INTERNATIONAL JOURNAL OF TRANSLATION
VOL. 22, NO. 1, JAN-JUN 2010

Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics

EDUARD HOVY
Information Sciences Institute, USA

JULIA LAVID
Universidad Complutense de Madrid, Spain

ABSTRACT

Corpus annotation—adding interpretive information into a collection of texts—is valuable for a number of reasons, including the validation of theories of textual phenomena and the creation of corpora upon which automated learning algorithms can be trained. This paper outlines the main challenges posed by human-coded corpus annotation for current corpus linguistic practice, describing some of the methodological steps required for this indispensable part of the research agenda of Corpus Linguistics in this decade. The first part of the paper presents an overview of the methodologies and open questions in corpus annotation as seen from the perspective of the field of Natural Language Processing. This is followed by an analysis of the theoretical and practical impact of corpus annotation in the field of Corpus Linguistics. It is suggested that collaborative efforts are necessary to advance knowledge in both fields, thereby helping to develop the kind of methodological rigour that would bring about a 'science' of annotation.

Keywords: Corpus annotation, tagging, Natural Language Processing, Computational Linguistics, annotation tools.

INTRODUCTION

Corpus annotation, sometimes called 'tagging', can be broadly conceptualized as the process of enriching a corpus by adding linguistic and other information, inserted by humans or machines (or a combination of them) in service of a theoretical or practical goal.

Neither manual nor automated annotation is infallible, and both have advantages. Manual annotation can be performed with relatively little corpus preparation and can be done inexpensively on quite complex phenomena. It is primarily a question of explaining to the annotators the desired annotation scheme and providing the data. But manual work is slow and limited to small results; mounting a large-scale annotation effort that covers tens of thousands of words is a significant undertaking. Automated annotation, on the other hand, requires a considerable investment in corpus preparation and the programming of the automated tagging system, especially if it is first trained on a seed corpus and then applied to a larger one. Its results may be of poor quality. But it is fast, and can produce outputs over very large corpora of all types in very little time.

The task of automating the annotation of various kinds of grammatical, lexical, and syntactic information has been achieved with a reasonable degree of accuracy with computational systems including part-of-speech taggers (with accuracy over 96% for English and some other languages) and parsers (with crossing bracket measure accuracy of over 92% for English and some other languages). However, the task of automating the annotation of higher levels of linguistic processing (e.g., semantic, pragmatic, or discourse categories) for use in applications such as Information Extraction (IE), Information Retrieval (IR), Automated Text Summarization, Machine Translation, among others, is a complex one. It requires manual annotation first, to produce a small corpus on which the computer algorithms can be trained. As explained in Section 1.1 below, this can only be achieved through a well-designed and reliable annotation procedure.

Though complex, this general human-followed-by-machine procedure has met with some success over the past decade in the Computational Linguistics (CL) / Natural Language Processing (NLP) communities around the world. The Penn Treebank (Marcus, Marcinkiewicz & Santorini 1993), for example, provides parse trees for 1 million words of English newspaper articles, and has enabled the training of several syntactic parsers, including (Charniak 2000; Klein and Manning 2003). Given this and other successes, there has been a flowering of interest among CL researchers to create and exploit annotated corpora of all kinds, including corpora tagged for semantic features such as verb and senses, preposition relations, and inter-noun relations, as well as for pragmatic features such as word sentiment and entailment.

Specifically, the current debate in these communities focuses on how to ensure the quality of the human-coded annotations. It is taken as axiomatic that any annotation must be performed by at least two, and usually more, people acting independently, so that their tagging decisions can be compared; if they do not agree with enough reliability then the whole project is taken to be ill-defined or too difficult. Some effort is therefore also devoted to investigating different measures of the annotation (and annotator) reliability, and metrics used include simple agreement, Krippendorff's alpha, variations of the Kappa measure (Cohen 1960; Krippendorff 2004; Artstein and Poesio to appear), and others. Though no overall reliability measure can give a complete story (Reidsma and Carletta 2008), one could characterize the recent developments in this area of NLP as slowly progressing toward a 'science' of annotation.

The position is quite different in the field of Corpus Linguistics. Here corpus annotation is not receiving the same attention as in NLP, despite its potential as a topic of methodological cutting-edge research both for theoretical and applied corpus studies (Lavid and Hovy 2008; Hovy and Lavid 2008). There are several reasons for this situation. First, there is not much concern for the reliability, validity, or consistency of the corpus annotation process, which is often considered to be the same as traditional corpus analysis by a single linguist (McEnery, Xiao and Tono 2006). Second, there is a lack of awareness of methodologies that would ensure consistency and reliability in the annotation process. Third, there is no clear picture yet of how to use the results of the annotation process for maximal benefit.

The goal of this paper is to call the corpus linguist's attention to the methodological challenge posed by corpus annotation as a topic of theoretical and applied impact in current corpus linguistic practice. In order to achieve this goal, the first part of the paper presents an overview of the methodologies, requirements, and open questions in corpus annotation as it is currently carried out in the field of NLP. The second part analyses the theoretical and practical impact of corpus annotation in Corpus Linguistics. The paper concludes with an outline of the basic steps towards what could be considered a 'science' a corpus annotation in this field.

1. ANNOTATION IN COMPUTATIONAL STUDIES

1.1. *The seven questions of annotation*

Corpus annotation can be viewed from the perspective of NLP as the process of transforming pure text into interpreted, extracted, or marked-up text. In early work, rules or computer programs to effect the transformations were built manually, while recent methodologies use machine learning techniques to acquire the transformation information automatically, in a process called 'training'. This methodology requires two principal stages: first, to have humans manually annotate texts (the 'training corpus') with the desired tags (i.e., the transformations); second, to train computer algorithms of various kinds on the corpus to perform the same job. In more detail, the annotation process for NLP consists of the following steps:

1. Identifying and preparing a selection of the representative texts as starting material for the 'training corpus' (sometimes called 'training suite').
2. Instantiating a given linguistic theory or linguistic concept, to specify the set of tags to use, their conditions of applicability, etc. This step includes beginning to write the annotator instructions (often called the Codebook or Manual).
3. Annotating some fragment of the training corpus, in order to determine the feasibility both of the instantiation and the annotator Manual.
4. Measuring the results (comparing the annotators' decisions) and deciding which measures are appropriate, and how they should be applied.
5. Determining what level of agreement is to be considered satisfactory (too little agreement means too little consistency in the annotation to enable machine learning algorithms to be trained successfully). If the agreement is not (yet) satisfactory, the process repeats from step 2, with appropriate changes to the theory, its instantiation, the Manual, and the annotator instructions. Otherwise, the process continues to step 6.
6. Annotating a large portion of the corpus, possibly over several months or years, with many intermediate checks, improvements, etc.
7. When sufficient material has been annotated, training the automated NLP machine learning technology on a portion of the training corpus and measuring its performance on the remainder (i.e., comparing its results when applied to the remaining text, often called the 'held-out data', to the decisions of the annotators).

8. If agreement is satisfactory, the technology can be applied to additional, unannotated, material of the same type, thereby assisting future analysis. If agreement is not satisfactory, the process repeats, possibly from step 2, or possibly from step 6 if more training data is required.

These steps are graphically represented in Figure 1, the generic annotation pipeline, in which 90% agreement is taken as the acceptability threshold. The three alternatives labelled Feedback indicate where attention might be paid should agreement not reach satisfactory levels.

