This year we made many changes to the standard flow of reviewing and decision-making. The modifications of the process were motivated by the desire to improve the quality of reviewing and at the same time minimize the demands on individual reviewers. This document contains a brief description of the main changes that proved useful in our experience and that future program chairs may consider for future conferences.

Much of our work was an effort to make specific and operationalize vague terms we often encounter in discussion on conference management and to clarify the expectations for the large number of people involved in the process of reviewing. Whenever possible, we asked for input from our team of 42 area chairs. Understandably, some of the specifics we adopted are open to discussion and future refinement.

The core changes were to recruit a large pool of reviewers, to involve more actively the area chairs who can be entrusted with greater responsibilities to mitigate problems with the larger reviewer pool and to rethink the organization of the field into areas.

Shortly before the conference took place, we sent out a survey to all conference reviewers and area chairs. In this report we include some of the feedback from this survey to support our recommendations or to draw attention to the aspects of our proposal that appear to be in a need for improvement.

NAACL HLT was relatively small compared to ACL and EMNLP but much of what we did is likely to be applicable to these larger conferences as well. We received 396 long papers and 284 papers. Of these, 100 long and 82 short papers were accepted. The reviewing and decisions for both types of paper happened in the same short time period (two months between submission and decision announcement).

**The purpose of reviewing**

The goal of reviewing is twofold: to allow the area and program chairs to choose the right papers, and to communicate to the authors of submitted papers why their paper was accepted or rejected.
NAACL HLT is a premier venue for presenting computational linguistics work. To maintain this status for the conference, we feel that the program chairs ought to select a program of papers presenting novel ideas, as well as papers presenting solid extensions of ongoing strands of work. The goal of reviewing is to find the right mix of papers that advance the field.

Of course, not all papers are perfect or completely flawed. Some statistics from NAACL HLT 2016 can put this statement in perspective. For this exercise we assume that reviewer scores are consistent and accurate in their evaluation of the papers.

As usual, along with their reviews, reviewers submitted a number of numeric scores for the paper, the most important of which is the overall recommendation. We used a standard 1-5 scale with the following explanation:

The range of possible value is a variant of what we used at NAACL HLT 2016:

- **5 = Exciting**: I would fight for this paper to be accepted.
- **4 = Strong**: I learned a lot from this paper.
- **3 = Borderline**: It has some merits but also some serious problems. I’m ambivalent about this one.
- **2 = Mediocre**: I would rather not see it in the conference.
- **1 = Poor**: I would fight to have it rejected.

Presumably, papers that are easy decisions are those that had three overall scores all greater than 3 (the great papers) and the papers that received three overall scores all smaller than 3. For long papers, these groups accounted for 9% and 10% of all submissions respectively. For short papers the respective percentages were 5% and 16%. In other words for almost 80% of the paper at least one reviewer found a worthy aspect and at least one found a problematic aspect.

Accepting only the top 10% of the papers is not reasonable, as it will preclude from presentation many interesting ideas that will invigorate the field. Moreover, reviewer identity and consistency will affect scores, leading to even greater frustrations over rejection decisions if such policy were to be adopted.

Our aim was to reduce the arbitrariness in decision making for that 80% of the papers. We attempted to do so by giving more responsibilities to the area chairs and asking them ultimately to make recommendations explaining why a paper should be or should not be presented in the conference, so that the authors of the paper understand the PC’s decision.
Area chair responsibilities

Computational linguistics conferences have grown. NAACL HLT received 700 submissions and this is a small number compared to ACL or EMNLP which handle about double this number of submissions. The program chairs select the area chairs who directly oversee the reviewing of a given number of papers. Area chairs have the most control over who reviews a given paper, which reviewers are asked to supply additional information in their reviews and how reviewers discuss the paper. They make a final recommendation about accept/reject to the program chairs.

Below are the instructions we used for each step of the process. We worked with 42 area chairs and 33 of them completed a survey a couple of weeks before the conference took place. Wherever relevant, we include results from the survey.

