This section should not exceed one page.

General information

1. Institution
Institution acting as the official secretary: Utrecht University

2. Main applicant *(see Additional information section for details)*
Main applicant: prof. dr. Isabel Arends

3. AI talent *(see Additional information section for details)*
AI talent: prof. dr. Massimo Poesio

4. Project title: Dealing with meaning variation in NLP

5. Summary of the proposal

Research plan: The project investigates variations in natural language use and interpretation, enabling Natural Language Processing (NLP) models to generate and understand text in challenging situations where misunderstandings are likely, including vague and ambiguous expressions, and language whose interpretation is highly subjective (e.g., offensive language). The AI Talent involved is Prof. Poesio, who is a leading NLP researcher with an extensive track record in research related to this project and interdisciplinary collaboration. The project contributes to challenges from AIREA-NL on NLP, AI Systems and Humans, and AI Systems & Society, and from AINed's focus area AI for Dutch Language.

Activities plan for embedding in the AINed/NL-AIC Ecosystem: Prof. Poesio will take a leading role in (1) training the next generation of AI researchers and professionals in the Netherlands on the topic of NLP, (2) bringing together academics and professionals interested in NLP across the traditional disciplinary boundaries and private and public sectors, (3) making Dutch citizens more aware of, involved in, and enthusiastic about, AI research, (4) supporting responsible AI, (5) increasing engagement from Utrecht University with AiNed-NL AIC. The plan explains Prof. Poesio's planned involvement with particular AiNed/NL AIC working groups, focus areas, ICAI and ELSA labs, and the regional AI Hub.

Plans for commitment: Utrecht University will fund 2 PhD students. A third PhD student will be funded by a company working on problematic language detection (to be agreed within one year of award), such as Rewire, Meta AI, or a Dutch media company, with all of whom there are existing links. The UU department of Information and Computing Sciences has a strong track record of establishing collaborations with private and public partners, including over 30 ongoing (co-)funded PhD students, and structures are in place to make us confident of success. Furthermore, we already have a PhD student co-funded by a media company working on a related topic. Additionally, Prof. Poesio has a strong track record in establishing collaborations with private partners on NLP.

6. Keywords
Natural Language Processing, variation, disagreement, subjectivity, vagueness

Research plan

*This section is limited to **6** pages, including footnotes, figure captions, and tables, and excluding literature references. When writing, cover the following criteria: appropriateness and quality of the research plan to carry out research on an AI aspect from the AIREA-NL agenda[1], incl. track record of the AI talent in this research.*

7. Description of the research plan

The meaning of natural language expressions varies, sometimes dramatically, along a number of dimensions. Key dimensions include **subjective bias** (e.g., what's funny / offensive for one person may not be for another; Akhtar et al, 2021; Almanea & Poesio, 2022; Kocon et al, 2021; Leonardelli et al, 2021) **ambiguity** (e.g., the question of what a pronoun like "he" refers to in a given context, Poesio & Artstein, 2005; Versley, 2008; Recasens et al, 2011; Passonneau et al, 2012; Plank et al, 2014; Pavlick & Kwiatkowski, 2019), **vagueness** (e.g., what data dimensions and thresholds do we apply when we call the weather "mild", or the condition of a patient "stable"? Van Deemter 2010, Douven et al. 2013), and **diachrony**, as when the meaning of an expression changes over the course of a conversation, or becomes temporarily constructed during the conversation (Brennan and Clark 1996, Pickering and Garrod 2006).

Natural Language Processing (NLP) is the area of Artificial Intelligence in which algorithms are studied that understand and/or produce text in ordinary languages, such as Dutch and English. Variations in meaning, as described in the previous paragraph, raise serious challenges for NLP, regardless of the research paradigm that is used (e.g., symbolic or sub-symbolic). These challenges arise from a scientific/technological point of view (e.g., How can systems learn how to interpret particular language expressions? How can these interpretations be evaluated?) and from an application point of view (e.g., What should a social media company do with a post that is offensive according to some people, but not according to others? How should a robot recognise when its interpretations are precise enough?). Industry and the scientific community have now recognized the challenge and started to study the problem (Plank et al, 2014; Aroyo & Welty, 2015; Akhtar et al, 2020; Uma et al, 2021b), often in projects led by Prof. Poesio and his team and/or his collaborators. Nonetheless, most of the fundamental questions still need to be addressed.

