On the Problem of Theoretical Terms in Empirical Computational Linguistics

Stefan Riezler*
Computational Linguistics
Heidelberg University, Germany

Philosophy of science has pointed out a problem of theoretical terms in empirical sciences. This problem arises if all known measuring procedures for a quantity of a theory presuppose the validity of this very theory, because then statements containing theoretical terms are circular. We argue that a similar circularity can happen in empirical computational linguistics, especially in cases where data are manually annotated by experts. We define a criterion of T-non-theoretical grounding as guidance to avoid such circularities, and exemplify how this criterion can be met by crowdsourcing, by task-related data annotation, or by data in the wild. We argue that this criterion should be considered as a necessary condition for an empirical science, in addition to measures for reliability of data annotation.

1. Introduction

The recent history of computational linguistics (CL) shows a trend towards encoding natural language processing (NLP) problems as machine learning tasks, with the goal of applying task-specific learning machines to solve the encoded NLP problems. In the following we will refer to such approaches as **empirical CL** approaches.

Machine learning tools and statistical learning theory play an important enabling and guiding role for research in empirical CL. A recent discussion in the machine learning community claims an even stronger and more general role of machine learning. We allude here to a discussion concerning the relation of machine learning and philosophy of science. For example, Corfield, Schölkopf, and Vapnik (2009) compare Popper's ideas of falsifiability of a scientific theory with "similar notions" from statistical learning theory regarding Vapnik-Chervonenkis theory. A recent NIPS workshop on "Philosophy and Machine Learning"[1] presented a collection of papers investigating similar problems and concepts in the two fields. Korb (2004) sums up the essence of the discussion by directly advertising "Machine Learning as Philosophy of Science."

In this article we argue that adopting machine learning theory as philosophy of science for empirical CL has to be done with great care. A problem arises in the application of machine learning methods to natural language data under the assumption that input–output pairs are given and do not have to be questioned. In contrast to machine learning, in empirical CL neither a representation of instances nor an association of

* Department of Computational Linguistics, Heidelberg University, Im Neuenheimer Feld 325, 69120 Heidelberg, Germany. E-mail: riezler@cl.uni-heidelberg.de.

1 http://www.dsi.unive.it/PhiMaLe2011/.

doi:10.1162/COLI_a_00182

instances and labels is always "given." We show that especially in cases where data are manually annotated by expert coders, a problem of circularity arises if one and the same theory of measurement is used in data annotation and in feature construction. In this article, we use insights from philosophy of science to understand this problem. We particularly point to the "problem of theoretical terms," introduced by Sneed (1971), that shows how circularities can make empirical statements in sciences such as physics impossible.

In the following, we will explain the problem of theoretical terms with the help of a miniature physical theory used in philosophy of science (Section 2). We will then exemplify this concept on examples from empirical CL (Section 3). We also make an attempt at proposing solutions to this problem by using crowdsourcing techniques, task-related annotation, or data in the wild (Section 4).

2. The Problem of Theoretical Terms in Philosophy of Science

In order to characterize the logical structure of empirical science, philosophy of science has extensively discussed the notions of "theoretical" and "observational" language. Sneed (1971)[2] was the first to suggest a distinction between "theoretical" and "non-theoretical" terms of a given theory by means of the roles they play in that theory. Balzer (1996, page 140) gives a general definition that states that a term is "*theoretical* in theory T iff every determination of (a realization of) that term presupposes that T has been successfully applied beforehand." Because there are no theory-independent terms in this view, an explicit reference to a theory T is always carried along when characterizing terms as theoretical with respect to T (*T-theoretical*) or non-theoretical with respect to T (*T-non-theoretical*). Stegmüller (1979) makes the notions of "determination" or "realization" more concrete by referring to procedures for *measuring* values of quantities or functions in empirical science:

