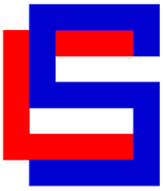
PaM 2020

Proceedings of the Conference on Probability and Meaning

Christine Howes, Stergios Chatzikyriakidis, Adam Ek and Vidya Somashekarappa (eds.)

Gothenburg and online
14–15 October 2020



Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

Introduction

This volume contains the papers presented at the CLASP midterm conference, Probability and Meaning (PaM2020) at the Department of Philosophy, Linguistics and Theory of Science (FLoV), University of Gothenburg, held on October 14-15th, 2020.

PaM brings together researchers interested in computationally relevant probabilistic approaches to natural language semantics and includes symbolic, machine learning and experimental approaches to this task, as well as hybrid models.

Papers were invited on topics in these and closely related areas, including (but not limited to) probabilistic approaches developed within a computational framework, the semantics of natural language for written, spoken, or multimodal communication, probabilistic type theoretic approaches to meaning, multimodal and grounded approaches to computing meaning, dialogue modelling and linguistic interaction, deep learning approaches and probability, syntax-semantics interface, alternative approaches to compositional semantics, inference systems for computational semantics, recognising textual entailment, semantic learning, computational aspects of lexical semantics, semantics and ontologies, semantic aspects of language generation, semantics-pragmatics interface.

This conference aims to initiate a genuine discussion between these related areas and to examine different approaches from computational, linguistic and psychological perspectives and how these can inform each other. It features 4 invited talks by leading researchers in these areas, and 18 peer-reviewed papers, 11 presented as long talks and 7 presented as posters.

We would like to thank all our contributors and programme committee members, with special thanks to CLASP for organising the virtual conference and our sponsors SIGSEM http://sigsem.org, the ACL special interest group on semantics, and the Swedish Research Council (Vetanskapprådet) for funding CLASP.

Christine Howes, Stergios Chatzikyriakidis, Adam Ek and Vidya Somashekarappa

Gothenburg

October 2020

Program Committee:

Rodrigo Agerri	University of the Basque Country
Anja Belz	University of Brighton
Emily M. Bender	University of Washington
Raffaella Bernardi	University of Trento
Jean-Philippe Bernardy	University of Gothenburg
Johan Bos	University of Groningen
Ellen Breitholtz	University of Gothenburg
Harry Bunt	Tilburg University
Aljoscha Burchardt	German Research Center for Artificial Intelligence (DFKI GmbH)
Nicoletta Calzolari	National Research Council of Italy
Rui Chaves	University of Buffalo
Alexander Clark	King's College London
Stephen Clark	Queen Mary University of London
Ariel Cohen	Ben-Gurion University of the Negev
Robin Cooper	University of Gothenburg
Philippe de Groote	Inria
Leon Derczynski	IT University of Copenhagen
Markus Egg	Humboldt University
Adam Ek	University of Gothenburg
Katrin Erk	University of Texas at Austin
Arash Eshghi	Heriot-Watt University
Jonathan Ginzburg	Université Paris-Diderot
Julian Hough	Queen Mary University of London
Elisabetta Jezek	University of Pavia
Richard Johansson	Chalmers University of Technology
John Kelleher	Technological University Dublin
Ralf Klabunde	Ruhr-Universität Bochum
Emiel Krahmer	Tilburg University
Shalom Lappin	University of Gothenburg
Staffan Larsson	University of Gothenburg
Vladislav Maraev	University of Gothenburg
Paul McKevitt	Ulster University
Louise McNally	Universitat Pompeu Fabra
Marie-Francine Moens	KU Leuven
Shashi Narayan	Google
Joakim Nivre	Uppsala University
Bill Noble	University of Gothenburg
Denis Paperno	Loria, CNRS
Anselmo Peñas	Universidad Nacional de Educación a Distancia
Manfred Pinkal	Saarland University
Massimo Poesio	Queen Mary University
Violaine Prince	University of Montpellier
Stephen Pulman	Oxford University
Matthew Purver	Queen Mary University of London
James Pustejovsky	Brandeis University

Program Committee (continued):

Christian Retoré	University of Montpellier
German Rigau	University of the Basque Country
Hannah Rohde	University of Edinburgh
Mehrnoosh Sadrzadeh	University College London
Asad Sayeed	University of Gothenburg
David Schlangen	University of Potsdam
Sabine Schulte im Walde	University of Stuttgart
Vidya Somashekarappa	University of Gothenburg
Tim Van de Cruys	University of Toulouse
Eva Maria Vecchi	University of Cambridge
Carl Vogel	Trinity College Dublin

Invited Speakers:

Heather Burnett, Laboratoire de Linguistique Formelle, Université de Paris
Stephen Clark, Department of Computer Science and Technology, University of Cambridge
Katrin Erk, Linguistics Department, University of Texas at Austin
Noah Goodman, Departments of Computer Science and Psychology, Stanford, CA

Invited talk 1: Heather Burnett

Social Signaling and Reasoning under Uncertainty: French “Écriture Inclusive”

Gender inclusive writing (“écriture inclusive" EI) has long been the topic of public debates in France. Examples of EI for the word “students" are shown in (1).

(1) a. étudiant·e·s (point médian)
b. étudiant.e.s (period)
c. étudiants et étudiantes (repetition)
d. étudiant(e)s (parentheses)
e. étudiant-e-s (dash)
f. étudiantEs (capital)
g. étudiant/e/s (slash)
h. étudiant- -e- -s (double dash)

These debates have amplified since the Macron government prohibited the use of the point médian (1a) in official documents in 2017 (Abbou et al. 2018). In addition to being a point of disagreement between feminists and anti-feminists, EI is also controversial among feminists: it has many variants (1), who often disagree on which variant should be used (Abbou 2017). In this talk, I argue that the source of many of these disagreements lies in the fact that French écriture inclusive has developed into a rich social signalling system: based on a quantitative study of EI in Parisian university brochures (joint work with Céline Pozniak (Burnett & Pozniak 2020)), I argue that writers use or avoid EI in part in order to communicate aspects of their political orientations. We show that these aspects involve writers’ perspectives on gender, but also stances on issues unrelated to gender, such as (anti)institutionalism and support for the Macron government. I then outline a research programme for studying this signalling system from a formal perspective: following Burnett (2019), I show how we can use probabilistic pragmatics to analyze EI’s contribution to writers’ political identity construction and the consequences that this has for its use as a tool for promoting gender equality and social change.

Invited talk 2: Katrin Erk

How to marry a star: Probabilistic constraints for meaning in context

Context has a large influence on word meaning; not only local context, like in the combination of a predicate and its argument, but also global topical context. In computational models, this is routinely factored in, but the question of how to integrate different context influences is still open for theoretical accounts of sentence meaning. We start from Fillmore’s "semantics of understanding", where he argues that listeners expand on the "blueprint" that is the original utterance, imagining the utterance situation by using all their knowledge about words and the world. We formalize this idea as a two-tier "situation description system" that integrates referential and conceptual representations of meaning.

A situation description system is a Bayesian generative model that takes utterance understanding to be the mental process of probabilistically describing one or more situations that would make a speaker’s utterance logically true, from the point of view of the listener.