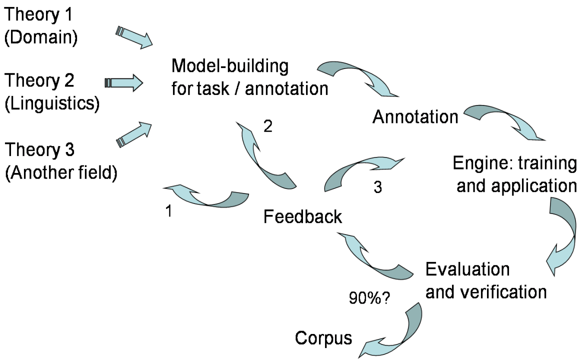


Figure 1. *The generic annotation pipeline*

The whole annotation process poses a number of interesting and thorny questions that remain under debate in the NLP community. The principal questions are the following.

1) Selecting the corpus

As discussed later (see Question 7), the point of creating a corpus is seldom just for a single use. But in order to be distributed to others, the raw material comprising the corpus should be unencumbered by copyright and should ideally be available to the whole community to facilitate comparison, extension, training, and evaluation of future systems.

Given the astonishing diversity of language across many different dimensions, the selection of material to be annotated includes some unexpectedly difficult issues. Probably no linguistic phenomenon is present in all genres, eras, subject areas, dialects, registers, etc. Depending on the nature of the annotation theory and its instantiation, it may be necessary to cover several genres—news, novels, academic prose, poetry, etc.—to ensure that at least a few instances of all the relevant phenomena are encountered. Or it may be sufficient to focus on a single genre, but to include exemplars in the corpus from different time periods and/or linguistic subgroups. Or simply to vary register across a single subject domain, era, and genre.

Manning and Schütze (1999) define a corpus as 'representative' of a phenomenon when what we find for the phenomenon in the sample corpus also holds for the general population or textual universe. For example, the Penn Treebank corpus, consisting of *Wall Street Journal* financial articles, includes the word "stock" hundreds of times, but never once to mean 'soup base'. It is therefore not representative of English word senses, though it is probably quite representative regarding grammar, since it contains over 15,000 different parse tree productions (grammar rules).

Ensuring representativeness of the phenomenon/a under consideration is not the only problem. Balance—the number of instances or exemplars of each variant of the phenomenon in question—is also important for annotation. While it is, after all, in the Zipfian nature of language that some variants of a phenomenon are far more prevalent than others, does this mean one should therefore annotate thousands of exemplars of one variant of the phenomenon against two or three of another—for example, ten thousand instances where "stock" means shares, against three where it means medieval restraining device? It seems natural to say that a balanced corpus is one in which variants are present in proportion to their natural occurrence in the domain, genre, etc., but that might mean that the corpus eventually includes so few exemplars of some variants that they cannot provide additional material for machine learning algorithms or for theoretical analysis. Yet artificially 'enriching' the corpus by including examples of the low-frequency variants distorts the verisimilitude of the corpus, and may in fact adversely affect some machine learning algorithms. This thorny problem remains under debate among corpus linguists (see Leech 2008; Kilgarriff and Grefenstette 2003).

2) Instantiating the theory

Every annotation instantiates some theory. The theory provides the basis for the creation and definition of the annotation categories as well as for the development of the annotation scheme and guidelines, a preliminary and fundamental task in the annotation process. The more complex the phenomena being annotated, the more complex the theory generally is, and hence the more complex the instructions to the annotators. Creating these instructions (the 'annotation manual' or 'codebook') is not a trivial task; for example, the Penn Treebank Codebook is three hundred pages long. Since the codebook is never complete before annotation starts, and is invariably developed and perfected during the annotation process—typically, whenever new variants, unforeseen by the theory, are encountered—, the growing and changing instructions pose problems for the annotation already completed, namely for ensuring that the annotation process remains internally consistent as the codebook develops over time.

At the outset, instantiating the theory encounters the problem that no theory is ever complete, and few if any are developed to an equal degree for all variants of the phenomena they address. Since theories tend to focus on some phenomena over others, uncertainty arises about exactly which categories to define as tags for annotation, how to define them exactly, and what to do with the residue not covered by the theory. The most common problems are that the theoretical categories are not exhaustive over the phenomena, or that they are unclear or difficult to define, often due to intrinsic ambiguity or because they rely too much on background knowledge, as is for example the case with discourse categories and theories.

Another open question is ensuring the stability of the annotation scheme. This can be achieved by working closely with annotators, watching what they do and constantly measuring inter-annotator agreement. In a stable situation, all annotators will make the same tagging decisions, namely the appropriate ones, for all exemplars of each category. This will be true even for new annotators who do not have the benefit of participating in the maturation of the process and the codebook. Although having the annotators repeatedly annotate a sample portion of the corpus until stability (agreement) is achieved is tedious, and suffers from the problem of annotator overtraining (see below), it is a good way of ensuring that the codebook is clear, at least for the variants encountered to date. Later modifications to the codebook should be made carefully, always ensuring that no earlier decisions are superseded.

A related question is how 'deeply' to instantiate the theory. Here there is a tradeoff between the desired theoretical detail or sophistication of categories and the practical attainability of a stable annotation. The general solution adopted in the NLP community is 'neutering' the theory: when the theory is controversial, when categories are hard to define, or when it appears impossible to obtain agreement, one can often still annotate, using a simpler, less refined, more 'neutral' set of terms/categories. For example, in the PropBank project (Palmer, Gildea & Kingsbury 2005), the original case roles associated with verb frames (e.g., *Agent, Patient, Instrument*) were 'neutered' to the PropBank roles (*arg0, arg1, argM*), where each *arg-i* relation is particular to its associated verb sense; there is no commonality, for example, across different verbs' *arg2* roles. In the OntoNotes verb-sense annotations (Pradhan, Hovy, Marcus, Palmer, Ramshaw & Weischededl 2007; Hovy, Marcus, Palmer, Ramshaw & Weischedel 2006; Palmer et al. 2005), the fine-grained verb senses originally obtained from WordNet (Fellbaum 1998) are merged into broader ones as needed to ensure inter-annotator agreements of over 85%. (When WordNet verb senses are taken unchanged from WordNet, inter-annotator agreement on the same texts drops to about 65%: too many WordNet senses are simply indistinguishable to annotators, given its level of definition and their level of ability.)

Examples of 'neutering' discourse categories can be found in the work on information structure phenomena (Baumann et al. 2004, Calhoun et al. 2005); discourse relations (Penn discourse treebank: Miltsakaki et al. 2004, Prasad et al. 2004, Webber 2005); opinion in discourse (MPQA Opinion Corpus: Wiebe et al. 2005), and clause-initial and thematic categories (Lavid & Hovy 2008; Lavid et al. 2009).

3) Selecting and training the annotators

A significant open question is the background and training of annotators. Some researchers claim that annotators should be experts (Kilgarriff 1999). Others propose training annotators just adequately for the task at hand (Hovy et al. 2006; Carlson, Marcu, and Okurowski 2003).

The problem becomes apparent when one considers extreme cases. On the one hand, if one trains the annotators so extensively that they have precise instructions for every possible conceivable variant and case, one will of course obtain 100% agreement. The problem is that in this case there would be no need for annotation, since one can simply encode the instructions as a computer program. However, it is patently impossible to fully prespecify the annotation of any nontrivial phenomenon—not only does no theory completely cover all the phenomena it seeks to explain, but language keeps evolving new cases and examples.