Objectives and sub-projects. The objective of this project is to carry out fundamental as well as applicable research on meaning variation in NLP along several dimensions of variation, exploring the interconnections between them and the implications for NLP research and applications. Two of the projects (P1, P2) carry out foundational research on theoretical linguistic theories and statistical tools for analysing variation; two projects (P3, P4) carry out in-depth empirical/computational research into areas of NLP in which variation has been shown to be prevalent, but which have so far resisted analysis using existing mathematical and computational models; the two remaining projects (P5, P6) look at how variation emerges along a temporal dimension, focusing on dialogue. These six projects will thus investigate closely related themes, allowing the researchers working on them to learn from each other and to closely collaborate in an interdisciplinary team.

The six PhD projects are the following:

Project 1: Theoretical foundations (1) Formal semantics for vagueness in interpretation
Abstract: Vagueness is the pervasive phenomenon in which words have imprecisely defined boundaries, which are applied differently in different contexts and by different people. For example, when a patient report describes a baby's blood pressure as "too high", or her condition as "stable", these terms are interpreted differently by different clinicians (Portet et al. 2009). This PhD project will study mathematical and computational models of uncertainty and vagueness. In recent years, the literature in this area has shifted away from 2-valued towards multi-valued models, based on modern versions of Zadeh-style fuzzy set theory, or on Gardenfors-style conceptual spaces (Douven et al. 2013), or on probabilistic models (Edgington 1997, Van Deemter 2010), but these models have rarely been tested with real data. This PhD project will use existing "big" datasets to find out which models predict and explain the data best.
Objective: This PhD project aims to deliver a computationally interpreted and empirically supported multivalued semantics for vagueness.
Area: This PhD project combines formal semantics with computational modelling. It will be a collaboration between Computing Science (profs. van Deemter, Poesio) and Linguistics (dr. Nouwen).

[1] https://www.nwo.nl/sites/nwo/files/documents/AIREA-NL%20AI%20Research%20Agenda%20for%20the%20Netherlands.pdf

Project 2: Theoretical foundations (2): Learning under disagreements between annotators

Abstract: In NLP, human judges, called annotators, are frequently needed to tell researchers what a given expression "means", by assigning the expression a label. When human judges disagree about a label (e.g., whether an utterance is offensive or not), these disagreements should be taken into account, as opposed to simply aggregating the values e.g., using reconciliation or majority voting. Such disagreements are now generally recognized to provide information rather than being noise (Aroyo& Welty, 2015; Plank et al, 2014; Uma et al, 2021b). We also need to recognize that uncertainties may originate from different sources: they may be due to semantic ambiguity, as in referential uncertainty (P3), or to subjective bias (P4). There is evidence that disagreements and their source should be taken into account when deciding which method to use for training models with such data, and evaluating such models (Reidsma & Carletta, 2008; Uma et al, 2021b). This PhD project will investigate differences between types of disagreement. The PhD candidate will assess whether the differences between various sources of disagreement (e.g., noise, ambiguity, and subjective bias) can be detected using statistical models. They will also investigate how to use datasets containing different types of variation to train and evaluate NLP models, starting from the state-of-the-art and the datasets created in DALI (Uma et al, 2021b) and subsequently in the 2021 Shared Task LeWiDi (Uma et al, 2021a); it will also leverage new datasets that have appeared since, e.g., for Natural Language Inference. Furthermore, the student will investigate whether the obvious candidates for soft evaluation metrics (cross-entropy, Kullback-Leibler divergence) apply to all these tasks. Finally, the student will study to what extent variations in *one person's* verbal behaviour can be understood mathematically in the same way as variations between different speakers.

Objectives: This PhD project aims to deliver:

- A statistical theory of disagreement that can be used to recognize and categorize disagreements,
- Forms of training and evaluation that are robust against disagreements between annotators.

Area: This PhD project would ideally be carried out by someone well-versed in information theory and the design and analysis of experiments with human participants. It will be a collaboration between Computing Science (dr. Gatt, prof. Poesio) and Linguistics (dr. Paperno).

Project 3: Empirical analysis of variation (1). Variation in coreference and reference

Abstract: Early research on disagreement was often motivated by findings about "anaphoric" referring expressions such as "he" or "she" (Poesio & Artstein, 2005; Versley, 2008; Recasens et al, 2011). But whereas methods for learning 'from crowds' have been successfully applied to other types of disagreements (Uma et al, 2021b), and substantial datasets now exist of multiple anaphoric judgments (Poesio et al,2019), computational models of referring expression *interpretation* do not yet exist that can effectively learn from such datasets. Training coreference models 'from crowds' has proven to be challenging to design, and there is no consensus over the question of how to test/evaluate interpretation models that model variation (a particularly interesting version of this problem takes place in dialogues (see projects 5 and 6)). This project will develop such models. It will also develop metrics that do justice to interpretative variation and use these metrics to test models. The development of these metrics will include a cognitive perspective, informed by the type of brain science that has investigated the processing of reference before (Van Berkum et al. 2007).