Invited talk 3: Stephen Clark

Grounded Language Learning in Virtual Environments

Natural Language Processing is currently dominated by the application of text-based language models such as BERT and GPT. One feature of these models is that they rely entirely on the statistics of text, without making any connection to the world, which raises the interesting question of whether such models could ever properly "understand" the language. One way in which these models can be grounded is to connect them to images or videos, for example by conditioning the language models on visual input and using them for captioning. In this talk I extend the grounding idea to a simulated virtual world: an environment which an agent can perceive, explore and interact with. More specifically, a neural-network-based agent is trained – using distributed deep reinforcement learning – to associate words and phrases with things that it learns to see and do in the virtual world. The world is 3D, built in Unity, and contains recognisable objects, including some from the ShapeNet repository of assets.

One of the difficulties in training such networks is that they have a tendency to overfit to their training data, so first we'll demonstrate how the interactive, first-person perspective of an agent provides it with a particular inductive bias that helps it to generalize to out-of-distribution settings. Another difficulty is that training the agents typically requires a huge number of training examples, so we'll show how meta-learning can be used to teach the agents to bind words to objects in a one-shot setting. Moreover, the agent is able to combine its knowledge of words obtained one-shot with its stable knowledge of word meanings learned over many episodes, providing a form of grounded language learning which is both "fast and slow".

Joint work with Felix Hill.

Invited talk 4: Noah Goodman

Reference, Inference, and Learning

A key function of human language is reference to objects and situations. Referential language grounds in stable semantic conventions, but flexibly depends on context. In this talk I will explore the computational mechanisms of referential language in the setting of language games. I will argue that many patterns of behavioral data can be explained by a combination of hierarchical learning for semantics – realized with the tools of deep neural networks – and recursive social reasoning for pragmatics – realized in the Bayesian rational speech acts (RSA) framework. I will consider phenomena of redundancy in reference, grounding semantics in vision, and adaptation under repeated interaction. Finally I will address a key puzzle for RSA (and other neo-Gricean theories): how can production be so quick and effortless if it depends on complex recursive reasoning?

Table of Contents

Conference Program

October 14th

Session 1

'Practical', if that's the word
Eimear Maguire

Personae under uncertainty: The case of topoi
Bill Noble, Ellen Breitholz and Robin Cooper

Dogwhistles as Identity-based interpretative variation
Quentin Dénigot and Heather Burnett

Session 2

Conditional answers and the role of probabilistic epistemic representations
Jos Tellings

Linguistic interpretation as inference under argument system uncertainty: the case of epistemic must
Brandon Waldon

Session 3

Linguists Who Use Probabilistic Models Love Them: Quantification in Functional Distributional Semantics
Guy Emerson

Fast visual grounding in interaction: bringing few-shot learning with neural networks to an interactive robot
José Miguel Cano Santín, Simon Dobnik and Mehdi Ghanimifard

October 15th

Poster session

Discrete and Probabilistic Classifier-based Semantics
Staffan Larsson

Social Meaning in Repeated Interactions
Elin McCready and Robert Henderson

Towards functional, agent-based models of dogwhistle communication
Robert Henderson and Elin McCready

Stochastic Frames
Annika Schuster, Corina Stroessner, Peter Sutton and Henk Zeevat

A toy distributional model for fuzzy generalised quantifiers
Mehrnoosh Sadrzadeh and Gijs Wijnholds

Generating Lexical Representations of Frames using Lexical Substitution
Saba Anwar, Artem Shelmanov, Alexander Panchenko and Chris Biemann

Informativity in Image Captions vs. Referring Expressions
Elizabeth Coppock, Danielle Dionne, Nathanial Graham, Elias Ganem, Shijie Zhao, Shawn Lin, Wenxing Liu and Derry Wijaya

October 15th (continued)

Session 4

How does Punctuation Affect Neural Models in Natural Language Inference
Adam Ek, Jean-Philippe Bernardy and Stergios Chatzikyriakidis

Building a Swedish Question-Answering Model
Hannes von Essen and Daniel Hesslow

Session 5

Word Sense Distance in Human Similarity Judgements and Contextualised Word Embeddings
Janosch Haber and Massimo Poesio

Short-term Semantic Shifts and their Relation to Frequency Change
Anna Marakasova and Julia Neidhardt

'Practical', if that's the word

Eimear Maguire
Laboratoire Linguistique Formelle (UMR 7110), Université de Paris
eimear.maguire@etu.univ-paris-diderot.fr

Abstract

Certain conditionals have something other than a clause as their consequent: their antecedent *if*-clauses are 'adverbial clauses' without a verb. We argue that they function in a way already seen for those with clausal consequents, despite lacking the content we might expect for the formation of a conditional. The use of the *if*-clause with sub-clausal consequents is feasible thanks to the fact that this function does not depend on the consequent content, and so is not impeded when the consequent does not provide a proposition, question or imperative. To support this we provide meaning rules for conditionals in terms of information state updates, letting the same construction play out in different ways depending on context and content.

1 Introduction

Biscuit conditionals are a subset of conditionals well-discussed for their deviance from typical hypothetical conditionals in terms of truth conditions, acceptability, and information conveyed. Within biscuit conditionals, there is a further metalinguistic subset like (1)[1] which are used to manage communication more directly:

(1) **Looks a bit lethargic** if you ask me. (*KP4 235*)

(2) we '**advertised**' it if that's the right term to the people at large that we were looking to acquire businesses (*ICE-GB S1B-065 078*)

In (1) the *if*-clause relates to *looks a bit lethargic*. Intuitively, (2)[2] does something quite similar, but the sub-utterance *advertised* is at least as intuitive a 'consequent' as the entire clause, despite being sub-propositional. (2) is an example of *if*-clauses used to 'condition' sub-clausal segments.

We will refer to this particular subset as *lexical hedges*, to distinguish them from metalinguistic hedges like (1) more generally. We distinguish them from the more general class by how they target a particular phrase or lexical item within the utterance rather than the whole sentential unit. This is the contrast between (1) and (2): the first targets the entire statement *Looks a bit lethargic*, while the second targets *advertised*.

They are not hypothetical conditionals, but given the form of their consequent neither are they trivially the same as other metalinguistic conditionals. To address this, we will demonstrate that they are the replication of a function observable for other metalinguistic conditionals, combined with incremental processing and a heterogeneous utterance representation. We first examine some characteristics of lexical hedge *if*-clauses in support of the argument that they lack a main clause consequent. We then identify a range of antecedent-consequent connections among *if*-constructions at increasing abstraction. Distinct information state update rules are provided for these, identifying the lexical hedge use as an extension of a function already found among conditionals with clausal consequents.

2 *If*-conditionals and internal coherence

There are competing approaches to the analysis of biscuit conditionals. One branch attempts to explain the differences from hypothetical conditionals through a fundamental semantic distinction (Iatridou, 1991; Siegel, 2006). This usually incorporates a version of the Performative Hypothesis (Ross, 1967), whereby the performance of speech acts is a part of clause structure. This prefix is at a different level in the structure depending on the

[1] British National Corpus via *SCoRE* (Purver, 2001)

[2] All examples from ICE-GB are cited via the data released at http://www.chiheelder.com/?attachment_id=144

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 1–7

type of conditional: for biscuit conditionals, this could be glossed as *if you are hungry, [I assert that] there are biscuits on the sideboard*. Biscuit conditionals are often called *speech-act conditionals* in reference to the intuition that they condition speech acts (or some aspect thereof) directly.