On the other hand, if one provides no training beyond a minimal task description, one relies on the annotators' background knowledge, preconceptions about language, and general educational level. This may differ quite considerably across annotators, and may prove such an obstacle to inter-annotator agreement as to doom the whole enterprise. Typically, when annotators are given relatively little training and a somewhat underspecified codebook, they spontaneously get together to discuss complex cases, and jointly develop some guidelines. Effectively, these guidelines become a hidden part of the codebook, but remain invisible to the annotation manager and to the eventual users at large. As a result, when other people try to duplicate the corpus, or add to it, they cannot achieve consistency—they simply don't have all the instructions. This sort of variability was encountered by various groups on trying to create additional Rhetorical Structure Theory discourse corpora similar to the initial work of (Carlson, Marcu, and Okurowski 2003). In this case, RST relations are simply not defined clearly and deeply enough.

The general approach taken is to use annotators who are reasonably similar in education and sophistication, and to use training and a fairly specific codebook to bring them into correspondence. Inevitably, though, one relies on the internal knowledge or intuitions of the annotators—and correctly so, since often the point of (exploratory) annotation is to discover people's native generalizations about the phenomenon in question.

4) Specifying the annotation procedure

Most annotation procedures start simply, but rapidly evolve to include several phases. The need for a phase of initial annotation to train the annotators and to help refine the codebook was mentioned earlier. Typically, annotators have weekly or biweekly meetings during which problematic cases are discussed, codebook problems are noted, the need for neutering is investigated, theoretical background is provided, etc. In some projects, annotation disagreements are brought to open discussion and jointly resolved (this is often called the 'reconciliation' stage); only cases for which the annotators simply cannot arrive at a mutually acceptable decision are brought to the attention of the manager or the theoretician. Reconciliation can greatly improve agreement scores, but comes at the cost of overtraining and even 'annotation drift' away from earlier procedures, and is hence not recommended, except during the initial training stages. Backing off occurs in cases of disagreement. Here there are several options: (1) making the option granularity coarser (neutering); (2) allowing multiple options; (3) increasing the context supporting the annotation decision; (4) annotating only the easy cases.

The selection of the exemplars to annotate can have a great impact on the speed and accuracy of the work. When annotating the senses of words, for example, the annotators can proceed top to bottom through each text, assigning to each relevant word its sense(s) from the sense inventory for each word, or they can skip from case to case, annotating all exemplars of one word completely before starting the next. The latter strategy, though it compromises on sentence context, is both far quicker and far more reliable: annotators need to hold in mind just one set of alternatives, and become astonishingly rapid and accurate. The former strategy was adopted in Farwell, Dorr, Habash, Helmreich, Hovy, Green, Levin, Miller, Mitamura, Rambow, Reeder and Siddharthan (2009); the latter, in OntoNotes.

For any complex task, full agreement is never possible, even with reconciliation. What should one do with the disagreements? One solution is so-called ‘adjudication’: an expert (or additional annotators) decides just the cases of residual disagreement. Who should do this adjudication—the theoretical expert (who brings in a lot of background knowledge not shared by the annotators, and possibly hence ‘pollution’ from overtraining), or additional annotators (who may simply continue to disagree)? Both views have strong cases. It is clearly advantageous to have the opinion of a single expert in the problematic cases, to ensure consistency for them. But this biases the difficult decisions to the opinion of just one person, precisely where annotation exposes complexity. An open question is how much disagreement can be tolerated before just redoing the annotation anew, or changing the theory or its instantiation radically. Another open question is whether the adjudicator(s) should see the annotation choices made by the annotators (which might bias their decision) or should not. None of these questions have been satisfactorily resolved.

Several annotation procedure heuristics have been used in projects. One approach follows the dictum to do the easy annotations first, allowing annotators to get used to the data and the task before tackling the harder cases. A common strategy is to ask annotators also to mark their level of certainty. There should be high agreement at high certainty (the clear cases). Some experience shows that for annotations that still include up to 50% disagreements, it pays to show new annotators possibly wrong annotations and have them correct them, instead of having them annotate anew.

5) Designing the annotation interface

The computer interface used for the annotation should, of course, facilitate speed and avoid bias. Ensuring this is less simple than it sounds. While it seems natural to implement pull-down menus, annotation proceeds much more quickly if the mouse is altogether avoided, and annotators simply use the fingers of one or both hands, with each key being bound to a different choice. (This has the advantage of limiting the number of options to about 8, which is commonly regarded as around the size of human short-term memory. But for theories with many more alternatives, it might imply that the same material has to be combed over several times, each time with a different set of choices in addition to 'none of the above'.)

A good way to maximize speed is to create very simple tasks, while trying to prevent annotator boredom. Bias or priming must be avoided by not presenting the choice options always in the same order, and certainly all options must be presented to the annotator on a single screen: scrolling is a very strong biasing factor.

For some tasks, annotating *en bloc*, i.e., presenting together a whole series of choices with expected identical annotation and allowing a single selection to tag them all simultaneously, is possible. This has obvious dangers, however.

The NLP and bioinformatics communities have developed numerous annotation interfaces and tools. The QDAP annotation centre at the University of Pittsburgh[1] performs annotation for hire for both NLP and Political Science clients. Also, NLP researchers have developed annotation standards that meet a number of criteria such as expressive adequacy, media independence, semantic adequacy, incrementality for new information in layers, separability of layers, uniformity of style, openness to theories, extensibility to new ideas, human readability, computational processability, and internal consistency (Ide and Romary 2004).

6) Choosing and applying the evaluation measures

Evaluation is a fundamental step in the annotation process. The underlying premise of annotation is that if people cannot agree enough, then either the theory is wrong (or badly stated or instantiated), or the annotation process itself is flawed. In any case, training of computer algorithms is impossible on inconsistent input.

But measuring agreement is a very complicated business. The reliability of the annotation is based both on intra-annotator agreement (measures stability: consistency of each annotator alone over time), and inter-annotator agreement (measures reproducibility: different annotators on the same problem). Different measures have been proposed and used for both. In a very nice study Bayerl and Paul (2007) describe various types of annotator 'drift', both individually over time and as a group.

The most frequent agreement measures used in the NLP community are simple agreement, various forms of Kappa (Cohen 1960), and Krippendorff's alpha (Krippendorff 2004). Simple agreement measures the percentage of cases that two (or more) annotators agree over all the decisions they have both made. It is used in OntoNotes (Pradhan et al. 2007; Hovy et al. 2006). It is easy to understand and quick to determine. Unfortunately, it might be misleading in the case of a very skewed distribution of variants (say, one choice occurs a thousand times as frequently as another: the fact that agreement is near-perfect might be less interesting than annotators' choices on just those rare cases).

To handle bias caused by chance agreement, Kappa is usually used. This measure discounts agreements that would happen by chance, such as with an overpreponderance of one choice over another, and hence favours those cases that are 'unusual'. Typically, in biomedical and statistical literature, Kappa scores of over 0.6 are considered trustworthy. But Cohen's Kappa cannot handle more than two annotators, and even Fleiss's Kappa cannot handle the tasks that allow multiple correct choices. A good reference for the relevant statistics is Bortz 2005 (in German).

