Lexical Tuning

Yorick Wilks and Roberta Catizone
University of Sheffield, UK

Abstract

The paper contrasts three approaches to the extension of lexical sense: what we shall call, respectively, lexical tuning; another based on lexical closeness and relaxation; and a third known as underspecification, or the use of lexical rules. These approaches have quite different origins in artificial intelligence(AI) and linguistics, and involve corpus input, lexicons and knowledge bases in quite different ways. Moreover, the types of sense extension they claim to deal with in their principal examples are different as well. The purpose of these contrasts in the paper is to establish the possibility of evaluating their differing claims by means of the current markup and test paradigm that has been recently successful in the closely related task of word sense discrimination (WSD). The key question in the paper is what the relationship of sense extension to WSD is, and its conclusion is that, at the moment, not all types of sense extension heuristic can be evaluated within the current paradigm requiring markup and test.

Introduction

The principal aim of this paper is to discuss what Lexical Tuning (LT) is, broadly defined and selectively practised, and to discuss its relationship to word sense disambiguation (WSD), with the aim of making it, too, quantitatively evaluable as WSD now is within the SENSEVAL regime (Kilgarriff 1998).

Automatic word-sense disambiguation (WSD) is now an established modular task within empirically-based computational linguistics and has been approached by a range of methods (Ide and Veronis, 1999) sometimes used in combination (Wilks and Stevenson, 1998a, 1998b). These experiments are already showing success rates at, or close to, the target ninety-five-per-cent levels attained by established modules like part of speech tagging in the mid-Nineties. Over a few text words Yarowsky has claimed success in the mid-Nineties (1995), and with systems that claim to deal with all text words , Sheffield (Wilks and Stevenson 1998b) and NMSU-CRL now also claim similar figures (REF).

These methods have included some, such as the use of the agent, object etc. preferences of verbs, that go back to those used in the earliest toy AI systems for WSD, such as (Wilks, 1972, 1978). Yet even those toy systems were set up with an explicit recognition that WSD was different in a key respect from tasks like part-of-speech tagging (POS): namely, that lexicons need to adapt dynamically in the face of new corpus input.

The contrast here is in fact quite subtle, as can be seen from the interesting intermediate case of semantic tagging: attaching semantic, rather than POS, tags to words automatically, a task which can then be used to do more of the WSD task (as in Dini et al., 1998) than POS tagging can, since the ANIMAL or BIRD versus MACHINE tags can then separate the main senses of "crane". In this case, as with POS, one need not assume any novelty in the tag set, in the sense of finding in the middle of the task that one needs additional tags. But one must also allow for novel assignments from the tag set to corpus words, for example, when a word like "dog" or "pig" was first used in a human sense. It is just this sense of novelty that POS tagging also has, of course, since a POS tag like VERB can be applied to what was once only a noun, like "ticket". This kind of assignment novelty, in POS and semantic tagging, can be premarked up with a fixed tag inventory, hence both these techniques differ from genuine sense novelty which, we shall argue, cannot be premarked in any simple way.

This latter aspect, which we shall call Lexical Tuning, can take a number of forms, including:

(a) adding a sense to the lexical entry for a word

(b) adding an entry for a word not already in the lexicon

(c) adding a subcategorization or preference pattern etc. to an existing sense entry

and to do any or all of these on the basis of inductive (corpus) evidence. (a) was simulated in the early work just referred to, and (b) was first attempted in Granger (1977). (c) is at first sight more problematical in that it could be argued that it cannot be defined in a theory-free way, since what can be added automatically to a lexical entry on the basis of corpus evidence necessarily depends on the structure of the lexicon to be augmented, e.g. the nature of the features the lexicon contains. This is undoubtedly correct, but the general notion of what lexical tuning is can still be captured in a non-trivial theory-free way by means of the "etc." above: the general notion proposed being one of a very general function mapping an existing lexicon and a corpus to a new (tuned) lexicon.

In practice, the three types above are neither exclusive nor exhaustive, although task (b) may be quite different in nature in that it excludes straightforward use of a well-known technique that appears under many names, such as "lexical rules" (Copestake and Briscoe 1991, Buitelaar 1997), and which strictly falls outside the function, described above, by which new senses of a word are induced from knowing not only a corpus but an existing lexical entry. This tradition of extending dictionary entries independently of corpus context can be seen as a direct inheritor of the generative linguistics tradition, in the sense in which that is now often contrasted with the corpus linguistics tradition. We shall argue below that this is not altogether fair, since LR researchers do often refer to and call upon corpora, but always that special set of corpora that should more properly be described as meta-corpora, namely the resources of facts about usage, such as (machine readable) dictionaries, thesauri and wordnets. Note that these are all machine readable and the difference here is not about computation, only about where one seeks one's evidence and to what extent all corpus forms can be accounted for in advance, and within lexical constructions.

Combining these three types of potential LT is also a source of potential confusion: if a word is unknown to a lexicon, then any computational system can see that immediately, but many would say (clinging firmly to the force of the word "homonymy") that the three main senses of "post" (post1 = mail; post2 = stake; post3 = job) are, in effect, different words, arbitrarily linked by English spelling. So, some would say that inferring a new sense of "post" (if using a lexicon from which one of the three above senses was missing) is identical to task (b) above that of interpreting new words, and not properly task (a), since one could not expect to induce a new, major, sense of "post" from its existing senses, by any system that could so extend senses in cases of so-called "regular polysemy" (Copestake and Briscoe, 1991).

This problem is independent of a more general one affecting tasks (a) and (c): namely, when does a new context for a word give rise to a description that should be deemed a new feature or new pattern, rather than a 'relaxed' version of an existing one. This is, like all forms of the problem, general learning problem and a matter in the end of arbitrary choice or parameter setting within an algorithm. In this part, we shall not always distinguish clearly between accomodating or interpreting a novel use, which night be unique and one-off, and covering a new sense, already well established, but not already in the lexicon in use. Obviously, the latter always start as the former, and the difference is only a quantitative one. A recent example in English that would repay corpus investigation, would be the verb target, as in:

The bomber targeted the city

and:

The Government targeted the poor

The Government targeted poverty.

