LiLT volume 14, issue 3 August 2016

Modal Sense Classification At Large

Paraphrase-Driven Sense Projection, Semantically Enriched Classification Models and Cross-Genre Evaluations

Ana Marasović‡⋆, Mengfei Zhou†⋆, Alexis Palmer⋆†,
Anette Frank‡ † ⋆,
‡*Research Training Group AIPHES*
†*Leibniz ScienceCampus "Empirical Linguistics and Computational Language Modeling"*
⋆*Department of Computational Linguistics*
Heidelberg University

Abstract*

Modal verbs have different interpretations depending on their context. Their sense categories – epistemic, deontic and dynamic – provide important dimensions of meaning for the interpretation of discourse.

Previous work on modal sense classification achieved relatively high performance using shallow lexical and syntactic features drawn from small-size annotated corpora. Due to the restricted empirical basis, it is difficult to assess the particular difficulties of modal sense classification and the generalization capacity of the proposed models.

In this work we create large-scale, high-quality annotated corpora for modal sense classification using an automatic paraphrase-driven projection approach. Using the acquired corpora, we investigate the modal sense classification task from different perspectives.

We uncover the difficulty of specific sense distinctions by investigat-

*This article represents an extension of prior work reported in Zhou (2015) and Zhou et al. (2015). It includes material of the published version in Zhou et al. (2015), with minor modifications.

ing distributional bias and reducing the sparsity of existing small-scale corpora used in prior work. We build a semantically enriched model for modal sense classification by designing novel features related to lexical, proposition-level and discourse-level semantic factors. Besides improved classification performance, closer examination of interpretable feature sets unveils relevant semantic and contextual factors in modal sense classification. Finally, we investigate genre effects on modal sense distribution and how they affect classification performance.

Our investigations uncover the difficulty of specific sense distinctions and how they are affected by training set size and distributional bias. Our large-scale experiments confirm that semantically enriched models outperform models built on shallow feature sets. Cross-genre experiments shed light on differences in sense distributions across genres and confirm that semantically enriched models have high generalization capacity, especially in unstable distributional settings.

1 Introduction

Factuality recognition (Saurí, 2008, de Marneffe et al., 2011, 2012, Saurí and Pustejovsky, 2012) is an important subtask in information extraction. Beyond the bare filtering aspects of veridicality recognition, classification of **modal senses** plays an important role in text understanding, plan recognition, and the emerging field of argumentation mining. Communication often revolves about *hypothetical, planned, apprehended or desired states of affairs*. Such "extra-propositional meanings" (Morante and Sporleder, 2012), or *intensional contexts* are often linguistically marked using modal verbs, adverbs, or attitude verbs,[2] as in (1) for hypothetical situations, or (2) for expression of *desires* or *requests*.

(1) a. He *must*'ve hurt himself.
    b. He has *certainly* found the place by now.
    c. We *anticipate* that no one will leave.

(2) a. We *must* solve this problem.
    b. It is *mandatory* to resolve this situation.
    c. I *want you* to solve this problem.

In the present work, we focus on modal verbs and their epistemic and non-epistemic meaning distinctions. Following Kratzer (1991)'s seminal work in formal semantics, recent computational approaches such as Ruppenhofer and Rehbein (2012) distinguish different modal 'senses',

[2]These are common strategies for encoding extra-propositional meaning in English and many other European languages; other grammatical mechanisms beyond these constructions are employed in the world's languages.

most prominently, *epistemic* (3.a), *deontic/bouletic* (3.b) and *circumstantial/dynamic* (3.c) modality.

(3) a. Geez, Buddha *must* be so annoyed!
(epistemic – possibility)

b. We *must* have clear European standards.
(deontic – permission/request)

c. She *can*'t even read them.
(dynamic – ability)

Modal sense tagging is typically framed as a supervised classification task, as in Ruppenhofer and Rehbein (2012). They manually annotated the modal verbs *must, may, can, could, shall* and *should* in the MPQA corpus of Wiebe et al. (2005). The obtained data set comprises 1330 instances. Individual lexical classifiers trained on this data set yield accuracies ranging from 68.7 to 93.5. While these accuracies seem high, there is a strong distributional bias in the training data. Due to the small data set size and its skewed distribution of senses, classifiers seem to overfit and hardly beat the majority baseline. It is thus an open question whether the obtained models have sufficient generalization capacity when applied to novel data with different sense distributions.

In this paper we reexamine the prior work on modal sense classification and show that indeed specific sense distinctions are difficult for state-of-the art models. This effect was obscured in previous work by the skewed distribution of modal senses in the small data sets.

In our work we aim to provide improved models for modal sense classification (i) that are based on larger and re-proportioned training data sets, (ii) that are linguistically insightful and (iii) that are generalizable across genres and robust against variations in sense distribution.

The main contributions and insights of our work are as follows:

i. Paraphrase-driven modal sense projection. We address the manual annotation bottleneck for modal sense labeling by devising a paraphrase-driven cross-lingual modal sense projection approach. We demonstrate that using this method we are able to create large volumes of sense-annotated data of very high quality with minimal effort.

ii. Semantically enriched models for modal sense classification. Using larger volumes of heuristically sense-annotated data, we demonstrate that specific modal sense distinctions are difficult to discriminate for state-of-the-art models.

Examples of difficult sense distinctions are dynamic vs. deontic readings of *can* (4.a), epistemic vs. dynamic readings of *could* (4.b) or epistemic vs. deontic readings of *should* (4.c).

(4) a. You *can* do this, if you want.
ability (dynamic) vs. permission (deontic)
b. He *could* have arrived in time.
possibility (epistemic) vs. ability (dynamic)
c. He *should* be aware of the issue.
possibility (epistemic) vs. obligation (deontic)

We investigate the effect of a semantically enriched feature space for modal sense classification and show that specific sense distinctions greatly benefit from semantic information relating to lexical, proposition-level and discourse-level aspects of meaning.

iii. Genre distinctions in modal sense distribution. Finally, we investigate the variability of modal sense distributions across genres and the generalization capacities of the induced classifier models in view of distributional variability. We construct a manually annotated corpus on the basis of the existing MASC corpus (Ide et al., 2008), sub-divided along various genre distinctions. We observe that certain genres and modal verbs are affected by variations in sense distribution. Experiments on this data set confirm that semantically enriched models outperform shallow feature spaces. Furthermore, we observe higher generalization capability of balanced training settings in view of sense shifts. Overall, our experiments corroborate that larger training data sets and semantically informed models support the induction of robust, scalable and highly generalizable modal sense classification models. By providing evidence for distributional variation across genres, our contribution lays the groundwork for adaptive classification models, to be explored in future work.

Overview. The structure of this article is as follows. We review related work in Section 2. In Section 3 we formulate our research questions. Section 4 outlines our paraphrase-driven approach for modal sense projection using parallel corpora and evaluates the quality of the induced annotations, after which, in Section 5, we present and analyze the data sets we are using as a basis for our empirical analyses: (i.) the manually annotated MPQA corpus used in Ruppenhofer and Rehbein (2012)'s work, (ii.) a large, automatically sense-tagged corpus we obtain from cross-lingual modal sense projection, and finally, (iii.) a manually labeled subcorpus of MASC, subdivided by different genres. Section 6 motivates and defines a semantic feature set for modal sense classification. Section 7 performs systematic classification experiments on the basis of the MPQA and the automatically constructed data set. We investigate various classification models, using different training data sets and feature inventories, evaluating in particular the

impact of semantic features on difficult sense distinctions in balanced and unbalanced training settings. Section 8 analyzes the occurrence of modal verbs and senses in different genres, and investigates the performance and robustness of different classifier models across genres. In Section 9 we summarize and discuss our findings and conclude with an outlook on future work.

2 Related Work

Factuality recognition. Events are presented in discourse as occurring with different degrees of factuality, or veridicality, such as *possible, probable* or *certain*. Such degrees of factuality, as well as their polarity (positive, negative) can be conveyed by modal verbs, adverbs and adjectives, as well as various types of attitude verbs. With FactBank, Saurí and Pustejovsky (2009) proposed an annotation scheme and an annotated resource that distinguishes 8 degrees of (non)factuality of events. Saurí (2008) and Saurí and Pustejovsky (2012) developed a rule-based system for factuality recognition trained on FactBank, including recognition of sources. de Marneffe et al. (2012) refined the annotations on FactBank and developed a machine learning classifier for event factuality using lexical and structural features, as well as approximations of world knowledge. Recent work in Lee et al. (2015) builds on and further improves upon this work. FactBank's focus is on notions of (degrees of) factuality or veridicality of events, and considers primarily epistemic uses of modal verbs. Related work in the biomedical domain (i.a. Light et al., 2004, Thompson et al., 2008, Morante and Daelemans, 2011, Szarvas et al., 2012) similarly focuses on the detection of factuality, hedging, and expressions marking evidence. In contrast, our work is concerned with the task of sense disambiguation of modal verbs, which imply non-factuality of their embedded verbs.

Annotating and classifying modal senses. Most relevant to our work is the current state of the art in automatic modal sense classification in Ruppenhofer and Rehbein (2012).

Ruppenhofer and Rehbein (2012) (henceforth, R&R) manually annotated modal verbs in the MPQA corpus of Wiebe et al. (2005). They build on the well-established inventory of modal sense categories of Kratzer (1991) in formal semantics: *epistemic, deontic/bouletic* and *circumstantial/dynamic* modality, illustrated in (3) above. R&R add three further categories: *concessive*, *conditional* and *optative*. Their annotation scheme proves reliable, with an inter-annotator agreement that ranges from $\mathcal{K}$=0.6 to 0.84 for the different modal verbs.

Rubinstein et al. (2013)'s modality annotation scheme is equally grounded in the categories of Kratzer (1981, 1991), but assumes various subdivisions, resulting in a fine-grained scheme of 8 categories. They also investigate various groupings of these 8 classes, up to a binary distinction of epistemic vs. non-epistemic modality. They measured very poor levels of inter-annotator reliability for the fine-grained classes, and substantial agreement for the binary distinction, at α=0.89.[3] Similar experience is reported by Cui and Chi (2013) for modality annotation on the Chinese Treebank. Nissim et al. (2013) propose a fine-grained hierarchical modality annotation scheme that can be applied cross-linguistically. It includes (subtypes of) factuality, as well as speaker attitude. They do not report annotation experiments. To our knowledge, with the exception of R&R's work, none of these recent annotation schemes has been used for computational tagging.

R&R's annotation scheme differs from an earlier annotation scheme for modality in Baker et al. (2010, 2012), who distinguish 8 categories. Next to *requirement, permissive, want* and *ability* that coincide with Kratzer's categories, they include the intensional categories *success, effort, intention* and *belief*. They measured precision in automatic tagging of 86.3% on a small data set of 249 modality-tagged sentences. Hendrickx et al. (2012) similarly take a broader perspective on modality, focusing more on sentiment and opinion.

The focus of our work is on modal verbs and their established sense inventories from formal semantics: *epistemic*, *deontic/bouletic* and *circumstantial/dynamic*. These categories correspond to Baker et al.'s 4 modal categories (with deontic split into *requirement* and *permissive*) and R&R's inventory, with regrouping of their additional categories.

Automatic modal sense classification and genre effects. Ruppenhofer and Rehbein (2012) perform classification experiments on the manually annotated MPQA corpus. Their classification model employs a mixture of target and contextual features, taking into account surface, lemma and PoS information, as well as syntactic labels and syntactic path features linking targets to their surrounding words and constituents. These features are able to capture very diverse contextual factors, but it is difficult to interpret their impact on distinguishing modal senses. The lexical classifiers yield accuracies between 68.7 and 93.5, depending on the verb.

Closer investigation of their data set, however, reveals a strong bias in sense distributions. As a consequence, the majority sense baselines are hard to beat: none of their classification models is able to beat the

[3]Krippendorff's α, see Krippendorff (2004)

majority baseline with uniform settings across all modal verb types.

Prabhakaran et al. (2012) present annotation and classification experiments on manually annotated data from different genres. Their annotation scheme is based on five categories from Baker et al. (2010): *Ability, Effort, Intention, Success*, and *Want*. They performed preselection based on an existing modality tagger and performed annotation using Amazon Mechanical Turk. Classification was performed with an SVM multiclass classification model based on lexical and shallow (no syntactic) features. They report 41.9 F_1 score on heterogeneous *gold* data consisting of newswire, letters, blog, and email genres. By contrast, 71.9 F_1 score was achieved for 4-fold cross-validation (CV) on a subset of homogeneous genre data consisting of *email only*. Using confidence values from annotator agreement in training, numbers rose to 44.0 F_1 score on *gold data*, and to 91.1 F_1 on within-domain training on *email*.

We thus observe clear overfitting (R&R) and cross-genre effects (Prabhakaran et al.) in state-of-the-art modal sense classification that call for careful analysis of data composition as well as the generalization capacities of specific classification models.

Finally, Passonneau et al. (2014) use a variety of stylistic features to model genre variation in the multi-genre MASC corpus.[4] Using principal component analysis with a set of 37 lexical, word class, and grammatical features (e.g. past tense), four components are identified as relevant for genre classification. These primarily involve typing over nouns and noun phrases (including named entities) but also reflect various attributes of the verb or the clause (e.g. adverbs or past participles). These results support our choice of MASC for studying cross-genre variation for modal senses.

Cross-lingual annotation projection. Cross-lingual annotation projection is a well-known technique of distant supervision that can be applied to acquire annotated training instances for an under-resourced language (the target language) from automatically labeled instances provided by an existing labeling system in another language (the source language) (Resnik, 2004). The annotations on the source language are projected via word alignment from the source to the target language sentences, building on the hypothesis that crucial, especially semantic, properties are shared between translated sentences in a parallel corpus.

This technique has been successfully applied for part-of-speech and named entity tagging (Yarowsky et al., 2001), temporal tagging (Spreyer and Frank, 2008), dependency parsing (Hwa et al., 2005), and

[4]See Section 5.3 for a more detailed discussion of MASC.

semantic role labeling (Padó and Lapata, 2009). In contrast to this line of work, we do not exploit existing and possibly noisy automatic classification models to project annotations to a novel target language. Our aim is to induce high-quality modal sense annotations from scratch, by exploiting unambiguous paraphrases of specific modal senses in a source language and projecting these senses to the target language. A similar technique has been applied for the induction of word sense annotations by Diab and Resnik (2002).

3 Research Questions and Overview

Our leading hypothesis is that modal sense classification can be improved by semantically enriched feature sets. To explore this hypothesis, we aim to find automated methods for extending the size of the currently available data sets, to reduce sparsity and distributional bias. An open question is also whether modal sense classification models are able to generalize across different genres.

We explore these questions in the following order:

Q1: How can we *resolve the manual annotation bottleneck* for modal sense classification, in view of acquiring large-size training corpora from different textual genres?

Q2: Can a *semantically grounded feature space* for modal sense classification improve on prior work in overall performance and robustness against variations in sense distribution?

Q3: Are there *genre differences in the distribution of modal senses*, and to what extent are they mirrored in the performance of different classification models?

In order to investigate modal sense classification from these multiple perspectives, we make use of existing annotated corpora and create novel annotated data sets. All data sets are described in Section 5.

Our models will be trained and evaluated on the primary existing data set annotated for modality, the **MPQA corpus** (cf. subsection 5.1). This **benchmark corpus** allows us to explore the *impact of semantic features* on modal sense classification in a standard setting and facilitates direct comparison to previous work. The existing MPQA corpus is small and shows a biased distribution of senses, so we need more annotated data to determine how well the alternative classification models scale. In order to create large training data sets without extensive manual annotation effort, we create an **automatically sense-tagged data set, EPOS** (cf. subsection 5.2). This dataset is created via cross-lingual projection (described in Section 4) using unambiguous

indicators of modal sense in a source language that is projected to our target language, English.