Speas and Tenny (2003) in particular have returned an updated edition of this theory of speech acts to more mainstream thinking, but we will not be taking such a directly syntactic approach here. Rather than a component of syntactic structure, we follow Ginzburg (2012a) in identifying illocutionary force as part of the semantics of certain lexical items, phrases and clause types, reflecting the action a speaker of the utterance believes themselves to have performed in uttering it.

The second group of approaches maintains that the differences between hypothetical and biscuit conditionals can be explained through pragmatic means (Franke, 2007; Biezma and Goebel, 2019). A pragmatic approach to biscuit conditionals is taken here: if possible, it is preferable to handle the differences through general principles and a unified analysis, rather than developing a semantic split. We will return to this when introducing the model used later.

Metalinguistic conditionals, including lexical hedges, are sometimes discussed as a subset of biscuit conditionals, since they too lack the intuitive link between the antecedent and consequent case found in hypothetical conditionals. Declerck and Reed (2001) recognise 'metalinguistic-P conditionals', which make a comment on "the form of the Q-clause [consequent-clause], on the choice of words in it or on the pronunciation of a word", but do not particularly propose an analysis, or very clearly distinguish them from 'speech condition-defining-P conditionals'. Elder (2015) makes a specific corpus case study of *if you like*, but otherwise classes lexical hedges amongst other metalinguistic conditionals which function as an 'illocutionary force hedge', while Quirk et al. (1985) include them among the class of other metalinguistic comments rather than discussing them in the context of other conditionals.

Dancygier (1992) distinguishes 'metatextual' conditionals (e.g. both (1) and (2)), from 'speech-act' conditionals (i.e. standard biscuit conditionals), although discussing the whole clause as the "consequent" in both cases. Dancygier's copious use of scare quotes indicates discomfort with assessing the entire clause as consequent, but the alternative that the consequent is the "focus" itself rather than the clause, is not explored. However, we consider the analysis of a comment on a single word as re-setting the entire utterance as a conditional to be unappealing, as will be briefly discussed in Section 2.1.

2.1 *If*-clause lexical hedges: features

The examples in this section were found via two sources: (i) a sample of 800 non-embedded *if*-clauses from the spoken data section of the BNC, where those associated with a non-clausal consequent were reviewed to identify those acting as lexical hedges; and (ii) among corpus study data from Elder (2015), found by reviewing the *if*-clauses classified as *Illocutionary Force Hedges*, where they were included among that class.

These *if*-clauses tend to appear adjacent to the hedged sub-utterance rather than at the beginning of the clause. Among the 41 examples identified, all but four have the *if*-clause directly adjacent to the focus-word or focus-phrase.

Lexical hedges can be contrasted with 'genuine' elliptical consequents. Where the consequent is sub-clausal, like in (3), its role in context may not be as a proposition:

(3) climate is just a little '**transient part**' *if you like* in this process *(ICE-GB S2A-043 044)*

(4) and then cut some bacon up, put that in saucepan just let it brown a bit [...] in a bit of fat, er soften onions, then put mince in, brown mince [...] erm **a bit of garlic** *if you like garlic (KB2 359–363)*

In (4) *a bit of garlic* essentially functions as an imperative on the basis of the previous instruction to "put mince in", and given the context could be expanded in interpretation to something like *put a bit of garlic in*. Unlike (4), for the first example to be elliptical we would need to posit the existence of an implicit clause that has no evidence anywhere elsewhere in the utterance. Either we treat the entire clause as the consequent, insist on a 'covert' conditional, or accept that *transient part* functions as a consequent item in its own right.

Moving these *if*-clauses to the clause boundary changes their interpretation. Consider (5):

(5) I'm sure you could all add to that list of kind of 'symptoms' **if you like** of waste

and inefficiency in organised society *(ICE-GB S2A-049 016)*

If the *if*-clause were fully pre-posed (*If you like, I'm sure...*) it would be interpreted as hedging the entire clause, not just *symptoms*. Given the preference for placing the *if*-clause adjacent to the target segment and the difference in interpretation, analysing the whole clause as consequent in (5) does not seem advantageous.

Nevertheless, we may consider (6):

(6) Is, is the a ⟨pause⟩ a danger Geoffrey Hoskin that the instability in the Soviet Union, if one can still call it, a Union, could affect us, could spill out across its borders? (*KJS 23*)

If this *if*-clause were external (*if one can still call it a Union, is there a danger...*), the potential issue with *union* would remain identifiable thanks to its explicit mention. In such a case it would be more reasonable to interpret the *if*-clause as associated with the whole utterance, akin to a modified (5). In contrast to (5) and its modification, the overall effect remains essentially the same as in the original, as a fault in a specific component of the utterance creates a fault in the utterance as a whole.

However, we should not over-generalise this. Re-simplifying the adjacent lexical hedge uses as therefore being conditions on the entire surrounding clause, on the grounds that full utterances are hedged in other cases, would be attempting to find the shared features in the wrong place. We can do better by recognising that *if*-clauses can be used to perform the same function at different levels – respecting both the similarity to self-repair of this particular use, and desire for a consistent analysis across *if*-clause uses.

2.2 Multi-purpose *if*-clauses

Acceptable use of a conditional generally requires some 'meaningful' link between antecedent and consequent (Douven, 2008; Skovgaard-Olsen et al., 2016). Biscuit conditionals are infamously a case where this fails, and we see this connection as something other than between the situations themselves, e.g. relevance of the consequent content.

The proposal that follows is based on the idea that for the utterance of a conditional to be acceptable, it must be possible to identify some 'meaningful' link between consequent and antecedent. On this basis we will walk through an increasing mentalisation around what this link is found to be, from a link between the antecedent and consequent cases, to between the antecedent case and some predication on the consequent content, and finally between the antecedent case and some predication on the consequent utterance or non-content aspect thereof. As the final case involves predication on a non-content aspect of the utterance, it can be applied to elements which do not provide a main clause or perform a dialogue move in their own right.

2.2.1 Model Set-up

As our formal framework we use Type Theory with Records (hereafter TTR) (Cooper, 2005, 2012; Cooper and Ginzburg, 2015), a model-theoretic rich type theory. A key notion in TTR is that of judgement, with $a : T$ indicating that object a is judged to be of type T. A record is a set of fields each with a label and a value: $a = v$ signifies that the value in field a is the object v. A record type is a set of labels and types such as a and T. Records can be judged to be of some record type on the basis of whether the values in the record's fields are of the type specified for the same labels in the record type, e.g. whether v is of type T.

To identify problems with the utterance itself, we need to engage with it as a whole rather than focusing directly on the semantic content. We use Ginzburg's (2012b) notion of a *Locutionary Proposition*, an utterance represented by its speech event and the classification of said event (e.g. that it is the utterance of a particular sentence). The notion of a proposition used here is an Austinian proposition, true or false depending on whether the situation in question is indeed of the given situation type. A very minimal example is given in (7): note the recognition of features for semantic content and phonology, and requirements on the context (a situation, speaker and location).