> What does it mean to say that a quantity (function) f of a physical theory T is T-theoretical?... In order to perform an empirical test of an empirical claim containing the T-theoretical quantity f, we have to measure values of the function f. But all known measuring procedures (or, if you like, all known theories of measurement of f-values) presuppose the validity of this very theory T. (page 17)

The "problem of theoretical terms" can then be stated as follows (see Stegmüller 1979): Suppose a statement of the form

$$x \text{ is a } P \tag{1}$$

where x is an entity and P is a set-theoretic predicate by which a physical theory is axiomatized. If this theory contains P-theoretic terms, then (1) *is not an empirical statement* because another sentence of exactly the same form and with exactly the same predicate is presupposed. An illustration of this concept can be given by Stegmüller (1986)'s miniature theory of an Archimedian Statics. Let us assume that this miniature theory is formalized by the set-theoretic predicate AS. The intended applications of the theory AS are objects $a_1, \ldots, a_n$ that are in balance around a pivot point. The theory uses

2 The following discussion of concepts of the "structuralist" or "non-statement view of theories" is based on works by Stegmüller (1979, 1986) and Balzer and Moulines (1996) that are more accessible than the original book by Sneed (1971). All translations from German are by the author.

two functions that measure the distance d of the objects from the pivot point, and the weight g. The central axiom of the theory states that the sum of the products $d(a_i)g(a_i)$ is the same for the objects on either side of the pivot point. The theory AS can then be defined as follows:

x is an AS iff there is an A, d, g *such that*:

1. $x = \langle A, d, g \rangle$,
2. $A = \{a_1, \ldots, a_n\}$,
3. $d : A \to \mathbb{R}$,
4. $g : A \to \mathbb{R}$,
5. $\forall a \in A : g(a) > 0$,
6. $\sum_{i=1}^{n} d(a_i)g(a_i) = 0$.

Entities that satisfy conditions (1) to (5) are called **potential models** of the theory. Entities that also satisfy the central axiom (6) are called **models** of the theory. An **empirical statement** is a statement that a certain entity is a model of the theory.

Stegmüller (1986) uses the miniature theory AS to explain the problem of theoretical terms as follows: Suppose we observe children sitting on a seesaw board. Suppose further that the board is in balance. Translating this observation into the set-theoretic language, we could denote by y the balanced seesaw including the children, and we would be tempted to make the empirical statement that

$$y \text{ is an } AS \tag{2}$$

In order to verify the central axiom, we need to measure distance and weight of the children. Suppose that we have a measuring tape available to measure distance, and suppose further that our only method to measure weight is the use of beam balance scales. Let us denote by z the entity consisting of the balanced beam scale, the child, and the counterbalancing measuring weight; then the validity of our measuring result depends on a statement

$$z \text{ is an } AS \tag{3}$$

Thus, in order to check statement (2), we have to presuppose statement (3), which is of the very same form and uses the very same predicate. That means, in order to measure the weight of the children, we have to presuppose successful applications of the theory AS. But in order to decide for successful applications of AS, we need to be able to measure the weight of the objects in such application. This epistemological circle prevents us from claiming that our original statement (2) is an empirical statement.

The crux of the problem of theoretical terms for the miniature theory AS is the measuring procedure for the function g that presupposes the validity of the theory AS. The term g is thus AS-theoretical. There are two solutions to this problem:

1. In order to avoid the use of AS-theoretic terms such as g, we could discard the assumption that our weight-measuring procedure uses beam balance scales. Instead we could use AS-non-theoretic measuring procedures such as spring scales. The miniature theory AS would no longer contain

AS-theoretic terms. Thus we would be able to make empirical statements of the form (2), that is, statements about certain entities being models of the theory *AS*.

2. In complex physical theories such as particle mechanics there are no simplified assumptions on measuring procedures that can be dropped easily. Sneed (1971) proposed the so-called **Ramsey solution**[3] that in essence avoids *AS*-theoretical terms by existentially quantifying over them.