The core modeling choice we address is the use of **semantic features**, both on their own and in combination with more standard features (cf. Section 6). With the new expanded datasets, we are also able to investigate to what extent different classification models can profit from being trained on *balanced* as opposed to *unbalanced* data. The results of these classification experiments are described in Section 7.

Finally, we investigate the applicability of the classification models built on MPQA and EPOS to corpus data stemming from various genres of text. To this end, we manually annotate portions of the English **multi-genre corpus, MASC** (cf. subsection 5.3) and evaluate our models on this new, multi-genre dataset. We analyze to what extent distributions of modal verbs and modal senses vary across genres and how these differences affect different classification models. These investigations are described in Section 8.

4 Paraphrase-Driven Modal Sense Projection

To address our first **research question Q1**, we propose a method for cross-lingual sense projection that will allow us to *automatically generate high-quality sense-labeled data* for different textual genres in large quantities, and thus to alleviate the manual annotation bottleneck.

Our approach exploits the paraphrasing behaviour of modal senses, which holds across modal verbs, modal adverbs and certain attitude verbs. As illustrated in (5) and (6), this paraphrasing behaviour can also be found across languages.

(5) a. He *may* be home by now. (possibility)
b. You *may* enter this building. (permission)
c. *May* you live 100 years. (wish)

(6) a. *Vielleicht* ist er schon zu Hause.
Maybe is he already at home.
b. Es ist *gestattet*, das Gebäude zu betreten.
It is permitted the building to enter
c. *Hoffentlich* werden Sie 100 Jahre.
Hopefully become you 100 years

Capitalizing on the paraphrasing capacity of such expressions, we apply a semi-supervised cross-lingual sense projection approach, similar to prior work in annotation projection by Diab and Resnik (2002).

(i) We establish a seed set of cross-lingual modal sense indicating paraphrases,

(ii) we extract modal verbs in context that are in direct alignment with one of the seed expressions in word-aligned parallel corpora, and

(iii) we project the sense label of the seed paraphrase to the aligned modal verb.

4.1 Annotation Scheme

As a starting point, we adopt the annotation scheme from Ruppenhofer and Rehbein, which is itself grounded in Kratzer's modal senses *epistemic, deontic* and *dynamic*. R&R add the novel categories *conditional, concessive* and *optative*. Examples are given in (7) (cf. R&R).

(7) *Should* anyone call, please take a message. (conditional)
But, fool though he *may* be, he is powerful. (concessive)
Long *may* she live!' (optative)

We subsume R&R's *conditional* and *concessive* as subtypes of *epistemic* and *optative* as a subtype of *deontic* modality.

Following R&R, we focused our annotation efforts on six modal verbs: *can*, *could*, *may*, *must*, *should*, and *shall*.[5] We established guidelines for annotation that are paraphrase-driven. They ask the annotators to consider four paraphrasing possibilities for modal verbs, in order to judge whether the usage of the modal verb conveys a:

possibility (epistemic) using a paraphrase such as: "someone is *likely/unlikely* to do something", or "something is *likely/possible/(im)probable* to happen/to be the case";

request (deontic) using as paraphrase: "*need to* do something" or "*it is required to* do something";

permission (deontic) using as paraphrase: "*allow/don't allow* somebody to do something", or

ability (dynamic) using the paraphrase "*be (un)able to* do something".

That is, for the annotation process, we split the category deontic into *permission* and *request*, and later merged these back to *deontic*. This was done in order to make the task more accessible also for non-experts. Examples are given below, together with the assigned sense labels.

(8) a. Sorry, I *must* have made a mistake. Possibility
Paraphrase: Sorry, *it is likely* that I have made a mistake.
b. You *should* go home now. Request
Paraphrase: You *need to* go home now.

[5] *might* has been established to be monosemous in R&R, so we excluded it from analysis and classification.

c. Yes, you *may* come in now. Permission
Paraphrase: Yes, you *are allowed to* come in now.

d. My father *can* run really fast. Ability
Paraphrase: My father *is able* to run really fast.

Since the dynamic sense implies epistemic possibility, as in (9.a), we asked the annotators to assign the stronger (dynamic) sense in case both readings are equally strong. They still assigned the epistemic sense in case they felt that this reading was stronger. An example of the latter case is (9.b).[6]

(9) a. Terrorists *can* now come into America and go to a gun show and, without even a background check, buy an assault weapon today.

b. We *could* take that oil-stained soil and those rusted factories, and create something new and beautiful.

The reliability of our annotation scheme was established via manual annotation of samples of the EPOS and MPQA data sets (see Table 1).

4.2 Method

Experimental setup. We assign modal senses to the English modals *can, could, may, must, should, shall*, which have been used in R&R's classification experiment. For projection, we choose German as the source language, and we project into English.

Seed selection. The seeds were manually selected from PPDB (Ganitkevitch et al., 2013). The major criterion, besides frequency of occurrence, was non-ambiguity regarding the modal sense. We chose 24 seed phrases. Examples are adverbs like *wahrscheinlich* (*probably* – epistemic), *hoffentlich* (*hopefully* – deontic), adjectives like *erforderlich* (*necessary* – deontic), verbs like *gelingen* (*succeed* – dynamic), *erlauben* (*admit* – deontic) or affixes such as *-bar* (*-able*) as in (*lesbar* (*readable*) – dynamic). The seed paraphrases are given in Appendix A.1, including the number of extracted instances per paraphrase and their estimated reliability. Reliability of seeds was evaluated in terms of accuracy, based on manual evaluation of 20 randomly-extracted instances for each seed. For projection we employed the word-aligned Europarl and OpenSubtitles parallel corpus provided by OPUS (Tiedemann, 2012). We used PostCAT (Graca et al., 2007) for word alignment.[7]

[6] Both examples are drawn from the MASC data set.

[7] We made use of PostCAT because its model optimizes the agreement between source-target and target-source alignments and hence size and quality of the intersective alignment, which is particularly important for our task. Graca et al. found that AER on the Hansard corpus benefits quite significantly from this method.

EPOS data (subset)	avg.
κ_class	0.62
κ_majclass_experts	0.83
κ_experts	0.87
accuracy	0.92
avg. sentence length	13

MPQA data (subset)	
κ(exp1, gold)	0.66
κ(exp2, gold)	0.77
κ(exp1, exp2)	0.78
κ (R&R, full data)	0.67
avg. sentence length	30

TABLE 1 Annotator agreement: EPOS (left) vs. MPQA (right) and accuracy of heuristic EPOS annotations

Projection and validation. We extracted 7,693 instances with direct alignment of modal sense paraphrase and modal verb. 70.6% were labeled epistemic, 12.5% deontic, 17.0% dynamic (cf. Section 5 for the complete distribution).

In order to assess the quality of the heuristically sense-labeled modal verbs we performed manual annotation on a balanced subset of the acquired data consisting of 420 sentences, using the guidelines discussed above. We performed two types of annotation, using the same guidelines: *Expert* annotation by two linguistically trained experts, and *Classroom* annotation by 36 students with linguistic background, who were divided into 6 groups. Each group received a subset of instances for independent annotation. The *Experts* also annotated a balanced subset of 103 instances from R&R's MPQA data set, in order to calibrate our annotation quality against the MPQA gold standard.

Results. Table 1 shows the agreement obtained for the heuristically sense-tagged instances.[8] Kappa is lower for *Classroom* at 0.62, while *Experts* achieve $\mathcal{K}$=0.87. When comparing the majority sense of *Classroom* against both *Experts*, we obtain $\mathcal{K}$=0.83, approaching *Expert* agreement. Evaluating the projected sense labels against ground truth,[9] we observe high accuracy of .92.

Agreement for the MPQA subset is lower compared to EPOS. We achieve moderate agreement (0.66, 0.77) against the gold standard and between annotators (0.78). In R&R, agreement averaged over the different modal verbs was 0.67. Thus, our annotation reliability is comparable. The main reason for the lower agreement score on MPQA

[8]We compute Fleiss' Kappa (Fleiss, 1971) between annotators within the 6 groups and report the average. We computed Cohen's Kappa (Cohen, 1960) between the majority vote from *Classroom* and *Experts* ($\mathcal{K}$_majclass_experts) and between experts ($\mathcal{K}$_experts).

[9]We chose the majority label of (majority_class, Expert1, Expert2) as ground truth for the heuristically tagged data.

	epistemic	deontic	dynamic	sum	distribution
must	11	**184**	0	195	15.4%
may	**133**	10	0	143	11.3%
can	2	116	**273**	391	30.8%
could	**158**	17	67	242	19.1%
should	26	**258**	0	284	22.4%
shall	0	**11**	2	13	1.0%
sum	330	596	342	1268	
distribution	26.0%	47.0%	27.0%		

TABLE 2 Distribution of senses and modals for MPQA data set

is that the heuristically annotated data stems from language varieties that differ considerably from the MPQA evaluation data (with average sentence length of 13 vs. 30 tokens).

5 Experimental Data Sets

In this section we describe in more detail the various data sets we use for experimentation. As motivated in Section 4.1, we adopt R&R's annotation scheme, with minimal adaptations.[10] For annotation, the guidelines remained stable across annotation tasks, with only small refinements made for later annotation efforts on MASC.

5.1 MPQA data set

This section analyzes the sense distributions for R&R's annotation of modal verbs in the MPQA Opinion Corpus (Wiebe et al., 2005). R&R annotated the 535 documents in the first release of MPQA, resulting in 1330 sense-annotated modal verbs. Table 2 shows the sense distribution over this data set, with our sense conflations as described above.[11] Summing over all verbs, the distribution of modal senses is not dramatically skewed. If, however, we look at the distribution of senses *per modal verb*, we see that in every case the distribution is dominated by a single sense. The percentage of instances for the predominant sense ranges from 65% (*could*) to 93% (*must*).

The predominant senses not only provide a very high baseline for the task, they also have a strong influence on classifiers trained from the MPQA annotations. This is especially true given the standard setting of training one classifier per modal verb.[12] Thus, it is an open question

[10] Recall that we regrouped R&R's *conditional, concessive and optative* categories into *epistemic* and *deontic*. The *deontic* category was split into *permission* and *obligation* in annotation, but merged to *deontic* for classification and data statistics.

[11] Also, in the table we omit *might* (62 instances), as they are monosemous.

[12] Given that the modal verbs have different ambiguity classes, the construction

whether the classifiers are able to generalize to novel textual genres or domains in case their sense distributions differ from those seen in training.

5.2 EPOS: the Europarl and OpenSubtitles heuristically labeled data sets

To our knowledge, R&R's modal sense annotation on MPQA is the largest existing data set so labeled that covers the full range of modal verbs in English.[13] We aim to build a larger modal sense-annotated corpus without extensive manual annotation, using cross-lingual sense projection (cf. Section 4). For this approach we selected the parallel German-English datasets from *Europarl* (Koehn, 2005) and from the OPUS *OpenSubtitles* corpus (Tiedemann, 2012). The German-English Europarl corpus contains roughly 1.5 million aligned sentences, and the German-English section of OpenSubtitles contains more than 5 million aligned sentences.

By applying our projection method, we obtained 7,693 instances of heuristically sense-annotated modal verbs (Table 3). We refer to this corpus henceforth as EPOS. Note that the extracted data set, because it is derived by alignment with a selected set of paraphrases, cannot be considered to represent a natural distribution. In fact, while this data set is similarly unbalanced as naturally occuring data, the predominant sense often differs. For example, whereas epistemic uses of *could* dominate the MPQA data, in EPOS, the predominant sense is dynamic. Overall, we see a great imbalance in the number of retrieved instances, with epistemic *must* and *may* and dynamic *can* in the range of 1,000-3,000 instances, deontic *may, must* and *should* and dynamic *could* in mid-ranges, while *shall*, epistemic *could, should, can* and deontic *can, could* are under-represented. For some modal verb senses, such as epistemic *could*, we did not find paraphrases with reliably accurate alignments.[14]

Overall, having more data to work with allows us to reduce not only the sparsity of the annotated data but also (to some extent) the class imbalance. In Section 7 we describe the creation of balanced training sets from the EPOS data for use in our classification experiments.

---

of lexical classifiers is the most obvious choice.

[13] See e.g. Rubinstein et al. (2013) for an overview.

[14] For the experimental data set used in Section 7, we added manually selected training instances: 28 for epistemic *could* and 8 instances for epistemic *should*.

	epistemic	deontic	dynamic	sum	distribution
must	**1630**	448	0	2078	27.0%
may	**3783**	165	0	3948	51.3%
can	17	10	**1215**	1242	16.1%
could	0	22	**83**	105	1.4%
should	0	**310**	0	310	4.0%
shall	0	4	**6**	10	0.1%
sum	5430	959	1304	7693	
distribution	70.6%	12.5%	17.0%		

TABLE 3 Distribution of senses and modals for EPOS data set

5.3 MASC: Modal senses in spoken and written genre corpora

The MPQA data set is extracted from news articles from a variety of sources, and the distributions in MPQA consequently reflect natural distributions for news texts (though of somewhat mixed styles).

The EPOS data set is based on corpora from parliamentary debates and movie subtitles, yet does not display a natural distribution.

The third data set from MASC was produced in order to investigate how well our classification models are able to generalize when applied to novel data that differs in textual genre and possibly sense distributions.

MASC – the Manually Annotated SubCorpus (Ide et al., 2008) of the Open American National Corpus – is a freely-available multi-genre corpus with manual (or manually-validated) annotations at multiple linguistic levels. In order to investigate the robustness of our classifiers, we build a test corpus with roughly 100 instances of modal verbs from each of the 19 genres in MASC. These consist of four spoken genres as well as 15 written genres (see Table 4 for an overview).

For each genre, we first identified all modal verbs from their part-of-speech tags, then we sorted the documents according to the number of modal verbs from our target set. Starting from the document with the highest number of modals, we added complete documents to the test corpus until the number of modal verbs was at least 100.[15] This resulted in overall 2,041 target instances.[16]

Table 4 shows the 19 genres and the number of instances annotated per genre, as well as the number of documents (or files, for some genres) selected in order to reach 100 modal targets. Additionally, we show the

[15] Exceptions are those genres in MASC with fewer than 100 modal targets: journal, newspaper, technical, travel guides, and telephone.

[16] 37 instances of *ought* were annotated but not further considered for classification, which leaves 2,004. An additional 38 instances were eliminated due to processing errors in classification, cf. Appendix A.4.

	#inst	#docs	#tok	len		#inst	#docs	#tok	len
blog	105	6	14985	33	newspaper	86	21	24206	28
email	108	7	12053	29	non-fiction	121	2	12535	30
essays	100	4	24586	35	technical	83	8	22737	33
ficlets	109	3	20721	16	travel	89	7	27934	26
fiction	139	2	20070	17	twitter	119	2	28629	19
govt-docs	122	3	15296	32					
jokes	106	7	18344	19	court-transcript	144	1	24602	23
journal	97	9	24677	34	debate-transcript	110	1	17991	23
letters	103	19	14469	24	face-to-face	149	2	18147	17
movie-script	102	4	29138	17	telephone	49	5	6829	30

TABLE 4 MASC per genre: no. of modals annotated, no. of documents/files selected, total no. of tokens in documents and average sentence length

total number of tokens in the selected documents, to give a sense of the relative density of modal verbs in the different genres. The distributions of senses and modal verbs appear in Table 5.

Note that the overall distribution of senses across all modal verbs in MPQA (Table 2) as opposed to MASC (Table 5) is different, with a predominance of *deontic* in MPQA vs. *dynamic* in MASC. There is also a change in the predominant sense of *could* which is *epistemic* in MPQA while it is *dynamic* across all genres represented in MASC.

A finer-grained analysis of the MASC genres with respect to varying distributions of both modal verbs and modal senses, and the influence of such variation on classification performance appears in Section 8.