The principal use, we are sure, is now the third, not the first, and this has been achieved by a form of meaning reversal, of "seeking out, to help", although the first sense probably persists, paradoxically, in the second example. Somehow, a new sense has been established for the third, not on the basis of semantics (poverty/the poor) but if we suggest, knowledge structures to do with the ideology of modern politics.

To summarise: this formulation of LT assumes we already have a human-created resource we shall call structure1, i.e. the lexicon we started with, perhaps together with an associated knowledge base or ontology. LT is thus the process or mapping function:

I: structure1 + corpus $\rightarrow$ structure2

which indicates that an earlier state of the structure itself plays a role in the acquisition, of which structure2 is then a proper extension (capturing new concepts, senses etc). This is a different

model from the wholly automatic model of lexicon acquisition often used in, say, TIPSTER related work (Lehnert et al, 1990), which can be written:

II: corpus → structure

Here one does not update or "tune" an existing lexicon but derives one directly and automatically from a corpus. There is no doubt process II. can be an effective tool, certainly in the case of unknown languages or domains, but the assumption made here about the quite different function I. is that we cannot understand the nature of the representation of meaning in lexicons, or elsewhere, unless we can see how to extend lexicons in the presence of incoming data that does not fit the lexicon we started with. The extension of representations, one might say, is part of an adequate theory of representation.

Evaluating WSD and its relationship to Lexical Tuning

A central issue in any application of empirical methods to computational linguistics is the evaluation procedure used, which is normally taken to consist in some form of experiment using premarked-up text divided into training and (unseen) test portions. Standard supervised learning for WSD involves an algorithm that selects tags to each text word (or more often each content, or open-class, word) corresponding to senses from a lexicon. Ideally, this process would result in one, and only one, tag per word, but it should at least reduce the set from what is in the lexicon.

Apart from the well-known problem of the difference between sense-sets (if we can continue to use that phrase unexamined, for the moment) for a given word in different lexicons — although they are not arbitrarily different, and that is a vital fact— there are problems concerned with subjects having difficulty assigning a corpus word occurrence to one and only one sense tag during the markup phase.

Kilgarriff (1993) has described such problems, though his figures suggest the difficulties are probably not as serious as he claims (see (Wilks, 1997)). However, we have to ask what it means to evaluate the process of Lexical Tuning as defined above . It seems to require annotating in advance a new sense in a corpus that does not occur in the reference lexicon. The clear answer is that, on the description of WSD markup given above, the sense extension (task (1) above: tuning to a new sense) *cannot* be pre-tagged and so no success rate for WSD can possibly exceed [100% MINUS the percentage of extended sense occurrences].

One question about Lexical Tuning that is not often discussed is made explicit by the last expression: what is the percentage of senses needing tuning in normal text? One anecdotal fact sometimes used is that, in any randomly chosen newspaper paragraph, each sentence will be likely to have an extended sense of at least one word, usually a verb, which is to say a use that breaks conventional preferences (Wilks 1972) and which might therefore be considered extended or metaphorical use, and which may or may not be in a standard lexicon. This is a claim that can be easily tested by anyone with a newspaper and a standard dictionary.

That, even if true, does not give us any firm figure to work with. However, it could suggest that any figure for basic WSD for general text of over 95% must be examined with great care, because it almost certainly cannot have been done by any method using pre-tagging (to a set of existing senses). The onus on anyone making a very high claim is to show what the alternative explanation of his high success figures is. Subsequent examination of machine WSD output for a posteriori "satisfactoriness" can never be a plausible measure: i.e. anything along the lines of "this is what our system gave as new sense contents for this corpus and we liked what we got"!

Another possibility, that will be considered in more detail later, is that novel sense might be detected by an occurrence that cannot be identified by the human tagger with any of the list of senses for the word. The problem here may be just an inadequate dictionary list, though novelty is always with respect to the state of a lexical structure. Also, this procedure will conflate regular novelty, that could have been produced by LR, with any other kind. However, none of these objections are insuperable and, indeed, (Kilgarriff 2000) used such a measure in an attempt to evaluate the Generative Lexicon (Pustejovsky 1995) approach to lexical novelty. On a small sample, Kilgarriff estimated the occurrence of novel senses at 2% over and above anything due to regular polysemy.

How then to evaluate Lexical Tuning claims?

If Lexical Tuning (alias LT) is a real phenomenon, it must be possible to evaluate it in some reasonable way. To make headway here, let us first set out possible basic paradigms or methods for sense extension and seek for clues as to how to evaluate them. One such early paradigm was set out in (Wilks 1978) under the title "Making preferences more active", and which was implemented at the "toy" levels of that period, though it may still be false as to the relationship of new senses to existing ones. Let us call that historical example:

Method A: It was based on the notion of:

i The cuing function (for LT) of the preference failure of a word W1 in a text (e.g. a verb used with an unexpected agent or object class);

ii The location of another word (sense) W2 in a knowledge structure, that defines how the world for that word sense normally is, and which has the right lexical preferences;

iii The substitution in the text representation of the "failed" word by a new, more satisfactory word sense W2 (in the lexicon) ;

iv The claim that W1 should have its lexicon extended by the structure for the appropriate sense of W2.

where such appropriate structure may mean its preferences, subcategorization patterns, semantic or other links, explanatory gloss etc.

