The Location of Words: Evidence from Generation and Spatial Description

David D. McDonald
Smart Information Flow Technologies (SIFT), Boston Massachusetts
dmcdonald@sift.info

Abstract

Language processing architectures today are rarely designed to provide psychologically plausible accounts of their representations and algorithms. Engineering decisions dominate. This has led to words being seen as an incidental part of the architecture: the repository of all of language's idiosyncratic aspects. Drawing on a body of past and ongoing research by myself and others I have concluded that this view of words is wrong. Words are actually present at the most abstract, pre-linguistic levels of the NLP architecture and that there are phenomena in language use that are best accounted for by assuming that concepts *are* words.

Introduction

David Marr did not think that it would be possible to find a Type I theory of natural language (1976). He expected that only a Type 2 theory would be possible: one where the interaction of the processes involved is its own simplest description—a theory just like protein folding, where the details of the structure and properties of the individual amino acids all contribute to the shape of the protein that is built from them. He was skeptical that deep principles about the nature of intelligence would be found to account for the information processing problem of language (which he took to be how we convert content into a one-dimensional form for sequential transmission and uptake).

He began to change his mind after looking at Mitch Marcus' thesis work, where the architecture of the Parsifal parser—its representation of incremental state during syntactic parsing—provided a principled account of two grammatical properties stipulated by Chomsky (1976): subjacency and the specified subject constraint (see Marcus 1981). Marcus' account explains those properties as a necessary consequence of his choice of computational architecture: notably his determinism hypothesis (language can be parsed by a simple mechanism without backtracking or parallelism) and his machinery for accounting for our reaction to garden-path sentences.

Words are one of the messiest, most idiosyncratic aspects of language. In this paper I will start an effort to simplify this mess, to make it more likely that we will be able to discover a Type 1 theory of word use in language rather than presume that words are our analog of protein folding that cannot be accounted for with anything less than all of their details.

Current computational models place words very close to if not part of the surface of language. The last stage of generation, the first for analysis. Syntactic grammars are lexicalized. Parsers work from the identity of a word back through the constructions it is part of. Many generators select words last.

This is incorrect. *Words are properly located at the most abstract, pragmatic levels of the language processing architecture*, albeit with some of their details (morphology, segmental structure) not relevant until very late levels. The rest of this paper will make the case for this assertion, showing that it makes for a simpler account of the productivity of spatial terms, word-sense disambiguation, and expressibility in generation.

Vague Spatial Terms

While actual locations and the spatial relationships among the objects they contain are specific, exact, and detailed, they cannot be described linguistically unless you back away from that exactness and drop much of the detail (consider describing the layout of your office or the path you take to drive home). If exact descriptions are required, we either observe the location directly or draw upon a map or a photograph to provide our representation.

What is the meaning of the words that we use in describing a scene or a route? I propose that *the mental representation of vague spatial words is something very close to the word itself*. These words contain enough information to guide you to pick out the right relation or location when you use them, while remaining vague enough to be applied creatively to an unbounded number of situations.

This is because, words like *middle* or *to the left of* do not denote probability distributions over anchored fields, the results of psychological experiments not withstanding. To assume that they do misses the point that language is used for a purpose in a context. The purpose provides the range of the interpretation, (*Put it over there <point>* means to take whatever salient thing you have and set it down in the indicated location.). Direct inspection of the context plus general knowledge or experience with the items involved grounds the interpretation. (If 'it' is boxes of books then 'there' is probably a region of the floor; if it is a vase of flowers then 'there' is usually a place in the center of a table.) Any remaining uncertainty is handled by follow-on conversation.

Sloman (2010) argues persuasively for this view of what spatial terms are. I came to it in collaboration with Rusty Bobrow of BBN while working on the problem that Beth Driver of NGA calls *Text to Sketch*. This the problem of how to get from a linguistic description, perhaps a message from an informant on site in an Afghani village, to the representational equivalent of a sketch of the layout and spatial relationships conveyed in the message. This sketch must capture as much spatial information from the text as possible in order to constrain its eventual grounding in a GIS system, but must also represent what is unknown or partially known so as to facilitate integrating other descriptions of the same place or searching for additional information.

Driver uses the example "*a bicyclist rode quickly past the White House and took the first right*". (Imagine it over the phone to a security guard in Washington D.C.). This text has enough information to let us make a sketch of what happened, but if you consult a map you will discover that it has eight satisfying models. We cannot render this description to a path on a map without additional information (which street the bicyclist was on and which way they were traveling). We need a representation that will capture the information that we do have, while being uncommitted as to the rest. In this case we need to represent the actor, the reference point, the fact that they made a turn on their egocentric axis to their right, and that the location of the turn was the first one with that affordance (bicycles don't need to stay on streets). I would argue that the best representation to use in this minimally committed model is the sense-disambiguated words ('right' as a direction rather than political philosophy) embedded in a matrix drawn directly from the linguistic structure. For example we don't have to render 'quickly' into a range of speeds, we just need to represent it compositionally as 'quickly for a bicycle' and elaborate it only if we get more information (racing bicyclist or tourist).