Whatever measure(s) is/are employed, the annotation manager has to determine the tolerances: when is the agreement good enough? The OntoNotes project uses 'the 90% agreement rule', basing it on the observation that if humans can agree on something at *N*%, systems will achieve (*N-10*)%; they consider 80% realistic for word senses. For many complex problems, 90% agreement is simply unreachable. Ultimately, the intended use of the annotated corpus is probably the single most important guideline. If the goal is to train machine learning systems, then there should be enough annotated data, at high enough agreement, to enable the systems to be trained to do their job effectively. If the goal is to identify all relevant phenomena, test the theory instantiation, and hence validate the underlying theory, then perhaps it doesn't matter what the agreement level is, as long as poor agreements are seriously investigated.

7) Delivering and maintaining the product

Once an annotated corpus has been created, the resource is obviously most useful when reusable by others. Various kinds of re-use are possible. Common in NLP is the development of new and better machine learning and application algorithms; the ability to train them on older corpora is invaluable for demonstrating their relative improvement over past algorithms. Sometimes, it is advantageous when annotating new theories and/or phenomena to re-use older corpora, so that the annotation systems can be superposed in so-called 'standoff' layers. This enables researchers to investigate the interactions between the various annotation schemes and theories.

Several technical issues must be considered for effective distribution and maintenance. Important is the ownership of the original corpus and the licensing required for distribution. If the original corpus is expensive or not otherwise directly distributable, the annotation layer may still be distributed separately, allowing potential users to approach the owner of the original corpus themselves.

Support and maintenance of the corpus may stretch over years. Since funders have historically been unwilling to provide the required funding long-term, such bodies as the Linguistic Data Consortium in the USA (LDC: www.ldc.upenn.edu) and European Language Resources Association (ELRA: www.elra.info) gather, maintain, and distribute annotated corpora. Annotation developers whose products are acceptable to these bodies facilitate the long-term impact of their work.

1.2. *Some examples of annotation in NLP*

Corpus annotation has been used in different subfields and areas of NLP, as well as in Political Science, Bioinformatics, and other text-related endeavours. Extensive annotation work has been carried out in the area of Information Extraction (IE) with the goal of identifying desired information in free-form text for a number of applications. The information which is usually extracted from texts includes organization names, types and symptoms of disease, people's opinions about products, etc. As the items to extract become more complex, defining what exactly to extract becomes harder. One has to move from pre-specified (hard-coded) rules to automated learning, and this requires annotation. An open issue in this area is how to determine the acceptability of annotation for IE. Figure 2 below (courtesy of Gully Burns) illustrates the role of annotation in the process of IE.

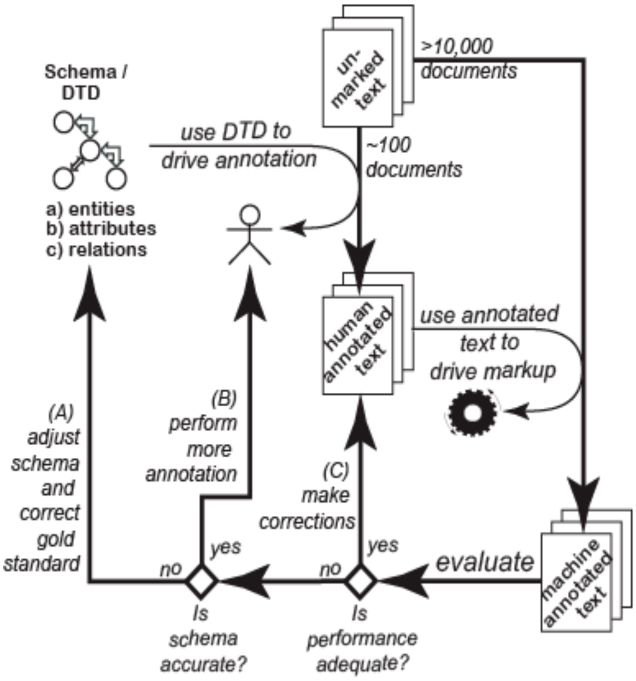
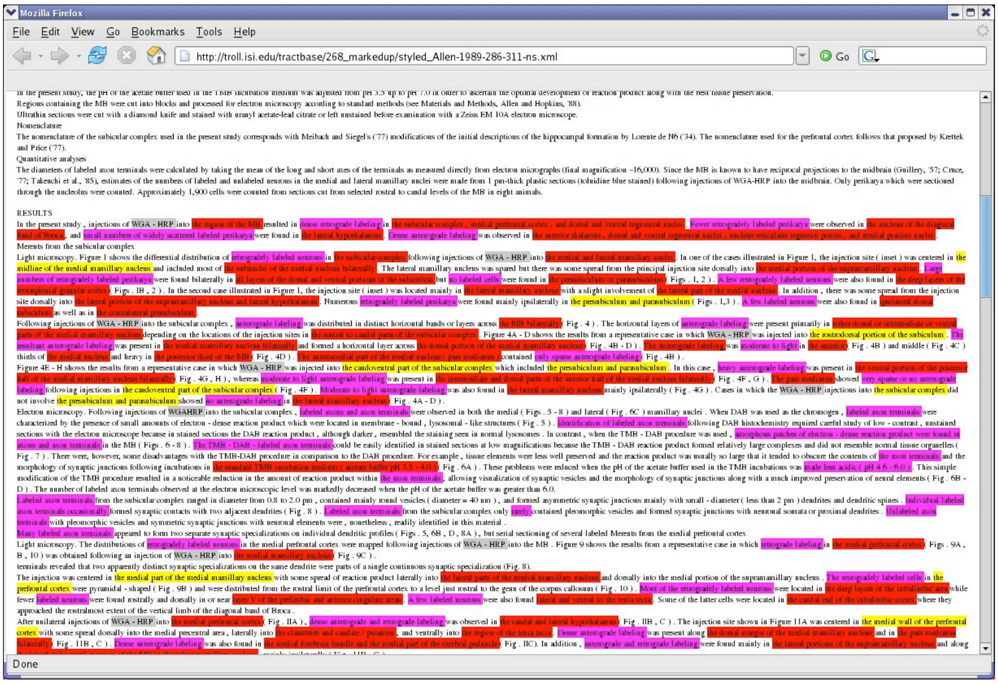


Figure 2. *The role of annotation in IE (figure courtesy of Gully Burns)*

Another area where corpus annotation is being used is in biomedical text mark-up. In this area, domain experts mark up texts to indicate the desired fields. This information is later used in a number of technological applications in the field of medicine. Figure 3 displays a text with biomedical text annotations; the system is described in Burns, Feng and Hovy (2007).


Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://troll.isi.edu/tractbase/268_markedup/styled_Allen-1989-286-311-ns.xml
Go
Done

In the present study, the pH of the acetate buffer used in the TMB incubation medium was adjusted from pH 3.5 up to pH 7.0 in order to ascertain the optimal development of reaction product along with the best tissue preservation.
Regions containing the MB were cut into blocks and processed for electron microscopy according to standard methods (see Materials and Methods, Allen and Hopkins, '88).
Ultrathin sections were cut with a diamond knife and stained with uranyl acetate-lead citrate or left unstained before examination with a Zeiss EM 10A electron microscope.
Nomenclature
The nomenclature of the subicular complex used in the present study corresponds with Meibach and Siegel's ('77) modifications of the initial descriptions of the hippocampal formation by Lorente de N6 ('34). The nomenclature used for the prefrontal cortex follows that proposed by Krettek and Price ('77).
Quantitative analyses
The diameters of labeled axon terminals were calculated by taking the mean of the long and short axes of the terminals as measured directly from electron micrographs (final magnification ~16,000). Since the MB is known to have reciprocal projections to the midbrain (Guillery, '57; Cruce, '77; Takeuchi et al., '85), estimates of the numbers of labeled and unlabeled neurons in the medial and lateral mamillary nuclei were made from 1 pm-thick plastic sections (toluidine blue stained) following injections of WGA-HRP into the midbrain. Only perikarya which were sectioned through the nucleolus were counted. Approximately 1,900 cells were counted from sections cut from selected rostral to caudal levels of the MB in eight animals.