Solution (1), where *T*-theoretical terms are measured by applications of a theory T', thus is the standard case in empirical sciences. Solution (2) is a special case where we need theory *T* in order to measure some terms in theory *T*. Gadenne (1985) argues that this case can be understood as a *tentative assumption* of theory *T* that still makes empirical testing possible.[4]

The important point for our discussion is that in both solutions to the problem of theoretical terms, whether we refer to another theory T' (solution (1)) or whether we tentatively assume theory *T* (solution (2)), we *require an explicit dichotomy between T-theoretical and T-non-theoretical terms*. This insight is crucial in the following analysis of possible circularities in the methodology of empirical CL.

3. The Problem of Theoretical Terms in Empirical CL

Most machine-learning approaches can be characterized as identifying a learning problem as a problem of estimating a prediction function $f(x)$ for *given* identically and independently distributed (i.i.d.) data $\{(x_i, y_i)\}_{i=1}^{N}$ of instances and labels. For most approaches in empirical CL, this prediction function can be characterized by a discriminant form of a function f where

$$f(x; w, \phi) = \arg\max_{y} F(x, y; w, \phi)$$

and where $w \in \mathbb{R}^D$ denotes a *D*-dimensional parameter vector, $\phi(x, y) \in \mathbb{R}^D$ is a *D*-dimensional vector of features (also called attributes or covariates) jointly representing input patterns x and outputs y (denoting categorical, scalar, or structured variables),

3 For the miniature theory *AS*, this is done by firstly stripping out statements (4)–(6) containing theoretical terms, achieving a **partial potential model**. Secondly statements (4) and (5) are replaced by a so-called **theoretical extension** that existentially quantifies over measuring procedures for terms like *g*. The resulting **Ramsey claim** applies a theoretical extension to a partial potential model that also satisfies condition (6). Because such a statement does not contain theoretical terms we can make empirical statements about entities being models of the theory *AS*.

4 Critics of the structuralist theory of science have remarked that both of the solutions are instances of a more general problem, the so-called Duhem-Quine problem, thus the focus of the structuralist program on solution (2) seems to be an exaggeration of the actual problem (von Kutschera 1982; Gadenne 1985). The Duhem-Quine thesis states that theoretical assumptions cannot be tested in isolation, but rather whole systems of theoretical assumptions and auxiliary assumptions are subjected to empirical testing. That is, if our predictions are not in accordance with our theory, we can only conclude that one of our many theoretical assumptions must be wrong, but we cannot know which one, and we can always modify our system of assumptions, leading to various ways of **immunity of theories** (Stegmüller 1986). This problem arises in Solution (1) as well as in Solution (2)

and F measures the compatibility of pairs (x, y), for example, in the form of a linear discriminant function (Taskar et al. 2004; Tsochantaridis et al. 2005).[5]

The problem of theoretical terms arises in empirical CL in cases where a single theoretical tier is used both in manual data annotation (i.e., in the assignment of labels y to patterns x via the encoding of data pairs (x, y)), *and* in feature construction (i.e., in the association of labels y to patterns x via features $\phi(x, y)$).

This problem can be illustrated by looking at automatic methods for data annotation. For example, information retrieval (IR) in the patent domain uses citations of patents in other patents to automatically create relevance judgments for ranking (Graf and Azzopardi 2008). Learning-to-rank models such as that of Guo and Gomes (2009) define domain knowledge features on patent pairs (e.g., same patent class in the International Patent Classification [IPC], same inventor, same assignee company) and IR score features (e.g., tf-idf, cosine similarity) to represent data in a structured prediction framework. Clearly, one could have just as well used IPC classes to create automatic relevance judgments, and patent citations as features in the structured prediction model. It should also be evident that using the same criterion to automatically create relevance labels *and* as feature representation would be circular. In terms of the philosophical concepts introduced earlier, the theory of measurement of relevance used in data labeling cannot be the same as the theory expressed by the features of the structured prediction model; otherwise we exhibit the problem of theoretical terms.

