Title
An Empirically Grounded Approach to Extend the Linguistic Coverage and Lexical Diversity of Verbal Probabilities

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 36(36)

ISSN
1069-7977

Authors
Engelmann, Chrsitine
Hahn, Udo

Publication Date
2014

Peer reviewed

An Empirically Grounded Approach to Extend the Linguistic Coverage and Lexical Diversity of Verbal Probabilities

Christine Engelmann Udo Hahn

Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena
Jena, Germany
engelmann.christine@uni-jena.de udo.hahn@uni-jena.de

Abstract

Linguistic expressions indicating uncertainty of states of knowledge or beliefs, such as *"possible"* or *"might suggest"*, are usually dealt with in the psycholinguistic community under the heading of 'verbal probabilities'. Despite a remarkable level of quantitative and experimental rigor, studies dealing with this phenomenon suffer from several methodological shortcomings: The selection of items under scrutiny usually lacks empirical justification besides subjective preferences, the items are often investigated in isolation, i.e. without sufficient linguistic context and focus is typically on only few word classes, usually adjectives and adverbs. Our study introduces a rigorous empirical, corpus-based criterion for the selection of relevant items, thus balancing the variety of word classes and, as a consequence, enlarging the lexical diversity dealt with in this area of research. We also collect preliminary evidence for the impact discourse context has on the properly adjusting verbal probabilities.

Keywords: verbal probabilities, epistemic modality, empirical semantics, uncertainty in language comprehension

Introduction

Our daily communication is full of linguistic signals to indicate lack of certainty or different degrees of belief in what we are saying. Choices of modal verbs (*"may"*), adjectives (*"possible"*), adverbs (*"probably'*) or lexical verbs (*"suggest"*), etc. are adequate means to calibrate the likeliness we attribute to a proposition we utter. The relevance of this phenomenon, commonly called *verbal probabilities* in the psychological community and *epistemic modality* in the linguistic community, has early been recognized by cognitive scientists who focus on the study of language comprehension (cf., e.g. Lichtenstein and Newman (1967)).

Still, the way these investigations have been carried out up until now suffers from several methodological shortcomings. First, the specific lexical items are collected with a considerable subjective bias mostly based on individual preferences. To the best of our knowledge, there is no study which justifies the selection of items under scrutiny by empirical criteria (e.g. distribution frequencies in a corpus). Given the long history of lexical association tasks in cognitive science, there is also no wonder that verbal probabilities are primarily studied without linguistic context. So many studies focus on the probability of *"possible"* or *"likely"* in complete isolation. Finally, the focus of previous work has predominantly been on few selected word classes, such as adjectives and adverbs, without paying equal attention to lexical verbs or nouns as carriers of probability information.

Our study introduces a rigorous empirical, corpus-based criterion for the selection of relevant items, thus balancing the variety of word classes and, as a consequence, enlarging the lexical diversity dealt with in this area of research. We also collect preliminary evidence for the impact discourse context has on the properly adjusting verbal probabilities.

Related Work

There is a long tradition and a vast amount of literature concerned with the translation of verbal into numerical probabilities (an extensive discussion is provided by Clark (1990)). The approaches are diverse, including e.g. the assignment of numbers to expressions (Lichtenstein & Newman, 1967; Reagan, Mosteller, & Youtz, 1989; Clarke, Ruffin, Hill, & Beamen, 1992), the assignment of expressions to numbers (Reagan et al., 1989), pair comparison (Budescu & Wallsten, 1985; Wallsten, Budescu, Rapport, Zwick, & Forsyth, 1986) and rank-ordering (Budescu & Wallsten, 1985). Usually, scales ranging from '0' to '1' or from 0% to 100% probability are employed. Despite the differences in methods results are relatively comparable (Reagan et al., 1989; Clarke et al., 1992). Teigen and Brun (2003) summarize the main findings by stipulating two claims—a high degree of similarity in the mean estimates between study groups, on the one hand, and a high degree of inter-individual variability within groups, on the other hand.