RESULTS
In the present study , injections of WGA - HRP into the region of the MB resulted in dense retrograde labeling in the subicular complex , medial prefrontal cortex , and dorsal and ventral tegmental nuclei . Fewer retrogradely labeled perikarya were observed in the nucleus of the diagonal band of Broca , and small numbers of widely scattered labeled perikarya were found in the lateral hypothalamus . Dense anterograde labeling was observed in the anterior thalamus , dorsal and ventral tegmental nuclei , nucleus reticularis tegmenti pontis , and medial pontine nuclei .
Merents from the subicular complex
Light microscopy. Figure 1 shows the differential distribution of retrogradely labeled neurons in the subicular complex following injections of WGA - HRP into the medial and lateral mamillary nuclei . In one of the cases illustrated in Figure 1, the injection site (inset) was centered in the midline of the medial mamillary nucleus and included most of the subnuclei of the medial nucleus bilaterally . The lateral mamillary nucleus was spared but there was some spread from the principal injection site dorsally into the medial portion of the supramamillary nucleus . Large numbers of retrogradely labeled perikarya were found bilaterally in all layers of the dorsal and ventral portions of the subiculum but no labeled cells were found in the presubiculum or parasubiculum (Figs . 1, 2) . A few retrogradely labeled neurons were also found in the deep layers of the retrosplenial granular cortex (Figs . IB , 2) . In the second case illustrated in Figure 1, the injection site (inset) was located mainly in the lateral mamillary nucleus with a slight involvement of the lateral part of the medial nucleus . In addition , there was some spread from the injection site dorsally into the lateral portion of the supramamillary nucleus and lateral hypothalamus . Numerous retrogradely labeled perikarya were found mainly ipsilaterally in the presubiculum and parasubiculum (Figs . 1,3) . A few labeled neurons were also found in ipsilateral dorsal subiculum as well as in the contralateral presubiculum .
Following injections of WGA - HRP into the subicular complex , anterograde labeling was distributed in distinct horizontal bands or layers across the MB bilaterally (Fig . 4) . The horizontal layers of anterograde labeling were present primarily in either dorsal or intermediate or ventral parts of the medial mamillary nucleus depending on the locations of the injection sites in the rostral to caudal parts of the subicular complex . Figure 4A - D shows the results from a representative case in which WGA - HRP was injected into the rostrodorsal portion of the subiculum . The resultant anterograde labeling was present in the medial mamillary nucleus bilaterally and formed a horizontal layer across the dorsal portion of the medial mamillary nucleus (Fig . 4B - D) . The anterograde labeling was moderate to light in the anterior (Fig . 4B) and middle (Fig . 4C) thirds of the medial nucleus and heavy in the posterior third of the MB (Fig . 4D) . The anteromedial part of the medial nucleus (pars medianus) contained only sparse anterograde labeling (Fig . 4B) .
Figure 4E - H shows the results from a representative case in which WGA - HRP was injected into the caudoventral part of the subicular complex which included the presubiculum and parasubiculum . In this case , heavy anterograde labeling was present in the ventral portion of the posterior half of the medial mamillary nucleus bilaterally (Fig . 4G , H) , whereas moderate to light anterograde labeling was present in the intermediate and dorsal parts of the anterior half of the medial nucleus bilaterally (Fig . 4F , G) . The pars medianus showed very sparse or no anterograde labeling following injections in the caudoventral part of the subicular complex (Fig . 4F) . Moderate to light anterograde labeling was also found in the lateral mamillary nucleus mainly ipsilaterally (Fig . 4G) . Cases in which the WGA - HRP injections into the subicular complex did not involve the presubiculum and parasubiculum showed no anterograde labeling in the lateral mamillary nucleus (Fig . 4A - D) .
Electron microscopy. Following injections of WGAHRP into the subicular complex , labeled axons and axon terminals were observed in both the medial (Figs . 5 - 8) and lateral (Fig . 6C) mamillary nuclei . When DAB was used as the chromogen , labeled axon terminals were characterized by the presence of small amounts of electron - dense reaction product which were located in membrane - bound , lysosomal - like structures (Fig . 5) . Identification of labeled axon terminals following DAB histochemistry required careful study of low - contrast , unstained sections with the electron microscope because in stained sections the DAB reaction product , although darker , resembled the staining seen in normal lysosomes . In contrast , when the TMB - DAB procedure was used , amorphous patches of electron - dense reaction product were found in axons and axon terminals in the MB (Figs . 6 - 8) . The TMB - DAB - labeled axon terminals could be easily identified in stained sections at low magnifications because the TMB - DAB reaction product formed relatively large complexes and did not resemble normal tissue organelles (Fig . 7) . There were, however, some disadvantages with the TMB-DAB procedure in comparison to the DAB procedure. For example , tissue elements were less well preserved and the reaction product was usually so large that it tended to obscure the contents of the axon terminals and the morphology of synaptic junctions following incubations in the standard TMB incubation medium (acetate buffer pH 3.5 - 4.0) (Fig . 6A) . These problems were reduced when the pH of the acetate buffer used in the TMB incubations was made less acidic (pH 4.6 - 6.0) . This simple modification of the TMB procedure resulted in a noticeable reduction in the amount of reaction product within the axon terminals , allowing visualization of synaptic vesicles and the morphology of synaptic junctions along with a much improved preservation of neural elements (Fig . 6B - D) . The number of labeled axon terminals observed at the electron microscopic level was markedly decreased when the pH of the acetate buffer was greater than 6.0.
Labeled axon terminals from the subicular complex ranged in diameter from 0.8 to 2.0 pm , contained mainly round vesicles (diameter = 40 nm) , and formed asymmetric synaptic junctions mainly with small - diameter (less than 2 pm) dendrites and dendritic spines . Individual labeled axon terminals occasionally formed synaptic contacts with two adjacent dendrites (Fig . 8) . Labeled axon terminals from the subicular complex only rarely contained pleomorphic vesicles and formed synaptic junctions with neuronal somata or proximal dendrites . Unlabeled axon terminals with pleomorphic vesicles and symmetric synaptic junctions with neuronal elements were , nonetheless , readily identified in this material .
Many labeled axon terminals appeared to form two separate synaptic specializations on individual dendritic profiles (Figs . 5, 6B , D , 8A) , but serial sectioning of several labeled Merents from the medial prefrontal cortex
Light microscopy. The distributions of retrogradely labeled neurons in the medial prefrontal cortex were mapped following injections of WGA - HRP into the MB . Figure 9 shows the results from a representative case in which retrograde labeling in the medial prefrontal cortex (Figs . 9A , B , 10) was obtained following an injection of WGA - HRP into the medial mamillary nucleus (Fig . 9C) .
terminals revealed that two apparently distinct synaptic specializations on the same dendrite were parts of a single continuous synaptic specialization (Fig. 8).
The injection was centered in the medial part of the medial mamillary nucleus with some spread of reaction product laterally into the lateral parts of the medial mamillary nucleus and dorsally into the medial portion of the supramamillary nucleus . The retrogradely labeled cells in the prefrontal cortex were pyramidal - shaped (Fig . 9B) and were distributed from the rostral limit of the prefrontal cortex to a level just rostral to the genu of the corpus callosum (Fig . 10) . Most of the retrogradely labeled neurons were located in the deep layers of the infralimbic area while fewer labeled neurons were found rostrally and dorsally in or near layer V of the prelimbic and anterior cingulate areas . A few labeled neurons were also found lateral and ventral to the tenia tecta . Some of the latter cells were located in the caudal end of the infralimbic cortex where they approached the rostralmost extent of the vertical limb of the diagonal band of Broca .
After unilateral injections of WGA - HRP into the medial prefrontal cortex (Fig . IIA) , dense anterograde and retrograde labeling was observed in the caudal and lateral hypothalamus (Fig . llB , C) . The injection site shown in Figure 11A was centered in the medial wall of the prefrontal cortex with some spread dorsally into the medial precentral area , laterally into the claustrum and caudate / putamen , and ventrally into the region of the tenia tecta . Dense anterograde labeling was present along the dorsal margin of the medial mamillary nucleus and in the pars medianus bilaterally (Fig . 11B , C) . Dense anterograde labeling was also found in the medial forebrain bundle and the medial part of the cerebral peduncle (Fig . IIC). In addition , anterograde and retrograde labeling were found mainly in the lateral portions of the supramamillary nucleus and along