The documents selected were manually labeled by two paid annotators, and curated by two expert annotators, using the same guidelines as used for MPQA and EPOS. During annotation,[17] the entire document was displayed, and the target modal verbs were highlighted.

Curation consisted of inspecting all cases for which the two annotators disagreed. The curators selected either the best label from the two or a third label. All uncertain cases were discussed between the two curators until a consensus could be reached.

Table 6 shows Cohen's Kappa (Cohen, 1960) between the two annotators and between each annotator and the curated annotations.[18] It became clear during curation that annotator 2 made a large number of errors, with many instances of a small number of error types. For example, this annotator had a strong bias toward the label Permission even in cases where that interpretation was clearly incorrect. These errors were corrected by the curation process, but they are reflected in the rel-

[17] Annotation was done using WebAnno (Yimam et al., 2013), available at https://www.ukp.tu-darmstadt.de/software/webanno/.

[18] Per-genre agreement figures appear in Appendix A.2, in Table 22.

	epistemic	deontic	dynamic	sum	distribution
must	29	**115**	0	144	7.3%
may	**217**	43	3	263	13.4%
can	88	72	**710**	870	44.3%
could	144	16	**251**	411	20.9%
should	27	**224**	0	251	12.8%
shall	3	**20**	0	23	1.2%
sum	508	490	964	1962	
distribution	25.8%	24.9%	49.0%		

TABLE 5 Distribution of senses and modals for MASC data overall

κ(anno1, anno2)	κ(anno1, curated)	κ(anno2, curated)
0.66	0.88	0.76

TABLE 6 Annotator agreement: MASC data set

atively low agreement seen between the two annotators. On the other hand, the high agreement between annotator 1 and the curated labels gives a sense of the potential high agreement achievable for this task, and this is confirmed by the agreement statistics reported in Section 4.

6 Semantic Features for Modal Sense Classification

Taking up our **research question Q2** from Section 5, we aim to develop a *semantically grounded feature space* for modal sense classification. Our aim is to develop a classifier with high performance that is robust against variations in sense distribution. Our hypothesis is that a semantically grounded feature space is able to generalize from surface or domain-specific lexical properties found with different textual genres. We thus expand the feature inventory used in prior work to incorporate semantic factors at various levels.

In the following we motivate and describe our semantic feature set, which we organize in seven groups. An overview of the proposed features is given in Table 7.[19]

VB: Lexical features of the embedded verb. The *embedded verb* in the scope of the modal plays an important role in determining modal sense. For instance, with the embedded verb *fly* in (10.a), we prefer a dynamic reading of *can*, whereas with *play* in (10.b) we find a deontic reading.

[19] In Zhou et al. (2015) our feature set included an additional feature group "Lexical aspect" (LA), following Friedrich and Palmer (2014). Significance tests showed that this group was not effective and thus can be omitted for overall simplicity.

(10) a. The children *can* **fly** (if they just believe, says Peter Pan)!
b. The children *can* **play outside**.

Building on the dependency parser[20] output, we access the `lemma` of the embedded verb and its `part-of-speech` tag in the sentence. We also extract whether the verb has a `particle` (e.g. *the plane could take* ***off***), and if yes, which.

In most cases, the relation between a modal verb and the embedded verb is given through the auxiliary (`aux`) dependency. Some other constructions that we capture through dependencies are the copula (`cop`) in examples like (11.a) and an open clausal complement (`xcomp`) in (11.b). Analysing misparses we noticed that some of them pass through an adverbial modifier clause (`advcl`) dependency. In case the embedded verb cannot be found, we return the value *none*. The relation between the embedded verb and its particle is given by the phrasal verb particle dependency (`compound:prt`).

(11) a. There *can be* no doubt about it.
b. It *can* be difficult to *decide* who are terrorists.

SBJ: Subject-related features. These features capture syntactic and semantic properties of the subject of the modal construction. In (12.a) a non-animate, abstract subject favors an epistemic reading for *could*, whereas with an animate subject (12.b), a dynamic reading is preferred. Other factors involve speaker/hearer/third party distinctions (13).

(12) a. The conflict *could* now move to a next stage. (epistemic)
b. He *could* now move to a next stage. (dynamic)

(13) a. I *must* be home by noon. (deontic only)
b. He *must* be home by noon. (deontic or epistemic)

We extract `person` and `number` of the subject and the `noun_type` (common, proper, pronoun). `person` is identified via personal pronoun features; the remaining features are extracted from POS tags.

The `countability` of the subject head is obtained from the Celex database (Baayen et al., 1996). We make use of the following columns from the database: lemma (`word`), countability (`C_N`) and uncountability (`Unc_N`) of a noun.[21] Since nouns can appear more than once with the same or different column values, we convert the output, so that every noun appears only once with a single countability value from

[20] Feature extraction is performed using Stanford's CoreNLP (Manning et al., 2014) and the Stanford dependencies (De Marneffe and Manning, 2008).

[21] Available at http://celex.mpi.nl/scripts/colstart.pl.

Embedded verb		
VB	lemma	lemma of head
	part-of-speech	POS of head
	particle	*up, off, on,...*
TVA	passive	true / false
	progressive	true / false
	perfect	true / false
	voice	active / passive
NEG	negation	true / false
WNV	WN sense $[0-2]$	WN senses (head+hypernyms)
	WN senseTop	top sense in hypernym hierarchy
Subject noun phrase		
SBJ	number	sg, pl
	person	1, 2, 3
	countability	count, uncount, ambig, none
	noun type	common, proper, pronoun
	WN sense $[0-2]$	WN senses (head+hypernyms)
	WN senseTop	top sense in hypernym hierarchy
	WN lexname active	person, artifact, event, ...
	WN lexname passive	person, artifact, event, ...
Sentence structure		
S	conjunct clause	true / false
	adjunct clause	true / false
	relative clause	true / false
	temporal mod.	true / false
Subject and Verb semantic features		
WN	all WN features of SBJ group and WNV features	

TABLE 7 Individual features and feature groups.

`{countable, uncountable, ambiguous, none}`. If the surface form of the subject head can be found in the above described countability database, countability of the surface form is used. Otherwise, the subject head is lemmatized and the countability of the lemma is assigned to it. Since entries in the Celex database are case-sensitive, the subject is lower cased if it is the first word in the sentence.

Lexical semantic features for the subject head are extracted from WordNet (Fellbaum, 1999) using NLTK.[22] Following Reiter and Frank (2010), we take the most frequent sense of the common or proper noun in WordNet (`subject_sense0`), add the direct hypernym of this sense, the direct hypernym of that hypernym, etc., resulting in features `subject_sense[1-3]`.

We also extract the topmost hypernym of `subject_sense0` in WordNet as `subject_sense_top` (e.g. *entity*). Finally, the name of the lexicographer file in WordNet, containing `subject_sense0`, is retrieved using two features, `lexname_active` and `lexname_passive`. If the subject appears in an active construction, the WordNet lexical filename is assigned to `lexname_active` and `lexname_passive` is set to *"none"*, and vice versa for the passive subject. POS is retrieved if there is no lexical filename accessible for the subject in WordNet. Using two features for lexical filename, with respect to the subject being in a passive or an active construction, captures the additional distinction of whether the subject is a deep subject or a deep object. This can be beneficial for sense distinctions as in (14).

(14) a. John *can* talk. (deontic or dynamic)
b. John *can* be talked about. (epistemic or deontic)

TVA: Tense/voice/grammatical aspect features. These features capture voice and grammatical aspect of the embedded verb complex. These characteristics of the embedded verb are important factors for modal sense disambiguation. (15.a) clearly favors an epistemic reading, as the event is in perfective aspect, and thus already completed, whereas deontic sense is favored with future events in indicative mood as in (15.b).

(15) a. He *must* have delivered this box already.
b. He *must* deliver this box tomorrow.

We capture both grammatical aspect and voice features using sequences of POS tags of the verbal complex, following Loaiciga et al. (2014). The boolean features `perfect` and `progressive` indicate the respective grammatical aspect; `voice` indicates active or passive voice.

[22] We use NLTK version 3.0.5 (Bird et al., 2009).

NEG: Negation. Negation is a semantic feature at the proposition level that can have reflections in modal sense selection. *Should*, e.g., seems to favor a deontic meaning when negated in (16.a). Also, negation can interact with disambiguation of epistemic vs. deontic readings depending on specific propositional or discourse context. In (16.b), the favored reading is deontic in the negative sentence.

(16) a. He *should* (not) have returned.
(epistemic/deontic (pos) vs. deontic (neg))
b. He *may* (not) drink more gin tonight.
(epistemic/deontic (pos) vs. deontic (neg))

The `negation` feature captures the presence or absence of negation in the modal construction. We use the dependency label `NEG` to identify negation. Along with negation (*not*), this modifier dependency also captures negation adverbials, such as (*never*).

WNV: Lexical semantic features of the embedded verb. This feature group encourages semantic generalization for lexical features of the embedded verb. It can play a role in interaction with other features, such as grammatical aspect, proposition-level features like negation, or the combined lexical semantic features described below (WN). The features in this group are parallel to the WordNet features described for the SBJ feature group above (minus `lexname_active` and `lexname_passive`), but apply to the embedded verb instead of the subject. If the embedded verb has a particle, then the WordNet sense of the phrase is extracted.

S: Features of sentence structure. When modals appear as part of a complex sentence, certain structural configurations can reflect rhetorical or temporal relations between the proposition modified by the modal and dependent clauses. An example are telic clauses that can favor specific modal sense readings, as in (17).

(17) a. You *must* buy more shares to make real money. (deontic)
b. You *could* use a short cut to save time. (epistemic/deontic)
c. You *can* take a leave to visit your brother (deontic)

We extract features from the constituent tree to capture such effects: whether the modal clause is conjoined to the main clause (`embedded ConjunctSentence`), whether it embeds adjunct clauses (and if so, the conjunction) (`adjunctSentence`), and whether it is in a relative clause (`relativeSentence`). Finally, `has_tmod` indicates the presence of a temporal modifier. It is easily extracted through the temporal modifier dependency (`nmod:tmod`).

WN: All WordNet features. This feature group aims to capture aspects of proposition-level semantics by combining semantic features

of the subject with those of the embedded verb. This feature group simply includes both the WordNet features described in SBJ and those in WNV. While this feature is indeed redundant when combined with other feature groups, we include it in the set of features in order to investigate the impact of combined subject-predicate semantic information in the ablation studies we describe later.

The intuition is that certain subject-predicate combinations may have a preference for certain modal senses. In (18), for example, *can* appears with a proposition that is subject to specific prescriptions or "laws": soldiers are subject to restrictions with respect to consuming alcohol.

(18) Soldiers *can* drink when off duty.

7 Classification Experiments and Results

Having addressed our **research question Q1**, the automated induction of high-quality modal sense tagged data in Section 4, and with the design of a semantically grounded feature space for modal sense classification in place (see Section 5), we are now in a position to experimentally investigate our **research question Q2**: Can a semantically grounded feature set improve the performance of modal sense classification, and to what extent does a semantic feature space generalize in view of distribution differences?

In order to investigate this question, we construct **contrasting classifier models** with different feature sets and different compositions of training data.

Our hypothesis is that classifiers trained only on the highly unbalanced MPQA data set will have difficulty separating the effect of distributional bias in the training data from the predictive force of their feature set. A classifier that follows the majority class in the training data will tend to neutralize the potential impact of its feature set. In order to separate the predictive force of different feature sets and the effect of different training data distributions, and also to alleviate sparsity inherent in the data, we evaluate different classifier models obtained from **different feature sets** and **different training data compositions**.

(i.) We extend the MPQA training set using **heuristically labeled instances** obtained from modal sense projection (cf. Section 4). In this way, sparsity can be considerably reduced.

(ii.) We also evaluate classifiers trained on perfectly **balanced data**. This will allow us to carve out the impact of the different feature sets in a non-biased training setting.

(iii.) Finally we measure the impact of **individual feature groups** via ablation (Section 7.3).

7.1 Experimental setup

Training data. We construct five different classifiers with respect to the datasets they are trained on. The first training dataset is the unbalanced MPQA used in R&R. We replicate their classifier and denote it with CL_M^{-b}.

The second training dataset is obtained from a balanced subset of our automatically acquired sense-tagged corpus EPOS. The composition of this dataset is shown in Table 9 (left-hand side). We denote the classifier trained on this blend of data with CL_E^{+b}.[23] [24]

Furthermore, to alleviate training data sparsity, we added the balanced EPOS subcorpus to the MPQA data. The resulting dataset is unbalanced, as MPQA is, but less skewed, due to the added instances. We denote the classifier trained on this dataset with CL_{ME}^{-b}.

We want to compare classifiers trained on the unbalanced datasets with classifiers trained on the balanced versions of the datasets. Therefore we balance the unbalanced training datasets by under- and oversampling. We perform a mixture of over- and undersampling, targeting about half the size of the larger class. That is, we consider the larger class, divide its size by two and undersample the larger class to meet this target class size, while the smaller class is upsampled to the target class size. The effects can be observed in Table 10. Given the unbalanced training set of MPQA in column CL_M^{-b}, we obtain a balanced training set in CL_M^{+b} as follows: for *must*, we select the larger class, deontic, with 149 instances, determine 70 as our target class size, and create training instances for CL_M^{+b} to meet this target class size.[25] We did not experiment with oversampling only. This might be useful in future explorations.

The classifier trained on the balanced MPQA dataset is denoted with CL_M^{+b}, and the classifier trained on the balanced blend of MPQA with

[23] In the following we refer to the balanced EPOS subcorpus as 'EPOS', provided the reference is clear from the context.

[24] We do not experiment with the unbalanced, full-fledged EPOS corpus, given that its distribution is not a natural one, but is determined on the selection of secure sense-indicating cross-lingual paraphrases. As seen in Tables 3 and 2, in comparison to MPQA the EPOS data set is missing certain senses (e.g., epistemic senses of *could, should* or *shall*), while it over-represents others, e.g., the epistemic sense of *must*.

[25] The target class size was manually chosen to approximate half of the larger class size, to the exception of *shall* which was upsampled to the larger class due to the small instance set. Note that zeros in the distribution tables mean that the corresponding sense is not established for the modal verb.

symbol	training dataset	(un)balanced training dataset
CL_M^{-b}	MPQA	unbalanced
CL_E^{+b}	EPOS	balanced
CL_{ME}^{-b}	blend of MPQA and EPOS	unbalanced
CL_M^{+b}	re-balanced MPQA	balanced
CL_{ME}^{+b}	re-balanced blend of MPQA and EPOS	balanced

TABLE 8 Subscripts on classifier names indicate the source of the training data. Superscripted $+b$ or $-b$ indicates balanced vs. unbalanced training set.

	CL_E^{+b} train			Full MPQA test		
	ep	de	dy	ep	de	dy
must	800	800	0	11	183	0
may	950	950	0	130	9	0
can	150	150	150	2	115	271
could	40	40	40	156	17	67
should	150	150	0	26	248	0
shall	0	5	5	0	11	2

TABLE 9 Heuristic ($+b$) and MPQA ($-b$) data

EPOS is denoted CL_{ME}^{+b}.

Ultimately we have five different training datasets and, therefore, five different classifiers. Their overview is given in Table 8.

Feature sets. Every classifier has three different configurations with respect to the feature set that represents the training data. The feature sets are R&R's shallow lexical and syntactic path features ($F_{R\&R}$) a feature set consisting of only our newly designed semantic features (F_{Sem}) and a combined set (F_{all}) consisting of both $F_{R\&R}$ and F_{Sem}. Five classifiers with three different feature sets make 15 different configurations.