Objectives: This PhD project aims to deliver:

- Insight into the processing of anaphoric disagreement in the brain
- Soft evaluation metrics for coreference
- Improved computational models of coreference resolution

Area: This PhD project will integrate computational modelling, with contributions from brain science. It will be a collaboration between computing scientists (dr. Gatt, prof. Poesio) and linguists (prof. Winter).

Project 4 Empirical analysis of variation (2). Subjectivity in the detection of problematic language.
[This PhD project is to be funded externally]:

Abstract: Even more than in relation to the *semantics* of language, variation in interpretation is particularly strong in the *pragmatics* of language use, for example when people are asked to judge whether an utterance is metaphoric or humorous (Simpson et al, 2019). This PhD project will focus on topics with a high societal relevance, such as disinformation and offensive/abusive language, where problematic use of language can be harmful to people. Judgments on whether a given utterance is problematic are notoriously subjective, where differences between judges can have difficult cultural, ethnic, and racial overtones (Akhtar et al, 2021; Almanea & Poesio, 2022; Kocon et al, 2022; Leonardelli et al, 2021). The project will develop models for detecting problematic language that take into account the fact that the labels involved can be controversial. It will use

accuracy metrics that take different interpretations of a potentially problematic expression into account (e.g., those developed in P2). Datasets for Arabic and English to study this phenomenon have recently been made available as part of the SEMEVAL 2023 Shared Task on Learning with Disagreement, which Prof. Poesio is co-organizing, but one of the objectives of the project is to work on Dutch, as datasets have recently become available such as DALC (Caselli et al, 2021) although in DALC disagreements are not preserved.
Objectives: This PhD project has both theoretical and practical objectives. In collaboration with media companies, it intends to create a dataset for the Dutch language, in which disagreements are preserved (e.g., an extension of DALC). It intends to answer the following questions:

- Is it possible to associate labels with such a dataset in an automatic fashion, by leveraging the stance of commentators on social media?
- What is the best way for evaluating a computer model of textual phenomena whose interpretation varies based on factors such as age, gender, and religious belief?
- What is the most appropriate method for a media platform to deal with content that some but not all judges consider to be problematic?

Area: This is a Computational Social Science project. Neural methods will be employed to train the models, whose architecture takes the diversity of opinion into account. It will be a collaboration between Computing Science (dr. Nguyen, prof. Poesio) and external partners (see Commitment Plan)

Project 5: Dialogue (1): Conflicting interpretations in dialogue
Abstract: Variation in interpretation becomes explicit in several respects in conversations (in person or online). First, all issues of uncertainty are magnified in a conversation given that language produced under time pressure is typically more uncertain. One of the effects of time pressure is that less attention is paid ensuring that expressions can be interpreted univocally, resulting in misunderstandings that often go undetected (Hirst et al, 1994; McRoy, 1998; Weigand, 1999). The utterances in the dialogue offer an explicit record of what expressions participants disagree upon and/or what type of interpretation they reach jointly, and provide a crucial diagnostic of which of these disagreements matter.

Misunderstandings between dialogue partners cause problems for all aspects of NLP research. The first problem is that specifying that an expression was interpreted in one way by one participant and in another way by the other participant is not possible with present annotation methods for lexical semantics, coreference, and reference with a few exceptions (Poesio et al, 2004). The second problem is that therefore, it is not possible to train models that can produce participant-specific interpretations; current models processing these conversations will incorrectly link together all these interpretations.
Objectives: This PhD project aims to deliver:

- Guidelines and datasets suitable for studying meaning negotiation and misunderstanding in Dutch dialogues, taking into account differences between linguistic subcultures.
- Models of utterance interpretation in Dutch dialogue that are aware of the possibility that participants may not interpret an expression in the same way.
- Spoken dialogue systems able to recognize misunderstandings and able to carry out strategies for repairing them.

Area: This project lies at the interface between corpus linguistics of conversations, NLP (coreference, reference), and conversational agents. Collaboration between Computing Science (dr. Nguyen, prof. Poesio) and Linguistics (prof. Sanders).