The main 1978 example was "My car drinks gasoline", which has a failed [human] agent preference, which is then (criterion i above) the trigger to locate a fact representable as

[cars use gasoline]

in a knowledge base about cars (ii and iii above), so that "use" can provide a plausible new sense of "drink" (iv above). However, this heuristic not wholly satisfactory, since it does not capture the idiomatic force of "drink" $\rightarrow$ "use a lot of" implicature of this usage. Moreover, the process must not just locate any action or process of cars associated with gasoline, for that will include "leak", as in

[cars leak gasoline].

We can suppose this is achieved either (or both) by assuming leaking gasoline is not described in a stereotypical car function knowledge base or that drink/use are linked by some underlying semantic structure (such as a shared type primitive or some degree of closeness, however defined, in a synonym/WordNet list classification) and in a way that drink/leak are not.

This location of a preference-satisfying KB entity to substitute for a failing semantic structure was called *projection* in 1978, and is the kind of inference that has played a substantial role in the later work of Pustejovsky and others under names like "coercion". The method illustrated above based on "preference failure" would apply only to verbs and adjectives, which were the grammatical types coded with preferences in that system, although another possibility set out in the 1978 paper was that either participant of the failed preference link could be substituted by something better fitting (ie. the verb or its agent): the sense extension proposed above is of the action because of what was in the knowledge base (KB), and within the standard AI assumption of knowledge-based processing, but one could also take the same example as ascribing a human quality to cars. However, the KB does not support any substitution based on the agent, because one would expect to locate in the car-KB forms like [person drive cars], but not any KB form like like [person drink gasoline], which is what would be needed to support an alternative, competing, tuning of "car".

Method A2: However, this last sort of possibility is the one that underlies a metonymic preference failure like

THE CHAIR opened the meeting.

Again we have agent-action failure, but now there is no KB support for any form with a value for ACTION satisfying [chair ACTION meeting) of the kind we saw for drink/use. However, in a standard KB, there would be support for

[person open meeting]

as part of a general knowledge structure for the conduct of meetings, which satisfy the preferences of the corresponding sense of "open". So, in this class of case as well, we might expect the same procedures to tune to a new sense of "chair" as "person" (who opens meetings).

Now let us contrast the above paradigm for sense extension with that used in recent CRL

work (REF), one intended as more fine grained than the "consumer driven" (Sergei Nirenburg's term) approach, or that of "final task" driven projects, such as the ECRAN project, namely that of carrying out a "final task" such as Information Extraction, before and after tuning a lexicon against a domain corpus and then seeing if Information Extraction results are improved. "Final task" here is to be contrasted with "intermediate tasks", such as WSD, which are often evaluated directly in competitions but which have no real NLP function outside some final task, one that serves a purpose for a language understander or consumer.

The CRL basic methodology (using the Mikrokosmos KB, which we shall call MK for short, Nirenburg and Raskin 1996) is quite different from A above. Let us (at the inevitable risk of error in summarising someone else's work) describe its two proposals as follows:

Method B1:

1. Locate preference failure of an occurrence of word W1 in the corpus
2. Seek the closest *existing* sense of W1 in the MK lexicon by relaxing the preference constraints of W1.
3. Consider later how to subdivide the expanded-relaxed occurrences of W1 to create a new sense if and when necessary, perhaps when the "expanded" occurrences form a new cluster, based on related relaxations, so that a new sense of W1 can be separated off in terms of a new set of constraints in the MK lexicon.

OR

Method B2:

1. Locate preference failure of an occurrence of word W1 in the corpus
2. Seek in the MK KB for a word sense W2 hierarchically below W1, but whose preferences are satisfied in the example.
3. Take W2 to be the sense of W1 in the given context.

It is not wholly clear in the context of the paper referred to whether B1 and B2 result in adaptations to the lexicon, which is what we are asking as the minimal, necessary, condition for anything to be called LT, so as to avoid including in LT all hapax occurrences of unusual conjunctions. However, these heuristics are of interest whether or not the lexicon is permanently adapted, as opposed to deriving a new sense representation for a word for immediate use. These methods make less radical use of world knowledge than A, but one which runs far less chance of making wrong extensions. The key notion in B1 is the search for a *closest existing sense* of the same word, which may well represent a core aspect of meaning extension missing from the earlier approach, and which will in any case be essential to task (c) (though it cannot, by definition, be used for task (b) which is that of the "unknown word"). It also cannot help in true homograph/homonym cases, like "post", where the approach A might stand a chance, but we proposed at the beginning to exclude consideration of such extension for now, or rather to accommodate it to task (b) and not (a).

Method B2 shows an interesting notion of preference breaking somewhat different from that of A: a canonical CRL example is:

He PREPARED the bread.

where the declared aim of the adaptation (REF) is to tune the sense of "prepare" , for this occurrence, to the appropriate sense of "bake", which is the verb in the Mikrokosmos KB for the preparation of bread and whose preferences fit a BREAD object as those of "prepare" do not. The process here is close to Method A in that a stored item in a KB licenses the tuning and, again like Method A, the result is the substitution of one word sense by the sense of another word. As with method A, this will only count as LT (on the strict definition used in this paper) if the lexicon is changed by this process so as to install "bake" as a sense of "prepare" and it seems this is not done in the CRL system.

However, the most interesting feature of the B method, is that the constraint satisfaction of "bake" is not passed up the hierarchy of actions and sub-actions. This is an idea going back to Grice (as a failure of the quantity maxim, Grice 1964), but one possibly original in lexical semantics: that a too general concept is semantically ill-fitting, just as a complete misfitting of a concept is. In preference terms, it means that the overly general is also a preference failure, quite contrary to the way that notion has usually been used to include subclasses of fillers, e.g. that to prefer a FOOD object is normally to accept a BREAD

object, given that bread is a kind of food. The current suggestion is equilvalent to: if a concept prefers BREAD, then FOOD would be ill-fitting.