Replicating R&R's modal sense classifier. We replicate R&R's classifier by reimplementing their feature set. They use a mixture of target and contextual features that take into account surface, lemma and PoS information, as well as syntactic labels and path features linking targets to surrounding words and constituents (cf. R&R, Table 5). Following R&R we use the Stanford parser for processing and induce maximum entropy models using OpenNLP[26] with default parameter settings.

We train one classifier per modal verb, using R&R's best feature

[26] `https://opennlp.apache.org`

	CL_M^{-b} train			CL_{ME}^{-b} train			CL_M^{+b} train			CL_{ME}^{+b} train			MPQA test		
	ep	de	dy	ep	de	dy	ep	de	dy	ep	de	dy	ep	de	dy
must	6	149	0	806	949	0	70	70	0	870	870	0	5	34	0
may	105	6	0	1055	956	0	50	50	0	999	1000	0	25	3	0
can	1	98	212	151	248	362	100	100	100	250	250	250	1	17	60
could	120	15	57	160	55	97	54	54	54	94	94	94	36	2	10
should	21	196	0	171	355	0	100	100	0	250	250	0	5	52	0
shall	0	9	1	0	14	6	0	10	10	0	15	15	0	2	1

TABLE 10 Cross-validation, one run: representative class distributions of training and test data.

setting (context feature window of 3 tokens left and right of target, target-specific features). Averaged accuracies for the replicated classifiers appear in Table 11 as CL_M^{-b} (feature set $F_{R\&R}$). Our scores are very similar to their published results, which appear in the same table in the column headed "R&R". R&R performed 10-fold cross-validation (CV) for evaluation. We perform 5-fold cross-validation instead.

Test data and evaluation. In order to compare to prior work, our test data is drawn exclusively from MPQA. For CL_E^{+b}, we evaluate on R&R's full data set; the composition of the test set appears in the right-hand side of Table 9. The other classifiers, $CL_M^{\pm b}$ and $CL_{ME}^{\pm b}$, are evaluated in a 5-fold CV setting, with testing on the naturally distributed MPQA instances. For each CV setting, only the training section is adapted, by addition of heuristic data, and/or balancing. Table 10 exemplifies one run of our cross-validation setting.

For $CL_M^{\pm b}$ and $CL_{ME}^{\pm b}$, for each fold, we split the (unbalanced) MPQA into 80% train and 20% test. For CL_M^{+b}, the training section is balanced, while the test section stays untouched. For CL_{ME}^{+b} the MPQA training section is balanced and the balanced data from EPOS is added. Note that the obtained training dataset is still balanced. Finally, for CL_{ME}^{-b} (unbalanced) we add EPOS to the training part without balancing it.

Baselines. For unbalanced classifiers, we compare to the most frequent sense (MFS) baseline, BL_{maj}, taking the majority sense of MPQA data. For balanced classifiers, we compare to the random baseline, BL_{ran}. Each modal verb has a certain number of possible sense classes, e.g. *must* can have the senses *epistemic* or *deontic*. The random baseline assigns equal probability to each class.

7.2 Discussion of Results

Table 11 compares the accuracy of classifiers trained on (un)balanced data, from different sources (MPQA, EPOS or blend), and with differ-

$F_{R\&R}$	R&R	CL^{-b}_{M}	BL_{maj}	CL^{-b}_{ME}	CL^{+b}_{M}	CL^{+b}_{ME}	CL^{+b}_{E}	BL_{ran}
must	93.50	**94.32**^{ME}	**94.32**	82.00	**76.25**	73.24	71.65	50.00
may	81.43	**93.57**	**93.57**	90.71	79.29	88.57^{M}	**92.14**^{M}	50.00
might		100.00	100.00	100.00	100.00	100.00	100.00	100.00
can	68.70	66.56	**69.92**	64.25	49.86	53.19	**57.84**^{M}	33.33
could		62.50	**65.00**	59.17	41.25	44.17	**49.17**	33.33
should	91.29	90.77	**90.81**	90.77	80.21	**85.83**^{M}	76.33^{ME}	50.00
shall		83.33	84.61	**90.00**	70.00	**90.00**	53.85	50.00
macro-avg	83.73	84.44	**85.46**	82.31	70.98	**76.63**	71.57	52.38
micro-avg		78.71^{ME}	**80.22**^{M,ME}	75.14	62.59	**66.40**^{M}	66.24^{M}	41.54

F_{Sem}	R&R	CL^{-b}_{M}	BL_{maj}	CL^{-b}_{ME}	CL^{+b}_{M}	CL^{+b}_{ME}	CL^{+b}_{E}	BL_{ran}
must	93.50	**94.32**^{ME}	**94.32**	87.07	**87.11**	86.03	85.05	50.00
may	81.43	92.86	**93.57**	90.00	83.57	**91.43**^{M}	90.00	50.00
might		100.00	100.00	100.00	100.00	100.00	100.00	100.00
can	68.70	66.33	**69.92**	62.73	**57.87**	55.54	54.76	33.33
could		**71.67**^{ME}	65.00	67.50	**61.25**	60.42	56.25	33.33
should	91.29	**92.91**	90.81	92.21	**87.27**	85.53	83.75	50.00
shall		83.33	**84.61**	83.33	63.33	**76.67**	69.23	50.00
macro-avg	83.73	**85.91**	85.46	83.36	77.20	**79.37**	77.00	52.38
micro-avg		**80.78**$^{ME}_{R}$	80.22^{ME}	77.36	72.52$_{R}$	**72.12**$_{R}$	70.29$_{R}$	41.54

F_{All}	R&R	CL^{-b}_{M}	BL_{maj}	CL^{-b}_{ME}	CL^{+b}_{M}	CL^{+b}_{ME}	CL^{+b}_{E}	BL_{ran}
must	93.50	**94.32**	**94.32**	92.78	84.48	**87.60**	85.57	50.00
may	81.43	**93.57**	**93.57**	92.14	87.86	**92.14**	**92.14**	50.00
might		100.00	100.00	100.00	100.00	100.00	100.00	100.00
can	68.70	66.06	**69.92**	64.76	53.73	59.64	**60.41**	33.33
could		**67.92**	65.00	63.33	60.00	**61.25**	56.25	33.33
should	91.29	**92.89**	90.81	91.48	86.57	**90.11**^{M}	88.34	50.00
shall		83.33	84.61	**90.00**	83.33	**83.33**	53.85	50.00
macro-avg	83.73	85.44	**85.46**	84.93	79.42	**82.01**	76.65	52.38
micro-avg		80.01$^{ME}_{R}$	**80.22**^{ME}	78.16$_{R}$	71.17$_{R}$	**74.98**$^{M}_{R,S}$	73.23$_{R,S}$	41.54

TABLE 11 Classifier accuracy for various training data and feature sets. See text for details.

ent feature sets ($F_{R\&R}$, F_{Sem} and F_{All}). We report results for individual classifiers (per modal verb) and macro- and micro-average across all verbs.[27] The two bold-faced numbers per table row indicate the best models for unbalanced and for balanced data. We test significance using McNemar's test ($p<0.05$) (McNemar, 1947). Within a row (comparing twice: within $+b$ and $-b$ classifiers), a superscript on a number indicates which classifier is significantly outperformed by the result. Across feature sets, we compare micro-averages and mark significance by subscripts (R=$F_{R\&R}$, S=F_{Sem}).

[27] Although our annotated datasets exclude non-ambiguous *might*, we include *might* in macro-averages for comparability with previous work.

We first discuss the classifiers trained on **unbalanced data**. With $F_{R\&R}$, CL_M^{-b} yields performance comparable to R&R's results, at 84.44% accuracy, 1.02pp (pp=percentage points) below the majority baseline. Individual lexical classifiers also approach R&R's performance, though never beating the baseline.[28]

Changing feature sets from $F_{R\&R}$ to F_{Sem} and F_{All}, classifier CL_M^{-b} for *could* and *should* is now able to beat the baseline. The effect is stronger for F_{Sem}, which reflects the impact of the semantic features. Interestingly, F_{Sem} outperforms $F_{R\&R}$, even though the classifiers learn **only** on the basis of semantic features. At the level of macro-average, CL_M^{-b} with F_{Sem} beats the majority baseline by 0.45pp and R&R's reconstructed classifier by 1.47pp. When we look at micro-averages, this same classifier significantly outperforms CL_M^{-b} with $F_{R\&R}$. The same is true for all three balanced classifiers using F_{Sem}. Combining the two feature sets (F_{All}) produces minimal differences for CL_M^{-b}, but yields stronger gains for CL_{ME}^{-b}. At the level of micro-average, all five classifier configurations using F_{All} significantly out-perform the respective classifiers with $F_{R\&R}$, and two of the balanced classifiers also out-perform F_{Sem}.

The addition of heuristically-tagged data in CL_{ME}^{-b} helps for some verbs, but hurts for others. Despite the larger training set size, individual classifier performances tend to drop, meaning they do not profit from the enlarged data set and the reduced training bias. Regarding feature sets, we observe that the purely semantic feature set F_{Sem} improves on $F_{R\&R}$, and F_{All} yields further improvement, yet all at lower levels compared to CL_M^{-b}. The gains are small and they are not significant for F_{Sem} compared to $F_{R\&R}$.

For classifiers trained on **balanced data**, the picture changes. Accuracies on balanced data are lower, reflecting the lack of distributional bias. But all results are well above the random baseline.[29]

Compared to CL_M^{+b} and CL_E^{+b}, we consistently observe the best results for CL_{ME}^{+b}, which mixes MPQA and EPOS data. This is significantly so with $F_{R\&R}$ and F_{All}. All semantically enriched models significantly outperform the balanced classifiers using $F_{R\&R}$, and CL_{ME}^{+b} and CL_E^{+b} using F_{All} significantly outperform their respective classifiers using only F_{Sem}. The best performance is obtained with F_{All}. CL_{ME}^{+b} with 82.01% on balanced mixed data closely approaches the performance of

[28] We report individual results, while R&R aggregated *may/might* and *shall/should*.

[29] All comparisons to the random baseline are significant except: CL_M^{+b} and CL_{ME}^{+b} with F_{Sem} for *should*, and anything involving *shall*.

the classifiers CL_{M}^{-b} and CL_{ME}^{-b} trained on biased training data using F_{All} and their majority baseline, with 3.43pp difference to CL_{M}^{-b} , and being very close to R&R's published results (-1.72pp).

Looking at **individual modal verb classifiers**, we observe interesting effects. *Can* and *could*, both with 3-fold sense distinctions and lowest performance overall, suffer the greatest loss in the balanced setting, in ranges of 41-57% for $\mathrm{F}_{R\&R}$. These verbs are hard to classify, and here we see a marked performance gain as the training data changes (from CL_{M}^{+b} to CL_{E}^{+b}, but significant only for CL_{E}^{+b}). Comparing F_{Sem} to $\mathrm{F}_{R\&R}$, we obtain better results overall, always well above 50% accuracy. With F_{All} we reach a range of 53-61%, achieving strong gains of +17pp for *could*, and about +15pp for *can*. We also note a rise for *should* with a +4pp gain over $\mathrm{F}_{R\&R}$. Across different feature sets, CL_{ME}^{+b} performs best, that is, the blend of MPQA and EPOS data is effective. Using only automatically acquired training data with CL_{E}^{+b} yields gains for some modal verbs, but does not achieve better performance compared to CL_{M}^{+b}.

To summarize, with increasingly refined models and a tendency of CL_{ME} outperforming CL_{M} on balanced training data, we obtain the following picture: semantic features contribute important information and reach their best performance with a mixture of training sets in balanced training situations. We also note that F_{Sem} and F_{All} both yield significant gains over $\mathrm{F}_{R\&R}$ for *could, must, should, can* and *may.*[30] A puzzling effect is the drop of performance that occurs when adding balanced training data to the unbalanced classifiers CL_{ME}^{-b}: the additional data harms rather than improves the results, and weakens the impact of semantic features. In Section 8 we will come back to this issue.

7.3 Impact of feature groups

A confusion analysis of the predictions made by CL_{E}^{+b} using $\mathrm{F}_{R\&R}$ yields some insight into the most difficult sense distinctions for specific modal verbs. Table 12 highlights the most prominent misclassification classes: for instance, deontic *can* is misclassified as *dynamic* in 107 cases; epistemic *could* is misclassified as dynamic in 53 cases, etc.

For a deeper analysis of the impact of our semantic features, especially for specific sense distinctions, we conducted a quantitative and qualitative evaluation by ablating individual feature groups (FGs) from the full feature sets F_{Sem} and F_{All}, for all balanced classifiers.

It turns out that for the modal verbs that exhibit prominent confusion classes in Table 12 we observe a significant performance drop

[30] Cross-feature set significance for individual verbs is not marked in Table 11 .

can	epistemic	deontic	dynamic	*could*	epistemic	deontic	dynamic
epistemic	1	0	1	epistemic	89	14	**53**
deontic	7	1	**107**	deontic	6	2	9
dynamic	**28**	**21**	223	dynamic	**29**	11	27

must	epistemic	deontic	*should*	epistemic	deontic
epistemic	5	6	epistemic	4	**22**
deontic	**49**	134	deontic	**45**	212

TABLE 12 Confusion analysis: R&R features, balanced EPOS training data. Row headings indicate gold label, column headings show predicted label

when omitting individual feature groups (FGs): Table 13 reports all configurations where omitting a particular FG yielded a significant accuracy loss. Most of the highly-significant ($p < 0.01$) FGs appear in conjunction with CL_E^{+b}, and primarily for the two FGs SBJ and TVA. The FGs with significance at $p < 0.05$ are distributed over the other classifiers and feature groups. In the following we analyze these cases in more detail.

Analysis. We define *gains* (or *rescues*) due to FG_x as cases in which including FG_x turns a wrong classification into a correct one, compared to a model that ablates FG_x. *Losses* record the opposite: a correct classification made without FG_x becomes incorrect when FG_x is active.

Table 14 summarizes the total numbers of classification gains and losses due to the significantly-influential configurations represented in Table 13. Overall, for both models F_{Sem} and F_{All} we observe **more gains than losses** due to the FGs SBJ, NEG, TVA, WNV and WN: 480 vs. 178 (37% losses, on average) for F_{Sem} and 218 vs. 66 (30% losses, on average) for F_{All}. For F_{All}, the gains/losses ratio is the best for the classifier trained on blended training data, where for F_{Sem}, training only on MPQA gives the best gains/losses ratio. For *must* there are only gains and no losses at all.

We observe different performance for correction of misclassifications for the different modal verbs, and we see clearly distinct contribution of FGs for the individual modal verb classifiers.

The most clear-cut positive effects are obtained for ***must***, with the highest number of gains (69/31 for $\text{F}_{Sem}/\text{F}_{All}$) and very few losses (13/0). Here, the FG TVA is the only one to show highest significance, leading to a majority of rescues of *deontic* readings that otherwise would be misclassified as *epistemic*. Only 3 rescues occur in the other direction (i.e., rescues of *epistemic* readings from misclassification as *deontic*).