Area chairs working in pairs This year we asked area chairs to work in pairs. Two area chairs shared the responsibility for managing the papers assigned to their area. The area chairs discussed each of the key decision they had to make. This minimizes the possibility for bias or arbitrariness in the decisions. The discussion required for coordinating the activities made area chairs articulate the rationale behind their choices, making it less likely that a decision was made simply for convenience in the interest of time.

94% of the 33 area chairs who responded to the survey said they were well-matched with their co-area chair. Only two reported that they were not well-matched with their co-area chair and that this made their work difficult.

Five of the area chairs who responded (15%) were area chairs for the first time. The pairing ensured that people working as ACs for the first time had guidance from their more experienced co-chair.

Defining conference areas As the conferences have grown, predicting area size and deciding on exact named areas has become increasingly difficult. In the past two years there have been last-minute area chair assignments, sometimes with poor match between area chair volunteer and the named area. An even bigger problem has been the prediction of the number of papers to be expected in the areas, leading to last minute recruiting of additional reviewers without particular criteria for reviewer competence beyond their willingness to participate in reviewing.

To deal with these issues and to account for the increased demand for area chair time under our changed review process, we limited the number of papers that could be assigned to an area to 40. There were 21 areas, so theoretically we were prepared to handle 840 papers, allowing for conference growth and differences in area popularity.
We used a list of keywords to pair area chairs, to assign papers to areas and to identify potential reviewers for the areas. The full list of keywords appears on the conference website. In addition to allowing flexibility in forming areas, it allowed for post-conference analysis which Owen Rambow presented at the opening session and the slides for which appear at the NAACL board website. These are also included as appendices in this report.

Area chairs chose their keywords from the provided list. The pairing of co-ACs was guided by keyword overlap. When assigning papers to areas, keywords shared by the two ACs had more weight, skewing the selection toward papers for which both had expertise. We did not assign descriptive names to areas. Of the area chairs who completed the survey, 52% felt that it was fine that there were no names for the areas and 33% said they would have preferred to have names for the areas but it was fine without names as well. Only 12% (4 area chairs) said that they found it irritating that areas had no names.

33% of the area chairs responded that their area was well-defined in terms of keywords. These were mostly areas that traditionally draw a large number of submissions (such as semantics and machine translation). 46% of area chairs responded that their area was not well-defined in terms of keywords but that that was ok. 21% of the area chairs responded that their areas were not well-defined in terms of keywords that that bothered them. Many traditional named areas also have a medley of papers (multilinguality, evaluation, applications), so comparison with other conferences is needed to determine if the number of mixed areas was larger than usual.

*Selecting reviewers for the track* Traditionally, area chairs have recruited the reviewers they wish to work with. There are two problems with this approach. First, as the number of submissions have grown, it has become increasingly difficult to recruit sufficient number of reviewers. Typically, here is a group of highly sought after reviewers, with many area chairs wanting to work with them. Second, this naturally leads to a “birds of a feather” effect, limiting breadth of points of view. We created a pool of potential reviewers and asked area chairs to bid on reviewers (with options “want”, “ok”, “don’t want”). We made sure that all pairs of area chairs had a fraction of the “prized” reviewers in their bidding pool. Both area chairs bid on all potential reviewers, and reviewers were assigned to an area only of both area chairs were at least ok with working with that reviewer.

We identified potential competent reviewers by analyzing the ACL anthology. We invited researchers who had published at least 3 papers (either short or long) in ACL, NAACL, EMNLP or COLING in the past 5 years or at least 5 papers in the same conferences in the last 10 years.
Each pair of co-chairs bid on 120 reviewers that had the most similar research expertise as indicated by their keyword overlap. Each potential reviewer was listed for bidding in four areas. We asked area chairs to look up the publication record of reviewers that they did not know by name. Many area chairs did do the look up and identified a number of reviewers that they would be happy to work with among those they did not know ahead of time. Other area chairs complained that the process is too time consuming and completed a default bid for researchers they did not know by name, either saying that they are ok with working with all of the reviewers in that category or that do not wish to work with any of the reviewers in this category. Area chairs were also free to recruit additional reviewers they wished to work with.