(7) *LocProp* for "I am here":

$$\left[\begin{array}{l} \text{sit} = u_0 \\ \text{sit-type} = \\ \left[\begin{array}{l} \text{phon} : \text{I am here} \\ \text{cat} = \text{V[+fin]} : \text{PoS} \\ \text{constits} = \{\text{I, am, here}\} : \text{set(sign)} \\ \text{dgb-params} : \left[\begin{array}{l} \text{spkr} : \text{Ind} \\ \text{l} : \text{Loc} \\ s_0 : \text{sit} \end{array}\right] \\ \text{cont} = \text{Assert(dgb-params.spkr, dgb-params.addr,} \\ \quad \left[\begin{array}{l} \text{sit} = s_0 \\ \text{sit-type} = \left[\begin{array}{l} c_0 : \text{in(dgb-params.spkr,} \\ \quad\quad \text{dgb-params.l)} \end{array}\right] \end{array}\right] : \text{IllocProp} \end{array}\right] \end{array}\right]$$

The type in *sit-type* can be composed with the help of linguistic resources known by the agent: in this case the types for a declarative clause and for the three lexical items used. We will use a and c to refer to the locutionary propositions of the antecedent and consequent. We will also reference their content X via the shorthand a.ct and c.ct, following the path c.sit-type.cont = Move(spkr, addr, X). In the case c is used to assert a proposition, like (7), this will be a path to the proposition.

We characterise the 'meaningful link' between antecedent and consequent as a question and satisfactory response following Biezma and Goebel (2019), most commonly that the consequent content satisfies an antecedent-based question[3] *if a.ct, what?* as follows:

(8) satisfy(c.ct, λx.if(a.ct,x))[4]

We use the *satisfy* relationship in (8) to indicate acceptability as a resolving answer as per Ginzburg (2012b): a potentially resolving answer which enables some desired outcome to be fulfilled.[5]

We set up a minimal dialogue representation as follows, based on the KoS framework from Ginzburg (2012b): the dialogue gameboard represents a single agent's understanding of the dialogue state at a given point, and tracks current questions under discussion (*QUD*), conversation history (*Moves*), as-yet-ungrounded utterances (*Pending*), and the common ground (*Facts*). We may define update rules for the gameboard as pairs of state types: the precondition on the state the board must be in for the rule to be applied, and the effects of applying the rule. For space, we will indicate the latest move by its content.

2.2.2 How to handle an *if*-conditional

The kinds of questions discussed in Section 2.2 can be partitioned into three groups, and following (8) (repeated here) can be represented as follows:

(9) a. *Content-based (simple):*
satisfy(c.ct, λx.if(a.ct, x))

b. *Content-based (complex):*
satisfy(c.ct, λx.if(a.ct, $f(x)$)),
for some predicate f

c. *Utterance-based:*
satisfy(c, λx.if(a.ct, $f(x)$), or
satisfy($c.z$, λx.if(a.ct, $f(x)$),
for some path z in c and f as above

The second and third groups could be merged – content is one of the fields of the consequent, and is therefore covered by the 'utterance-based' category. However, we treat it separately as it is the linking case: the surrounding question has been made more complex, but the required element is still the same as in the content case. The third group is essentially a generalisation beyond the content to other aspects, and does not need to involve the content of the consequent at all. In the rest of this section, we will see these play out in different ways: the first for hypothetical conditionals, the second for typical biscuit conditionals, and the third initially for metalinguistic uses on full-clause consequents, then on sub-clausal units.

The idea that at least some biscuit conditionals provide a condition on the felicity of the consequent has been frequently raised (e.g. Sweetser, 1990), and this targeting of felicity conditions is indeed something which naturally scales down from complete utterances to term choice. When we progress to explicitly metalinguistic cases, we will use *groundability* as a general predicate.

The dialogue state update for hypothetical conditionals can now be given as follows:[6]

(10) if she disappeared I'd be worried all time *(KB1 527)*

(11)
$$\left[\begin{array}{l}\text{pre}: \\ \quad \left[\begin{array}{l}\text{LatestMove} = \\ \quad \text{Assert(spkr, if}(a.\text{ct}, c.\text{ct})) : \text{LocProp} \\ \quad c_q : \text{satisfy}(c.\text{ct}, \lambda x.\text{if}(a.\text{ct}, x)) \\ \quad \text{QUD} = [?\text{if}(a.\text{ct}, c.\text{ct}) \mid \text{rest}] : \text{poset(Question)}\end{array}\right] \\ \text{effects}: \\ \quad \left[\begin{array}{l}\text{QUD} = \text{pre.QUD.rest} : \text{poset(Question)} \\ \text{Facts} = \text{pre.Facts} \cup \text{if}(a.\text{ct}, c.\text{ct})\end{array}\right]\end{array}\right]$$

The speaker has asserted the conditional *if a, c*, and the agent finds it satisfies the constraint that the consequent content is a satisfactory answer to the simplest of the three question types. A related

[3] Conditionals have also been proposed to express functions from situations of the antecedent to those of the consequent (see Cooper). Some relation as a core meaning is worth consideration, given a connection is evidently leveraged for non-hypothetical cases. For now however, we content ourselves with the coherence constraint on the connection.

[4] This simple gloss is used for the content of a conditional to let us reference the antecedent and consequent easily.

[5] e.g. when asking *Where are they going?* to learn a destination, a potentially resolving answer either provides a location or states that there is no such destination: a resolving answer will provide the actual destination.

[6] Satisfaction of the question is expected to have other effects e.g. inferring support for a 'meaningful link' in the form of a new topos if one was not already known (see e.g. (Breitholtz, 2014) for discussion of the role of topoi in identifying non-logical connections in dialogue).

issue is raised to QUD, as an agent may accept or reject the assertion's content. This is a general rule for assertions, and means that explicit agreement or disagreement with the most recent assertion will be coherent with respect to the QUD. For the example, we could gloss this as *is it so that if she disappeared, spkr would be worried all the time?*. In enacting the rule above, the assertion is accepted, adding it to *Facts* and removing the now-resolved issue of *?if(a.ct, c.ct)* from *QUD*.

Although we may also include a general rule that any assertion should address a question on *QUD*, we require a connection specifically between antecedent and consequent. They should still be considered with respect to each other if separated by distance, as in the retrospective addition of an *if*-clause to a speaker's own assertion or to that of another speaker. Antecedent-consequent coherence is required even when the original assertion is already recognisable as addressing another live issue.

Where the consequent content fails as a resolving answer to the direct content-based question, as in the case of biscuit conditionals, we must reevaluate the question with respect to other potential relationships between antecedent and consequent.

The rule given in (13) is for this case, and specifically where we can additionally determine that the consequent holds outside of the conditional. It is commonly noted that biscuit conditionals are used to convey their consequent, but this is not always so. Compare "If you want a huge lie, G.W. Bush and Condoleezza Rice are married" (from Siegel (2006)) and "If you want a huge lie, there are political leaflets on the table": it takes further reasoning to determine whether the consequent is itself the sought-after lie, or true information which will help the addressee to find one (e.g. pre-existing knowledge that the consequent is false, a topos that politicians are dishonest). The specifics of the predicate in the question will depend on a more complex combination of lexical content, reasoning, and recognition of utterance goals than we can hope to approach here: we fall back on some notion of *relevance*, a general case associated strongly enough with biscuit conditionals that it is one of their alternative names.