A particularly strong effect is seen for TVA, which avoids misclas-

verb	**FG**	**feature set used**	**impact** CL_M^{+b}	CL_{ME}^{+b}	CL_E^{+b}
must	TVA	F_{Sem}		6.70^{**}	10.31^{**}
		F_{All}		7.73^{**}	8.25^{**}
	WNV	F_{Sem}	5.67^{*}		
	WN	F_{Sem}	6.18^{*}		
may	SBJ	F_{Sem}		7.86^{*}	
	WN	F_{All}	5.72^{*}		
can	SBJ	F_{Sem}			6.95^{**}
		F_{All}			2.83^{*}
	TVA	F_{Sem}			3.09^{**}
	NEG	F_{Sem}	2.57^{*}		
	WNV	F_{Sem}	5.14^{*}		
	WN	F_{Sem}	5.65^{*}		
		F_{All}		4.63^{*}	
could	SBJ	F_{Sem}			17.08^{**}
		F_{All}		5.42^{*}	8.75^{**}
should	SBJ	F_{Sem}	7.77^{**}		14.49^{**}
		F_{All}			8.54^{**}
	NEG	F_{Sem}		2.47^{*}	
	WN	F_{Sem}	5.30^{*}		6.36^{**}
		F_{All}			8.73^{**}

**: $p<0.01$; *: $p<0.05$

TABLE 13 Accuracy loss by FG omission. Third column specifies from which feature set we ablate.

feature set	classifier	gains	losses	ratio
F_{Sem}	CL_M^{+b}	216	104	48.1%
	CL_{ME}^{+b}	35	4	11.4%
	CL_E^{+b}	229	70	30.6%
	TOTAL	480	178	37.1%
F_{All}	CL_M^{+b}	10	2	20.0%
	CL_{ME}^{+b}	81	35	43.2%
	CL_E^{+b}	127	29	22.8%
	TOTAL	218	66	30.3%

TABLE 14 For each combination of feature set and classifier, the total number of gains and losses for significantly influential feature groups

sification of up to 11% of all instances of ***must*** as *epistemic*. All cases follow the pattern in (19.a): the embedded verb is neither in past tense nor perfective aspect, and we prefer a deontic interpretation, whereas perfective aspect in (19.b) indicates epistemic usage.

(19) a. [..] whoever is on the other side is the evil that must **be** destroyed [..]

b. The event must **have** rocked the halls of power [..]

should displays similar sense ambiguities and confusion patterns, but here the picture is less clear: as with *must* we obtain rescues of *deontic* readings, but here the WN features are most effective, jointly with SBJ. In contrast to *must*, we observe a mixture of gains (152 for WN, 57 for SBJ) and losses (49 for WN, 7 for SBJ), where the latter are due mostly to over-correction.

For ***could***, with a 3-way sense ambiguity, the SBJ feature group is the only one showing significant influence. Most rescues to an *epistemic* reading are due to including SBJ features. We also observe rescues of *dynamic* readings that would have been classified as *epistemic* instead. On the losses side, we observe upwards of 40% of losses as opposed to gains for both F_{Sem} and F_{All}.

SBJ features apparently capture a preference for inanimate, abstract subjects for *epistemic* as opposed to deontic (or dynamic) readings, as with *the message* or propositional anaphora in (20.a,b), which triggered rescues to *epistemic*. The same pattern is observed with ***should*** (20.c).

(20) a. "**the message** *could* not be clearer."

b. [..] officials said **this** *could* prompt industries to change behavior ....

c. [..] if **that** *should* prove necessary, De Winne will [..] pilot the space ship.

Finally, ***can*** is our most difficult case. We obtain moderate gains by rescues of *dynamic* readings from *epistemic/deontic*, with small contributions made by 5 of our 7 feature groups. In each case, though, the gains are small, showing no clear patterns, and combined with up to 50% losses. This means we are still lacking precise features that can differentiate epistemic and dynamic readings with *can*.

Overall, the ablation analysis confirms the design of our semantic feature set. Feature groups relating to tense and aspect properties of the embedded verb, negation, abstractness of the subject and semantic features of the embedded verb yield significant effects on classification performance, and the observed effects on specific instances confirm the linguistic intuitions underlying the semantic feature space.

We additionally performed a second ablation experiment in which we first ranked the feature groups according to their impact and then added them one by one, testing classifier performance after each addition. The performance curve obtained from this experiment shows that all feature groups indeed contribute to classification performance with minimal redundancy. Thus the best combination of FGs is to include them all.

8 Modal Sense Classification across Genres

We now address our third research question (**Q3**): *Are there genre differences in the distribution of modal senses*, and *to what extent are they mirrored in the performance of classification models*? To address this question, we first examine differences in modal sense distributions across genres, and then investigate the performance of different classifier models applied to data from various genres.

8.1 Genre differences in sense distributions

The data extracted from MASC and annotated with modal senses (as described in Section 5.3) consists of approximately 100 modal targets for each of the 19 genres represented in MASC. To explore differences between these genres and the distribution of modal verbs and modal senses, we compute Kullback-Leibler divergence (Kullback and Leibler, 1951) between normalized distributions for each pair of genres.

Kullback-Leibler divergence, denoted $D_{KL}(P||Q)$ (1.1), measures the difference between two distributions P and Q. The measure is asymmetric, measuring the information gain that arises when revising a prior probability distribution Q to the posterior probability distribution P. It is also known as the *relative entropy* of P with respect to Q. The value of $D_{KL}(P||Q)$ is ≥ 0, and equal to 0 when P is equal to Q.

$$D_{KL}(P||Q) = \sum_i P(i) \cdot \log \frac{P(i)}{Q(i)} \,. \tag{1.1}$$

Even though KL divergence is not a metric, it can serve as a metric for quantifying the distance between two genres, determined on the basis of specific types of distributions. We looked at three different types of distributions: (i) distribution of modal senses per genre, (ii) distribution of modal verbs per genre, and (iii) distribution of modal sense per modal verb per genre. For each such comparison, we present a heat-map and some interpretation.

As we want to compare any two genres in a non-directed way, we use a symmetrized version of KL divergence. For discrete probability distributions P and Q a symmetric version of KL divergence is defined

by

$$D_{sym}(P||Q) = D_{KL}(P||Q) + D_{KL}(Q||P) \tag{1.2}$$

This measure is well defined if $Q(i) = 0$ if and only if $P(i) = 0$, for all i. As this is not the case for every pair of genres, we occasionally smooth a distribution to avoid zero values.

Genre subgroupings in MASC

The genres represented in MASC are diverse and wide-ranging. In order to produce more meaningful and cohesive heat maps, we arrange the genres according to the clusterings produced by Passonneau et al. (2014). They induce subgroups of genres using hierarchical clustering with a set of 37 linguistic features; we adopt these clusters straightforwardly. Two of the obtained clusters are singletons; these are the genres `travel guides` and `technical` texts. The remaining four clusters are described below, using the informal names suggested by Passonneau et al. (2014).

Spoken genres. This cluster covers the four spontaneously spoken genres in MASC. They include transcriptions of informal interactions, both phone calls (`telephone`) and recordings of participants jointly solving some task (`face-to-face`).[31] The other two genres consist of transcriptions of more formal spoken interactions: parliamentary debates (`debate-transcripts`) and courtroom proceedings (`court-transcripts`).

Offline-interactive genres. The four genres contained in this cluster have in common that they often consist of non-spontaneous interaction between participants. These are `letters`, `emails`, `spam`,[32] and `tweets`. Though they are grouped with other offline interaction genres, the letters contained in MASC are fundraising letters, with different characteristics from the personal letters that may first come to mind.

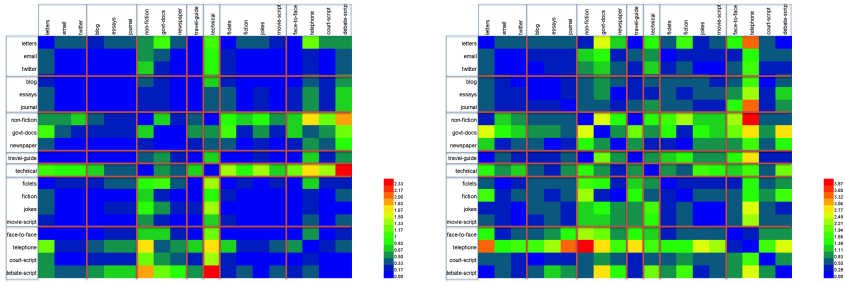
Discursive genres. This cluster (the largest) consists of discursive, non-fictional texts. They are generally long form texts, and many correspond to traditional text genres: `blog`, `essay`, `journal`, `non-fiction`, `government documents`, and `newspaper` texts.

Story-telling genres. The final cluster contains four genres, all of them fictional. They range from `fiction` (excerpts from several longer fictional works), to `ficlets` (very short fictional pieces of roughly 5-20 sentences), and the non-prose forms of `jokes` and `movie scripts`.

Although we do not see very strong patternings for the genre subgroups, they are nonetheless helpful for interpreting the results of the

[31] Passonneau et al. (2014) group these two together.

[32] Not included in our set of genres.



(a) (b)

FIGURE 1 Differences between genres with respect to: (a) modal sense distributions and (b) modal verb distributions

distributional analysis, given that we are comparing the characteristics of nineteen different genres.

Analysis of distributions

We investigate differences in modal sense distributions from several different perspectives.

Distributions of modal senses. First we look at how the three modal senses pattern across genres. Table 5 (Section 5, page 17) provides counts of the modal senses for MASC overall, and Table 16, page 37 indicates the most frequent sense for each verb, for each genre. The heat-map in Figure 1 (a) illustrates the degrees of difference between genres, using KL divergence as defined in Equation (1.2). Finally, the detailed distributions appear in the Appendix A.3, as Table 23.

For most genre pairs, the modal senses have very similar distributions. Four genres in particular stand out as showing higher divergence from a number of the other genres: `non-fiction`, `technical`, `telephone`, and `debate-transcript`. Interestingly, `non-fiction` and `technical`, and `telephone` and `debate-transcript` each show quite low divergence with each other. Three of these stand out against other genres by featuring a distinct predominant sense (`non-fiction` and `technical`: epistemic; `telephone`: deontic; `debate-transcript` stands out with a very low proportion of epistemic sense (cf. Appendix A.3)

Distributions of modal verbs. Next, we look at how the individual modal verbs are distributed in different genres. Table 15 shows, for each genre, the two most frequent modal verbs and their portion of the distribution (for all annotated modal targets for that genre). *Can* and *could* frequently occur together as the top two modals, but the proportion of targets they capture varies wildly. Two genres stand out

as the only with *should* in their top two: `twitter` and `telephone`.

	#1 modal	#2 modal		#1 modal	#2 modal
blog	can(45.7%)	may(19.0%)	newspaper	could(34.9%)	can(26.7%)
email	can(41.7%)	could(20.4%)	non-fiction	can(60.2%)	may(25.7%)
essays	can(35.0%)	may(24.0%)	technical	could(30.5%)	may(30.5%)
ficlets	can(41.3%)	could(33.9%)	travel	can(59.6%)	may(22.5%)
fiction	could(32.6%)	can(23.3%)	twitter	can(52.6%)	should(17.2%)
govt-docs	could(32.0%)	may(22.1%)			
jokes	can(54.3%)	could(18.1%)	court-transcript	can(48.6%)	could(22.2%)
journal	can(40.0%)	could(18.9%)	debate-transcript	can(75.0%)	could(9.1%)
letters	can(67.0%)	could(12.6%)	face-to-face	can(46.3%)	could(25.5%)
movie-script	can(50.5%)	could(16.2%)	telephone	should(42.6%)	can(29.8%)

TABLE 15 MASC data: per genre, the two most frequent modal verbs, with % of occurrence

Although this table presents a fairly consistent picture regarding which modals are most frequent, the heat map (Figure 1 (b)) reveals that the distributions of modal verbs vary considerably across genres.

Taken together with the fact that distributions for modal *senses* vary quite little across genres (cf. Figure 1 (a)), this suggests that genres vary with respect to the preferred lexical realization of particular modal senses.

Distributions of modal senses per modal verb. Turning to the distributions of modal senses per modal verb and how they vary across genres, we display heat maps for each modal verb.[33] Each map shows the divergence between sense distributions for the given verb, for each pair of genres.

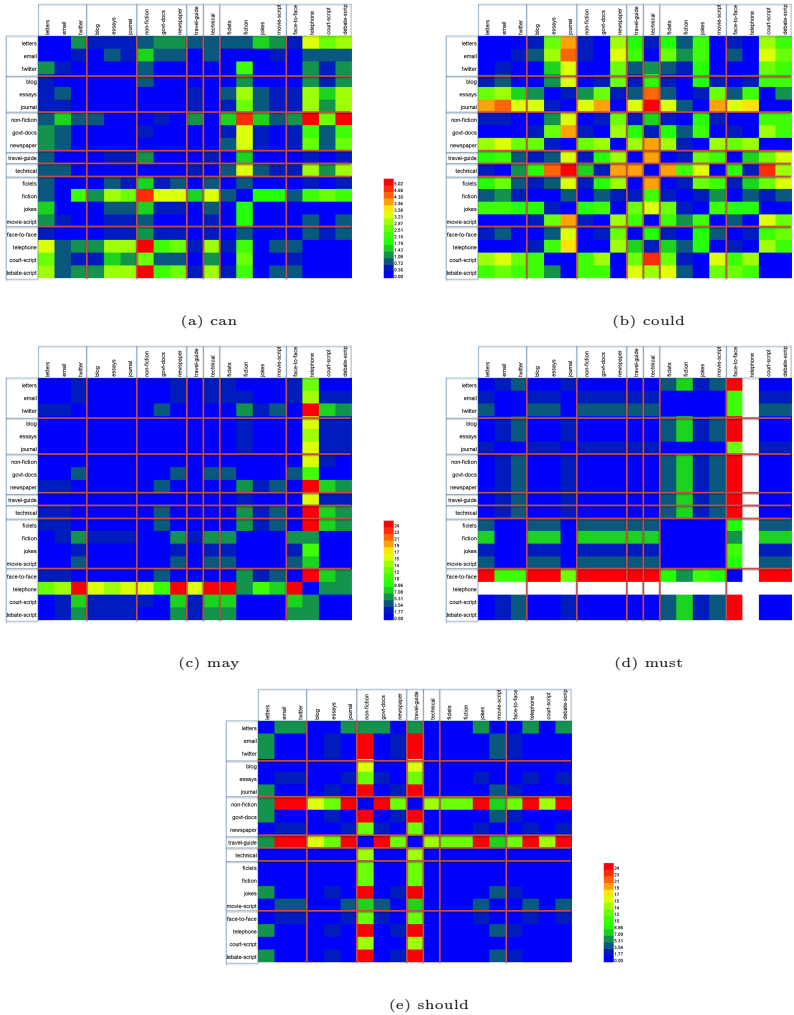
For *may* and *must* (Figures 2 (c) and (d)), most genre pairs have highly similar sense distributions, with only a small amount of divergence showing for the spoken genres and for `fiction`.[34]

The heat map for *should* (Figure 2 (e)) again shows a mostly low-divergence picture, with the notable exceptions of `non-fiction` and `travel-guides`. These two genres are similar to each other and highly divergent from the other genres.

Can presents yet another different picture. Here Figure 2 (a) is the only verb for which we see something like a block differentiating one of the genre subgroups, as the spoken genres at the right-hand side of the figure show high similarity with each other and higher divergence with the other genres. Both `fiction` and `non-fiction` here show a slightly higher degree of divergence.

[33] *Shall* is excluded due to the extremely small number of occurrences.

[34] The blank lines for *must* and `telephone` indicate no value, given that we had no instances for this modal verb and genre combination.



(a) can

(b) could

(c) may

(d) must

(e) should

FIGURE 2 Sense distribution differences for every modal verb.

Finally, the heat map for *could* shows a high amount of variance for many genre pairs (Figure 2 (b)). Still, several subgroup blocks of homogeneous sub-genres are noticeable here: offline interaction in the upper left, *non-fiction* and `government documents` in the center, and the grouping of `movie script/face-to-face/telephone` to the right.