Project 6: Dialogue (2): Semantic alignment in dialogue
Abstract: A well-established collection of findings in psycholinguistics tells us that speakers adapt to each other in many ways over the course of a dialogue (Pickering and Garrod 2006). For example, when participants in an experiment decided to call a previously unfamiliar (highly abstract) shape a “violinist”, their dialogue partners tend to henceforth call it a violinist as well (Brennan and Clark 1996). Similarly, if the climate in a room has been described as “cold”, then this usage will tend to set a standard that will be used when other rooms are described as well (i.e., if the temperature in that other room is the same or lower, then this room counts as cold as well, cf. project P1). Although such alignment/entrainment phenomena have been studied experimentally, no formal or computational models have been proposed so far. The aim of this PhD project is to give these ideas a computational expression by means of a chatbot model that aligns with its users in naturalistic ways. The resulting models will be tested experimentally in Turing-test like settings. If successful, this work will lead to theories (i.e., models) of alignment that are far more explicit and detailed than any existing theory of these phenomena.

Objectives: This PhD project aims to deliver:

- Computational models of lexical alignment in dialogue based on existing psycholinguistic findings.
- Experimental evaluation of the resulting models in terms of the naturalness of the resulting dialogues.

Area: This is classic computational modelling, in which computing scientists (profs. Poesio, van Deemter) collaborate with psycholinguists and linguists (prof. van den Bosch).

Management: The PhD projects will all be (co)-supervised by Prof. Poesio in collaboration with other UU faculty. Prof. Poesio has extensive experience managing research teams. Management of the project will be conducted in such a way that broader cross-fertilization between Computing Science and Linguistics will be maximized. We anticipate weekly individual supervision meetings of each PhD student with their supervisors, who will normally hail from different academic disciplines, and we plan regular research presentations at all involved UU departments. Project participants will also take part in the weekly NLP group meetings at Computing and Information Sciences, where research and teaching in NLP are discussed.

8. Appropriateness of research plan with AIREA-NL agenda

The Artificial Intelligence Research Agenda for the Netherlands (AIREA-NL) strongly encourages collaborations across disciplines. In addition to strengthening the Dutch AI community with the arrival of Prof. Poesio, the proposed research will help to bind together a group of researchers who are already at Utrecht University, but spread out over different Faculties (Science and Humanities), strengthening existing ties and creating an enhanced new community of transdisciplinary researchers in Natural Language Processing. At the heart of this new collaboration will be researchers at the Natural Language Processing group at the department of Information and Computing Sciences (Faculty of Science), and researchers at the department of Languages, Literature and Communication (Utrecht Institute of Linguistics, Faculty of Humanities), the latter of which includes a substantial number of social scientists (e.g., psycholinguistics).

According to the Artificial Intelligence Research Agenda for the Netherlands (AIREA-NL) "*Advancing NLP is an AI challenge by itself*". The project addresses the following AIREA-NL research challenges:

(1) "how to deal with the rich variation and cultural differences in language use and communication at the personal and group level in a data efficient manner." Differences in interpretation between people are often socially determined, for example when words have different connotations across different age groups, ethnic groups, and so on. These issues are relevant across the project as a whole and are addressed directly in PhD project P4. Outcomes will enable media companies and others to understand, track, and police problematic language more effectively.

(2) "how to optimise interactive language-based systems in extremely large, non-stationary state and action spaces". NLP datasets are almost invariably extremely large. Additionally, by focussing on the temporal dimension, PhD projects 5 and 6 address the problems posed by non-stationary interpretation in dialogue.

(4) *"how to achieve naturalness in generated speech, responses, and narratives, using persona-based, emotional, and knowledge grounded content generation and understanding."* Choosing words that have the contextually and socially appropriate connotations is a key aspect of achieving naturalness in generated text. This issue is addressed in PhD projects P1, P4, P5.

Furthermore, the project contributes to AIREA-NL's grand challenges, in particular the following Research Questions:

- AI Systems and Humans: *RQ-3.1 "How can humans and AI systems productively interact and understand each other's behaviour in context?"* These issues are studied throughout the project because differences in the interpretation of text are a key obstacle to effective interaction.
- AI Systems & Society: *RQ-4.1. "How do we ensure that everyone benefits from AI? RQ-4.3. How do we design value-sensitive, norm-aware AI systems?"* Avoiding misunderstandings (e.g., Projects P1, P3, and P6) will enable chatbots to use vague and polysemous words in ways that are understood by their clients. An important aspect of being value-sensitive is using and interpreting language in a culturally sensitive way, for example by avoiding language that one's audience is likely to view as impolite (Project 4), and by understanding the cultural implications of a dialogue partner's use of language.