These observations have led researchers to focus on the inherent vagueness of probability expressions. The core of such investigations is the modeling of verbal probabilities as fuzzy concepts and their subsequent characterization as membership functions over the probability scale (Wallsten & Budescu, 1995). In this respect, probabilities can be assigned values ranging from '0', if they are not included in the concept, to '1', if they are perfect exemplars of the concept. The vagueness of a specific expression is then represented by location, range and shape of the membership function. Recently, this approach has been adapted in a study by Bocklisch, Bocklisch, Baumann, Scholz, and Krems (2010). The authors describe a two-step procedure which includes direct estimations from participants of minimal, maximal and best corresponding probability values, as well as data analysis in terms of membership function construction. Furthermore, considerable work has been carried out on factors that might influence the interpretation and choice of verbal probabilities, e.g. extra-linguistic context (Brun & Teigen, 1988),

prior probabilities (Juanchich, Teigen, & Villejoubert, 2010) or speaker's perspective (Smits & Hoorens, 2005).

A fundamental drawback of almost all of the previous approaches is related to the biased selection of uncertainty expressions that are presented in judgment tasks. The choice is usually subjective, often based on individual preferences. On the one hand, lexical items are selected that are deemed most commonly used and conventional to describe the expression of probability (Reagan et al., 1989; Clarke et al., 1992). There is, however, a lack of solid empirical data, such as corpus-based frequency information, to justify this choice. Due to this selection bias, on the other hand, the focus is mainly on a very restricted subclass of parts of speech, namely adjectives like *"possible"*, adverbs like *"probably"* or nouns like *"chance"*. Only very lately this focus has switched to other word classes such as modal auxiliaries (Teigen & Filkuková, 2013). Hence, the stimulus material used in translation studies cannot be understood to properly reflect the actual language use. From a linguistic perspective this seems to be problematic since a multitude of alternative expressions—which all can be subsumed under the general notion of epistemic modality—may serve to qualify a proposition in terms of probability. Examples are lexical verbs such as *"believe"* or more complex phrases such as *"remains to be shown"*.

Method

We conducted an online questionnaire study to gather probability judgments for a representative set of verbal probabilities in a special sub-domain of language use, scientific discourse. This approach follows studies by Scott, Barone, and Koeling (2012) and de Marneffe, Manning, and Potts (2012), who looked at summarizing evaluations for uncertainty expressions with regards to annotation processes in biomedical texts. In the following, we will discuss the construction of the study material, the recruitment of participants and a preliminary data analysis.

Material

For the selection of a wide range of verbal probabilities we made use of previous annotation efforts in our reference domain, the life sciences. We resorted to the BIOSCOPE Corpus (Vincze, Szarvas, Farkas, Móra, & Csirik, 2008) which comprises more than 20,000 sentences (taken from radiology reports, biological journal articles and their abstracts) annotated for negation and speculative words together with their scope, as well as the Meta-Knowledge enrichment of the GENIA Event Corpus (Kim, Ohta, Tateisi, & Tsujii, 2003) provided by Thompson, Nawaz, McNaught, and Ananiadou (2011) which includes annotations for 36,858 biological events with regards to several dimensions such as polarity or certainty level. The exact clue words indicating values for these dimensions were directly marked. For reasons of compatibility of these resources, we focused for both corpora on the abstracts of scientific articles only.

We first extracted the word forms marked for speculation and certainty (e.g. *"suggests"*, *"suggested"*) and then normalized them to their base forms (e.g. *"suggest"*). Expressions not primarily indicating probability (e.g. *"ability"*) were sorted out. Finally, we calculated the relative frequency (h) for the base forms in each of the two corpora separately and selected 19 expressions with the highest relative frequency in both corpora (see Table 1; first two column blocks). Although expressions indicating low probability were rare in the corpora, we included three of them (see Table 1; third column block) to also populate extreme positions at the lower side of the probability scale.