Reviewers were assigned to two areas where the ACs had indicated they want to or are ok to work with that reviewer. In this way, each reviewer could bid on up to 80 submissions.

The area chairs overwhelmingly (98%) liked the setting in which a pool of reviewers if presented to them. In addition, 69% were able to find many good reviewers in the list presented to them; 25% liked the option of choosing from a preexisting pool of reviewers but had trouble finding appropriate reviewers. The main reason for the difficulty some ACs felt in finding appropriate reviewers was that the ACs did not know many of the researchers in their bidding pool. This was identified as a difficulty by 32% of the area chairs. This is not surprising. As the field has grown many people do not necessarily know everyone with similar research interests. It seems reasonable to us to expect that ACs would invest a couple of hours in reviewer selection, as this approach would make the review process more inclusive. In the future, this issue may need to be brought up at AC selection time.

Assigning reviewers to papers After the conference submission deadline, reviewers cast their bids on papers in their two tracks. A number of reviewers (11%) commented that they saw fewer relevant papers than usual and found the process more tedious than in previous conferences. This low percentage suggests that the approach was sufficiently accepted by the reviewers. We wished to find three or four papers for each reviewer, so having many more appropriate papers was not necessary.

After the reviewer bids, we ran the automatic assignment of reviewers in START. Reviewers could be assigned to papers if they indicated they want or that they could review a paper. After that, area chairs were asked to make adjustments to the automatic assignments. The adjustments had to take into account the seniority of the reviewers, any know bias in terms of topic or approach for solutions or any other relevant information. This step is extremely important but often overlooked in the rush to kick start the reviewing period. Leaving a paper with three less experienced reviewers or with reviewers with different research philosophy or
taste can make a huge difference in the final outcome. For this reason, it is important
for area chairs to know the reviewers they are working with.

70% of the ACs reported that it was easy for them to match the papers with
reviewers. The other 30% reported some difficulties in finding suitable reviewers
for some papers in their track. Again, there is no data from prior years for
comparison to decide if this percentage was higher than the usual difficulties due to
papers with unusual combination of topics.

12% of reviewers said they had some papers outside their area of expertise, and 2%
said all their papers were outside their areas of expertise, or of no interest to them.
69% of the reviews were written by reviewers who had bid indicating that they
want to review the paper; 29% of the reviews were written by reviewers who had
bid indicating they are ok with reviewing the paper. The average reviewer load was
3 papers, which included a mix of long and short submissions. Only 43 reviewers
had more than four papers to review.

Leading the post-review and author response discussion After the review period, we
asked area chairs to read all reviews and contact reviewers who did not provide
enough information to justify their numerical recommendations to either accept or
reject the paper. In some cases, the reviewers did not supply the necessary
information and we recruited last minute reviewers to provide additional reviews.
In some cases, the area chairs identified the additional reviewer themselves, in
other cases we helped them identify potential reviewers from our pool with good
keyword match.

The post-review discussion was meant to address concerns raised in the reviews
and/or in the author response.

Meta review with final recommendation Finally, the area chairs wrote a meta-review
to the program chairs, briefly explaining why the paper should be accepted or
rejected. Both area chairs had to agree on the recommendation and its justification
or explicitly indicate that they disagree. In preparing their meta-review, they took
the author response into account when deciding how damning a given weakness of
the paper is. For papers recommended for acceptance, area chairs had to briefly
state the contribution of the paper that makes with worth presenting at NAACL HLT.
We did not make the meta review available to the authors in order to allow the ACs
to write freely, for example comparing to other papers or mentioning reviewers by
name.