Text to sketch is a large task that will require considerable time and resources. The biggest need is for a tractable way to represent the potentially encyclopedic background knowledge that is required. To say *bicycle* is to potentially evoke all that you know about bicycles. We are usually uncomfortable thinking of such vast stores as part of lexical semantics, but words are the access points to this knowledge and it may be impossible to draw a clean line. The information processing question is how this store is selectively accessed.

Parsing route descriptions

As a small step on this project, I have begun to develop a word-level conceptual model and semantic grammar for my Sparser language understanding system (McDonald 1992, 1993). At this point it can instantiate models for concrete descriptions like this one.[1]

> "*From the north or east, drive to the junction of US 20 and MA 102, just south of the Lee exit of the Massachusetts Turnpike, and then travel 2.5 miles west on MA 102 to South Lee. Turn left onto Meadow Street, cross the Housatonic River, then turn right immediately onto Beartown Mountain Road. Follow this road uphill past the forest boundary (1.5 miles) to a fork (2.5 miles). At the fork, turn right (downhill) onto Beartown Road and drive another 4.0 miles to Benedict Pond.*"

Sparser gets its semantic grammar by projection from its conceptual model. Every category in the model is associated with at least one way of expressing it, represented schematically by a TAG tree family. To construct the grammar, the parameters of the categories are mapped to syntactic categories in the schemas, and the categories of the parameters' type restrictions become the semantic labels of the grammar. This ensures that the semantic grammar will be linguistically well-founded, permits syntactic rules to mix with the semantic, and integrates the model tightly with the grammar. Texts are analyzed incrementally left to right at syntactic and conceptual levels simultaneously.

In this instance, the model has categories for points of the compass, following a path by 'driving', road junctions, turns in paths, and so on. Overall, route description are conceptualized in the obvious way as an ordered set of waypoints connected by descriptions of the action you

[1] From "50 Hikes in Massachusetts" White & Brady, The Countryman Press, Woodstock VT.

have to take to move from one to the next. That is what a text like this is taken to mean, and it is what Sparser produces as a result of its analysis.

This meaning is represented as a network of instantiated categories ('individuals'). This is a shift in type—from words or phrases to Sparser's internal representation of what it knows—but not a shift in content or information: the categories are essentially just primed words. The model contains the notion of a 'junction', of a 'street', and so on. It deliberately does not convert these lexical concepts to more primitive terms and does no canonicalization beyond merging synonyms. We sometimes call this capital letter semantics, with the connotation that all we have done is change the notation, but I would argue that for at least the case of spatial terms this is exactly what we want.

We know a great deal about a sortal category like 'junction' or any other primed word. Part of that knowledge is linguistic: we know the ways that it patterns and the differential effects those alternatives achieve.[2] Most of what we know about the category is hard to think of as its 'meaning' (in any ordinary sense) because it is encyclopedic and idiosyncratic. In a full robotic architecture there would be a connection between the category highway-junction and tactics for recognizing one while driving; there could be case memory of salient junctions that it has encountered, and so on.

To some extent there is more than this just within Sparser as well. Categories are organized into a subsumption lattice, and at the higher nodes of that lattice there are associated axioms (CLOS methods) that carry information such as for any instance of PTRANS ('drive'), the thing driving ('you') changes location. This is the sort of knowledge that could be (but is not yet) be used to control a simulation. In principle there would also be schematic world knowledge to fill in gaps, but until that is available, my focus has to be on relatively complete texts like this example.

But even in a text that makes no substantial inferential demands to fill in missing pieces, there are still interesting ontological issues. If we took all the propositions we recovered from the text (e.g. **move-to(junction-1)**) and entered them into an unstructured knowledge base they would be contradictory: a person cannot have moved to two different places as the same time. Some sort of temporal indexing must be added to the literal content we glean from the text.

I have found the SNAP/SPAN ontological framework of BFO[3] to be the most useful way to conceptualize the necessary indexing (Grenon & Smith 2004). Briefly, it is a partitioning approach. Enduring entities (a location, 'you' as you follow the hiking guide, your left ear) are registered in an ontology (A-box) using terms from the SPAN taxonomy. The 'snapshots' of the values of these entities' properties at a particular moment are registered with terms from the SNAP taxonomy.