As we noted, Method B2 is not LT if the lexical entry for "prepare" is not altered by the acceptance of "He prepared the bread", but this is mere definition. Relaxation to "higher classes" can, however, be explicitly marked in a lexicon, and is therefore LT, as would be the case with "The Chair opened the meeting" example, if viewed as relaxation to accept PHYSOBJ agents and not just HUMAN ones. There is always a price to pay in relaxation accounts of tuning because once a preference is relaxed it cannot subsequently be used to select as a constraint.

Consider the following two sentences:

> The whole office waited for the boss to arrive
>
> The two men cleaned the offices as ?they waited for the janitor to arrive.

One cannot both relax the lexical entry for "wait" so as to accommodate its agent in the first sentence and use the standard preferences of "wait" for [human] agents to resolve ?they in the second. This point is an argument not only against relaxation but against any method for deriving preferences by corpus analysis (Grishman 1986, Resnik, 1993) in any simple manner since both sentences could well be attested in the same corpus.

There will naturally be disputes about how widely this kind of quantity restriction can be enforced: one might also say that preparing bread is a sequence of subactions, including mixing and leaving-to-rise (rather like Schank scripts of old, Schank and Abelson, 1977); in which case the type BREAD is the proper object for all of them including "prepare", so that the B methods can never be called in because there is no preference failure to trigger them.

Method B1 should lead to a quite different interpretation of this example: on B1 "prepare bread" (if deemed preference breaking as CRL claim) should lead to a relaxation to an *existing* sense of "prepare" (and not "bake" at all), yet what is that existing sense?

Is the car/drink example (Method A) one of lexical extension when compared to the B methods; which is to say, do we want to deem "use" a sense of "drink" in the context of a car's consumption of gasoline and retain that modification in a lexicon, or is it simply a one off metaphor? Identifying this as a possible extension is a necessary but not sufficient condition for a full LT lexicon modification which requires further confirming instances of entities of a machine type drinking fuel-like liquids, e.g. steam engines drinking water, aeroengines drinking kerosene and so on. This is a different type of extension from the B-type examples involving possible relaxations of agents and objects of verbs already fixed in hierarchies. Both A and B type extensions, if real, are different from what others are calling regular polysemy, in that they cannot be precoded in lexical entries by rules or any similar method.

Closest sense heuristics and text markup

The CRL approach measures success, at least initially, by human mark up to the closest existing lexical sense (though see below on "Chateaubriand"). This may make it possible to achieve a generally acceptable type of evaluation procedure for lexical tuning (whether or not one adapts the lexicon, in the face of any particular example, need not affect the use of the procedure here) if there can be inter-subjective agreement on what is a lexically closest sense in a training text. That would then be the phenomenon being tested, along with the general (and attested) ability to assign a sense to a text word when the sense in the lexicon is used, though the human marker should also obviously have the choice of declining to mark a closest sense, given a particular state of the lexicon, if he believes it inappropriate in the context. If LT is to be evaluated in such a way, a marker will have to be able to indicate closest sense separately from appropriate sense.

Examples can be produced (due in this case to Steve Helmreich) within the well-known Restaurant Metonymy example paradigm to suggest that the extended sense to be constructed by this Method B1, leading to the closest existing sense, may not always be appropriate.

Consider:

> The Chateaubriand wants a drink

where "Chateaubriand" is lexically coded both as a steak (what the diner ordered) and an Eighteenth Century French politician of that name. The latter may well be chosen (by an algorithm, though not by a human marker, of course) as the closest sense (since it satisfies the [human] agent

constraint on "want") but the extended or relaxed sense should actually be related to steak, the first sense.

Restaurant Metonymies (RMs), though attested, have perhaps played too strong a role in the field, given their infrequency in real life. Proper name RMs could perhaps be dismissed as a tiny subclass of a tiny subclass and a proper subject only for AI. Perhaps the closest sense heuristic can be saved by some careful analysis of "the" in the last example; it is always the cue for a Restaurant Metonymy, but rarely in politics, and we shall assume in what remains that the heuristic can be saved in some such way. After all, there need be no similar problem here with "standard" RMs that are not proper names, as in:

> The lasagna wants a drink.

Pustejovsky's position on lexical expansion

In The Generative Lexicon (1995, TGL for short) Pustejovsky (JP for short) sets out a position that has features in common with work already described, but offers a distinctive view of the lexicon and in particular its underspecification in crucial respects; and the aspect that will concern us in this paper is whether or not that underspecification is any form of LT as described here, as implying the augmentation of the lexicon in the face of sense novelty in a corpus. It seems JPs position is that his key class of examples does not imply the creation of a new sense from an existing one in the face of corpus evidence, but is rather the incorporation of a prestored ambivalence within a lexical entry. That this can be misunderstood can be seen from an attack on JPs TGL by Fodor and LePore (FL for short, Fodor and Lepore, 2000) in which they attribute to him a sense ambiguity for such examples, and indeed an unresolvable one.

They claim that JP's:

> He baked a cake.

is in fact ambiguous between JP's "create" and "warm up" aspects of "bake", where baking a cake yields the first, but baking a potato the second. JP does not want to claim this is a sense ambiguity, but a systematic difference in interpretation given by inferences cued by features of the two objects, which could be labels such as ARTIFACT in the case of the cake but not the potato:

> But in fact, "bake a cake" is ambiguous. To be sure, you can make a cake by baking it; but also you can do to a (preexistent) cake just what you can do to a (preexistent) potato: viz. put it in the oven and (non creatively) bake it. (FL. p.7)

¿From this FL conclude that "bake" must be ambiguous, since "cake" is not. But all this is absurd and untrue to the simplest facts about cakes, cookery and English. Of course, warming up a (preexistent) cake is not baking it; who ever could think it was? That activity would be referred to as warming a cake up, or through, never as baking. You can no more bake a cake again, with the other (warm up) interpretation, than you can bake a potato again and turn it into a different artifact. The only obvious exception here might be "biscuit", whose etymology is, precisely, "twice cooked", though not baked.