Expression	h	**Expression**	h	**Expression**	h
suggest	.231	*putative*	.015	*no evidence*	.0008
may	.185	*propose*	.014	*unlikely*	.0008
can	.131	*think*	.013	*cannot*	.0004
indicate	.113	*seem*	.012		
appear	.058	*unknown*	.010		
might	.026	*possibly*	.009		
could	.024	*imply*	.008		
likely	.023	*potentially*	.008		
possible	.020	*hypothesis*	.008		
potential	.017				

Table 1: Frequency-ordered List of Epistemic Modal Expressions Indicating Probability

In natural language use, verbal probabilities are often not interpretable without embedding into appropriate linguistic context. This is partly captured in some studies by including modifications such as *"**very** likely"* or *"**almost** certain"* (Lichtenstein & Newman, 1967; Reagan et al., 1989). We extended our study material by including further examples for certain selected expressions such as modifications by degree adverbs (*"strongly suggest"*), passive voice constructions (*"has/have been suggested"*), co-occurrences of modal auxiliaries and lexical verbs (*"suggest + may"*), and the embedding of probability nouns as arguments of specific verbs (*"investigate + hypothesis"*, *"support + hypothesis"*). Finally, we included the neutral expression *"examine whether"*. This resulted in a total of 27 verbal probabilities for our investigation.

Item Construction

As opposed to adjectives or adverbs, most of the verbal probabilities in our study cannot reasonably be presented in isolation (consider judging the probability for the modal auxiliary *"may"* without further linguistic information). We therefore constructed our questionnaire items as sentences containing the expression under scrutiny as well as a central proposition. This proposition is modified by the expression in that there is a certain probability of it being true. We mainly resorted to the original sentences in the two corpora, slightly modifying them in some cases to avoid cumbersome length or confusing anaphoric expressions (e.g. *"**these results** suggest"*).

Brun and Teigen (1988) already point out that the embedding sentence content can influence the probability interpretation of the expressions under investigation. In an attempt to balance such effects, we arbitrarily constructed three different sentences for each expression. We further included a test item

containing no verbal probability at all. This item served to investigate whether participants really only judged the probability expressions provided, or whether they (wrongly) referred to the content of the statement itself. The prediction was thus that the test item would consistently receive the highest probability rating possible. Deviations from this assumption might indicate that such a participant should be excluded from further analysis.

Data Collection

To collect our data, we conducted an online-questionnaire study using the software package SoSci Survey.[1] In this study, participants were randomly presented one of the three sentences for each verbal probability, so that, on the whole, they had to provide 28 judgments (including the test item). The order of presentation for the different expressions was randomized. We partly followed the mentioned approaches in membership function construction in that we did not only collect a single probability value for each expression, but rather tried to capture the inherent vagueness of verbal probabilities by asking participants to give ratings for probability ranges. The most representative value for an expression was then derived by calculating the midpoint of these ranges. This procedure was validated during a pre-test where we asked whether participants thought that the midpoint of a suggested range would optimally represent a verbal probability, if only one value had to be selected. We thus obtained measures for four ratio-scaled variables: direct measures for LOWER BOUNDARY and UPPER BOUNDARY, as well as indirect measures for VAGUENESS (the range between both boundaries) and OPTIMAL PROBABILITY (the midpoint between the boundaries).

To provide these measures in the study, participants were asked to rate the probability they think the author of the sentence attributes to a (scientific) statement as signaled by the use of certain cue words (the verbal probabilities). To ease orientation the relevant cues were marked in red, whereas the modified statement appeared in blue. Below the sentence the participants were presented the probability scale ranging from 0% to 100% probability (with divisions into 10% units) that the statement is true. The subjects were asked to use two sliders to indicate the range of probabilities they deemed appropriate. Figure 1 depicts the general experimental set-up.

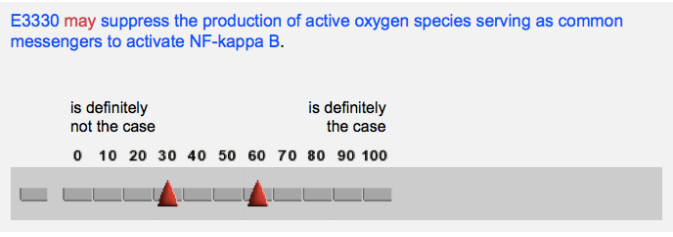

Figure 1: Questionnaire Set-up

[1] https://www.soscisurvey.de

Participants

We recruited participants by email distribution of the link to the online survey. Target subjects were persons with a biomedical or linguistic education background, since these subjects typically serve as text annotators in our reference domain and were thus assumed to provide representative judgments concerning our area of language use. We contacted selected biomedical as well as linguistic institutions in the German and English speaking region and received responses from 97 subjects, of which ten only completed a fraction of the survey and one person stated to have systematically misunderstood the task. The final group of subjects taking part in our study ($n = 86$, in total; 43 women and 41 men (two participants had omitted this information); mean age 31.70 years with standard deviation 10.73; 65 German, 16 English speaking subjects and five subjects with other mother tongues) can thus be viewed as an *ad hoc* sample.