Figure 3. *Text with biomedical mark-up*

The annotation of semantic categories is being the object of extensive research in the NLP community, with numerous projects working on semantic annotation. Figure 4 below presents some of these projects:

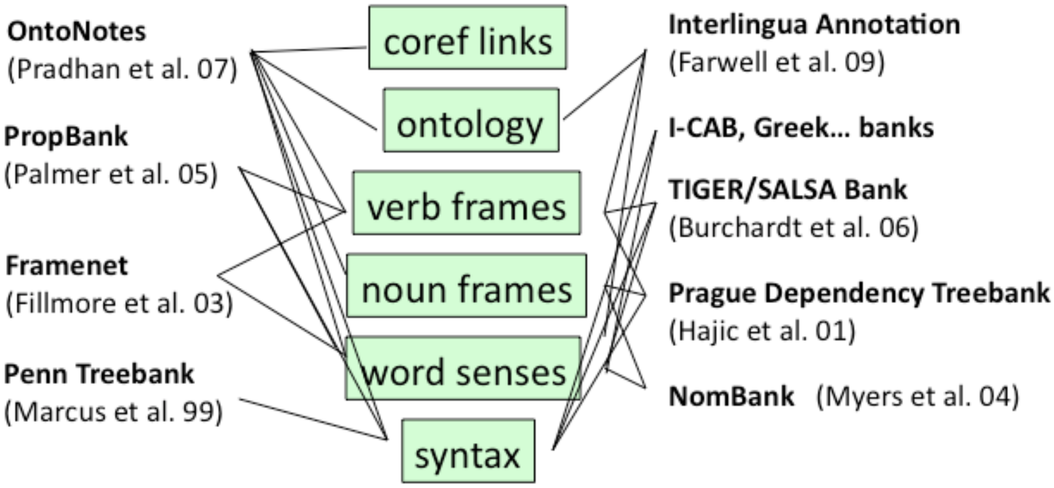


Figure 4. *Some semantic annotation projects*

Other recent annotation projects are Time-ML (Pustejovsky, Palmer & Meyers 2005) http://www.timeml.org/ and the annotation of the phenomenon of subjectivity and opinion in texts (MPQA: subjectivity / 'opinion' (Wiebe and Mihalcea 2006) http://www.cs.pitt.edu/mpqa). In Europe two important national efforts are the Prague dependency treebank (http://ufal.mff.cuni.cz/padt/PADT_1.0/docs/index.html) and the SALSA German semantic annotation corpus (http://www.coli.uni-saarland.de/projects/salsa/page.php?id=research-salsa1; Burchardt, Erk, Frank, Kowalski, Padó and Pinkal 2006). Work on corpus annotation is also being developed in Japan, with two ministries (MIC & METI) planning next 8 years' NLP research with annotation taking up an important part.

2. ANNOTATIONS IN CORPUS LINGUISTICS

In the Corpus Linguistics community, corpus annotation (both automatic and human-coded) is perceived as an activity that enriches and gives 'added value' to a corpus, since it adds linguistic information that can then be used for a variety of purposes, such as theoretical investigation, the creation of dictionaries or lexical databases, further corpus compilation, etc. The perceived advantages of corpus annotation in linguistics lie in three main aspects: reusability, stability, and reproducibility.

In terms of reusability, an annotated corpus is a more valuable resource than a raw corpus since it can be reused by other researchers for additional purposes. In terms of reproducibility, an annotated corpus records linguistic analysis explicitly, thus exposing the analysis (and its underlying theory) to scrutiny and critique. In terms of stability, an annotated corpus provides a standard reference resource, a stable base of linguistic analysis so that successive studies can be compared and contrasted on a common basis. (McEnery et al. 2006).

In spite of these advantages, and the clear need for a well-founded scientific methodology that would ensure the reliability of corpus annotation, best practice in corpus annotation remains a desideratum in Corpus Linguistics. As stated by Geoffrey Leech (2005): "Best practice in corpus annotation is something we should all strive for—but which perhaps few of us will achieve".

Two principal reasons account for this situation. First, there is a lack of awareness of the need and the methodologies that would ensure consistency and reliability in the annotation process, which is often considered the same as traditional corpus analysis by a single linguist (McEnery et al. 2006). Second, there is no clear picture yet of how to use the results of the annotation process for maximal benefit.

In this paper we claim that best practice in corpus annotation must become part of the research agenda of Corpus Linguistics in this decade, due to its theoretical, methodological, and practical impact on the field. We outline the nature of this impact in the following subsections.

2.1. *Theoretical impact of corpus annotation*

When used as a mechanism to test aspects of linguistic theories empirically, annotation has both theoretical and methodological impact. At least two aspects of theories can be tested through corpus annotation: theory formation and theory redefinition. In addition, corpus annotation makes it possible to enrich theories with quantitative information.

1. Theory formation: One of the first steps in the annotation process is the creation of an annotation scheme (the Codebook or Manual) which instantiates a linguistic theory or model of the behaviour of a linguistic phenomenon. For example, in order to annotate information structure (IS) categories such as topic, theme, or focus, one has to design an annotation scheme with tags and choice-specific definitions corresponding to these categories and their subtypes. When confronted with the data, annotators may however identify new (heretofore

theoretically unanticipated) aspects of these categories which they cannot handle. If this is the case, the original theoretical coverage of the phenomenon falls short and will have to be extended to cover the phenomena observed in the data.

2. Theory redefinition: During the annotation process, annotators may consistently disagree on certain exemplars or types of exemplars. This may be considered to be an indication that the theoretical definition of this case lacks clarity and accuracy, and requires redefinition (or even further theorizing).
3. Enriching theories with quantitative information: The result of the annotation process allows one to determine the relative frequency of each subtype of the phenomenon being studied empirically. It is not uncommon that theories devote a great deal more attention to relatively rare phenomena than to (equally interesting) common ones.