When reading a hiking guide, the thing whose properties change is the tacit 'you' who is carrying out (mentally) the directives for what to do, and what changes is 'your' location. Operationally in Sparser, this means that there are a different, distinct objects representing the location of the single constant object 'you', one for each 'move' step in the directions.

When contemplating a route, there is no need to go any deeper that the words of the text to supply the terms to represent what happens, e.g. **cross(Housatonic-River)**. Only when we actually do the hike or look at the region on Google Maps will we learn, e.g. how wide the river is or how long it takes to get there from the previous waypoint. Our direct perception (or the sensors on Google's cars) provides the detail when we need it. This is certainly the case for a machine, whether it is also true for people remains to be seen.

Word senses

Words in isolation typically will have several different senses, and it is specific senses that are my candidate 'primitives'. This puts a premium on having mechanisms for word-sense disambiguation. Given that we are working with a semantic grammar, my experience has been that these possibilities can be markedly reduced through the use of associated context sensitive rewrite rules. The word *travel*, for example, has 9 senses in WordNet but when it is followed by a constituent labeled as a distance (*2.5 miles west on MA 102*) it definitively means 'move'.

The types of participants associated with an instance of a word will often lead to a specialization of the associated concept, moving it to (or in some modes creating) a more specific position in the subsumption lattice. Consider that the word *junction*, when it appears with two highways as it does here (*the junction of US 20 and MA 102*) becomes an instance of a highway-junction, which would have concrete recognition criteria to, say, Google's self-driving car, and quite different from, e.g., the junction of two rivers.

This type-driven specification of concepts from their context depends on a knowledge-rich analysis and a precise (i.e. accurate) parser such as Sparser or the comparable precise language understanding systems developed by James Allen, Peter Clark, or the DELPH-IN group.

Another approach is to reconsider whether the fine-sense divisions that we see in WordNet are really there. The spatial word/concept *across* has two readings in WordNet. But both would be accommodated by the extensive analysis that Len Talmy has provided (2000, pg. 187), which in any

[2] For instance, there is the pattern "<path-1>'s junction with <path-2>', which is best used when path-1 is more salient than path-2, perhaps because it is the topic of that segment of the text.

[3] "Buffalo Foundational Ontology" – http://www.ifomis.org/bfo

event would be more valuable to a reasoning system than the lexical semantic relationships provided by WordNet.

Matthew Stone (2003) takes an even more radical approach, asserting that content words should be modeled as simple terms in a semantic vocabulary that only take on meaning when they are used in a particular situation. (*I like coffee* goes to the beverage when saying what you want to drink; but on a playground it could refer to a soccer team if one of them is named 'Coffee'.) This allows for the enormous creativity that people exhibit in their language use, though it does raise the requirements for a knowledge intensive analysis still further.

Indirect Evidence from Generation

Language is much harder to study than Marr's vision problems. We see only the surface of a rich system, and with few exceptions our computational models do not penetrate deep into it. Most parser's stop their analysis as soon as they construct a logical form and map words into predicates, and we have absolutely no idea what the starting point of the generation process is.[4] To make scientific progress, we must triangulate from what we can observe to constrain the representations and mechanisms that we cannot. In this section we will look at evidence from accounts of speech errors and of expressibility.

Morphology and content words

One of the lexical phenomena that needs an account is the patterning of speech errors. Not every rearrangement is possible, and this calls for an explanation. Below are two examples of word-exchange speech errors from the MIT corpus; the first is a full-word error, the second is a mixed-form error. Underlines indicate the parts that are in the wrong place.

(1) *The whole country will be covered to a foot of one depth with dung beetles*

(2) *Oh, that's just a back trucking out.*

Word exchanges are set apart from other classes of speech errors by the larger distances that separate the exchanged words and by the fact that they rarely involve words of different form classes. Garrett (1975) outlined a multi-level architecture for language generation based on the principle that items had to be present on the same level in order for them to exchange positions.

Garrett's reasoning influenced my early work on language generation, and was reflected in the design of my surface realization system Mumble (McDonald 1979, Meteer et al. 1987). Mumble uses a variation on Joshi's Tree Adjoining Grammar (McDonald & Meteer 1990) that separates morphological realization from the choice of lemma. Morphology is only represented in the linguistic structure—the 'slots' into which words are placed during the assembly of the TAG derivation tree that drives the process of surface realization. When the words at the leaves of the surface structure tree are read out, they then receive the morphological form that their functional context dictates.