In line with current trends in NLP and with the AINED Focus area "AI for Dutch Language", the project will look at a variety of languages, prominently including Dutch. For example, PhD projects P4 and P5 will focus in part on Dutch, looking at connotations of Dutch words and expressions for different social groups.

9. Literature references

Akhtar, S., Basile, V. & Patti, V. (2020). Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. https://arxiv.org/abs/2106.15896 .

Almanea, D., and Poesio, M. (2022). ArMIS - The Arabic Misogyny Corpus with Annotator Subjective Disagreements. Proc. of Conf. Learning Resources and Evaluation (LREC).

Aroyo, L. & Welty, C. (2015). Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine* 36(1).

Brennan, S.E. and Clark, H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 22: 1482-1493.

Caselli, T., Schelhaas, A., Weultjes, M., Leistra, F., van der Veen, H., Timmerman, G., and Nissim, M. 2021. "DALC: the Dutch Abusive Language Corpus". Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), ACL.

Del Tredici, M. & Fernández R. (2018) The road to success: Assessing the fate of linguistic innovations in online communities. Proc. of Int. Confer. on Computational Linguistics (COLING). 1591-1603

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proc. of North American Chapter of the Association for Computational Linguistics (NAACL), pages 4171–4186.

Douven, I., Decock, L., Dietz, R., and Égré P.(2013). Vagueness: A conceptual spaces approach. *Journal of Philosophical Logic* 42 (1), 137-160.

Edgington, D. (1997). Vagueness by Degrees. In R. Keefe & P. Smith (eds.), *Vagueness: A Reader*. MIT Press (1997).

Hirst, G., McRoy, S., Heeman, P., Edmonds, P. and Horton D. (1994). Repairing conversational misunderstandings and non-understandings. *Speech Communication*, 15 (1994), 213-230

Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T. & Kazienko, P. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5).

Leonardelli, E., Menini, S., Palmero Aprosio, A., Guerini, M., & Tonelli, S. (2021). Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. Proc. of Empirical Methods in Natural Language Processing (EMNLP).

McRoy, S.W. (1998). Preface - Detecting, repairing and preventing Human-Machine Miscommunication. *International Journal of Human-Computer Studies*, 48: 547-552.

Passonneau, R.J., Bhardwaj, V., Salleb-Aouissi, A., and Ide, N. (2012). Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):219–252.

Pavlick, E. & Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *Proc. of NAACL*.

Pickering, M. and Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation* 4: 203-228.

Plank, B., Hovy, D., and Sogaard, A. (2014). Linguistically debatable or just plain wrong? Proc. European Chapter of the Association for Computational Linguistics (EACL).

Poesio, M., Delmonte, R., Bristot, A., Chiran, L., and Tonelli, S. (2004). The VENEX corpus of anaphora and deixis in spoken and written Italian. University of Venice report.

Poesio, M., and Artstein, R. (2005). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. Proc. of ACL Workshop on Frontiers in Corpus Annotation, p.76–83.

Portet, F., Reiter, E, Gatt, A, Hunter, J., Sripada, S., Freer, Y., and Sykes, C. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence* 173 (7-8), p.789-816.

M. Recasens, E. Hovy, and M. Antonia Marti (2011). Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Reidsma, D. & Carletta, J. (2008). Reliability measurement without limits. *Computational Linguistics* 34, 319–326. doi: 10.1162/coli.2008.34.3.319

Schlechtweg, D., Schulte im Walde, S., and Eckmann, S. (2018). Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. Proc. of North American Chapter of the Association for Computational Linguistics (NAACL), 169–174

Simpson, E., Do Dinh, E., Miller, T. and Gurevych, I. (2019). Predicting humorousness and metaphor novelty with Gaussian process preference learning. Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, 5716–5728.

Trott, S. and Bergen, B. (2021). RAW-C: Relatedness of ambiguous words in context (a new lexical resource for English). Proc. of the Association for Computational Linguistics (ACL), 7077–7087.

Uma, A., Fornaciari, T., Dumitrache, A., Miller, T., Chamberlain, J., Plank, B., Simpson, E., Poesio, M. (2021). Learning with Disagreements. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021).

Van Berkum, J.J.A. Koornneef, A.W., Otten, M., and Nieuwland, M.S. (2007). Establishing reference in language comprehension: An electrophysiological perspective. Brain research 1146, 158-171.

van Deemter, K. (2010) *Not Exactly: in Praise of Vagueness*. Oxford University Press.

Weigand, E. (1999). Misunderstanding: The standard case. *Journal of Pragmatics*, 31 (6), 763-785