Explorative Data Analysis

Preliminary data analysis showed that not all participants had worked on the judgment task as expected. The following problematic points had to be considered before summarizing and presenting the data:

- Two cases were repeatedly marked as outliers, in that these participants had judged a considerable amount of expressions (e.g. *"propose"*, *"can"*) to indicate a probability of 0%. Both sets were thus excluded from further analysis.
- An unexpectedly large group of participants had only moved one of the sliders and thus not provided ratings of probability ranges. To handle such cases we referred to findings in our pre-test, where similar rating strategies had been observed. Since pre-test subjects had indicated not to have read instructions carefully enough and thus only provided one optimal probability value, in the actual study we also treated such single values as measures for OPTIMAL PROBABILITY, disregarding the other variables for the subjects in question.
- Nearly 40% of our participants had not provided probability values of 100% for the test item, i.e. the one containing no verbal probability at all, with a safe probability value of 100%. Of these, 19 subjects had provided values for OPTIMAL PROBABILITY of less then 100% and 14 subjects had omitted the test item.
  As excluding these participants would have resulted in a considerable loss of data, we first tested whether the response frequencies were due to the non-applicability of the task to the test item (asking subjects to rate the marked expressions might have led to confusion for a sentence that did not include any red markings). We tested whether subject groups (made operational in terms of the independent nominal variable ANSWER BEHAVIOR ON TEST ITEM: 100% – $<$ 100% – n/a) differed with regard to ratings on all other items. We here looked only at participants with a biomedical background (57 participants), since only these subjects would have sufficient knowledge to systematically restrict judgments to the actual content of the state-

ment. In this biomedical subgroup, 11 subjects had provided values for OPTIMAL PROBABILITY of less then 100% and 8 subjects had omitted the test item.

Significance tests were conducted for all 27 verbal probabilities separately, using One-Way ANOVA for 11 expressions where the condition of normal distribution for residuals was satisfied, and non-parametric Kruskal-Wallis test for the other 16 expressions. Tests were conducted on a Bonferroni-corrected significance level of $\alpha = .002$. For none of the verbal probabilities we found a significant main effect of ANSWER BEHAVIOR ON TEST ITEM on OPTIMAL PROBABILITY, suggesting that participants' deviant ratings of the test item were not an indicator for systematically judging statement content instead of the meaning of the probability expression. We thus kept the data for the participants in question.

- For the three items containing low probability expressions (see Table 1; 3rd column block) rating distributions were striking, since (in contrast to all other expressions) they were bimodal or binormal, with an unexpectedly large amount of high probability ratings (see Figures 2, 3 and 4). Visual inspection suggested the existence of two participant groups with different rating strategies which we made operational in terms of an independent nominal variable ANSWER BEHAVIOR ON LOW PROBABILITY ITEMS (consistently $< 50\%$ – mixed answers – consistently $\geq 50\%$). 39 subjects had consistently provided values for OPTIMAL PROBABILITY of less then 50%, the ratings of 15 subjects were consistently equal to or higher than 50%, and 32 subjects had given mixed results.