This problem can also arise in scenarios of manual data annotation. One example is data annotation by expert coders: The expert coder's decisions of which labels to assign to which types of patterns may be guided by implicit or tacit knowledge that is shared among the community of experts. These experts may apply the very same knowledge to design features for their machine learning models. For example, in attempts to construct semantic annotations for machine learning purposes, the same criteria such as negation tests might be used to distinguish presupposition from entailment in the labeling of data, and in the construction of feature functions for a classifier to be trained and tested on these data. Similar to the example of automatic data annotation in patent retrieval, we exhibit the problem of theoretical terms in manual data annotation by experts in that the theory of measurement used in data annotation and feature construction is the same. This problem is exacerbated in the situation where a single expert annotator codes the data and later assumes the role of a feature designer using the "given" data. For example, in constructing a treebank for the purpose of learning a statistical disambiguation model for parsing with a hand-written grammar, the same person might act in different roles as grammar writer, as manual annotator using the grammar's analyses as candidate annotations, and as feature designer for the statistical disambiguation model.

The sketched scenarios are inherently circular in the sense of the problem of theoretical terms described previously. Thus in all cases, we are prevented from making empirical statements. High prediction accuracy of machine learning in such scenarios indicates high consistency in the application of implicit knowledge in different roles of a single expert or of groups of experts, but not more.

This problem of circularity in expert coding is related to the problem of reliability in data annotation, a solution to which is sought by methods for measuring and enhancing inter-annotator agreement. A seminal paper by Carletta (1996) and a follow-up survey

5 In this article, we concentrate on supervised machine learning. Semisupervised, transductive, active, or unsupervised learning deal with machine learning from incomplete or missing labelings where the general assumption of i.i.d. data is not questioned. See Dundar et al. (2007) for an approach of machine learning from non-i.i.d. data.

paper by Artstein and Poesio (2008) have discussed this issue at length. Both papers refer to Krippendorff (2004, 1980a, page 428) who recommends that reliability data "have to be generated by coders that are widely available, follow explicit and communicable instructions (a data language), and work independently of each other. . . . [T]he more coders participate in the process and the more common they are, the more likely they can ensure the reliability of data." Ironically, it seems as if the best inter-annotator agreement is achieved by techniques that are in conflict with these recommendations, namely, by using experts (Kilgarriff 1999) or intensively trained coders (Hovy et al. 2006) for data annotation. Artstein and Poesio (2008) state explicitly that

> experts as coders, particularly long-term collaborators, [. . .] may agree not because they are carefully following written instructions, but because they know the purpose of the research very well–which makes it virtually impossible for others to reproduce the results on the basis of the same coding scheme . . . . Practices which violate the third requirement (independence) include asking the coders to discuss their judgments with each other and reach their decisions by majority vote, or to consult with each other when problems not foreseen in the coding instructions arise. Any of these practices make the resulting data unusable for measuring reproducibility. (page 575)

Reidsma and Carletta (2007) and Beigman Klebanov and Beigman (2009) reach the conclusion that high inter-annotator agreement is neither sufficient nor necessary to achieve high reliability in data annotation. The problem lies in the implicit or tacit knowledge that is shared among the community of experts. This implicit knowledge is responsible for the high inter-annotator agreement, but hinders reproducibility. In a similar way, implicit knowledge of expert coders can lead to a circularity in data annotation and feature modeling.

4. Breaking the Circularity

Finke (1979), in attempting to establish criteria for an empirical theory of linguistics, demands that the use of a single theoretical strategy to identify and describe the entities of interest shall be excluded from empirical analyses. He recommends that the *possibility* of using T-non-theoretic strategies to identify observations be made the defining criterion for empirical sciences. That is, in order to make an empirical statement, the two tiers of a T-theoretical and a T-non-theoretical level are necessary because the use of a single theoretical tier prevents distinguishing empirical statements from those that are not.