This structural property of Mumble provides an account of why exchange errors don't mix lexical form classes: they are simply never present at the same time during realization. Example 1 involves an error in the mapping of words to positions (substitution nodes) in the derivation tree. In example 2 the grammatical function that the words will have is already established in the surface tree. Whatever word ends up in the gerund slot will be given and '*+ing*' suffix (*trucking*).

The morphologically realized features that Garrett found to be stranded by mixed-form exchanges included number, tense, derivational suffixes, gerundive, comparative, and possessive. It is no accident that these are also the phenomena that Mumble treats as late-operating morphological effects, though this does also simplify the earlier stages of generation: At those levels, words can be treated by just their identity. Word-form is not relevant until 'the last minute' when words are queued up to be uttered. Earlier stages just select words and indicate their functional roles.

The use of a TAG just by itself provides an account of why people speak so grammatically.[5] Every elementary base tree in a TAG is grammatical, as is every adjunct tree in Mumble's formulation. Because these are the only elements available in our account from which to assemble an utterance, the Mumble architecture does not provide the ability to produce an ungrammatical text.

Problems with planning

Just because Mumble can not say anything ungrammatical does not imply that what it says will be sensible. It doesn't even imply that it will work—that the combination of elements passed to Mumble will be lexically compatible. Total success (the usual state of affairs with people) depends on operations in the 'upstream' components of the generator. Using Levelt's terms (1989) these are macroplanning, where the content and intent of what will be said are established, and microplanning, where the form that this content will take is established.

[4] See. For example. Wilks (1990/2003) or McDonald 1994. So-called 'generation' from logical forms is a trivial problem compared with what people do when the talk.

[5] This deliberately ignores restarts, which are probably the result of monitoring and replanning, and downstream motor problems such as stuttering or perseveration.

Evidence of upstream problems comes from instances where the speaker 'walks themselves into a corner' and is unable to express what they are trying to say. There are a few examples in the speech error corpora that seem to be instances of this. In this example the speaker winced as they uttered the final weird phrase.

(3) ... *we'll look at some children where it appears that some of these plans become talkable about.*

There are other instances that may well reflect the same kind of problem but the results are fluent if unusual and possibly unique. For example:

(4) *I would really like you guy over for dinner, so let me know whether for you it is better before or after Florida.* (i.e. before or after your vacation in Florida)

(5) *When we get it down like that it will stay clean for a couple of inches* (i.e. shoveling snow from the sidewalk that thoroughly will leave it free from ice for the time it takes to accumulate two inches of new snow)

Federica Busa and I studied these (1994) and concluded that they worked (i.e. the speaker found their way out of their dilemma and their listeners understood them perfectly) because first of all they occurred in a situation that was mutually well understood, and then that the speakers were able to draw on the 'coercion' machinery of Pustejovsky's Generative Lexicon (1995). This allowed the speakers to for instance, take a measurement (two inches of snow) and convert it in context to a duration, or take a location (Florida) and have it understood as an event (vacationing in Florida).

Expressibility

But utterances like that are actually quite rare. Rare enough that those of us inclined to do so collect them. Marie Meteer (1992) referred to this as the problem of *expressibility*: how is it that microplanners are not continually talking themselves into corners; how is it that the text plans that they formulate are virtually always expressible?

The question is how do the upstream components of the generator select lexical and syntactic resources that are compatible. This is not a simple matter. The lexical paradigms of English are not always complete: one can not assume a priori that the concept you want to express will have, e.g., a realization as an adverbial form. Consider the examples in Figure 1, adapted from Meteer 1992 pg. 50. On the face of it there are four possible outcomes, the head can be either a noun or a verb, and the modifier can be either an adjective or an adverb. Only three of these work. A generation system that made its word choices late in microplanning (the vast majority) could find that it has, e.g., selected the head but has no compatible modifier.

Expression	Construction ('decide')
"*quick decision*"	<result> + <quick>
"*decide quickly*"	<action> + <quick>
"*important decision*"	<result> + <important>
* "*decide importantly*"	<action> + <important>

Figure 1: Constraints on expressibility: To say that there was a decision and it was important, you are forced to use the noun form because there is no adverbial form for *important* as there is for *quick*

As Charlie Greenbacker and I describe in our 2010 paper, there are a number of ways that expressibility has been dealt with in the past. The two common ones are lookahead and revision. These are engineering solutions, however, and are not psychologically plausible.

Our conclusion is that expressibility is possible because knowledge of what words and other lexical resources[6] are available for realizing any particular 'chunk' of mental content is available at the earliest (most abstract, least linguistic) moment in the generation process.