2.2. *Practical impact of corpus annotation*

Corpus annotation can have a practical impact in several fields. We mentioned earlier several applications of annotated corpora in the NLP community, where they serve as training material for computational systems. Other areas of application in the linguistics community, virtually unexplored to date, are language teaching, contrastive linguistics, and translation.

In the area of language teaching and, more specifically, the teaching of translation, there are a number of possible exploitation activities to be carried out with students. First, it is possible to work with students as annotation checkers (Lavid 2007). Here students are provided an annotation or coding scheme designed by experts on a given linguistic phenomenon and requested to annotate a training text. The experts use their results to check whether the proposed tag definitions can account for all exemplars in the data. The outcome of the annotation process may lead to the creation of new tags to cover those cases, which are not accounted for the original theory (theory formation), and/or to the redefinition of the existing ones when they are not clear or accurate enough to ensure annotator agreement (theory redefinition). The changes to the original coding scheme will help language experts to create a model of the behaviour of the linguistic phenomenon, which can then be applied to a larger set of texts. Using a frequency table, and carrying out the annotation on different text types and genres, it is also possible to obtain quantitative information on the frequency of occurrence of certain features of the linguistic phenomenon under study.

Second, it will also be possible to teach students aspects of the behaviour of a given linguistic phenomenon by using as a 'gold standard' an annotation scheme and a corpus previously annotated by experts. The students will be able to study examples of tags assigned to exemplars in the annotated corpus, as an illustration of how a given theoretical notion is applied to actual data. This will allow them to discuss difficult cases, and later to reproduce the annotations in a new set of texts.

Similar activities can be carried out in the context of the contrastive analysis or translation classrooms. In the case of contrastive analysis, it is possible to design two separate annotation schemes for a given linguistic phenomenon in each language (see Schultz 2009), or, if there are commonalities at the more general level, one could design one common core annotation scheme for both languages and then an extended one for each individual language, as it is currently done in the CONTRANOT Project[2].

Table 1 below illustrates the initial core tagset and the extended tagset declaration of the annotation scheme for the notion of Thematic Field in English and Spanish.

Table 1. *Initial Core and Extended Tagsets for Thematic Field*

Annotation layer	**Thematic field**	
Definition		Complex functional zone in clause-initial position serving a variety of clausal and discourse functions
Unit		Scope of Thematic field depends on the realisation and the position of the Subject in the main clause
CORE ANNOTATION SCHEME		
Tags:	TH	Thematic Head
	PH	Pre-Head
	TE	Thematic Equative
	PT	Predicated Theme
EXTENDED ANNOTATION SCHEME		
Tags:	TH-SNom	Thematic Head-Simple Nominal
	TH- CNom	Thematic Head-Complex Nominal
	TH- SCir	Thematic Head-Simple Circumstance
	TH- CCir	Thematic Head-Complex Circumstance
	TH- Vinf	Thematic Head- Verbal inflection
	TH- Clause	Thematic Head-Clause
	PH-Cir	PH-Vlex Circumstance
	PH-Vlex	PH-lexical part of verb

	PH-Se (pronominal)	PH-'se-pron'

In the area of translation, corpus annotation may serve a variety of purposes. A translation corpus annotated with translation equivalences may be the basis for the development of machine translation systems that learn those equivalences using machine-learning algorithms. Farwell et al. (2009) describe the use of interlingual word sense tags that enable translation systems to learn to select appropriate lexical translations.

If the purpose is to teach translation patterns to students, one can use the results of preliminary intensive quantitative corpus analysis on a set of translations to derive characteristic translation strategies, which can then form the basis for the annotation of a translation corpus. Table 2 below illustrates the results of the corpus analysis of the clausal thematic patterns in a bilingual corpus of original and translated fiction texts in English and Spanish in both directions of translation (Lavid in press).

Table 2. *Distribution of translation patterns in English and Spanish (after Lavid in press)*

	Eng-Spa	%	Spa-Eng	%	Total
SV <=> V (S)	32	50	32	50	64
Circumstantial fronted	7	29.1	17	70.8	24
Change of Subject	3	75	1	25	4
It/there	3	33.3	6	66.6	9
Reorganization of TS	14	50	14	50	28
Total	59		70		129

As shown in Table 2, the most common strategy in both directions of translation is preserving the language-specific word order in declarative sentences: *Subject ^ Verb* for English and *Verb (Subject)* for Spanish. This is followed by the *Reorganization of the thematic structure* in both languages and directions of the translation, and then the high frequency of *Circumstantial* elements in initial position. These translation patterns can form the basis of an annotation scheme that may be used for teaching purposes or for the annotation of a different translation corpus. Annotation of different genres in translation might indicate systematic differences in certain phenomena, such as the well-known florid style of the business letter in French or the relative absence of pronominal subjects in Japanese.

3. CONCLUSION: TOWARDS A 'SCIENCE' OF CORPUS ANNOTATION

In view of the theoretical and the practical impact of corpus annotation on Corpus Linguistics, we propose some basic steps toward what could be considered a 'science' of annotation:

1. Developing a clear description of the phenomenon being studied, including a possible pre-theoretical description of the types and ranges of the classes and values observed.
2. Providing a clear motivation for the corpus that is going to be annotated, as well as a demonstrated understanding of the effects of possible bias.
3. Designing a proper annotation procedure, including the good training of the annotators, independent annotation by at least two people, controlled inter-annotator discussions, and attention to the environment (interfaces, order of presentation of choices, etc.).
4. Developing and applying a proper evaluation, with a clear understanding of the value and problems of agreement measures.
   Given the increasing confluence of the interests of corpus linguists and computational linguists, we are entering a new era of corpus building, with need for annotated corpora and annotation experts. But the current absence of a generalised standard practice and solid methodology hampers the trustability of results. The NLP community generally is not very concerned with theoretical linguistic soundness. The Corpus Linguistics community does not seem to seek 'reliability' in the annotation process and results. There is, therefore, ample need for collaborative efforts to advance in both camps: enriching computational studies with theory, and enriching corpus studies with experimental methodologies, which ensure the 'reliability' in the annotation process.

NOTES

[1] ATLAS.TI: annotation toolkit (www.atlasti.com), QDAP annotation center at U of Pittsburgh (www.qdap.pitt.edu).

[2] The CONTRANOT project aims at the creation and validation of contrastive English-Spanish functional descriptions through text analysis and annotation. The project is financed by the Spanish Ministry of Science and Innovation under the I+D Research Projects Programme (reference number FFI2008-03384), with Prof. Julia Lavid as team leader of the

research group. Part of the work reported in this paper is being carried out in the framework of this project as well as thanks to the financial support of the BSCH-UCM grant awarded to Prof. Lavid's research group 930175, so we gratefully acknowledge the financial support provided by these authorities.

REFERENCES

Artstein, R. and M. Poesio. 2008. Inter-coder agreement for Computational linguistics (survey article). *Computational Linguistics,* 34/4, 555–596.

Bayerl, Petra Saskia & Paul, Karsten Ingmar. (2007). Identifying sources of disagreement: generalizability theory in manual annotation studies. *Computational Linguistics*, 33/1, 3-8.