A deeper analysis of divergences in *predominant senses* the modal verbs take in different genres is given in Table 16. For several genres

	must	may	can	could	should	shall
letters	deontic	epistemic	dynamic	**epistemic**	**ep/de**	–
email	deontic	epistemic	dynamic	**epistemic**	deontic	–
twitter	deontic	epistemic	dynamic	dynamic	deontic	deontic
blog	deontic	epistemic	dynamic	dynamic	deontic	deontic
essays	deontic	epistemic	dynamic	dynamic	deontic	–
journal	deontic	epistemic	dynamic	dynamic	deontic	**epistemic**
non-fiction	deontic	epistemic	dynamic	dynamic	**epistemic**	–
govt-docs	deontic	epistemic	dynamic	**epistemic**	deontic	deontic
newspaper	deontic	epistemic	dynamic	dynamic	deontic	–
travel-guides	deontic	epistemic	dynamic	dynamic	deontic	–
technical	deontic	epistemic	dynamic	**epistemic**	deontic	–
ficlets	deontic	epistemic	dynamic	dynamic	deontic	–
fiction	**epistemic**	**ep/de**	dynamic	dynamic	deontic	deontic
jokes	deontic	epistemic	dynamic	dynamic	deontic	**epistemic**
movie-script	deontic	epistemic	dynamic	**epistemic**	deontic	deontic
face-to-face	**epistemic**	epistemic	dynamic	dynamic	deontic	–
telephone	–	**deontic**	dynamic	**ep/dy**	deontic	–
court-transcript	deontic	**deontic**	dynamic	dynamic	deontic	deontic
debate-transcript	deontic	**ep/de**	dynamic	dynamic	deontic	–

TABLE 16 Predominant sense of individual modal verbs in distinct genres. Departures from a verb's dominant sense overall are marked in bold face.

we observe a switch of predominant sense for one or several modal verbs: for `letters, email, govt-docs, technical, movie-script` towards epistemic *could*, for `non-fiction` with a switch to epistemic *should*, for `fiction` and `face-to-face` towards epistemic *must* and `telephone, court-transcript` for deontic *may*. The most frequent changes in predominant sense are seen for *could*, and the most stable modal verb is *can*.

With the exception of `travel guides`, all the genres that stand out in the heat-maps in Figure 2 (`fiction, non-fiction, movie-script, face-to-face, telephone`) go along with a change of predominant sense for at least one modal verb. In some, but not in all cases, the observed changes in predominant sense are reflected by high variances in overall sense distributions for the respective genres, compared to other genres (e.g. `non-fiction` for *should* or `face-to-face` for *must*). This means that in addition to variances that result in a complete switch of predominant sense, there are also weaker divergences across genres.

Summary. By examining the distributions across the genres of MASC, we indeed see a wide variance from genre to genre in the distributions of modal verbs, as well as the distributions of senses per modal verb. We note certain tendencies for some genres to behave similarly to each other, in overall sense distribution and sense per modal verb, but also that there are specific (sub)genres that do not cohere with the groups proposed by Passonneau et al. (2014). Some modal verbs are

relatively stable in sense distribution for many genre pairs (or within-genre groups), while especially *could* (and to a lesser degree *can*) shows strong variational behaviour. We also observe a considerable number of changes in predominant sense for certain modal verbs and genres, most prominently for *could, may* and *must*. These observations are relevant for Q3, in that we must expect variation in sense distributions – to different degrees for different verbs – if we apply modal sense classifiers to out-of-domain data from novel genres. Thus we next turn to the second part of question Q3, and explore to what extent such distributional differences influence automatic modal sense classification.

8.2 Experimental settings

In this set of experiments we are concerned with the question of how robust the classifiers described in Section 7 are against distributional variation in the test data. To this end, we apply the models trained on MPQA and EPOS to the different genre sub-corpora of MASC and evaluate their performance for each genre. The MASC data is only used for testing.[35]

Classification setup. In Section 7 we investigated the performance of classifiers trained on five distinct training datasets using three different feature sets. Out of these 15 classifiers we chose four models for evaluation on MASC testing data.[36]

$\mathbf{CL}^{\pm b}_{ME}$ using $\mathbf{F}_{All}$: Concerning sources and size of the training data, we chose two models of type CL_{ME}, trained on the blend of MPQA and EPOS. This is because we expect that *reducing sparsity and distributional bias* is advantageous for the generalization power of a classifier when applied to variationally distinct genres, especially when changes of predomiant sense occur. The results in Table 11 showed that the classifiers induced from *balanced* training data are usually outperformed by the *unbalanced* classifiers. Still, with the complete feature set F_{All}, we obtained largely comparable results. Thus, on the assumption that a classifier trained on balanced data generalizes better when changes in sense distribution occur in the test data, we decided to investigate the relative performance and robustness of the respective classifiers, $\mathrm{CL}^{\pm b}_{ME}$ with F_{All}.

$\mathbf{CL}^{-b}_{M}$ using $\mathbf{F}_{R\&R}$ vs. $\mathbf{F}_{All}$: using classifier CL^{-b}_{M} with $\mathrm{F}_{R\&R}$ gives us insight into the robustness of R&R's original unbalanced classifier

[35] A more direct way to investigate this question is to train classifiers for distinct (MASC) genres and to test them against others. As this involves creation of larger training data sets, we postpone this to future work. See our discussion in Section 9.

[36] The four different models per modal verb are all trained on the best-performing fold from our cross-validation experiments.

trained on MPQA, when applied to data from different genres, and constitutes a strong baseline. Given the significant improvement of CL_M^{-b} with F_{All} on MPQA, we also investigated this model in order to evaluate the contribution of the semantic features. Moreover, comparing CL_M^{-b} and CL_{ME}^{-b} on the same (strong) feature set F_{All} should show to what extent reduction of sparsity alone contributes to the overall results.

Finally, we compare the classifiers to the random (BL_{ran}) or majority baselines (BL_{maj}) (cf. Section 7.1).

The individual modal verb classifiers obtained for the four classifier models are evaluated on all 19 genres: `blog`, `court-transcript`, `debate-transcript`, `email`, `essays`, `face-to-face`, `ficlets`, `fiction`, `govt-docs`, `jokes`, `letters`, `movie-script`, `newspaper`, `non-fiction`, `technical`, `telephone`, `travel-guides`, `twitter`.

Loss of instances. Some of the MASC genres differ considerably from standard text types, also with respect to their sentence structure. We thus occasionally encounter a modal verb target in a context that poses challenges for feature extraction. 38 sentences had to be deleted due to such pre-processing errors.[37]

8.3 Classification results

The classification results for the different classifiers on MASC are presented in four tables. Each cell in a table represents the classifier performance for a particular combination of genre and modal verb. For each cell we indicate which of the other three classifiers it outperforms at a significant level ($p < 0.05$) for that genre/verb combination. The classifiers and the symbols used to represent them are shown in Table 17. Table 18 compares the overall results for all models, and Tables 19 and 20 contrast the results obtained for the CL_{ME} and CL_M models for all individual genres and modal verb classifiers.[38]

Effect of training data and feature sets. We first compare the overall performance of the different classifiers: $CL_{ME}^{\pm b}$ using the full feature space, and the original classifier, CL_M^{-b}, with two feature settings, R&R's original features and again the full feature set. Table 18 states the aggregated micro-average results for each model. Both classifiers trained on the extended data set, $CL_{ME}^{\pm b}$, outperform CL_M^{-b} using R&R's feature set. CL_M^{-b} with F_{All} is surprisingly competitive, and

[37] An overview of the number of instances per modal verb and genre that were annotated vs. used for classification is given in the Appendix A.4.

[38] The last two columns in Tables 19 and 20 present micro-averages over all verbs for individual genres. The rightmost column excludes *shall*, for which we have low instance counts.

classifier	symbol
CL_{M}^{-b} with $F_{R\&R}$	$\bullet$
CL_{M}^{-b} with F_{All}	$\star$
CL_{ME}^{-b} with F_{All}	$\circ$
CL_{ME}^{+b} with F_{All}	$\diamond$

TABLE 17 Symbols for different classifiers used for reporting significance results. Each symbol indicates a model that is significantly outperformed.

classifier	micro-average
CL_{M}^{-b} with $F_{R\&R}$	71.67
CL_{M}^{-b} with F_{All}	72.38
CL_{ME}^{-b} with F_{All}	$\mathbf{74.36}^{\bullet\diamond}_{\star}$
CL_{ME}^{+b} with F_{All}	71.87

TABLE 18 Aggregated micro-average results for classification on MASC

beats the balanced model trained on extended data, CL_{ME}^{+b}, using the same feature set, F_{All}. However, the difference is not significant. By contrast, CL_{ME}^{-b} using all features significantly outperforms all competitor models. Note that CL_{ME}^{-b} with F_{All} performing significantly better than CL_{M}^{-b} with F_{All} confirms that reducing sparsity is important for overall performance. If we compare CL_{M}^{-b} using all features compared to using only R&R's, we see that the former achieves better results. Thus, in this configuration, and across various genres, semantic features contribute, while not significantly.

Looking at individual modal verb classifiers, we see that CL_{M}^{-b} with R&R's features achieves significantly better micro-average than CL_{ME}^{+b} ($\diamond$), for *can*, *shall* and *should*. It is significantly better than CL_{ME}^{-b} ($\circ$) only for *can*. For *could* and *may* both classifiers, $CL_{ME}^{\pm b}$, are significantly better than the original classifier ($\bullet$). By contrast, identical results are obtained for *could* and *may* when R&R's feature set is replaced with all the features for the original classifier, CL_{M}^{-b}. This means that it is only the classifier for *can* that does not profit from reduced sparsity. Moreover, for *can*, CL_{M}^{-b} with all features, F_{All}, is significantly better compared to using only R&R's features ($\bullet$), in terms of micro-average. This shows that semantic features are specifically important for identifying the correct sense for *can*.

Across all models, the micro-averages for CL_{ME}^{+b} are stronger for *may* and *could* compared to CL_{ME}^{-b}, while CL_{ME}^{-b} is better for *must*. For *can* and *should*, CL_{M}^{-b} with F_{All} achieves highest micro-averages.

F_{All} + CL^{+b}_{ME} ($\diamond$)	must	may	can	could	should	shall	micro-avg.	-shall
letters	87.50	81.82	76.81	69.23	50.00	-	76.70	
email	75.00	77.78	80.00	**72.73**	90.00	-	79.63	
twitter	60.00	100.00	70.00	**69.23**	95.00	100.00	76.52	76.32
blog	100.00	90.00	64.44	**69.23**	94.12	100.00	77.23	77.00
essays	77.78	94.74	64.71	**65.00**$^{\bullet}_{\star}$	81.82	-	**74.19**	
journal	**86.67**	**93.75**	63.16	**44.44**	100.00	0	69.47	70.97
non-fiction	100.00	**96.55**	52.94	**83.33**	0	-	68.14	
govt-docs	80.00	**74.07**	65.00	**79.49**	81.25	30.00	72.13	**75.89**
newspaper	66.67	100.00	56.52	**53.33**	72.73	-	65.88	
travel-guides	100.00	95.00	79.25	28.57	100.00	-	80.90	
technical	100.00	100.00	71.43	60.00	**88.89**	-	79.01	
ficlets	77.78	100.00	74.42	**64.87**$^{\bullet}_{\star}$	84.62	-	73.83$^{\bullet}_{\star}$	
fiction	**86.67**	**64.29**	62.50	**68.89**$^{\bullet\diamond}_{\star}$	90.00	50.00	**73.19**$^{\bullet}_{\star}$	**73.53**$^{\bullet}_{\star}$
jokes	**88.89**	100.00	67.27	**63.16**	77.78	0	69.47	70.21
movie-script	**83.33**	**100.00**	62.75	50.00	**69.23**	25.00	64.36	65.98
face-to-face	**100.00**	**100.00**	69.12	**57.89**	77.14	-	69.39	
telephone	-	100.00	69.23	27.27	78.95	-	63.64	
court-transcript	66.67	**65.38**$^{\bullet}_{\star}$	55.88	**68.75**$^{\bullet}_{\star}$	88.89	100.00	**63.83**	**63.04**
debate-transcript	100.00	**75.00**	63.64	**50.00**	100.00	-	67.05	
micro-avg.	83.33	**87.83**$^{\bullet}_{\star}$	66.74	**62.86**$^{\bullet\diamond}_{\star}$	84.06	41.67	71.87	72.25
macro-avg.	85.38	**89.91**	66.79	**60.29**	80.02	50.63		

F_{All} + CL^{-b}_{ME} ($\circ$)	must	may	can	could	should	shall	micro-avg.	-shall
letters	**100.00**	81.82	85.51$^{\diamond}$	**76.92**	50.00	-	84.47$^{\diamond}$	
email	75.00	77.77	**86.67**	68.18	**100.00**	-	83.33	
twitter	**70.00**	100.00	70.00	**69.23**	**100.00**	100.00	78.26	78.07
blog	100.00	90.00	**82.22**$^{\diamond}$	38.46	94.12	100.00	81.19	81.00
essays	88.89	94.74	67.65	50.00$^{\bullet}_{\star}$	**90.91**	-	**74.19**	
journal	**86.67**	**93.75**	**76.32**	27.78	100.00	0	**71.58**	**73.12**
non-fiction	100.00	**96.55**	**58.82**	66.67	0	-	**70.80**	
govt-docs	90.00	**74.07**	**70.00**	64.10	**100.00**	80.00	**75.41**	75.00
newspaper	83.33	100.00	**69.57**	46.67	**81.81**	-	**69.41**	
travel-guides	100.00	95.00	**83.02**	14.29	100.00	-	**82.02**	
technical	100.00	100.00	71.43	72.00	88.89	-	**82.72**	
ficlets	**88.89**	100.00	79.07	**64.87**$^{\bullet}_{\star}$	84.62	-	**76.64**$^{\bullet}_{\star}$	
fiction	**86.67**	57.14	71.88	48.89	90.00	**100.00**	68.84	68.38
jokes	77.78	100.00	78.18$^{\diamond}$	47.37	**100.00**	0	**73.68**	**74.47**
movie-script	**83.33**	**100.00**	70.59	68.75	61.54	50.00	**71.29**$^{\diamond}$	**72.16**
face-to-face	50.00	100.00	76.47	55.26$^{\bullet}$	**80.00**	-	**72.11**	
telephone	-	**100.00**	76.92	36.36	**100.00**	-	77.27	
court-transcript	66.67	**65.38**$^{\bullet}_{\star}$	57.35	53.13	88.89	100.00	60.99	60.14
debate-transcript	100.00	**75.00**	65.15	**50.00**	100.00	-	68.18	
micro-avg.	**86.11**	87.45$^{\bullet}_{\star}$	73.17$^{\diamond}$	55.34$^{\bullet}_{\star}$	89.64$^{\diamond}$	70.83$^{\diamond}$	**74.36**$^{\bullet\diamond}_{\star}$	**74.41**$^{\bullet\diamond}_{\star}$
macro-avg.	85.96	89.54	73.52	53.63	84.78	66.25		

TABLE 19 Classifier accuracy for MASC genres: Balanced and unbalanced, blended training data.

The overall best micro-average results for *could* are obtained with CL^{+b}_{ME}, showing substantial gain of 62.86 compared to 39.81 and 39.32 obtained by CL^{-b}_{M} (using $F_{R\&R}$ and F_{All}) – thus, additional training data and balancing proves to be very beneficial for *could*. For *can*, the best micro-average results are obtained with CL^{-b}_{M} or CL^{-b}_{ME}, both with F_{All}. This means that semantic features are profitable, and that adding additional data is generally useful.