(12) you can put carrots in it if you want *(KB4 206)*

(13)
$$\left[\begin{array}{l} \text{pre}: \left[\begin{array}{l} \text{LatestMove} = \\ \quad \text{Assert(spkr, if}(a.\text{ct}, c.\text{ct}))\rangle : \text{LocProp} \\ \text{c}_{q1} : \neg\text{satisfy}(c.\text{ct}, \lambda x.\text{if}(a.\text{ct}, x)) \\ \text{c}_{q2} : \text{satisfy}(c.\text{ct}, \lambda x.\text{if}(a.\text{ct}, \text{rel}(x))) \\ \text{QUD} = [?c.\text{ct} \mid \text{rest}] : \text{poset(Question)} \end{array}\right] \\ \text{effects}: \left[\begin{array}{l} \text{QUD} = \text{pre.QUD.rest} : \text{poset(Question)} \\ \text{Facts} = \text{pre.Facts} \cup c.\text{ct} \end{array}\right] \end{array}\right]$$

This time, the consequent content does not provide a satisfactory answer to the most direct *if*-based question, which we may gloss for (12) as *if you want, what?* However, the antecedent and consequent cases can be related by including predication in the question – the second case on our list at the beginning of this subsection. Re-framing it, we might identify the potential issue resolved by the consequent as *if you want, what is relevant?* – that the addressee *can* make the wanted addition to the recipe.

Given that we have failed to draw a direct relation between the antecedent and consequent cases, this conditional can be treated as a vehicle for conveying the consequent (having already said that we are dealing with the set of biscuit conditionals where the consequent holds). The relevant proposition raised to *QUD* for potential acceptance is simply the consequent, rather than the entire conditional. We can still compose an asserted conditional, but a link is no longer identified directly between the antecedent and consequent cases themselves: the explicit conditional content is a side-effect produced in pursuit of the actual purpose of the utterance, and not necessarily worth keeping.

The third question set described in (9) also uses predication to relate *if*-clause and consequent, but beyond taking communicative issues into account for the content, it deals with the consequent utterance. We may predicate on an aspect of the utterance other than content, or on the utterance itself. Rather than managing information, this usage manages the groundability of the consequent utterance itself, the content of the *if*-clause flagging a potential issue with the consequent utterance (e.g. appropriateness). We are no longer interpreting a grounded dialogue move, but evaluating one still pending.

(14) if I might say so disabled people were treated oddly in those days *(HDM 275)*

(15)
$$\left[\begin{array}{ll} \text{pre}: & \left[\begin{array}{l} \text{LatestPending} = \\ \quad \text{Assert(spkr, if}(a.\text{ct}, c.\text{ct})) : \text{LocProp} \\ \text{c}_{q1} : \neg\text{satisfy}(c.\text{ct}, \lambda x.\text{if}(a.\text{ct}, x)) \\ \text{c}_{q2} : \text{satisfy}(c, \lambda x.\text{if}(a.\text{ct}, \text{grndble}(x))) \\ \text{QUD} = [?a.\text{ct} \mid \text{rest}] : \text{poset(Question)} \end{array}\right] \\ \text{effects}: & \left[\begin{array}{l} \text{Facts} = \text{pre.Facts} \cup a.\text{ct} \\ \text{LatestMove} = \text{Assert}(c.\text{ct}) : \text{LocProp} \\ \text{QUD} = [?c.\text{ct} \mid \text{pre.QUD.rest}] : \text{poset(Question)} \end{array}\right] \end{array}\right]$$

The speaker has (provisionally) asserted a conditional for which the agent can interpret the *if*-case as making the consequent utterance groundable, guided by the content of the antecedent, as in the mention of speaking in (14). If the antecedent is not the case, there is a problem with the pending utterance.

In accepting the antecedent case, removing it from *QUD* and adding it to *Facts*, the consequent utterance is affirmed as groundable. The consequent itself can now be treated as an ordinary move: the result-state of this rule is similar to one where an assertion equivalent to the consequent had been made, and the active issue is whether to accept it: *is it so that disabled people were treated oddly in those days?*

Having predicated on something other than the content, we can do the same for items with content that cannot be used to compose a conditional. These should be considered in relation to repair behaviours: the *if*-clause flags a potential reparandum within an utterance, which in the non-*if*-clause case cannot or should not be grounded. In (9) the utterance-based questions were divided into two subtypes: those associated with the entire targeted utterance, and those associated with an aspect of the targeted utterance e.g. its phonetic realisation (e.g. *if I'm saying that right*). Here we use an example that does not take issue with a specific aspect of the target word.

(16) there are two principles if you like in the theological field *(F86 211)*

(17)
$$\left[\begin{array}{ll} \text{pre}: & \left[\begin{array}{l} \text{LatestPending} : \textit{pend} \mid \textit{if-cond} \\ \text{c}_q : \text{satisfy}(c, \lambda x.\text{if}(a.\text{ct}, \text{grndble}(x))) \\ \text{QUD} = [?a.\text{ct} \mid \text{rest}] : \text{poset(Question)} \end{array}\right] \\ \text{effects}: & \left[\begin{array}{l} \text{LPend} = \text{pre.LPend}.\textit{pend} \\ \qquad \mid \text{pre.Lpend}.\textit{if-cond}.c : \text{LocProp} \\ \text{Facts} = \text{pre.Facts} \cup a.\text{ct} \\ \text{QUD} = \text{pre.QUD.rest} : \text{poset(Question)} \end{array}\right] \end{array}\right]$$

The previous full-clause case flagged a potential issue with an otherwise standalone utterance: these perform the same function for a sub-utterance. If the flagged element is ungroundable it will require repair. If it is groundable, as here, it can be integrated into the larger utterance as usual, and the flag dismissed.

In the *effects* of both (15) and (17), the *if*-case has been included in the common ground *Facts*. However, in (15) this enables us to recognise that the intended assertion is of the consequent, and its acceptance becomes the active issue. In (17) the utterance is not yet complete: although the pending contribution of the 'conditional' can again be re-evaluated as the consequent only (in this case, incorporating the flagged phrase into the overall pending utterance), a new conversational move has not yet been completed, and there is not yet a new asserted proposition to stage via *QUD*.

3 Conclusion

The most common semantics for conditionals is founded on *if* as restricting the scope of another (potentially covert) operator (Kratzer, 1986). However, we have seen that *if*-clauses may be associated with only a constituent, and a conditional effect introduced by *if* can evidently arise without participating in a process via a main clause.

Although we cannot provide extended discussion, the above provides more motivation for seeking another way to replicate the restrictor theory's advantages. One appeal of the restrictor theory is that it evades proofs by Lewis (1975) showing that one could have a semantics for conditionals that reflects the intuitive (and eventually empirically verified) judgement that a conditional is as probable as its consequent given its antecedent (Stalnaker, 1970), or a semantics whereby conditionals express propositions, but not both. One alternative for avoiding this problem is to take on a trivalent semantics, and although Lewis considered it too extreme a solution, use of a trivalent semantics for conditionals remains an active albeit non-mainstream area. In addition to being interesting in its own right, the phenomenon addressed here hopefully provides food for thought in what a semantics for conditionals needs to accommodate.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 665850.