Let us call Finke's requirement the criterion of **T-non-theoretical grounding**.[6] Moulines (see Balzer 1996, page 141) gives a pragmatic condition for T-non-theoreticity that can be used as a guideline: "Term $\bar{t}$ is T-non-theoretical if there exists and acknowledged method of determination of $\bar{t}$ in some theory T' different from T plus some link from T' to T which permits the transfer of realizations of $\bar{t}$ from T' into T."

Balzer (1996) discusses a variety of more formal characterizations of the notion of T-(non-)theoretical terms. Although the pragmatic definition cited here is rather informal, it is sufficient as a guideline in discussing concrete examples and strategies to break the circlularity in the methodology of empirical CL. In the following, we will exemplify how this criterion can be met by manual data annotation by using naive coders, or by

6 Note that our criterion of T-non-theoretical grounding is related to the more specific concept of **operationalization** in social sciences (Friedrichs 1973). Operationalization refers to the process of developing **indicators** of the form "X is an a if Y is a b (at time t)" to connect T-theoretical and T-non-theoretical levels. We will stick with the more general criterion in the rest of this article.

embedding data annotation into a task extrinsic to the theory to be tested, or by using independently created language data that are available in the wild.

4.1 *T*-non-theoretical Grounding by Naive Coders and Crowdsourcing

Now that we have defined the criterion of *T*-non-theoretical grounding, we see that Krippendorff's (2004) request for "coders that are widely available, follow explicit and communicable instructions (a data language), and work independently of each other" can be regarded as a concrete strategy to satisfy our criterion. The key is the requirement for coders to be "widely available" and to work on the basis of "explicit and communicable instructions." The need to communicate the annotation task to non-experts serves two purposes: On the one hand, the goal of reproducibility is supported by having to communicate the annotation task explicitly in written form. Furthermore, the "naive" nature of annotators requires a verbalization in words comprehensible to non-experts, without the option of relying on implicit or tacit knowledge that is shared among expert annotators. The researcher will thus be forced to describe the annotation task without using technical terms that are common to experts, but are not known to naive coders.

Annotation by naive coders can be achieved by using crowdsourcing services such as Amazon's Mechanical Turk,[7] or alternatively, by creating games with a purpose (von Ahn and Dabbish 2004; Poesio et al. 2013).[8] Non-expert annotations created by crowdsourcing have been shown to provide expert-level quality if certain recommendations on experiment design and quality control are met (Snow et al. 2008). Successful examples of the use of crowdsourcing techniques for data annotation and system evaluation can be found throughout all areas of NLP (see Callison-Burch and Dredze [2010] for a recent overview). The main advantage of these techniques lies in the ability to achieve high-quality annotations at a fraction of the time and the expense of expert annotation. However, a less apparent advantage is the need for researchers to provide succinct and comprehensible descriptions of Human Intelligence Tasks, and the need to break complex annotation tasks down to simpler basic units of work for annotators. Receiving high-quality annotations with sufficient inter-worker agreement from crowdsourcing can be seen as a possible litmus test for a successful *T*-non-theoretical grounding of complex annotation tasks. Circularity issues will vanish because *T*-theoretical terms cannot be communicated directly to naive coders.

4.2 Grounding by Extrinsic Evaluation and Task-Related Annotation

Another way to achieve *T*-non-theoretical grounding is **extrinsic evaluation** of NLP systems. This type of evaluation assesses "the effect of a system on something that is external to it, for example, the effect on human performance at a given task or the value added to an application" (Belz 2009) and has been demanded for at least 20 years (Spärck Jones 1994). Extrinsic evaluation is advertised as a remedy against "closed problem" approaches (Spärck Jones 1994) or against "closed circles" in intrinsic evaluation where system rankings produced by automatic measures are compared with human rankings which are themselves unfalsifiable (Belz 2009).

7 `http://www.mturk.com`.

8 See Fort, Adda, and Cohen (2011) for a discussion of the ethical dimensions of crowdsourcing services and their alternatives.