  Since high probability ratings for the mentioned expressions were counter-intuitive, we again compared groups with regard to ratings on all other items. Significance tests were conducted for 24 verbal probabilities, using One-Way-ANOVA (10 expressions) as well as Kruskal-Wallis test (14 expressions) on a Bonferroni-corrected significance level of $\alpha = .002$. A significant main effect of ANSWER BEHAVIOR ON LOW PROBABILITY ITEMS on OPTIMAL PROBABILITY was only found for *"unknown"* ($F(2,78) = 10.942, p = .000, partial\ \eta^2 = .22$). *Post hoc* comparisons between groups via Mann-Whitney-U tests showed that participants that had consistently attributed probabilities equal to or higher than 50% (median = 60.00) differed significantly in their ratings for *"unknown"* from participants rating low probability items consistently lower than 50% (median = 30.00) ($U = 88.50, p = .000, r = -.53$), as well as from those providing mixed answers (median = 40.00) ($U = 111.50, p = .006, r = -.41$). Based on these findings, for the analysis of low probability items we excluded all ratings equal to or higher than 50%. A discussion on why we might have found such deviations, while the rest of the ratings are comparable, is provided below. For the analysis of *"unknown"*, we, finally, excluded all participants that had consistently provided high ratings for low probability items.

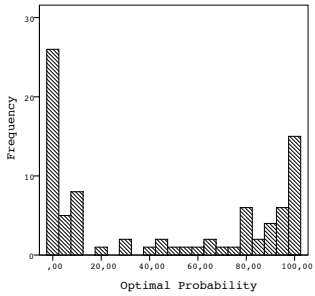

Figure 2: Distribution of Ratings for *"cannot"*

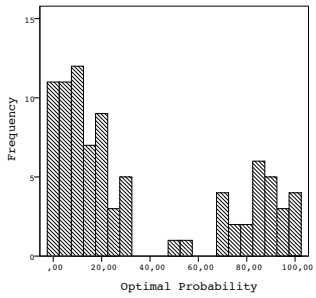

Figure 3: Distribution of Ratings for *"no evidence"*

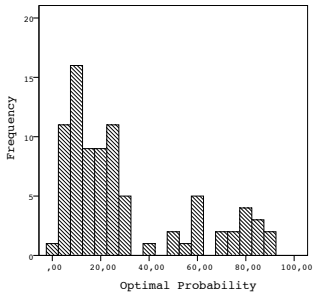

Figure 4: Distribution of Ratings for *"unlikely"*

Results

Probability Values

We now present the rating results for the set of expressions we investigated. Table 2 lists mean values[2] for our four ratio-scaled variables together with the valid cases of participants providing these values (as determined in the exploratory data analysis). To illustrate these results, Figure 5 includes mean values for LOWER and UPPER BOUNDARY (lower and upper bars in the diagram, respectively), as well as OPTIMAL PROBABILITY (circles) and VAGUENESS (length of the lines). The diagram depicts that the 27 ranges we found nearly cover the whole of the probability scale and that two neighboring (relative to their position as indicated by OPTIMAL PROBABILITY) expressions always overlap.

To compare our results to previous findings we looked at expressions that were also dealt with in other studies. Since our item material includes only a few adjectives and adverbs this intersection set is admittedly very small. Reagan et al.

[2] All calcluations were done using the IBM SPSS Statistics (version 21) software package.

Expression	OPTIMAL PROBABILITY	*n*	VAGUENESS	LOWER BOUNDARY	UPPER BOUNDARY	*n*
cannot	07.00	45	08.93	03.21	12.14	28
no evidence	11.81	58	18.54	02.62	21.22	41
unlikely	15.79	63	19.78	05.56	25.33	45
unknown	33.86	66	30.63	17.92	48.54	48
might	47.24	85	28.67	31.17	59.83	60
may	50.47	86	27.21	34.26	61.48	61
possibly	51.01	84	23.39	37.46	60.85	59
could	51.31	84	25.42	36.44	61.86	59
investigate + hypothesis	52.90	81	35.86	33.44	69.31	58
possible	56.16	86	25.90	42.46	68.36	61
has/have been suggested	56.76	85	28.67	39.50	68.17	60
examine whether	56.78	79	41.07	35.46	76.73	55
suggest + may	57.06	85	28.20	41.97	70.16	61
think	59.16	83	26.44	44.92	71.36	59
potentially	59.94	85	27.70	45.08	72.69	61
putative	60.43	82	23.62	47.41	71.03	58
seem	60.82	85	24.92	47.70	72.62	61
potential	61.71	85	26.67	48.50	75.17	60
appear	63.71	85	23.11	50.33	73.44	61
suggest	65.30	84	27.70	51.15	78.85	61
propose	66.06	85	25.08	52.95	78.03	61
likely	68.31	83	21.67	57.33	79.00	60
imply	72.53	85	21.83	61.67	83.50	60
support + hypothesis	73.06	85	22.50	60.17	82.67	60
can	73.35	85	26.67	59.83	86.50	60
indicate	75.76	86	19.84	65.08	84.92	61
strongly suggest	81.14	83	18.83	72.00	90.83	60