Area chairs were not given quotas for how many papers to accept or reject. We did
not have any target acceptance rate in mind either and accepted all papers for which
there was a compelling justification for why they should be accepted.
Program chairs’ responsibilities

Selecting area chairs  With the increased responsibility of area chairs, it is clear that program chairs should be careful in their selection of area chairs. We had several requirements for people we invited to serve as area chairs. All of the area chairs met stricter requirements for participation on the field than that we used to get a list of potential reviewers. With only a couple of exceptions, the area chairs were among researchers who had published at least 3 long papers at NAACL, ACL or EMNLP in the past 5 years or at least 5 long papers in the same three conferences in the last 10 years. We also excluded from consideration researchers who had served as area chairs the previous year. We began our selection by initially inviting future area chairs with expertise matching the traditional named areas from 2015.

Selecting a pool of qualified reviewers  Our attempt was to get as large a reviewer pool as possible while at the same time ensuring that the reviewers can make judgments in line with the standard for the conference. The process worked well and it can be fine-tuned in future years. The pool will also change over the years, as new conferences contribute to the pool.

Making final decisions  We worked closely with the area chair when deciding the final accept or reject decision for each paper. We read all meta-reviews. When these did not give a clear indication for why the work in the paper should we presented in the conference, we contacted the area chairs for further clarification. In several cases the area chairs read their assigned papers to make their final recommendations and in a few cases we solicited additional reviews.

Our goal was to be as inclusive as possible, without compromising the quality of the conference. We read the original reviews and the author response for any long paper recommended for reject that also received at least one overall score of 4. We did the same for any short paper recommended for reject with at least one overall recommendation of 3.

Proposed *ACL Board Responsibilities

There were several conference policies which we felt are best addressed by the NAACL and ACL boards rather than by the program chairs, to ensure consistency over the years, as well as consistency across conferences each year.

- Clarify meaning of “double submission”, including TACL.
Double submission TACL and subsequent NAACL paper: make clear that submission to TACL counts as a double submission.

Double submissions in general: make clear that double submissions must always be declared at time of submission, and if one happens after submission, then the PC co-chairs must be notified by email. An undeclared double submission is ground for immediate rejection.

Pre-publications (ArXiv and others): make the policy clear: non-anonymous pre-publication is allowed anywhere (including personal websites)/is allowed only on ArXiV/is not allowed at all. Clarify whether this needs to be indicated in the submission. Clarify policy about non-anonymous pre-publication after submission. Then clarify to reviewers what to do with this information (presumably, do not follow links indicated by the authors when they google something).

- Clarify that *all* aspects of the formatting guidelines are binding (font size, spacing of title, line spacing, citation format), and any violation of any element is grounds for immediate rejection. This is because of fairness considerations.

**APPENDIX: Analysis of Keywords**

We used a system of keywords to characterize AČs’ and reviewers’ expertise, as well as papers. The idea was that there are several orthogonal dimensions to describing research, and by having areas that are defined by one area at a time we run the risk of not being able to find good reviewers.

**CORE DESCRIPTION 1: LINGUISTIC TARGET OF STUDY**

1) Discourse
2) Pragmatics
3) Semantics (can’t choose this --- have to choose subtype(s) --- more than one ok)
   a) Semantics -- lexical semantics
   b) Semantics -- distributional semantics
   c) Semantics -- formal semantics
   d) Semantics -- propositional semantics
   e) Semantics -- event semantics
   f) Semantics -- extra-propositional semantics
   g) Semantics -- grounded semantics (e.g. vision)
   h) Semantics -- ontologies
4) Syntax
5) Morphology
6) Phonology and phonetics
7) Prosody

CORE DESCRIPTION 2: APPLICATION TASK: WHAT DO YOU WORK ON?