References

María Biezma and Arno Goebel. 2019. Being pragmatic about biscuits.

Ellen Breitholtz. 2014. *Enthymemes in Dialogue: A micro-rhetorical approach*. Ph.D. thesis, University of Gothenburg.

Robin Cooper. *Type theory and language: From perception to linguistic communication*. Published: In prep.

Robin Cooper. 2005. Austinian truth, attitudes and type theory. *Research on Language and Computation*, 3(2-3):333–362.

Robin Cooper. 2012. Type theory and semantics in flux. In Ruth Kempson, Tim Fernando, and Nicholas Asher, editors, *Philosophy of Linguistics*, Handbook of the Philosophy of Science, pages 271 – 323. North-Holland, Amsterdam.

Robin Cooper and Jonathan Ginzburg. 2015. TTR for natural language semantics. In Chris Fox and Shalom Lappin, editors, *Handbook of Contemporary Semantic Theory*, 2 edition, pages 375–407. Blackwell, Oxford.

Barbara Dancygier. 1992. Two metatextual operators: Negation and conditionality in English and Polish. *Annual Meeting of the Berkeley Linguistics Society*, 18(1):61–75.

Renaat Declerck and Susan Reed. 2001. *Conditionals: A Comprehensive Empirical Analysis*, volume 37 of *Topics in English Linguistics*. DeGruyter.

Igor Douven. 2008. The evidential support theory of conditionals. *Synthese*, 164(1):19–44.

Chi-Hé Elder. 2015. *On the forms of conditionals and the functions of 'if'*. Ph.D. thesis, University of Cambridge.

Michael Franke. 2007. The pragmatics of biscuit conditionals. In *Proceedings of the 16th Amsterdam Colloquium*, pages 91–96.

Jonathan Ginzburg. 2012a. *The Interactive Stance*. Oxford University Press.

Jonathan Ginzburg. 2012b. *The Interactive Stance*. Oxford University Press.

Sabine Iatridou. 1991. *Topics in Conditionals*. Ph.D. thesis, Massachusetts Institute of Technology.

Angelika Kratzer. 1986. Conditionals. *Chicago Linguistics Society*, 22(2):1–15.

David Kellogg Lewis. 1975. Adverbs of Quantification. In Edward L. Keenan, editor, *Formal Semantics of Natural Language*, pages 3–15. Cambridge University Press.

Matthew Purver. 2001. SCoRE: A tool for searching the BNC. Technical Report TR-01-07, Department of Computer Science, King's College London.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. Longman, USA.

John Robert Ross. 1967. *Constraints on Variables in Syntax*. Ph.D. thesis, MIT.

Muffy E. A. Siegel. 2006. Biscuit conditionals: Quantification over potential literal acts. *Linguistics and Philosophy*, 29(2):167–203.

Niels Skovgaard-Olsen, Henrik Singmann, and Karl Christoph Klauer. 2016. The relevance effect and conditionals. *Cognition*, 150:26–36.

Peggy Speas and Carol Tenny. 2003. Configurational properties of point of view roles. In *Asymmetry in Grammar*, volume 1: Syntax and Semantics, pages 315–344. John Benjamins Publishing Company, Amsterdam.

Robert Stalnaker. 1970. Probability and Conditionals. *Philosophy of Science*, 37(1):64–80.

Eve Sweetser. 1990. *From etymology to pragmatics: The mind-body metaphor in semantic structure and semantic change*. Cambridge University Press.

Personae under uncertainty: The case of topoi

Bill Noble
bill.noble@gu.se

Ellen Breitholtz
ellen.breitholtz@ling.gu.se

Robin Cooper
robin.cooper@ling.gu.se

Centre for Linguistic Theory and Studies in Probability
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg

Abstract

In this paper, we propose a probabilistic model of social signalling which adopts a persona-based account of social meaning. We use this model to develop a socio-semantic theory of conventionalised reasoning patterns, known as *topoi*. On this account the social meaning of a topos, as conveyed in a argument, is based on the set of ideologically-related topoi it indicates in context. We draw a connection between the role of personae in social meaning and the *category adjustment effect*, a well-known psychological phenomenon in which the representation of a stimulus is biased in the direction of the category in which it falls. Finally, we situate the interpretation of social signals as an update to the information state of an agent in a formal TTR model of dialogue.

1 Introduction

Consider the (somewhat dramatic) Example 1, from Lavelle et al. (2012), a corpus of dialogues where participants are instructed to resolve a moral dilemma. The subjects are asked to decide, based on limited information, who out of four passengers in a hot air balloon to sacrifice in order to save the other three.

Example 1

39 C: Well I'm not throwing a kid out [I just couldn't cope with it].

42 A: And the other thing is I mean what what what she achieves er in her life if she becomes as famous as famous as Mozart erm will go on er [forever]=

45 A: So I mean the person it seems like the person with least value is the .

48 B: [she's] pregnant.

51 B: [So you're] killing two people instead of one.

52 C: Yhh and another thing is would he be able to pilot the balloon if his wife is overboard?

Apart from communicating semantic content, arguments often implicitly evoke a *topos*, a pattern of reasoning the speaker draws on to warrant their argument. For example, the argument against sacrificing the pregnant woman (1-51) relies on a topos such as *if you have to choose between killing n and m people and $m < n$ then choose m*.

Upon recognizing the evoked topos, an interlocutor may draw certain conclusions about the speaker, namely that they are the *kind of person* who reasons in this way. Given that information, the interlocutor may, in turn, choose to frame their arguments in a way that appeals to the kind of person they infer the speaker to be.

Such topoi can be seen as signals conveying *social meaning* by association with *personae* or stereotypical categories of people (Eckert, 2012). The use of personae as a *semantic medium* in a theory of social meaning is analogous to how possible worlds (Lewis, 1970), infons, or situation types (Barwise and Perry, 1983) are used in truth-theoretic accounts of propositional meaning. Just as declarative sentences restrict the set of possible worlds or situation types, the social meaning of a social signal restricts the personae attributed to the speaker. Recent work by Burnett (2017), for example, uses game theoretic modelling

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 8–16

to formalise social meaning in terms of personae. In contrast to Burnett (2017), who considers dialectical variables orthogonal to semantic content, we consider the social meaning of topoi in argumentation, following Breitholtz (2014). We develop a probabilistic model that formalises the relationship between topoi and personae through Bayesian inference and integrate this account into a formal TTR model of dialogue by defining an update to the information state of an agent.

2 Personae, topoi, and social meaning

In this section, we give background on the rhetorical and sociolinguistic phenomena we seek to model.

2.1 Personae

The variationist branch of sociolinguistics is interested in the construction of linguistic style through the use of *linguistic variables* (Hudson, 1996). A variable is any axis along which an individual's language may differ from someone else in the same community. Linguistic variables can be found at all levels of linguistic analysis, including phonetics (e.g., accent), prosody, lexical choice, morphology, and syntax.

Some of the earliest work in variationist sociolinguistics, for example, studies phonetic variations different groups of speakers on the island of Martha's Vineyard (Labov, 1963). This *first wave* of variationist sociolinguistics (Eckert, 2012), is principally concerned with variation across macrosociological categories such as race, class and gender .