Table 2: Probability Ratings and Valid Cases in Detail

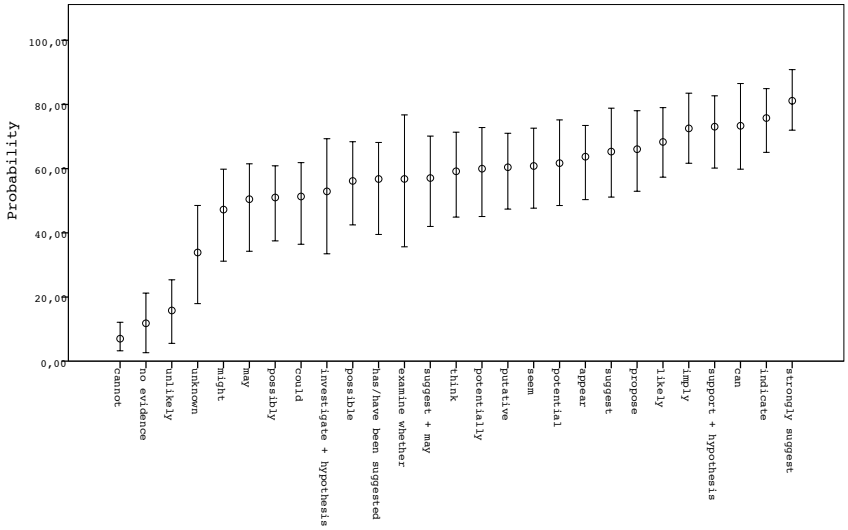

Figure 5: Mean Probability Values for LOWER and UPPER BOUNDARY, OPTIMAL PROBABILITY and VAGUENESS

(1989) already compare their results with previous studies. We include parts of their summary for the adjectives *"unlikely"*, *"possible"* and *"likely"* in Table 3 and add our values for comparison. In contrast to Table 2 we here give median, not mean values for the expressions.

Whereas Lichtenstein and Newman (1967) (here referred to as **LN**) give optimal probability (OP) values, Reagan et al. (1989) (referred to as **R+**) and Wallsten et al. (1986) (referred to as **W+**) provide measures for lower (LB) and upper boundaries (UB). The data show that the ordering of the expressions is identical over all studies and that values for OPTIMAL PROBABILTIY are largely comparable with a maximal difference of 12.5% (*"possible"* in our study *vs.* Reagan et al. (1989)). Probability ranges are also quite similar with the only exception relating to the ratings for *"possible"* in Wallsten et al. (1986), where the range is located considerably lower down the scale.

	LN	R+			W+		Present Study		
	OP	OP	LB	UP	LB	UP	OP	LB	UP
unlikely	16	15	10	25	2	30	15	0	20
possible	49	40	40	70	1	55	52.50	40	70
likely	75	70	65	85	59	90	70	60	80

Table 3: Comparison of Results with Previous Studies

Conclusions

Although there is only little overlap between the verbal probabilities in our and previous studies, the observed similarities regarding the three adjectives suggest that comparable results can be obtained when expressions are presented in isolation as well as in sentential context. This can be seen as generally validating our approach in extending the investigation to verbal probabilities in the wider sense, such as lexical verbs (e.g. *"think"*) and modal auxiliaries (e.g. *"may"*). It should, of course, be noted that the semantic content of such expressions is certainly not restricted to the indication of probability. Especially in scientific writing there are quite a few other factors influencing the choice of terms, such as persuasion strategies, avoiding direct responsibility, or the indication of evidential status. Such motivations are often investigated as hedging phenomena (Hyland, 1998). Some tendencies in our data might be explained by referring to this notion as well.