1) Language understanding (in general)
2) Morphological analysis
3) Tagging
4) Chunking
5) Syntactic parsing
6) Semantic parsing
7) Word sense disambiguation
8) Textual entailment
9) Semantic similarity
10) Information extraction
11) Named entity recognition
12) Relation extraction
13) Event detection
14) Question answering
15) Knowledge acquisition
16) Information retrieval
17) Language generation
18) Summarization
19) Machine translation
20) Anaphora resolution
21) Dialog structure and analysis of conversations
22) Determining discourse relations and text organization (including argumentation mining)
23) Discourse semantic parsing
24) Dialogue and interactive systems
25) Image/video description generation
26) Integrating language and other modalities
27) Sentiment/opinion/emotion analysis
28) Belief/factuality/modality/irrealis analysis
29) ASR and other spoken language processing
30) Word segmentation in spoken utterances
31) Text categorization (of words, sentences and longer texts)
32) Spelling and/or grammar correction
33) Text quality prediction
34) Style analysis
35) Predicting speaker/writer characteristics
36) Lexicon and paraphrase induction
37) Mathematical models of language
38) Medical applications (public health/medical diagnosis/etc)

RESEARCH GOALS: WHY ARE YOU WORKING ON IT? (Choose if applicable)

1) End-user application building
2) Corpus creation and evaluation
3) Linguistic theories for NLP
4) Machine learning for NLP
5) Theoretical linguistics (theories of human language)
6) Cognitive modeling and psycholinguistic research
7) Applications to social sciences and humanities (including but not limited to political sciences/sociolinguistics/digital humanities)

APPROACHES: HOW ARE YOU WORKING ON IT? (Choose if making contribution in this area)

1) Machine learning
   a) Machine learning -- topic modeling
   b) Machine learning -- deep learning
   c) Machine learning -- Bayesian models
   d) Machine learning -- kernel methods
   e) Machine learning -- structured prediction
   f) Machine learning -- generative models
   g) Machine learning -- discriminative models
   h) Machine learning -- graphical models
   i) Machine learning -- representation learning
   j) Machine learning -- semi-supervised learning
   k) Machine learning -- unsupervised learning
   l) Machine learning -- other
2) Exploiting multilingual resources
3) Modeling linguistic knowledge (e.g. grammars)
4) Graph methods
5) Algorithm development for NLP
6) Corpus/data analysis (statistical or manual)
7) User evaluation

YOUR DATA: LANGUAGES

1) Morphologically rich languages
2) Low-Resource Languages
3) Dead languages
4) Chinese NLP
5) Arabic NLP
6) Indian languages NLP

YOUR DATA: GENRES

1) Chat and email (private unedited written dialog)
2) Literature
3) News
4) Social media (non-private unedited text)
   a) Twitter
   b) Blogs
   c) Discussion forums
   d) Other social media
5) Spoken dialog
6) Other spoken genres
7) Search log analysis
8) Encyclopedia
9) Biomedical texts

We now show acceptance rates for keywords with 10 or more submissions to the long paper track, which had an overall acceptance rate of 25%. The numbers on the categories indicate the number of submissions with that category.

LINGUISTIC AREAS

“NO LINX” means that no linguistic area was chosen. We note that such papers had a below average acceptance rate.
NLP AREAS
METHODS
GOALS

METHODS

Long Papers

ML: graphical models 12
ML: representation learning 52
Multilingual resources 35
ML: deep learning 90
ML: structured prediction 28
ML: semi-supervised learning 27
ML: unsupervised learning 32
ML: Bayesian models 11
Modeling language knowledge 32
NOMETH 57
Crowd sourcing 26
Graph methods 16
Algorithms for NLP 49
Corpus/data analysis 74
ML: other 44
ML: topic modeling 16

GOALS
LANGUAGES

We, unfortunately, did not have English as an option, which explains the large number of NO LANGUAGE choices.
GENRES
Long Papers

Chat and email 13
Spoken genres 11
NOGENRE 164
Discussion forums 13
News 72
Twitter 27
Other social media 28
Social media 24
Blogs 14
Literature 18
Biomedical texts 13
Encyclopedia 19

Other spoken genres 13
Long Papers