An example of an extrinsic evaluation in NLP is the evaluation of the effect of syntactic parsers on retrieval quality in a biomedical IR task (Miyao et al. 2008). Interestingly, the extrinsic set-up revealed a different system ranking than the standard intrinsic evaluation, according to F-scores on the Penn WSJ corpus. Another example is the area of clustering. Deficiencies in current intrinsic clustering evaluation methods have led von Luxburg, Williamson, and Guyon (2012) to pose the question "Clustering: Science or Art?". They recommend to measure the usefulness of a clustering method for a particular task under consideration, that is, to always study clustering in the context of its end use.

Extrinsic scenarios are not only useful for the purpose of evaluation. Rather, every extrinsic evaluation creates data that can be used as training data for another learning task (e.g., rankings of system outputs with respect to an extrinsic task can be used to train discriminative (re)ranking models). For example, Kim and Mooney (2013) use the successful completion of navigation tasks to create training data for reranking in grounded language learning. Nikoulina et al. (2012) use retrieval performance of translated queries to create data for reranking in statistical machine translation. Clarke et al. (2010) use the correct response for a query to a database of geographical facts to select data for structured learning of a semantic parser. Thus the extrinsic set-up can be seen as a general technique for T-non-theoretical grounding in training as well as in testing scenarios. Circularity issues will not arise in extrinsic set-ups because the extrinsic task is by definition external to the system outputs to be tested or ranked.

4.3 Grounded Data in the Wild

Halevy, Norvig, and Pereira (2009, page 8) mention statistical speech recognition and statistical machine translation as "the biggest successes in natural-language-related machine learning." This success is due to the fact that "a large training set of the input–output behavior that we seek to automate is available to us *in the wild*." While they emphasize the large size of the training set, we think that the aspect that the training data are given as a "natural task routinely done every day for a real human need" (Halevy, Norvig, and Pereira 2009), is just as important as the size of the training set. This is because a real-world task that is extrinsic and independent of any scientific theory avoids any methodological circularity in data annotation and enforces an application-based evaluation.

Speech and translation are not the only lucky areas where data are available in the wild. Other data sets that have been "found" by NLP researchers are IMDb movie reviews (exploited for sentiment analysis by Pang, Lee, and Vaithyanathan [2002]), Amazon product reviews (used for multi-domain sentiment analysis by Blitzer, Dredze, and Pereira [2007]), Yahoo! Answers (used for answer ranking by Surdeanu, Ciaramita, and Zaragoza [2008]), reading comprehension tests (used for automated reading comprehension by Hirschman et al. [1999]), or Wikipedia (with too many uses to cite). Most of these data were created by community-based efforts. This means that the data sets are freely available and naturally increasing.

The extrinsic and independent aspect of data in the wild can also be created in crowdsourcing approaches that enforce a distinction between data annotation tasks and scientific modeling. For example, Denkowski, Al-Haj, and Lavie (2010) used Amazon's Mechanical Turk to create reference translations for statistical machine translation by monolingual phrase substitutions on existing references. "Translations" created by workers that paraphrase given references without knowing the source can never lead to the circularity that data annotation by experts is susceptible to. In a

scenario of monolingual paraphrasing for reference translations even inter-annotator agreement is not an issue anymore. Data created by single annotators (e.g., monolingual meaning equivalents created for bilingual purposes [Dreyer and Marcu 2012]), can be treated as "given" data for machine learning purposes, even if each network of meaning equivalences is created by a single annotator.

5. Conclusion

In this article, we have argued that the problem of theoretical terms as identified for theoretical physics can occur in empirical CL in cases where data are not "given" as commonly assumed in machine learning. We exemplified this problem on the example of manual data annotation by experts, where the task of relating instances to labels in manual data annotation and the task of relating instances to labels via modeling feature functions are intertwined. Inspired by the structuralist theory of science, we have defined a criterion of *T*-non-theoretical grounding and exemplified how this criterion can be met by manual data annotation by using naive coders, or by embedding data annotation into a task extrinsic to the theory to be tested, or by using independently created language data that are available in the wild.