$F_{R\&R}$ + CL_M^{-b} (•)	must	may	can	could	should	shall	micro-avg	-shall
letters	**100.00**	81.82	75.36	69.23	50.00	-	81.55	
email	75.00	77.78	84.44	**72.72**	**100.00**	-	83.33	
twitter	**70.00**	100.00	68.33	61.54	90.00	100.00	77.39	77.19
blog	100.00	**95.00**	75.56$^{\diamond}$	38.46	**100.00**	100.00	**83.17**	**83.00**
essays	**100.00**	94.74	**70.59**	45.00	90.91	-	66.67	
journal	**86.67**	87.50	60.53	27.78	100.00	0	69.47	70.97
non-fiction	100.00	93.10	55.88	50.00	0	-	66.37	
govt-docs	**100.00**	70.37	**70.00**	58.97	**100.00**	90.00$^{\diamond}$	**75.41**	73.21
newspaper	**100.00**	100.00	**69.57**	46.67	**81.82**	-	67.06	
travel-guides	100.00	95.00	69.81	**42.86**	100.00	-	76.40	
technical	100.00	100.00	71.43	**76.00**	88.89	-	81.48	
ficlets	66.67	100.00	79.07	37.84	84.62	-	61.68	
fiction	60.00	50.00	84.38	33.33	**93.33**	**100.00**	61.59	61.03
jokes	66.67	100.00	67.27	36.84	**100.00**	0	71.58	72.34
movie-script	75.00	60.00	72.55	**62.50**	**69.23**	**75.00**	69.30	69.07
face-to-face	50.00	100.00	73.53$^{\diamond}$	47.37	74.29	-	69.39	
telephone	-	0	84.62	36.36	89.47	-	81.82	
court-transcript	**100.00**	42.31	**72.06**	50.00	**100.00**	100.00	57.45	56.52
debate-transcript	100.00	50.00	77.27$^{\diamond}$	**50.00**	100.00	-	77.27	
micro-avg.	80.56	82.51	76.61$^{\diamond}_{\circ}$	39.81	89.24$^{\diamond}$	83.33$^{\diamond}$	71.67	71.52
macro-avg.	**86.11**	78.82	72.75	49.66	84.87	70.63		

F_{All} + CL_M^{-b} (⋆)	must	may	can	could	should	shall	micro-avg.	-shall
letters	**100.00**	81.82	**89.86**$^{\diamond}$	61.54	**50.00**	-	**85.43**	
email	75.00	77.77	93.33	59.09	100.00	-	**84.26**	
twitter	**70.00**	100.00	**78.33**	38.46	**100.00**	100.00	**79.13**	**78.95**
blog	100.00	**95.00**	**82.22**$^{\diamond}$	30.77	94.12	100.00	81.12	81.00
essays	**100.00**	94.74	67.65	15.00	**90.91**	-	67.74	
journal	80.00	87.50	**76.32**	22.22	100.00	0	68.42	69.89
non-fiction	100.00	93.10	55.88	50.00	0	-	67.26	
govt-docs	**100.00**	70.37	**70.00**	56.41	**100.00**	**100.00**$^{\diamond}$	74.59	72.32
newspaper	**100.00**	100.00	**69.57**	36.67	**81.82**	-	67.06	
travel-guides	100.00	95.00	79.25	14.29	100.00	-	79.78	
technical	100.00	100.00	71.43	72.00	88.89	-	**82.72**	
ficlets	66.67	100.00	**81.40**	18.92	84.62	-	59.81	
fiction	46.67	50.00	**90.63**$^{\diamond}$	31.11	86.67	**100.00**	61.59	61.03
jokes	66.67	100.00	**81.82**$^{\diamond}$	36.84	**100.00**	0	72.63	73.40
movie-script	66.67	60.00	**74.51**	56.25	**69.23**	**75.00**	69.31	69.07
face-to-face	50.00	100.00	**88.24**$^{\diamond}_{\circ}$	34.21	**80.00**	-	**72.11**	
telephone	-	0	**92.31**	**54.55**	**100.00**	-	**84.09**$^{\diamond}$	
court-transcript	**100.00**	42.31	66.18	37.50	**100.00**	100.00	58.87	57.97
debate-transcript	100.00	50.00	**80.30**$^{\diamond}_{\circ}$	25.00	100.00	-	**76.14**	
micro-avg.	80.56	82.51	**78.21**$^{\bullet}_{\circ}$	39.32	**90.04**$^{\diamond}$	**83.88**$^{\diamond}$	72.38	72.25
macro-avg.	84.54	78.82	**78.38**	39.52	**85.59**	**71.88**		

TABLE 20 Classifier accuracy for MASC genres: unbalanced MPQA training data, two different feature sets.

Effect of balancing training data. The individual micro-averages for every modal verb show that the classifier induced from balanced training data, CL_{ME}^{+b} (◇), is significantly outperformed by the classifier trained on unbalanced data, CL_{ME}^{-b} (○), for three modal verbs (*can, should, shall*). However, the converse is found for *could*. Overall, as seen in Table 18, the unbalanced classifier clearly beats its balanced equivalent in overall micro-average.

We further analyzed this behaviour by investigating differences in the distribution of modal senses as well as changes of predominant sense occurring between training and testing data, and how this affects classification results for the different models.

If the training dataset has a distribution of senses that is similar to the testing dataset, we expect that the classifier generalizes well to the test data, whereas it may suffer if the distributions change. To verify this assumption, we computed heat maps that compare the distributions of senses for the individual modal verbs in the unbalanced and balanced training datasets with the distribution of senses found in the MASC test data. However, the differences we obtained for balanced vs. unbalanced training sets did not correlate with the observed performance differences.

Still, it is possible that despite differences in the distributions of senses in the training and testing datasets, the most frequent sense stays the same. If the most frequent sense in training and testing data is constant, a classifier trained on unbalanced data will tend to perform better, even though the distributions differ. Instead, if there is no stable predominant sense, a classifier with a balanced training regime should perform better. And indeed, there is a change in predominant sense for 12 out of 16 genres for which the classifier with the balanced training set outperforms its unbalanced counterpart. In fact, from the 22 cases for which a change of predominant sense occurs, the balanced classifier gains an absolute increase of 2.18pp in accuracy. These results confirm that with changes of predominant sense between training and testing datasets, a classifier trained on a differently skewed distribution will suffer from performance losses.

Further effects on classifier behaviour can be due to the nature of the added training data. If the distributions we find in EPOS closely correspond to the distributions we find in specific genres of MASC, the classifiers trained on the unbalanced dataset can profit more from the training data than the balanced classifiers. Indeed, by using EPOS as additional training data, CL_{ME} includes instances of genres that are close to several of the genres we find in MASC, especially spoken genres. While we could not explicitly test this assumption, the results on `movie-scripts` point in this direction. Recall also that the results in Section 7 clearly showed that adding more data from distinct genres, such as EPOS, can harm the performance of an unbalanced classifier when evaluated on in-domain testing data.[39]

Thus, even though the classifier using unbalanced training data gives

[39]This effect was stronger, though, on models using shallow features.

overall best results compared to its balanced counterpart, there are situations where training and testing data show strong divergences in sense distributions that can harm performance. We observed clear effects of this kind when changes in predominent sense occur. In such cases, the balanced classifier model performs more robustly.

Analysis of performance for different genres. Examining micro-average results for individual genres across all four classifier types, classifier CL^{-b}_{ME} with F_{All} stands out with 11 out of 19 genres yielding best results, averaged over all modal verbs.[40] CL^{-b}_{M} with F_{All} follows with 7 genres out of 19. Among these genres, four were determined to involve cross-genre distributional variance within the MASC dataset: `non-fiction, travel-guides, face-to-face` and `debate-transcript` (cf. Section 8.1).

The offline-interactive genres (`letters, email, twitter`) profit most from semantic features in CL^{-b}_{M} (closely followed by CL^{-b}_{ME}). The discursive and story-telling genres (with the exception of `blog`) yield best results with CL^{-b}_{ME}, i.e., with semantic features and extended training data. Also `non-fiction` and `travel-guides` profit most from the unbalanced classifier with extended training base and semantic features. The spoken genres obtain best results for CL^{-b}_{M} with F_{All} and CL^{-b}_{ME}. Here, semantic features and extended training data help. In contrast, for `court-transcript`, which exhibits a predominant sense distinct from its training data, we obtain better results with CL^{+b}_{ME}. A similar pattern is observed for `fiction`.

Comparing the results for individual modal verb classifiers per genre, we observe that difficult genres for *could* are `travel-guides` and `journal`, with 3 out of 4 classifiers not beating the random baseline. The original classifier CL^{-b}_{M}, with both feature sets, performs worse than random for *may* when evaluated on `court-transcript` and `telephone`. All four classifiers for *should* evaluated on `non-fiction` are outperformed by the random baseline.

Reasons for the low performance of these individual classifiers on some genres can be various: a change of predominant sense between training and testing data, a small number of instances or unsuccessful feature extraction, due to sentence complexity or difficulties in the automatic annotation. We did not deeply analyze these effects, but confirmed they are present in the data.

We observed that for example in `travel-guides`, for *could* only one instance is annotated for the *dynamic* reading. Classifiers trained on

[40] We mark best results across all four tables with bold face if at least one classifier outperforms another.

the unbalanced datasets, CL_{ME}^{-b} and CL_{M}^{-b}, have *epistemic* as the most frequent sense in their training datasets. This can explain their poor performance on this specific genre and modal verb combination. Also, the data for *may* and `telephone` contains only one instance, (21.a), which can be easily misclassified, because it is hard to perform extraction of dependency features on this kind of language. Similarly, the test set for *should* in `non-fiction` contains only two instances. Also short sentences can be problematic, such as (21.b), from `court-transcript`. Finally, in the test data for *could* in `journal` we find examples where difficult sense distinctions can result in problems for classification. In (21.c) *dynamic* reading is annotated in the gold data, but the classifiers assign the *epistemic* reading. For this example, *epistemic* sense can be justified too, but *dynamic* was judged stronger by the annotators, and was assigned in accordance with the guidelines.

(21) a. no i had them walk out of my class and not say their name anything and i finally got to where i go okay i 'm Debbie Moore you know *may* i ask who you are and what you are in my classroom for

b. You *may*.

c. That incredible empire dominated the world and I imagine it did not seem possible to people living then that anyone *could* loosen that empire's grip on humanity.

Summary. Evaluated across all genres of the MASC corpus, the classifier model trained on the unbalanced blend of MPQA and EPOS with the full, semantically enriched feature set achieves significantly better micro-average than all the other classifiers it is contrasted with. It outperforms the unbalanced classifier that is only trained on MPQA using the same feature set. This clearly shows that extending the training set contributes to the overall performance of the classifier. Comparing results to the balanced version of this classifier shows that for certain genres the unbalanced distributions in the extended training set can be harmful, and that balancing the training data can improve results especially in case of changes of predominant sense. Contrasting the classifier models trained only on the MPQA data with R&R's features vs. all features we find that the semantic features improve results, yet not significantly. This differs from our results for in-domain evaluation, in Section 7.

Individual micro-averages for modal verbs show that the classifier with balanced training sources is outperformed by its unbalanced counterpart. Deeper analysis shows this behaviour occurs when the most frequent sense that is seen in training corresponds to the one found in the

testing data. Our results suggest that the sense distributions in EPOS can be profitable for some genres in MASC with the unbalanced CL^{-b}_{ME}, but we also see clear improvements of the balanced classifier with certain genres, for instance `fiction` and `court-transcript`, which were shown to exhibit divergence of sense distributions and changes of predominant sense for different genres represented in MASC. So, even though the classifier with unbalanced training sources yields overall best results, we provided evidence that re-proportioning training data to overcome distributional bias is profitable in case of strong distributional differences. A further, clear result of our analysis is that to achieve good performance, within or across genres, the addition of semantic features is crucial.

From the analysis of classifier performance for diverse genres we learn that some genres and modal verb combinations are difficult for some or most of the classifiers. Reasons can be a change in the most frequent sense from the training to the testing data, unsuccessful feature extraction due to processing difficulties, a small number of instances, or general difficulties in sense distinctions. We also saw that some genre groups profit more from semantic features alone, while others profit from both features and (un)balanced additional training sources.

In response to our research question Q3: *Are there genre differences in the distribution of modal senses, and to what extent are they mirrored in the performance of classification models?*, we can conclude that by analyzing the distribution of modal senses across different genres, we find that (i.) there are considerable differences in overall distribution of senses for some genres and (ii.) that for some modal verbs we find a great variation in sense distribution across genres. We also confirmed that (iii.) in case of distribution differences that involve a change in predominant sense, classifiers trained on balanced data sets can be beneficial for classifier performance. Overall, classifiers profit most from unbalanced natural sense distributions with extended training data of diverse sources, and their performance is best when using semantic features.

Our results compare favourably against prior work of Prabhakaran et al. (2012), who experienced great losses when applying their classifiers on non-homogeneous data (cf. Section 2).[41] A natural next step from our current findings is to take further advantage of the manually labeled portions of the multi-genre corpus MASC to perform systematic *domain* or *genre adaptation*. This could be done by using sampling

[41] Though we cannot compare directly to their work given that their annotation scheme and experimental data differs from ours.

techniques that re-proportion training data to approximate the distribution of out-of-domain target genres. Another option is to instead adapt model parameters, as proposed in Bloodgood and Vijay-Shanker (2009), who adapt the cost factor of SVM classifier models to the estimated distribution of out-of-domain evaluation data.

9 Conclusions

Modality is an important aspect of discourse analysis relating to argumentation, planning, and reasoning in intensional contexts. In this work, we focused on the classification of modal verb senses that carry different types of intensional meaning: epistemic, deontic and dynamic modality.

Prior work in Ruppenhofer and Rehbein (2012) established a well-founded annotation scheme for modal senses, provided manually annotated resources and induced lexical modal sense classifiers. Yet, due to the small data set and strong distributional bias in the data, the classifiers could hardly beat the most frequent sense baseline, and it was unclear whether the model generalized to variations in meaning distribution.

In our work, we address this issue, and provide answers to three research questions relating to (i.) overcoming the sparsity of annotated resources, (ii.) the design of a semantic classification model that achieves significant performance improvements and proves to be robust against variation in sense distributions, and (iii.) gaining insights into the variability of modal sense distributions across different genres, via annotation of a considerable portion of the MASC corpus, followed by investigation of the extent to which distributional differences may affect classifier models in an out-of-domain evaluation experiment.

In response to these questions, we have made various contributions and obtained a number of insights.

We applied a *paraphrase-driven projection* method for the *automatic induction of sense-labeled training data using parallel corpora*. The senses of modal verbs bear an ideal level of granularity for sense projection using a small set of paraphrases. Using carefully selected paraphrase seeds, the induced annotations yield high accuracy of 0.92. This method can be applied on a broad scale and for many languages, given that bitexts are available in large quantities. Recent work in Marasović and Frank (2016) applied the same method to automatically acquire a large dataset for modal sense classification for German.

We designed a *linguistically motivated semantic feature space for modal sense classification* that is effective in reducing misclassifications.

We obtain high performance gains especially for modal verbs with difficult sense distinctions and variable sense distributions (most prominently *can* and *could*). The high generalization capacity of these models is confirmed in balanced and unbalanced training settings. Even in isolation, the novel semantic features achieve competitive performance, outperforming traditional feature sets.

Finally we investigated the *variability of modal sense distributions across various genres*, by manually annotating a portion of the multi-genre corpus MASC. The overall sense distribution across 19 genres diverges considerably from MPQA, but only specific genres or genre-groups in MASC diverge from each other in their overall sense distribution. In our annotated data, the sense distribution per modal verb shows quite individual behaviour, across and also within genre groups. In our annotated data sections of MASC, 12 out of 19 genres exhibit a change in predominant sense for at least one modal verb. When applying lexical classifiers trained on MPQA to all genres of MASC, classifiers trained on unbalanced corpora with extended data sizes and using semantic features significantly outperform the model from prior work, as well as balanced classifier variants. The balanced classifiers, however, prove to be robust against changes in predominant sense and outperform classifiers trained on unbalanced data in such configurations.

Our analyses on MASC give some new insights, compared to the experiments on MPQA. While the addition of semantic features is favourable in both experiments, the addition of training data from EPOS was harmful for classifiers trained in an unbalanced setting and evaluated on in-domain data from MPQA. When applying the classifier to diverse genres from MASC, however, additional data proves effective. This shows that additional training data needs to be selected from appropriate sources.