The second wave of variationist study was interested in more fine-grained social categories, sometimes referred to as *personae* (Eckert, 2012). A persona is a widely recognised social category which is available as a reference point for the expression of social identity in a given community. For example, Eckert (1989, 2008) identifies the personae of "jock" and "burnout" as central to the social semiotic system of an early-2000s Detroit-area high school. Through their dress, behaviour, and linguistic style, students signal identification with or distance from the established personae.

Third-wave sociolinguistics considers the role of variation in the expression of social meaning, rather than merely reflective of social categories (Eckert, 2012). Personae are the semantic common ground that makes communicating social meaning possible. In a given speech community, a linguistic *variant* (the expression of a linguistic variable) constitutes a *social signal* in virtue of its association with one or more personae. Speakers identify themselves as ideologically aligned with a given persona by adopting variants associated with it. This is referred to as *projecting* a persona. Speakers typically do not identify uniquely with one persona, however. Each individual constructs a unique style, mixing and matching variants associated with different personae in a process Eckert (2000) refers to as *bricolage*.

While previous work assumed that linguistic variables were orthogonal to propositional meaning, third-wave sociolinguistics acknowledges that that separation is not always possible. Eckert (2008) writes that her view of linguistic style "precludes the separation of form from content, for the social is eminently about the content of people's lives". In the following section we present *topoi*, a pragmatic phenomenon that play a role in semantic content, but that we argue can also be viewed as a constituent of linguistic style.

2.2 Topoi

Argumentation and reasoning in dialogue is predominantly *enthymematic*, that is, it partly relies on what is "in the mind" (*ἐνθύμημα*) of the listener (Breitholtz, 2014). Aristotle referred to the principles of reasoning which enthymematic arguments are based on as the *topoi* of the arguments. For Aristotle, a topos was a "place" or "field", where a public speaker or a participant in a dialectic debate could find ideas on which to build his argument.

In the 20th century the idea of topos has been taken up in linguistics by Ducrot (1980) and Anscombre et al. (1995) who suggest that every link between a statement and another statement, or between a statement and (for example) an exhortation in discourse is a topos and that topoi are thus essential to any theory of semantics beyond the sentence, as well

as important for contextual interpretation of lexical meaning. One of the leading ideas in Ducrot's take on topoi, is that topoi are not part of factual knowledge about the world, but part of "ideology", that is the agent's conception of acceptable ways to make inferences. This does not mean that topoi are unrelated to facts—for example, a topos of gravity is not likely to be unrelated to the way gravity works. However, it is clear that a large number of topoi are related to ethical considerations such as what is good or beneficial, and these cases are clearly ideological in the sense that they are relative to context.

For example, Ducrot discusses different ways of arguing about giving tips. One individual might encourage another to give a tip to a porter who "carried the bags all the way here", while someone else might advise against it, for the reason that the porter is already paid to carry bags, and why should you pay someone for something they are already paid to do? This is an example of how different topoi may apply in one situation, and lead to inconsistent results or conclusions. Which topoi we appear to draw on while making an argument in a given situation thus gives our interlocutors information of an ideological nature. This is true both in situations where we reason from a context (a set of premises present in a context) to a conclusion, and when we have a particular conclusion in mind that we argue for. In the first case, applying different topoi might lead to different conclusions, but in both cases the implicit ideological information conveyed might differ depending on the topos used.

We argue that topoi, in virtue of their ideological association, constitute social signals that contribute to the persona projected by a dialogue participant, much like use of particular linguistic variants. Topoi are an attractive subject of study as social signals since, unlike social variables like physical appearance or pronunciation, they may be extracted from written text or transcribed dialogues.

3 Two probabilistic models of social meaning

In this section, we develop a simple probabilistic model that associates topoi with personae. In particular, we model how the use of a topos by one agent results in an update to another agent's model of their persona. Since we restrict our attention to a single utterance, we refer to the listener, whose internal state is updated, as *Self* and the speaker, who evoked the topos, as *Other*.

We present two versions of the model. In the *first-order model*, Self models Other as a simple categorical probability distribution over personae. In the *second-order model*, Self represents Other as a Dirichlet distribution over *possible* categorical distributions over personae.

In both cases, the event being modelled is the same: Other (O) invokes a topos (τ) in a dialogue with Self (S). Then, Self's updates their model of Other as a result of that social signal.

Unlike the social signaling game from Burnett (2017)'s, which is based on rational speech acts (Frank and Goodman, 2012), we do not assume any level of social recursion in the speaker; that is, the speaker does not consult a model of the listener's model of themself when producing an utterance.

We assume that each agent has access to a set of personae, $\Pi = \{\pi_1, ..., \pi_K\}$, and topoi, $\Psi = \{\tau_1, ..., \tau_N\}$. A probability distribution φ_π is assigned to each persona π such that:

$$\varphi_\pi(\tau) = P_S(\tau \mid \pi). \tag{1}$$

The probability given by $\varphi_\pi(\tau)$ is the likelihood that someone projecting π will evoke τ. This distribution models the idealogical association between topoi and personae—it is what gives the topoi their social meaning. For now, we assume that Π, Ψ and $\boldsymbol{\varphi}$ are shared community resources.

We begin with the first-order model as a demonstration of the setting and, after discussing its weaknesses, move on to the second-order model.

3.1 First-order model

In the first-order model, Self models Other as categorical probability distribution over personae. Let $\theta_{S,O}$ be S's model of O; that is, the probability, according to S, that O will project the persona π:

$$\theta_{S,O}(\pi) = P_S(\pi \mid O) \tag{2}$$

When O evokes τ, S updates their prior model of O accordingly. Intuitively, S learns that O is more likely to project personae that are likely to evoke τ:

$$\Delta_1(\theta_{S,O}, \varphi, \tau) = \lambda\pi.\frac{\varphi_\pi(\tau) \cdot \theta_{S,O}(\pi)}{\sum_{\pi'} \varphi_{\pi'}(\tau) \cdot \theta_{S,O}(\pi')} \tag{3}$$

In Bayesian terms, the update function gives the posterior distribution of $\theta_{S,O}$, given τ:

$$\begin{aligned}\frac{\varphi_\pi(\tau) \cdot \theta_{S,O}(\pi)}{\sum_{\pi'} \varphi_{\pi'}(\tau) \cdot \theta_{S,O}(\pi')} =& \frac{P(\tau \mid \pi) \cdot P_S(\pi \mid O)}{P_S(\tau)} \\ =& P_S(\pi \mid \tau, O)\end{aligned}$$

To make the situation more concrete, consider again utterance 51 from Example 1. Among the topoi elicited by this utterance is the assumption that, given the choice, it's always better to kill fewer people. Let's call this utterance τ_3 and let $\Delta_1(\theta_{S,O}, \varphi, \tau_3) = \hat{\theta}_{S,O}$.

We may imagine any number of personae associated with τ_3, but most relevant are those personae based on different kinds of moral reasoning. In this case, S believes that the *humanist* and *cold rationalist* personae give some prior probability to the evoked topos (see figure 1). Self updates their model of Other in proportion to the product of the likelihood of the topos given the persona, and the persona's prior probability for Other.