Baumann, S., Brinckmann, C., Hansen-Schirra, S., Kruijff, G-J., Kruijff-Korbayová, I., Neumann, S. and E. Teich. 2004. Multi-dimensional annotation of linguistic corpora for investigating information structure. In *Proceedings of the NAACL/HLT. Frontiers in Corpus Annotation.,* Boston, MA.

Calhoun, S., Nissim, M., Steedman, M., and Jason Brenier. 2005. A framework for annotating information structure in discourse. In *Frontiers in Corpus Annotation II: Pie in the Sky, ACL2005 Conference Workshop*, Ann Arbor, Michigan, June 2005.

Bortz, J. 2005. Statistik: Für Human- und Sozialwissenschaftler, Berlin: Springer.

Burchardt, A., K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In Proceedings of the LREC-2006 Conference, Genoa, Italy.

Burns, G.A.P.C., D. Feng, and E.H. Hovy. 2007. Intelligent approaches to mining the primary research literature: techniques, systems, and examples. In A. Kelemen, A. Abraham, Y. Chen, and Y. Liang (eds), *Computational Intelligence in Bioinformatics*. Springer Verlag, series Studies in Computational Linguistics.

Carlson, L., Marcu, D. and M.E. Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Jan van Kuppevelt and Ronnie Smith (eds.), *Current directions in discourse and dialogue*. Kluwer Academic Publishers.

Charniak, E. 2000. A Maximum-Entropy-Inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference* (pp- 132-139), Seattle (Washington).

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20/1, 37–46.

Farwell, D.L., B.J. Dorr, N. Habash, S. Helmreich, E.H. Hovy, R. Green, L. Levin, K. Miller, T. Mitamura, O. Rambow, F. Reeder, and A. Siddharthan. 2009. Interlingual annotation of multilingual text corpora and FrameNet. In H.C. Boas (ed.), *Multilingual FrameNets in Computational*

Lexicography: Methods and Applications, 287–318. Berlin: Mouton de Gruyter.

Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database*, Cambridge, MA: MIT Press, http://www.cogsci.princeton.edu/˜wn.

Fillmore, C.J., C.R. Johnson, and M.R.L. Petruck. 2003. Background to Framenet. *International Journal of Lexicography*, 16/3, 235–250.

Hajic, J., B. Vidová-Hladká, and P. Pajas. 2001. The Prague Dependency Treebank: annotation structure and support. *Proceeding of the IRCS Workshop on Linguistic Databases*, 105–114.

Hovy, E.H., M. Marcus, M. Palmer, L. Ramshaw and R. Weischedel. 2006. OntoNotes: the 90% solution. In *Proceedings of theHLT- NAACL-2006 Conference*, 57–60.

Hovy, E.H. and J. Lavid. 2008. Corpus Annotation: Framework and Hands-on Experience, Tutorial presented at the *6th Language Resources and Evaluation Conference* (*LREC 2008),* Marrakech, Morocco.

Ide, N., Romary, L. 2004. International standard for a linguistic annotation framework. *Journal of Natural Language Engineering*, 10/3–4, 211–225.

Kilgarriff, A. 1999. 95% replicability for manual word sense tagging. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway. 277-278.

Kilgarriff, A. and G. Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics,* 29/3, 333-348.

Klein, D. and C.D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics* (pp. 423–430), Vol. 1, Morristown, NJ.

Krippendorff, K. 2004. Reliability in content analysis: some common misconceptions and recommendations. *Human Communication Research*, 30(3), 411– 433.

Lavid, J. 2007. Web-based bilingual text resources for the learner: an innovative experience through the Virtual Campus. In M. Martinez-Cabeza, N. McLaren and Luis Quereda (eds.), *Estudios en honor de Rafael Fente Gómez*, 293–303, Editorial Universidad de Granada.

Lavid, J. in press. Contrasting choices in clause-initial position in English and Spanish: a corpus-based analysis. In Elizabeth Swain, (ed.), *Thresholds and Potentialities of Systemic Functional Linguistics: Applications to other disciplines, specialised discourses and languages other than English*. Trieste: Edizioni Universitarie.

Lavid, J. and E.H. Hovy. 2008. Why do we need annotation and what can we learn from it: insights from Natural Language Processing. Poster paper presented at *ICAME 2008 Conference*, Ascona, Switzerland.

Lavid, J., Arús, J. and L. Moratón. 2009. Thematic features in English and Spanish: an SFL analysis of two newspaper genres. Paper presented at the *21st European Systemic Functional Linguistics Conference and Workshop*, Cardiff University, United Kingdom, 8-10 July.

Leech, G. 2005. Adding linguistic annotation. In M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, 17–29, Oxford: Oxbow Books.

Leech, G. 2008. New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf and C. Biewer (eds.), *Corpus Linguistics and the Web*, 133–150, Amsterdam & New York: Rodopi.

Miltsakaki, E., Prasad, R., Joshi, A., and Bonnie Webber. 2004. Annotating discourse connectives and their arguments. In *NAACL/HLT Workshop on Frontiers in Corpus Annotation, Boston.*

Prasad, R., Miltsakaki, E., Joshi, A. and Bonnie Webber. 2004. Annotation and data mining of the Penn Discourse TreeBank. In *ACL Workshop on Discourse Annotation,* Barcelona, July 2004.

Manning, C.D. and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marcus, M., M.A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19/2, 313–330.

McEnery, A., R. Xiao, and Y. Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. New York: Routledge.

Meyers, A., R. Reeves, C Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The NomBank project: an interim report. *Frontiers in Corpus Annotation*, *Workshop in conjunction with HLT/NAACL.*

Palmer, M., D. Gildea, and P. Kingsbury. 2005. The proposition bank: a corpus annotated with semantic roles. *Computational Linguistics.* 31/1, 71–106

Pradhan, S., E.H. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2007. OntoNotes: a unified relational semantic representation. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC-07)*, 517–524.

Pustejovsky, J., M. Palmer and A. Meyers. 2005. Merging PropBank, NomBank, TimeBank, Penn Discourse Treebank. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, 5–12, Ann Arbor, Michigan.

Reidsma, D.and J. Carletta. 2008. Reliability measurement without limits, *Computational Linguistics* 34/3, 319–326.

Schulz, Anke. 2009. Contrasting choice: annotation of an English-German corpus. Paper presented at the *21st European Systemic-Functional Linguistics Conference and Workshop*, Cardiff University.

Webber. Bonnie. 2005. A short introduction to the Penn Discourse TreeBank. In *Copenhagen Working Papers in Language and Speech Processing*.

Wiebe, J., Breck, E., Buckley, C., Cardie, C., Davis, P., Fraser, B. Litman, D., Pierce, D., Riloff, E., Wilson, T., Day, D. and Mark Maybury. 2003. Recognizing and organizing opinions expressed in the world press. In *Working Notes of the AAAI Spring Symposium in New Directions in Question Answering*, Palo Alto (California), 12–19.

Wiebe, J. and R. Mihalcea**.** 2006. Word sense and subjectivity. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics. (COLING-ACL 2006),* Sydney, Australia.

Prof. Dr. Julia Lavid
Dept. of English Philology I
Faculty of Philology
Universidad Complutense de Madrid, Spain
E-mail: <lavid@filol.ucm.es>

Dr. Eduard Hovy
Information Sciences Institute
University of Southern California, USA
E-mail: <hovy@isi.edu>