We believe that this comes about because we make a mental record of everything that we hear—all of the different ways that we have learned can be used to refer to our stock of concepts. We have implemented this using the combination of Sparser and Mumble as a bidirectional system, starting on the analysis side, keeping records in the conceptual model in the form of a synchronous TAG (Shieber & Schabes 1991). We draw from that stock of known ways that a concept or set of concepts have been expressed to assemble a derivation tree to drive Mumble. This research is still in its earliest stages.

Concluding Remarks

We have tried to show by argument to the best explanation that the representation of spatial relations and the mechanisms of language generation are based on the direct use of words at the most abstract level. Much remains to be done. Just what a word is at these levels needs to be made precise: is "word" shorthand for linguistic resources of any type? Are concepts literally words or is there just a tight association between mental and linguistic units?

More evidence needs to be gathered and more implementation needs to be done before the assertions of this paper can be accepted. Nevertheless I believe that research along the lines outlined here holds the promise of

[6] Prosodic tunes, idioms, conventional phrasings, syntactic constructions, etc.

providing a principled account of the lexical aspects of language.

References

Chomsky, Noam (1976) “Conditions on Rules of Grammar”, Linguistic Analysis 2:303.

Garrett, Merrill (1975) “The Analysis of Sentence Production” in Bever (ed.) The Psychology of Learning and Motivation vol. 9. Academic Press.

Grenon, Pierre & Barry Smith (2004) “SNAP and SPAN: Towards Dynamic Spatial Ontology” Spatial Cognition and Computation (4)1, 69-104.

Levelt, Willem (1989) Speaking: From Intention to Articulation. MIT Press.

Marcus, Mitchell (1981) “A computational account of some constraints on language” in Joshi, Webber & Sag (eds.) Elements of Discourse Understanding, Cambridge.

Marr, David (1976) Artificial Intelligence – a personal view, AIM 355, available at http://courses.csail mit.edu/6.803/pdf/marr.pdf.

McDonald, David, D & Charles F. Greenbacker. 2010. “‘If you’ve heard it, you can say it’ towards an account of expressibility” In Proceedings of the 6th International Natural Language Generation Conference, Trim, Co. Meath, Ireland, 185-189.

McDonald, David (1979) “Steps toward a Psycholinguistic Model of Language Production”, MITAI working paper 193.

McDonald, David (1991) “Issues in the Choice of a Source for Natural Language Generation” Computational Linguistics 19:1, 191-197.

McDonald David (1992) A*n Efficient Chart-based Algorithm for Partial-Parsing of Unrestricted Texts,* proceedings 3d Conference on Applied Natural Language Processing (ACL), Trento, Italy, 193-200.

McDonald, David (1993) “The Interplay of Syntactic and Semantic Node Labels in Partial Parsing”, in the proceedings of the Third International Workshop on Parsing Technologies, August 10-13, 1993 Tilburg, The Netherlands, pp. 171-186; revised version in Bunt and Tomita (eds.), Recent Advances in Parsing Technology, Kluwer Academic Publishers, pgs. 295-323.

McDonald, David & Federica Busa (1994) “On the Creative Use of Language: the Form of Lexical Resources: 7th Intl. Workshop on Natural Language Generation, Kennebunkport, Maine, 81-89.

McDonald, David & Marie Meteer (1990) “The Implications of Tree Adjoining Grammar for Generation”, Proceedings 1st Intl. Workshop on Tree Adjoining Grammar, Dagstuhl, Germany.

Meteer, Marie W. (1992) Expressibility and the Problem of Efficient Text Planning, Pinter, London.

Meteer, Marie, David McDonald, Scott Anderson, David Forster, Linda Gay, Alison Huettner & Penelope Sibun. 1987. Mumble-86: Design and Implementation, TR #87-87 Dept. Computer & Information Science, UMass., September 1987, 174 pgs.

Pustejovsky, James (1995) The Generative Lexicon, MIT Press.

Shieber, Stuart & Yves Schabes (1991) Generation and synchronous tree-adjoining grammar, *Computational Intelligence*, 7(4), 220-228.

Sloman, Aaron (2010) http://www.cs.bham.ac.uk/projects/cosy/research/papers/spatial-prepositions.

Stone, Matthew (2003) “Knowledge representation for language engineering”, in Farghaly (ed.) *Handbook for Language Engineers*, CSLI Publications, 299-366.

Talmy, Leonard (2000) Toward a Cognitive Semantics, Vol. 1, MIT Press.

Wilks, Yorick (1990/2003) “Where am I coming From: The Reversibility of Analysis and Generation in Natural Language Processing. Originally presented 1990 at the Intl. Generation Workshop, Pittsburg, Revised and reprinted 2003 in Nirenburg, Somers, & Wilks (eds.) Reading in Machine Translation, MIT.