATR so we shall be able to build on that system.

Details of the proposed system

The software package which we propose to build will be written in C++, and when compiled it will be made freely available to any user for use on the internet or for downloading to a PC. It will include the following facilities:

- An interface with a **database of facts** which is stored in text form so that users can inspect it visually. This database will hold all the declarations for individual nodes and relationships (with the exception of the 'isa' relationship, which is built-in because of its role in inheritance).
- A **screen** interface which allows the user:
    - to select some node in the database for display as the centre of a small network of closely related nodes (where 'closely' may be defined by the user);
    - to add, delete and edit nodes and relationships;
    - to define queries by adding a 'query' variable to the network which the system then tries to instantiate;
    - to 'step' through queries watching changes to the network as they happen.
- An inference engine based on the Word Grammar definitions of:
    - **Multiple Default Inheritance:**
        - **If:** A isa B and
        - X of B = M and not:
            - Y of C = N where:
                - Y = X or Y isa X and
                - C = B or C isa B and
                - N ≠ M
        - **then:** X of A = M.
    - **Isa:**
        - **If:** A isa B and
          B isa C
        - **then:** A isa C.
- An elementary **'spreading activation'** model which will simply select candidates for testing by the inference engine (e.g. all the stored morphemes whose letters overlap with those in the word whose identity is being queried, and in which these letters occur in the same order). In this project this activation will involve on/off values which will spread a specified distance through the network, but in later versions these values can be replaced by numbers.

At least in language processing it is unusual to combine spreading activation with default inheritance, so this project will also test the feasibility of this idea. The default logic presented above faces well known problems of speed, because whenever it is possible to infer a default value, a check has to be run for more specific facts which might override the default. This requires a complex search through the entire database, which is unrealistic psychologically as well as unpromising computationally. However, the addition of spreading activation solves this problem because all the facts that might be relevant will be active, so the search can be restricted only to a few of the most active facts.

Reference List

Brown,D., G.Corbett, N.Fraser, A.Hippisley, and A.Timberlake. 1996. Russian noun stress and network morphology. *Linguistics* 34, 53-107.

Corbett,G. and N.Fraser. 1993. Network morphology: a DATR account of Russian nominal inflection. *Journal of Linguistics* 29, 113-142.

Creider,C. and R.Hudson. 1999. Inflectional Morphology in Word Grammar. *Lingua* 107, 163-187.

Evans,R. and G.Gazdar. 1996. DATR: A language for lexical knowledge representation. *Computational Linguistics* 22, 167-216.

Fang,A.C. 2000. A lexicalist approach towards the automatic determination for the syntactic functions of prepositional phrases. *Natural Language Engineering* 6, 183-201.

Fraser,N. and G.Corbett. 1995. Gender, animacy and declensional class assignment: a unified account for Russian. *In Year Book of Morphology 1994*. G.Booij and J.v.Marle, eds. pp. 123-150. Dordrecht: Kluwer.

Gazdar,G. 1992. Paradigm Function Morphology in DATR. *In Sussex Papers in General and Computational Linguistics*. L.Cahill and R.Coates, eds. pp. 43-53. Brighton, Sussex: University of Sussex.

Hudson,R. 1973. An 'item-and-paradigm' approach to Beja syntax and morphology. *Foundations of Language* 9, 504-548.

Hudson,R. 1976. Regularities in the lexicon. *Lingua* 40, 115-130.

Hudson,R. 1977. The power of morphological rules. *Lingua* 42, 73-89.

Hudson,R. 1984. *Word Grammar.* Oxford: Blackwell.

Hudson,R. 1990. *English Word Grammar.* Oxford: Blackwell.

Hudson,R. 2000. *I amn't. *Language* 76, 297-323.