Our suggestions for *T*-non-theoretical grounding are related to work on grounded language learning that is based on weak supervision in the form of the use of sentences in naturally occurring contexts. For example, the meaning of natural language expressions can be grounded in visual scenes (Roy 2002; Yu and Ballard 2004; Yu and Siskind 2013) or actions in games or navigation tasks (Chen and Mooney 2008, 2011). Because of the ambiguous supervision, most such approaches work with latent representations and use unsupervised techniques in learning. Our suggestions for *T*-non-theoretical grounding can be used to avoid circularities in standard supervised learning. We think that this criterion should be considered a necessary condition for an empirical science, in addition to ensuring reliability of measurements. Our negligence of related issues such as validity of measurements (see Krippendorff 1980b) shows that there is a vast methodological area to be explored, perhaps with further opportunity for guidance by philosophy of science.

Acknowledgments

We are grateful for feedback on earlier versions of this work from Sebastian Padó, Artem Sokolov, and Katharina Wäschle. Furthermore, we would like to thank Paola Merlo for her suggestions and encouragement.

References

Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Balzer, Wolfgang. 1996. Theoretical terms: Recent developments. In Wolfgang Balzer and C. Ulises Moulines, editors, *Structuralist Theory of Science. Focal Issues, New Results*. de Gruyter, pages 139–166.

Balzer, Wolfgang and C. Ulises Moulines, editors. 1996. *Structuralist Theory of Science. Focal Issues, New Results*. de Gruyter.

Beigman Klebanov, Beata and Eyal Beigman. 2009. From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503.

Belz, Anja. 2009. That's nice ... what can you do with it? *Computational Linguistics*, 35(1):111–118.

Blitzer, John, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 440–447, Prague.

Callison-Burch, Chris and Mark Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL-HLT 2010*

Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 1–12, Los Angeles, CA.
Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):1–6.
Chen, David L. and Raymond J. Mooney. 2008. Learning to sportscast: A test of grounded language learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML'08)*, pages 128–135, Helsinki.
Chen, David L. and Raymond J. Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI'11)*, pages 859–866, San Francisco, CA.
Clarke, James, Dan Goldwasser, Wing-Wei Chang, and Dan Roth. 2010. Driving semantic parsing from the world's response. In *Proceedings of the 14th Conference on Natural Language Learning (CoNLL'10)*, pages 18–27, Uppsala.
Corfield, David, Bernhard Schölkopf, and Vladimir Vapnik. 2009. Falsificationism and statistical learning theory: Comparing the Popper and Vapnik-Chervonenkis dimensions. *Journal for General Philosophy of Science*, 40:51–58.
Denkowski, Michael, Hassan Al-Haj, and Alon Lavie. 2010. Turker-assisted paraphrasing for English-Arabic machine translation. In *Proceedings of the NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 66–70, Los Angeles, CA.
Dreyer, Markus and Daniel Marcu. 2012. HyTER: Meaning-equivalent semantics for translation evaluation. In *Proceedings of 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2012)*, pages 162–171, Montreal.
Dundar, Murat, Balaji Krishnapuram, Jinbo Bi, and R. Bharat Rao. 2007. Learning classifiers when the training data is not IID. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07)*, pages 756–761, Hyderabad.
Finke, Peter. 1979. *Grundlagen einer linguistischen Theorie. Empirie und Begründung in der Sprachwissenschaft*. Vieweg.
Fort, Karën, Gilles Adda, and K. Bretonnel Cohen. 2011. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.
Friedrichs, Jürgen. 1973. *Methoden empirischer Sozialforschung*. Opladen, Westdeutscher Verlag, 14th (1990) edition.
Gadenne, Volker. 1985. Theoretische Begriffe und die Prüfbarkeit von Theorien. *Zeitschrift für allgemeine Wissenschaftstheorie*, XVI(1):19–24.
Graf, Erik and Leif Azzopardi. 2008. A methodology for building a patent test collection for prior art search. In *Proceedings of the 2nd International Workshop on Evaluating Information Access (EVIA)*, pages 60–71, Tokyo.
Guo, Yunsong and Carla Gomes. 2009. Ranking structured documents: A large margin based approach for patent prior art search. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'09)*, pages 1,058–1,064, Pasadena, CA.
Halevy, Alon, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24:8–12.
Hirschman, Lynette, Marc Light, Eric Breck, and John D. Burger. 1999. Deep read: A reading comprehension system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 325–332, College Park, MD.
Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT-NAACL'06)*, pages 57–60, New York, NY.
Kilgarriff, Adam. 1999. 95% replicability for manual word sense tagging. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 277–278, Bergen.
Kim, Joohyun and Raymond J. Mooney. 2013. Adapting discriminative reranking to grounded language learning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 218–277, Sofia.
Korb, Kevin. 2004. Introduction: Machine learning as philosophy of science. *Minds and Machines*, 14(4):1–7.
Krippendorff, Klaus. 1980a. *Content Analysis. An Introduction to Its Methodology*. Sage, third (2013) edition.