JP has resisted augmentation of the lexicon, though other researchers would probably accept it and this difference may come down to no more than the leaving of traces in a lexicon and what use is made of them later, and where augmentation of the lexicon would be appropriate if such cases became statistically significant. However, we can still ask whether underspecification is just language-specific lexical gaps?

Let us look at the key Pustejovsky example in a new way: the bake cake/bread/potato examples may not draw their power from anything special to do with baking but with lexical gaps and surplus in English connected with cake and bread. Suppose we adopt, just for a moment, a more Wierzbickian approach to baking and assume as a working hypothesis that there is only one, non-disjunctive, sense of bake and it is something like:

> "to cook a food-substance X in a heated enclosed space so as to produce food-substance Y"

Thus we have, for X and Y for our standard food substances in English:

> **potato** [potato, baked potato]
>
> **bread** [dough, bread]
>
> **cake** [cake mixture, cake]
>
> **pie** [pie, pie]

as well as:

> **fish** [fish, (baked) fish]
>
> **ham** [ham, baked ham]

There is no mystery here, but looking at a range of well-known substances can take us out of the rather airless zone where we discuss the relationship of "bake" and "prepare" away from all data, and without considering in parallel "roast", "boil", "grill" etc. We would argue that there is no pressing need to gloss the implicit structure here as a disjunction of senses or aspects of "bake". It is simply that the lexical repertory of English varies from food to food, thus

> We bake ham and get baked ham
> We bake dough and get bread
> We bake cake mixture and get cake
> We bake (a) potato and get a (baked) potato

There is no reason to believe that these cases fall into two classes, the creative and non-creative at all: it simply that we have words in English for baked dough (bread) and baked cake mixture (cake) but not a word for a baked potato. If we did have such a word, baking a potato would seem more creative than it does. Contrast Kiswahili (or Japanese), which has a word for uncooked rice (mchele) and a word for cooked rice (wali). In English

> We cooked rice

does not seem creative but rather matter of mere heating since there is only the same word for the object when transformed. But, on an underspecification a pproach to Kiswahili:

> We cooked wali/mchele

are two sentences (if all in Kiswahili) bearing two differing interpretations of "cook", only one of them TELIC, and hence

> We cooked rice

should also be ambiguous/underspecified/disjoined in interpretation in English. But this is surely not true, indeed absurd, and a language cannot have its verb semantics driven by its lexical gaps for nouns! If this analysis is plausible there is no disjunction present at all in baking cakes and potatoes either, and if by chance "dough" meant dough or bread in English (as is surely the case in some language) this whole issue could never have arisen.

We should not exaggerate the differences between the main approaches discussed so far: all subscribe to

i. sense is resolvable by context

ii. we can create/extend sense in context by various procedures

but not all to

iii. the methods of (ii) are close to WSD and lead naturally to lexical adaptation/tuning

iv. the adaptation produced by (ii) leaves some record in the lexicon.

Generalising the contrast of LT approaches with Lexical Rules (LR)

Lexical Tuning (LT) is closely related to, but rather different from, a group of related theories that are associated with phrases like "lexical rules" (LR); all of the latter seek to compress lexicons by means of generalisations, and we take that to include DATR (Evans and Gazdar 1996), methods developed under AQUILEX (Copestake and Briscoe 1991), as well as Pustejovsky's TGL discussed above and Buitelaar's more recent research on underspecified lexicons (1997). LR we take to be any approach, such as Pustejovsky or Briscoe, in the tradition of Givon that seeks to extend lexical entries independently of corpora. To take a classic example, lexical entries for animals that can be eaten can be contracted and marked only ANIMAL, given a rule that extends on demand to a new sense of the word marked MEAT. This is an oversimplification of course, and problems arise when distinguishing between eatable and uneatable animals (by convention if not in strict biology). Very few want to extend "aardvark" with MEAT though there is no logical reason why not, and an ethnographic survey might be needed for completeness in this area; foods are simply not universal.

All this work can be traced back to early work by Givon (1967) on lexical regularities, done, interestingly to those who think corpus and MRD research began in the 1980s, in connection with the first computational version of Websters Third at SDC in Santa Monica under John Olney in 1966. It can also can be brought under the heading "lexical compression" whether or not that motive is made explicit. Givon became interested in what is now called "systematic polysemy", as distinguished from homonymy which is assumed to be unsystematic: his key examples were "grain"

which is normally given a count noun or PHYOBJ sense in a (diachronic) dictionary and cited earlier than the mass noun sense of "grain in the mass". This particular lexical extension can be found in many nouns, and resurfaced in Briscoe and Copestake's "grinding rule" (1989) that added a mass substance sense for all animals, as in their "rabbit all over the road" example. The argument was that, if such extensions were systematic, they need not be stored individually but could be developed when needed unless explicitly overridden. The paradigm for this was the old AI paradigm of default reasoning: Clyde is an elephant and all elephants have four legs BUT Clyde has three legs, and the latter fact must take precedence over the former inference. It has been some thing of a mystery why this foundational cliche of AI was greeted later within computational linguistics as remarkable and profound.

Gazdar's DATR (Evans and Gazdar, 1996) is the system that makes lexical compression the most explicit, drawing as it does on fundamental notions of science as a compression of the data of the world. The problem has been that language is one of the most recalcitrant aspects of the world and it has proved hard to find generalisations above the level of morphology in DATR; those to do with meaning have proved especially elusive. Most recently, there has been an attempt to generalise DATR to cross-language generalisations which has exacerbated the problem. One can see that, in English, Dutch and German, respectively, "house", "huis" and "Haus" are the "same word", a primitive concept DATR seems to require. But, whereas "house" has a regular plural, "Haus" ("Haeuser") does not, so even at this low level, significant generalisations are very hard to find.