Most of the expressions which received high probability values (e.g. *"strongly suggest"*, *"imply"*) are terms that indicate (strong) direct evidence for an observation, making this observation highly likely. Expressions located in the middle of the scale might be described as indicating indirect evidence (e.g. *"may"*, *"seem"*) or simply lack of knowledge (e.g. *"examine whether"*). The fact that we received quite divergent ratings for low probability items can also be regarded as evidence that there are different aspects of meaning lending themselves for assessment. Besides probability there is also the notion of speaker certainty. In this view, erroneously high values, e.g. for *"unlikely"*, might indicate that some participants judged certainty (the author is 100% certain that something is not the case), instead of probability (there is a probability of 0% that something is the case). It could be the case that presenting verbal expressions in sentential context might lead to such interferences. Task descriptions should thus be clear enough for participants to provide the correct judgment. Based on work regarding membership function construction our study also includes the evaluation of semantic vagueness by requiring participants to provide ranges of probability values rather than exclusively having them attribute single values. With regard to the above-mentioned variety of semantic aspects this seems to be a legitimate approach.

Acknowledgment. This work has partially been funded by the *German Ministry for Education and Research* [*Bundesministerium für Bildung und Forschung–BMBF*] within the GERONTOSYS funding initiative in collaboration with the *Jena Centre for Systems Biology of Ageing* (JENAGE) under grant no. 0315581.

References

Bocklisch, F., Bocklisch, S. F., Baumann, M. R. K., Scholz, A., & Krems, J. F. (2010). The role of vagueness in the numerical translation of verbal probabilities: a fuzzy approach. In *CogSci 2010—Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 1974–1979).

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes*, *41*(3), 390–404.

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes*, *36*(3), 391–405.

Clark, D. A. (1990). Verbal uncertainty expressions: A critical review of two decades of research. *Current Psychology: Research & Reviews*, *9*(3), 203–235.

Clarke, V. A., Ruffin, C. L., Hill, D. J., & Beamen, A. L. (1992). Ratings of orally presented verbal expressions of probability by a heterogeneous sample. *Journal of Applied Social Psychology*, *22*(8), 638–656.

de Marneffe, M.-C., Manning, C. D., & Potts, C. (2012). Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, *38*(2), 301–333.

Hyland, K. (1998). *Hedging in Scientific Resarch Articles*. John Benjamins.

Juanchich, M., Teigen, K. H., & Villejoubert, G. (2010). Is guilt 'likely' or 'not certain'? Contrast with previous probabilities determines choice of terms. *Acta Psychologica*, *135*(3), 267–277.

Kim, J.-D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus: A semantically annotated corpus for biotextmining. *Bioinformatics*, *19*(Suppl 1), i180–i182.

Lichtenstein, S., & Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science*, *9*(10), 563–564.

Reagan, R. T., Mosteller, F., & Youtz, C. (1989). Quantitative meanings of verbal probability expressions. *Journal of Applied Psychology*, *74*(3), 433–442.

Scott, D. S., Barone, R., & Koeling, R. (2012). Corpus annotation as a scientific task. In *LREC 2012—Proceedings of the 8th International Conference on Language Resources and Evaluation* (pp. 1481–1485).

Smits, T., & Hoorens, V. (2005). How probable is probably? It depends on whom you're talking about. *Journal of Behavioral Decision Making*, *18*, 83-96.

Teigen, K. H., & Brun, W. (2003). Verbal expressions of uncertainty and probability. In D. Hardman & L. Macchi (Eds.), *Thinking. Psychological Perspectives on Reasoning, Judgment and Decision Making* (pp. 125–145). Wiley.

Teigen, K. H., & Filkuková, P. (2013). Can $>$ will: Predictions of what can happen are extreme, but believed to be probable. *J. of Behavioral Decision Making*, *26*(1), 68–78.

Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2011). Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, *12*, 393–411.

Vincze, V., Szarvas, G., Farkas, R., Móra, G., & Csirik, J. (2008). The BIOSCOPE corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, *9*(Suppl 11), 9–17.

Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: general principles and empirical evidence. *Knowl. Eng. Rev.*, *10*(1), 43–62.

Wallsten, T. S., Budescu, D. V., Rapport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *J. Experimental Psychology*, *115*, 348–365.