Krippendorff, Klaus. 1980b. Validity in content analysis. In Ekkehard Mochmann, editor, *Computerstrategien für die Kommunikationsanalyse*. Campus, pages 69–112.

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

Miyao, Yusuke, Rune Saetre, Kenji Sagae, Takuya Matsuzaki, and Jun'ichi Tsujii. 2008. Task-oriented evaluation of syntactic parsers and their representations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT'08)*, pages 46–54, Columbus, OH.

Nikoulina, Vassilina, Bogomil Kovachev, Nikolaos Lagos, and Christof Monz. 2012. Adaptation of statistical machine translation model for cross-lingual information retrieval in a service context. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, pages 109–119, Avignon.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*, pages 79–86, Philadelphia, PA.

Poesio, Massimo, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for Internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, 3(1):Article 3.

Reidsma, Dennis and Jean Carletta. 2007. Reliability measurements without limits. *Computational Linguistics*, 34(3):319–326.

Roy, Deb K. 2002. Learning visually grounded words and syntax for a scene description task. *Computer Speech and Language*, 16:353–385.

Sneed, Joseph D. 1971. *The Logical Structure of Mathematical Physics*. D. Reidel.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*, pages 254–263, Edinburgh.

Spärck Jones, Karen. 1994. Towards better NLP system evaluation. In *Proceedings of the Workshop on Human Language Technology (HLT'94)*, pages 102–107, Plainsboro, NJ.

Stegmüller, Wolfgang. 1979. *The Structuralist View of Theories. A Possible Analogue of the Bourbaki Programme in Physical Science*. Springer.

Stegmüller, Wolfgang. 1986. *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie. Band II: Theorie und Erfahrung*. Springer.

Surdeanu, Mihai, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to rank answers on large online QA collections. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, pages 719–727, Columbus, OH.

Taskar, Ben, Dan Klein, Michael Collins, Daphne Koller, and Christopher Manning. 2004. Max-margin parsing. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*, pages 1–8, Barcelona.

Tsochantaridis, Ioannis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. 2005. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 5:1453–1484.

von Ahn, Luis and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'04)*, pages 319–326, Vienna.

von Kutschera, Franz. 1982. *Grundfragen der Erkenntnistheorie*. de Gruyter.

von Luxburg, Ulrike, Robert C. Williamson, and Isabelle Guyon. 2012. Clustering: Science or art? In *Proceedings of the ICML 2011 Workshop on Unsupervised and Transfer Learning*, pages 1–12, Bellevue, WA.

Yu, Chen and Dana H. Ballard. 2004. On the integration of grounding language and learning objects. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI'04)*, pages 488–493, San Jose, CA.

Yu, Haonan and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 53–63, Sofia.