Most crucially, there can be no appeals to meaning from the concept of "same word": "town" (Eng.) and "tuin" (Dut.) are plainly the same word in some sense, at least etymologically and phonetically, and may well obey morphological generalisations although now, unlike the "house" cases above, they have no relation of meaning at all, as "tuin" now means garden in Dutch, unless one is prepared to move to some complex historical fable about towns and gardens being related "spaces surrounded by a fence". There has been no attempt to link DATR to established quantitative notions of data compression in linguistics, like Minimum Description Length (Risannen 1984), which gives a precise measure of the compaction of a lexicon, even where significant generalisations may be hard to spot by eye or mind.

The systems which seek lexical compression by means of rules, in one form or another, can be discussed with particular attention to Buitelaar, since Briscoe and Pustejovsky differ in matters of detail and rule format but not in principle. Buitelaar continues Pustejovsky's campaign against the "unstructured list" view of lexicons: viewing the senses of a word merely as a list as dictionaries are said to do, in favour of a clustered approach, one which distinguishes "systematic polysemy" from mere homonymy (like the ever present senses of "bank").

Clustering a word's senses in an optimally revealing way is something no one could possibly object to, and the problem here is the examples Buitelaar produces, and in particular his related attack on WSD programs (including the present authors) as assuming a list-view of sense, is misguided. As Nirenburg and Raskin (1997) have pointed out in relation to Pustejovksy, those who criticise list views of sense then normally go on in their papers to describe and work with the senses of a word as a list, and Buitelaar continues this tradition. Moreover, it must be pointed out that opening any modern English dictionary, especially one for learners like LDOCE, shows quite a complex grouping of the senses it contains and not a list at all.

Buitelaar's opening argument against standard WSD activities rests on his counter-example where multiple senses of "book" must be kept in play and so WSD cannot be done: the example is:

A long book heavily weighted with military technicalities, in this edition it is neither so long nor so technical as it was originally.

Leaving aside the question of whether this is a sentence, let us accept that Buitelaar's list of possible senses (and glosses) of "book" is a reasonable starting point (with our numbering added):

(i) the information content of a book (military technicalities);

(ii) its physical appearance(heavily weighted),

(iii) and the events involved in its construction (long) (ibid. p. 25).

The issue, he says, is to which sense of "book" does the "it" refer, and his conclusion is that it cannot be disambiguated between the three.

This seems to us quite wrong, as a matter of the exegesis of the English text: "heavily weighted" is plainly metaphorical and refers to content (i) not the physical appearance (ii) of the book. We have no trouble taking "long" as referring to the content (i) since not all long books are physically large; it depends on the print size etc. On our reading, the "it" is univocal (to sense (i)) between the senses of "book". However, nothing depends on an example, well or ill-chosen, and it may well be that there are indeed cases where more than one sense must remain in play in a word's deployment; poetry is often cited, but there may well be others, less peripheral to the real world of the *Wall Street Journal*.

The main point in any answer to Buitelaar must be that, whatever is the case about the above example, WSD programs have no trouble capturing it: many programs, and certainly that of (Stevenson and Wilks, 1998a) that he cites and its later developments, work by casting out senses and are perfectly able to report results with more than one sense still attaching to a word, just as many part-of-speech taggers result in more than one tag per word in the output. Historians of the AI approach to NLP will also remember that Mellish (1983), Hirst (1987) and Small (1988) all proposed methods by which polysemy might be computationally reduced by degree and not in an all or nothing manner. Or, as one might put it, underspecification, Buitelaar's key term, is no more than an implementation detail in any effective tagger!

Let us turn to the heart of Buitelaar's position: the issue of systematicity (one within which other closely related authors' claims about lexical rules can be taken together). Buitelaar lists clusters of nouns (e.g. blend, competition, flux, transformation) that share the same top semantic nodes in some structure like a modified WordNet (which would be act/evt/rel in the case of the list just given).

Such structures, he claims, are manifestations of systematic polysemy. But what is one to take that to mean, say by contrast with Levin's (1993) verb classes where, she claims, the members of a class share certain syntactic and semantic properties and, on that basis, one could in principle predict additional members? That is simply not the case here: one does not have to be a firm believer in natural kinds to see that the members of the cluster above (i.e. blend etc.) have nothing systematic in common, but are just arbitrarily linked by the same "upper nodes" in Wordnet. Some such classes are natural classes, as with the one Buitelaar gives linked by being both animate and food (all of which, unsurprisingly, are animals and are edible, at least on some dietary principles), but there is no systemic relationship here of any kind. Or, to coin a phrase, one might say that the list above is just a list and nothing more!

In all this, we intend no criticism of his useful device, derived from Pustejovsky, for showing disjunctions and conjunctions of semantic types attached to lexical entries, as when one might mark something as (act AND relation), or an animal sense as (animate OR food). This is close to older devices in artificial intelligence such as multiple perspectives on structures (in Bobrow and Winograd's KRL 1977), and so on. Showing these situations as conjunctions and disjunctions of types may well be a superior notation, though it is quite proper to continue to point out that the members of conjuncts and disjuncts are, and remain, in lists!

Finally, Buitelaar's (1998) proposal to use these methods (via CoreLex) to acquire a lexicon from a corpus may also be an excellent approach, and one of the first efforts to link the LR movement to a corpus. It would probably fall under type II. acquisition (as defined earlier), and therefore not be LT, which rests essentially on structural modification by new data. Our point here is that that method (capturing the content of e.g. adjective-noun instances in a corpus) has no particular relationship to the theoretical machinery described above, and is not different in kind from the standard NLP type II. projects of the 1980s like Autoslog (Lehnert et al. 1990), to take just one of many possible examples. In a small but suggestive experiment, Kilgarriff (2000) found it possible to accommodate Generative Lexicon structures, to dictionary senses for which their had not been precoded, for only some of the cases required it.