Our work also contributes a substantial amount of annotated data. All our annotated resources will be publicly available.[42] They include the automatically annotated EPOS data set from Europarl and OpenSubtitles, comprising 7,693 instances. The set of seed paraphrases employed for projection are listed in Appendix A.1. Also the annotations on 19 genres of MASC, a total of 2,041 instances, will be publicly available. To our knowledge this is the largest manually annotated corpus of this kind.

Though in this work we made considerable progress, there are remaining open issues and various avenues for future work.

[42]The resources can be downloaded at `http://projects.cl.uni-heidelberg.de/modals`.

Most evidently, we have not yet fully addressed the issue of cross-genre evaluation and adaptation. A next step is to adapt classifier models to the (estimated) distributions of specific target genres, following Bloodgood and Vijay-Shanker (2009)'s approach. The individual data set sizes for the 19 genres are modest. On their basis, however, and using the cross-lingual projection methods we developed, it will be possible to investigate cross-genre model adaption and evaluation on a broader scale.

Not all genres are equally suitable for deeper semantic processing, due to noise in preprocessing layers or lack of coverage of the semantic resources employed. This can be addressed by employing more light-weight distributional semantic models, such as neural networks, which do not rely on pre-processing. Recent work in Marasović and Frank (2016) show that convolutional neural networks are able to improve on manually crafted feature-based approaches and are easily portable to novel languages, while preserving semantic factors that were found to be relevant in the present work.

Currently, we build lexically specific local classifier models. A single classifier model for all modal verbs could also be explored. Further extensions will include source and target role assignment. Possible application areas for our work are argumentation mining (Becker et al., 2016) and sentiment analysis (Benamara et al., 2012) or detection of speculative language. Lexical modal sense classification also has immediate applications in Machine Translation (Baker et al., 2012).

Acknowledgments

This work was partially funded by the Leibniz Science Campus *Empirical Linguistics and Computational Language Modeling*, funded by the Leibniz Association under grant no. SAS-2015-IDS-LWC and by the Ministry of Science, Research, and Art (MWK) of the state of Baden-Württemberg, as well as the German Research Foundation as part of the Research Training Group *Adaptive Preparation of Information from Heterogeneous Sources* (AIPHES) under grant No. GRK 1994/1.

We thank Yannick Versley for providing us with word alignments for Europarl and OpenSubtitles using PostCAT. We thank Annemarie Friedrich for advice on feature extraction, preparations for the MASC data, as well as her contributions to Zhou et al. (2015). We also thank our annotators: Stefan Gorzitze and Nils Feldhus, as well as students of the Semantics course at ICL WS 2014/15. We are grateful to Éva Mújdricza-Maydt for supporting us in using WebAnno.

References

Baayen, Harald R., Richard Piepenbrock, and Leon Gulikers. 1996. CELEX2. Philadelphia: Linguistic Data Consortium.

Baker, Kathryn, Michael Bloodgood, Bonnie J. Dorr, Chris Callison-Burch, Nathaniel W. Filardo, Christine Piatko, Lori Levin, and Scott Miller. 2012. Modality and Negation in SIMT Use of Modality and Negation in Semantically-Informed Syntactic MT. *Computational Linguistics* 38(2):411–438.

Baker, Kathryn, Michael Bloodgood, Bonnie J Dorr, Nathaniel W Filardo, Lori Levin, and Christine Piatko. 2010. A Modality Lexicon and its use in Automatic Tagging. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC-2010)*, pages 1402–1407.

Becker, Maria, Alexis Palmer, and Anette Frank. 2016. Clause types and modality in argumentative microtexts. In *Workshop on Foundations of the Language of Argumentation (in conjunction with COMMA 2016)*. Potsdam, Germany.

Benamara, Farah, Baptiste Chardon, Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. 2012. How do negation and modality impact on opinions? In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 10–18. Jeju, Republic of Korea: Association for Computational Linguistics.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly Media, Inc.

Bloodgood, Michael and K Vijay-Shanker. 2009. Taking into account the differences between actively and passively acquired data: The case of active learning with support vector machines for imbalanced datasets. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 137–140.

Cohen, J. 1960. A coefficient for agreement for nominal scales. *Education and Psychological Measurement* 20(1):37–46.

Cui, Yanyan and Ting Chi. 2013. Annotating Modal Expressions in the Chinese Treebank. In *Proceedings of IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM)*, pages 24–32. Potsdam, Germany.

De Marneffe, Marie-Catherine and Christopher D Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. Association for Computational Linguistics.

de Marneffe, Marie-Catherine, Christopher D. Manning, and Christopher Potts. 2011. Veridicality and Utterance Understanding. In *Proceedings of the Fifth International Conference on Semantic Computing at IEEE 2011*, pages 430–437. IEEE.

de Marneffe, Marie-Catherine, Christopher D. Manning, and Christopher Potts. 2012. Did It Happen? The Pragmatic Complexity of Veridicality Assessment. *Computational Linguistics* 38(2):301–333.

Diab, Mona and Philip Resnik. 2002. An Unsupervised Method for Word Sense Tagging using Parallel Corpora. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, pages 255–262. Philadelphia, Pennsylvania, USA.

Fellbaum, Christiane. 1999. *WordNet*. Wiley Online Library.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378–382.

Friedrich, Annemarie and Alexis Palmer. 2014. Automatic prediction of aspectual class of verbs in context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL-2014)*.

Ganitkevitch, Juri, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-HLT 2013)*, pages 758–764. Atlanta, Georgia.

Graca, Joao V., Kuzman Ganchev, and Ben Taskar. 2007. Expectation maximization and posterior constraints. In *Advances in Neural Information Processing Systems (NIPS)*.

Hendrickx, Iris, Amália Mendes, and Silvia Mencarelli. 2012. Modality in Text: a Proposal for Corpus Annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 1805–1812. Istanbul, Turkey.

Hwa, Rebecca, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Natural Language Engineering* 11(3):311–325.

Ide, Nancy, Collin Baker, Christiane Fellbaum, and Charles Fillmore. 2008. MASC: The manually annotated sub-corpus of American English. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-2008)*.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 2005 Machine Translation Summit*, pages 79–86.

Kratzer, Angelika. 1981. The Notional Category of Modality. In H. J. Eikmeyer and H. Rieser, eds., *Words, worlds, and contexts: New approaches in word semantics*, pages 38–74. Berlin: de Gruyter.

Kratzer, Angelika. 1991. Modality. In A. von Stechow and D. Wunderlic, eds., *Semantics: An International Handbook of Contemporary Research*, pages 639–650. Berlin: de Gruyter.

Krippendorff, Klaus. 2004. Measuring the reliability of qualitative text analysis data. In *Annenberg School for Communication Department Papers*. University of Pennsylvania.

Kullback, Solomon and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22(1):79–86.

Lee, Kenton, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP-2015)*, pages 1643–1648. Lisbon, Portugal.

Light, Mark, Xin Y. Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *Proceedings of BioLINK 2004*.

Loaiciga, Sharid, Thomas Meyer, and Andrei Popescu-Belis. 2014. English-French Verb Phrase Alignment in Europarl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Manning, Christopher D, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Marasović, Ana and Anette Frank. 2016. Multilingual Modal Sense Classification using a Convolutional Neural Network. In *First Workshop on Representation Learning for NLP*. Berlin, Germany.

McNemar, Quinn. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2):153–157.

Morante, Roser and Walter Daelemans. 2011. Annotating modality and negation for a machine reading evaluation. In *CLEF 2011 Labs and Workshop Notebook Papers*.

Morante, Roser and Caroline Sporleder. 2012. Modality and Negation: An Introduction to the Special Issue. *Computational Linguistics* 38(2):223–260.

Nissim, Malvina, Paola Pietrandrea, Andrea Sanso, and Caterina Mauri. 2013. Cross-linguistic annotation of modality: a data-driven hierarchical model. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 7–14. Potsdam, Germany.

Padó, Sebastian and Mirella Lapata. 2009. Cross-lingual annotation projection of semantic roles. *Journal of Artificial Intelligence Research* 36(1):307–340.

Passonneau, Rebecca J., Nancy Ide, Songqiao Su, and Jesse Stuart. 2014. Biber Redux: Reconsidering Dimensions of Variation in American English. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING-2014)*.

Prabhakaran, Vinodkumar, Michael Bloodgood, Mona Diab, Bonnie Dorr, Lori Levin, Christine D. Piatko, Owen Rambow, and Benjamin

Van Durme. 2012. Statistical modality tagging from rule-based annotations and crowdsourcing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 57–64. Jeju, Republic of Korea.

Reiter, Nils and Anette Frank. 2010. Identifying Generic Noun Phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010)*, pages 40–49. Uppsala, Sweden.

Resnik, Philip. 2004. Exploiting hidden meanings: Using bilingual text for monolingual annotation. In A. Gelbukh, ed., *Computational Linguistics and Intelligent Text Processing*, no. 2945 in Lecture Notes in Computer Science, pages 283–299. Springer.

Rubinstein, Aynat, Hillary Harner, Elizabeth Krawczyk, Daniel Simonson, Graham Katz, and Paul Portner. 2013. Toward fine-grained annotation of modality in text. In *Proceedings of IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM)*, pages 38–46. Potsdam, Germany.

Ruppenhofer, Josef and Ines Rehbein. 2012. Yes we can!? Annotating the senses of English modal verbs. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 1538–1545.

Saurí, Roser. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, Brandeis University.

Saurí, Roser and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation* 43(3):227–268.

Saurí, Roser and James Pustejovsky. 2012. Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text. *Computational Linguistics* 38(2):262–299.

Spreyer, Kathrin and Anette Frank. 2008. Projection-based Acquisition of a Temporal Labeller. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP '08)*. Hyderabad, India.

Szarvas, György, Veronika Vincze, Richard Farkas, György Mora, and Iryna Gurevych. 2012. Cross-Genre and Cross-Domain Detection of Semantic Uncertainty. *Computational Linguistics* 38(2).

Thompson, Paul, Gilua Venturi, John McNaught, Simonetta Montemagni, and Sophia Ananiadou. 2008. Categorising modality in biomedical texts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-2008)*.

Tiedemann, Jörg. 2012. Parallel Data, Tools and Interfaces in OPUS. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, eds., *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218. Istanbul, Turkey. ISBN 978-2-9517408-7-7.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation* 39(2-3):165 – 210.

Yarowsky, David, Grace Ngai, and Richard Wicentowski. 2001. Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *Proceedings of HLT 2001, First International Conference on Human Language Technology Research*.

Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL-2013)*.

Zhou, Mengfei. 2015. *Cross-lingual Semi-Supervised Modality Tagging*. Bachelor thesis, Department of Computational Linguistics, Heidelberg University.

Zhou, Mengfei, Anette Frank, Annemarie Friedrich, and Alexis Palmer. 2015. Semantically Enriched Models for Modal Sense Classification. In *Proceedings of the EMNLP 2015 Workshop LSDSem: Linking Models of Lexical, Sentential and Discourse-level Semantics*. Lisbon, Portugal.

Appendix A.1: Seed Paraphrases

Seed paraphrase terms for modal sense projection (see Section 4 for details).

	epistemic			deontic			dynamic		
		#inst	acc		#inst	acc		#inst	acc
must	scheinen	63	0.90	brauchen	281	0.95			
	sicher	387	0.95	bedürfen	87	0.95			
	bestimmt	353	0.95	benötigen	17	0.95			
	wahrscheinlich	73	0.95	unbedingt	45	0.95			
	wohl	732	0.90	erforderlich	18	0.95			
	vielleicht	12	0.95						
	sicherlich	10	0.90						
may	vielleicht	3202	0.95	gestatten	89	0.95			
	möglicherweise	267	0.95	erlauben	67	0.95			
	wohl	106	0.95	hoffentlich	9	0.67			
	womöglich	63	0.95						
	eventuell	61	0.95						
	wahrscheinlich	58	0.95						
	sicher	14	0.95						
	vermutlich	12	1.0						
can	bestimmt	17	0.88	erlauben	8	1.0	schaffen	1145	0.90
				gestatten	2	1.0	gelingen	53	0.90
							-bar	17	0.82
could				erlauben	20	0.95	schaffen	70	0.90
				gestatten	2	1.0	gelingen	6	1.0
							-bar	7	0.86
should				brauchen	93	0.95			
				lieber	92	0.95			
				besser	108	0.90			
				erforderlich	10	0.90			
				hoffentlich	7	0.71			
shall				gestatten	2	1.0	gelingen	5	0.80
				erlauben	2	1.0	schaffen	1	1.0

TABLE 21 German seeds for sense projection. Number of retrieved instances and manual evaluation (accuracy on random samples of up to 20 instances).

Appendix A.2: Agreement on MASC per genre

Per-genre inter-annotator agreement for the labeled MASC data (see Section 5.3 for details).

MASC data	κ(anno1, anno2)	κ(anno1, curated)	κ(anno2, curated)
blog	0.76	0.86	0.81
email	0.71	0.86	0.86
essays	0.72	0.88	0.82
ficlets	0.77	0.93	0.81
fiction	0.64	0.90	0.75
govt-docs	0.69	0.93	0.73
jokes	0.75	0.90	0.82
journal	0.53	0.84	0.59
letters	0.74	0.91	0.74
movie-script	0.68	0.84	0.80
newspaper	0.72	0.88	0.80
non-fiction	0.52	0.83	0.69
technical	0.60	0.85	0.71
travel-guides	0.72	0.94	0.78
twitter	0.61	0.84	0.77
court-transcript	0.60	0.87	0.71
debate-transcript	0.68	0.82	0.88
face-to-face	0.51	0.87	0.61
telephone	0.57	0.97	0.68

TABLE 22 Annotator agreement: MASC data

Appendix A.3: Modal sense distributions in MASC genres

Overall modal sense distributions for the labeled MASC data (see Section 5.3 for details.

	#	epistemic	deontic	dynamic
letters	103	0.22	0.11	**0.67**
email	108	0.22	0.29	**0.49**
twitter	116	0.23	0.30	**0.47**
blog	105	0.27	0.25	**0.48**
essays	100	0.33	0.23	**0.44**
journal	95	0.33	0.24	**0.43**
non-fiction	113	**0.50**	0.10	0.40
govt-docs	122	0.36	**0.38**	0.26
newspaper	86	0.35	0.26	**0.39**
travel-guides	89	0.26	0.17	**0.57**
technical	82	**0.60**	0.13	0.27
ficlets	109	0.15	0.18	**0.67**
fiction	138	0.23	0.31	**0.46**
jokes	105	0.15	0.30	**0.55**
movie-script	99	0.25	0.27	**0.48**
face-to-face	149	0.18	0.22	**0.60**
telephone	47	0.13	**0.51**	0.36
court-transcript	144	0.14	0.35	**0.51**
debate-transcript	88	0.06	0.31	**0.63**

TABLE 23 Overall distribution of modal senses for each genre in the annotated MASC data.

Appendix A.4: Instances

MASC modal verb instances excluded from test set due to pre-processing issues (see Section 8.2 for details).

	genre	annotated	processed
can	blog	48	45
	court-transcript	70	68
	essays	35	34
	face-to-face	69	68
	ficlets	45	43
	fiction	33	32
	jokes	58	55
	telephone	14	13
	twitter	61	60
could	technical	26	25
	telephone	12	11
may	essays	24	19
	jokes	4	2
must	blog	6	5
	jokes	15	9
	newspaper	7	6
should	court-transcript	10	9
	essays	12	11
	face-to-face	36	35
	technical	10	9
	telephone	20	19
	movie-script	5	4
ought	debate-transcript	22	0
	journal	2	0
	non-fiction	8	0
	telephone	2	0
	twitter	3	0

TABLE 24 Overview of genres (per modal verb) for which number of annotated instances and instances left after processing differ.