Vagueness

The critique of the broadly positive position on WSD in this paper, and its relationship to LT, comes not only from those who argue (a) for the inadequacy of lexical sense sets over productive lexical rules (as above) but also from proponents of (b) the inherently *vague* quality of the difference between senses of a given word. We believe both these approaches are muddled *if* their proponents conclude that WSD is therefore fatally flawed as a task.

The vagueness issue is an old one, one that, if

taken seriously, must surely result in a statistical or fuzzy-logic approach to sense discrimination, since only probabilistic (or at least quantitative) methods can capture real vagueness. That, surely, is the point of the Sorites paradox: there can be no rational or qualitative criterion (which would include any quantitative system with clear limits: e.g. tall = over 6 feet) for demarcating "tall", "green" or any inherently vague concept.

If, however, sense sets/lists/inventories are to continue to play a role in language processing and understanding, vagueness can mean no more than highlighting what all systems of WSD must have, namely some parameter or threshold for the assignment to one of a list of senses versus another, or setting up a new sense in the list. Talk of vagueness adds nothing specific to help that process for those who want to assign words, on some quantitative basis, to one sense rather than another; sense tuning means seeing what works and fits our intuitions.

Vagueness would be a serious concept only if the whole sense list for a word (in rule form or not) was abandoned in favour of statistically-based clusters of usages or contexts. There have been just such approaches to WSD in recent years (e.g. Bruce and Wiebe, 1994, Pedersen and Bruce, 1997, Schuetze & Pederson, 1995) and the essence of the idea goes back to Sparck Jones 1964/1986) but such an approach would find it impossible to take part in any competition like SENSEVAL (Kilgarriff, 1998) because it would inevitably deal in nameless entities which cannot be marked up for.

Vague and Lexical Rule based approaches also have the consequence that all lexicographic practice is, in some sense, misguided: on that view, dictionaries for such theories are fraudulent documents that could not help users whom they systematically mislead by listing senses. Fortunately, the market decides this issue, and it is a false claim. Vagueness in WSD is either false (the last position) or trivial, and known and utilised within all quantitative methodologies.

Lexical rules and pre-markup

Can the lexical rules approach to some of the phenomena discussed here be made evaluable, using some conventional form of pre-markup, in the way that we saw is difficult for straightforward LT of new senses, but which may be possible if LT makes use of some form of the "closest sense" heuristic? The relevance of this to the general WSD and tuning discussion is that the very idea of pre-markup would presumably require that all lexical rules are run, so that the human marker can see the full range of senses available, which some might feel inconsistent with the "data compression" motivation behind lexical rules. However, this situation is no different in principle from POS tagging where a language, say English, may well have a tag meta-rule that any word with N in its tag-lexicon could also have the tag ADJ (but not vice versa). Clearly, any such rule would have to be run before pre-markup of text could be done, and the situation with senses is no different, though psychologically for the marker it may seem so, since the POS tag inventory can usually be kept in memory, whereas a sense inventory for a vocabulary cannot.

Conclusion: which of these methods lead to evaluation?

What is the conclusion here on the relationship of lexical extension, in whatever form, to the task of WSD, given that the thrust of the paper has been to see if new evaluable methods of WSD apply to LT, and can be adapted to make it evaluable too? It is clear that the LR approach, at least as represented by Buitelaar, sees no connection and believes WSD to be a misleading task. And this is not a shocking result, for it only brings out in a new way the division that underlies this paper, and which is as old as the generative vs. corpus linguistics divide, one that has existed for decades but was effectively disguised by the denial by the dominant Chomskyan generative paradigm that anything akin to corpus linguistics existed.

Our reply to this is that Buitelaar's examples do not support his attack on WSD, since underspecification is largely a misnomer. Corpora could be premarked for the senses coded in such a lexicon, if treated as real disjunctions, but there is no way of knowing which of these are attested or attestable in data, and we argued that aspects of the key example "bake" are not in fact related to sense distinction or polysemy phenomena at all.

On the other hand, the method A (1978) phenomena are impossible to premark and therefore could be tested only within a final task like IE, IR or MT. The relaxation phenomena of method B (1997), on the other hand, could possibly be premarked for (and then tested as part of a WSD program) but by doing so do not constitute ex-

tended sense phenomena, in the full sense of LT, at all, since by relaxing to an existing sense one denies a new sense is in operation. In the B2 type cases, with data like that of the LR researchers, the extension of "prepare" to "bake" (of bread) should result in the representation of "bake" being added as possible sense of "prepare" (by analogy with Method A) whether or not this effects a one-off or permanent (LT) adaptation.

A further empirical possibility would be to remove a principal sense from a lexicon for a set of words, and investigate "negative markup" (along the lines of (Kilgarriff 2000)) to see whether and when markers were unwilling to assign a sense, although a problem here might be their unconscious knowledge of the missing sense. In parallel, it would be possible to deploy a large general, WSD program (like Wilks and Stevenson 1998b) to see if, with a pruned lexicon, it was unable to assign a sense in the same set of cases. An alternative might be to apply a number of independent WSD programs for such cases and see if the missing senses correspond to significant disagreement in the programs' output.

There is some evidence for the positive evaluation of tasks like WSD and LT within what we have called "final" (as opposed to intermediate) tasks: within the SPARKLE project Grefenstette (1998) produced a form of lexical augmentation that improved overall information retrieval precision and recall by a measurable amount. It is most important to keep some outcome like this in mind as an active research goal if the markup paradigm is seen to be impossible for LT, because our aim here is to separate clearly evaluable approaches from weaker notions of computer-related lexical work.

Acknowledgments

The authors are much indebted to comments from Evelyne Viegas, Steve Helmreich and Wim Peters, but the errors are their own, as always.The work has been supported by the UK EPSRC grant MALT, and the EC IST HLT projects ECRAN and NAMIC.

Bibliography

Bobrow, D.G. and Winograd, T.(1977) An Overview of KRL, a Knowledge Representation Language, Cognitive Science 1, pgs.3-46.

Bruce, R. and Wiebe, J. (1994) Word-sense disambiguation using decomposable models, Proc. ACL-94.

Buitelaar, P. (1998) Corelex: Systematic Polysemy and Underspecification, Dept. of Computer Science, Brandeis University, Boston.

Buitelaar, P.(1997) A Lexicon for Underspecified Semantic Tagging, Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics: What, why and how?", April, pgs 25-33, Washington, D.C..

Copestake, A. and Briscoe, T.(1991) Lexical operations in a unification-based framework, Proceedings of ACL SIGLEX Workshop, Berkley.

Dini, L., di Tommaso, V. and Segond, F. (1998) Error-driven word sense disambiguation. In Proc. COLING-ACL98, Montreal.

Evans, R. and Gazdar, G.(1996) DATR: A Language for Lexical Knowledge Representation, Computational Linguistics, vol. 22, no.2, pgs. 167-216.

Fodor, J. and Lepore E. (2000). The emptiness of the Lexicon: critical reflections on J. Pustejovsky's The Generative Lexicon. In Bouillon and Busa (eds.) Meaning and the Lexicon. New York: Crowell.

Givon, T. (1967) Transformations of Ellipsis, Sense Development and Rules of Lexical Derivation. SP-2896, Systems Development Corp., Sta. Monica, CA.

Granger, R.(1977) FOULUP: a program that figures out meanings of words from context, Proceedings of the Fifth Joint Conference on AI.

Grefenstette, G. (1998) The Problem of Cross Language Information Retreival in Cross Langague Information Retreival, G.Grefenstette (Ed.), Luwer Academic Publishers, Boston.

Grishman, R. (1986) Analyzing Language in Restricted Domains: Sublanguage Description and Processing, Eds. Grishman, R. and Kittredge, R, Lawrence Erlbaum, Hillsdale, New Jersey.

Hirst, G. (1987) Semantic Interpretation and the Resolution of Ambiguity, CUP, Cambridge, England.

Kilgarriff, A.(1998) SENSEVAL, An Exercise in Evaluating Word Sense Disambiguation Programs, Proceedings of the First International conference on Language Resources and Evaluation,Granada, Spain.

Kilgarriff, A. (2000) Generative lexicon meets corpus data: the case of non-standard word uses, In Word meaning and Creativity, Cambridge University Press, (Ed.) P. Bouillon and F.Busa.

Lehnert, W., Cardie, C., Fisher, D., McCarthy, J. and Riloff, E. (1990) Description of the (CIRCUS) System as Used for MUC-4, Proceedings of the Fourth Message Understanding Conference (MUC-4), pgs. 282-288, Morgan Kaufmann.

Levin, B.(1993) English Verb Classes and Alternations, Chicago, Il.

Mellish, C. (1983) Incremental semantic interpretation in a modular parsing system. In Karen Sparck-Jones and Yorick A. Wilks (eds.) Automatic Natural Language Parsing, Ellis Horwood/Wiley, Chichester/NYC.

Nirenburg, S., Beale, S., Mahesh, K., Onyshkevych, B., Raskin, V., Viegas, E., Wilks, Y., and Zajac, R. (1996) Lexicons in the Mikrokosmos Project, Proceedings of the Artificial Intelligence and Simulated Behavior Workshop on Multilinguality in the Lexicon, Brighton, UK.

Nirenburg, S. and Raskin., V. (1997) Ten choices for lexical semantics. Research Memorandum, Computing Research Laboratory, Las Cruces, NM.

Pedersen, T. and Bruce, R. (1997) Distinguishing Word Senses in Untagged Text, Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, pp. 197-207, Providence, RI.

Pustejovsky, J. (1995) The Generative Lexicon, MIT Press: Cambridge, MA.

Resnik, P.(1993) Selection and Information: A Class-based Approach to Lexical Relationships, University of Pennsylvania.

Risannen J. (1984) Universal coding, information, prediction and estimation, IEEE Transactions on Information Theory, IT-30(4), pp.629-636.

Schank, R. and Abelson R.P. (1977). Scripts, Plans, Goals and Understanding, Hillsdale NJ: Lawrence Erlbaum.

Schutze, H. and Pederson, J. (1995) Information Retrieval based on Word Sense, Proc. Fourth Annual Symposium on Document Analysis and Information Retrieval. Las Vegas, NV.

Small, S., Cottrell, G., and Tanenhaus, M. (Eds.) (1988) Lexical Ambiguity Resolution, Morgan Kaufmann: San Mateo, CA.

Sparck Jones, K. (1964/1986) Synonymy and Semantic Classification. Edinburgh UP: Edinburgh.

Wilks, Y. (1968) Argument and Proof. Cambridge University PhD thesis.

Wilks, Y.(1972) Grammar, Meaning and the Machine Analysis of Language, Routledge, London.

Wilks, Y.(1978) Making Preferences More Active, Artificial Intelligence, vol.11, no.3, pgs.197-223.

Wilks, Y. (1980) Frames, Semantics and Novelty. In D. Metzing (ed.), Frame Conceptions and Text Understanding, Berlin: de Gruyter.

Wilks, Y. (1997) Senses and Texts. Computers and the Humanities.

Wilks, Y. and Stevenson, M. (1998a) The Grammar of Sense: Using part-of-speech tags as a first step in semantic disambiguation, Journal of Natural Language Engineering, 4(1), pp. 1-9.

Wilks, Y. and Stevenson, M. (1998b) Optimising Combinations of Knowledge Sources for Word Sense Disambiguation, Proceedings of the 36th Meeting of the Association for Computational Linguistics (COLING-ACL-98), Montreal,