In this first-order model, $\theta_{S,O}$ has two possible interpretations:

1. It represents Self's uncertainty about which persona Other projects (but Self assumes that Other uniquely projects one persona).

2. It represents Self's belief about Other's persona tendencies—i.e., their *bricolage* (but no uncertainty is modelled).

Both of these interpretations have drawbacks. The (false) assumption that each person projects a unique persona results in inconsistency when an agent observes an interlocutor evoke both τ_1 and τ_2 that don't appear in any of the same personae. However, if $\theta_{S,O}$ instead represents Self's take on Other's bricolage of personae, the lack of uncertainty leaves the Bayesian belief revision given by Equation 3 unfounded. To simultaneously account for bricolage and uncertainty, we must add a second layer of analysis to the agent model.

3.2 Second-order model

In the second-order model, we assume that Self attributes some particular distribution over personae to Other, but that their representation captures uncertainty about exactly what distribution it is. Thus, instead of a prior over personae, S's model of O is a prior over distributions over personae. For this, we use a Dirichlet distribution parametrized by K-dimensional positive real-valued $\boldsymbol{\alpha}_{S,O}$.

The Dirichlet distribution is a probability density function defined as follows:

$$f(\theta; \boldsymbol{\alpha}_{S,O}) = \frac{1}{B(\boldsymbol{\alpha}_{S,O})} \prod_{i=1}^{K} \theta(\pi_i)^{\alpha_{S,O,i}}$$

where the domain, θ, is defined on the K-simplex—the space of all possible categorical probability distributions in $\mathbb{R}^K$. Unlike the parameter for a categorical distribution, there is no requirement that $\boldsymbol{\alpha}_{S,O}$ sum to 1. In general, higher overall values for $\boldsymbol{\alpha}_{S,O,i}$ tend to produce flatter distributions, whereas lower values favour sparser ones. For this reason, the Dirichlet parameter is sometimes referred to as a *concentration parameter*.

A higher relative value for a given $\alpha_{S,O,i}$ means the Dirichlet is biased in favour of θ's that assign a high probability to π_i. In fact, by integrating over θ, we arrive again at the marginal probability that S assigns a given persona for O:

$$\begin{aligned}P_S(\pi_i \mid O) &= \int \mathcal{D}(\theta; \boldsymbol{\alpha}_{S,O}) \theta(\pi_i) d\theta \\ &= \frac{\alpha_{S,O}, i}{\sum \boldsymbol{\alpha}_{S,O}}\end{aligned} \tag{4}$$

As before, Self updates their model of Other based on the topos they evoked. This time, Self interprets τ by way of a particular persona—the persona *projected* by the social signal. We define the persona projected by τ (according to S) as the as the most likely persona, given the topos and Self's model of Other. This is given by Bayes rule and Equa-

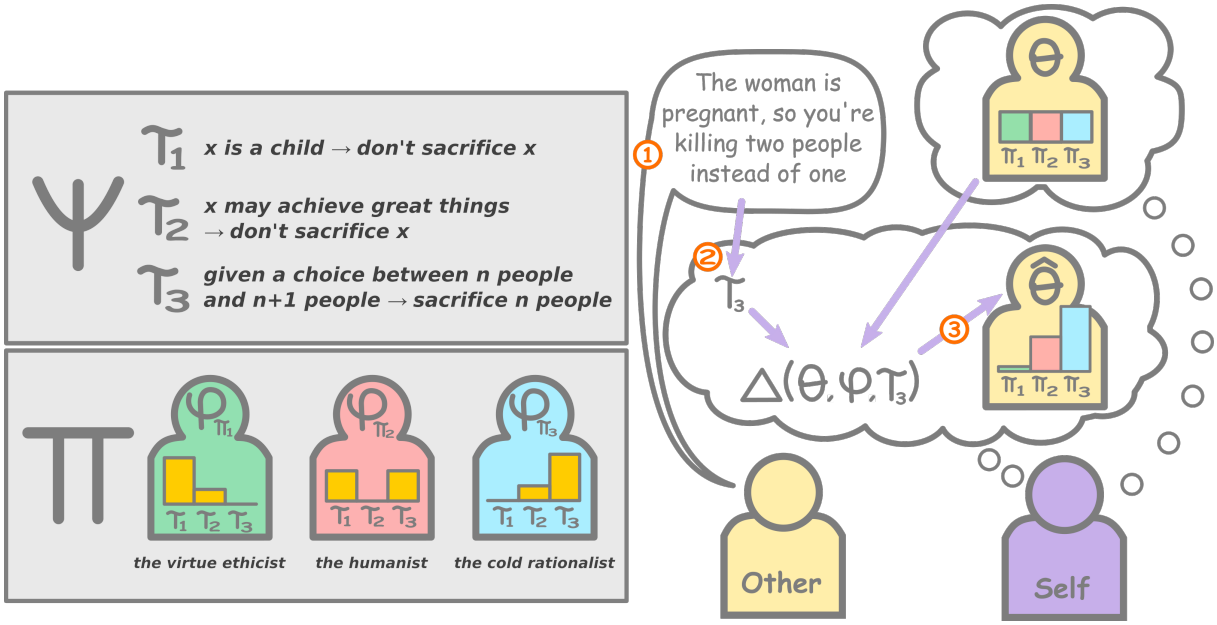


Figure 1: Using the shared topoi, personae, and topos distribution for personae (left), Self updates their representation of Other (right) as follows: (1) Other utters 1-51. (2) Self interprets 1-51 as evoking τ_3. (3) Other applies the update function from Equation 3, incorporating their prior model of Other, the topos distributions for personae, and the evoked topos.

tion 4:

$$\begin{aligned}\text{Proj}(\alpha_{S,O}, \varphi, \tau) &= \underset{i \leq K}{\text{argmax}}\, P(\pi_i \mid \tau) \\ &= \underset{i \leq K}{\text{argmax}}\, P(\tau \mid \pi_i) \cdot P_S(\pi_i \mid O) \\ &= \underset{i \leq K}{\text{argmax}}\, \varphi_{\pi_i}(\tau) \cdot \frac{\alpha_{S,O,i}}{\sum \boldsymbol{\alpha}_{S,O}}\end{aligned} \tag{5}$$

Now let $\text{Proj}_S(\alpha_{S,O}, \varphi, \tau) = \hat{\pi}$. The projected persona is used to update S's model of O as follows:

$$\Delta_2(\alpha_{S,O,i}, \hat{\pi}) = \begin{cases} \alpha_{S,O,i} + 1 & \text{for } \pi_i = \hat{\pi} \\ \alpha_{S,O,i} & \text{otherwise} \end{cases} \tag{6}$$

Note that the updated model O is equal to the Bayesian posterior distribution, given that $\hat{\pi}$ was observed. This is a result of the conjugacy of the Dirichlet distribution over the categorical. For proof, let $\Delta_2(\boldsymbol{\alpha}_{S,O}, \tau) = \hat{\boldsymbol{\alpha}}_{S,O}$ in the following:

$$\begin{aligned} D(\theta, \hat{\boldsymbol{\alpha}}_{S,O}) &= \int \prod_{i=1}^{K} \theta(\pi_i)^{\hat{\alpha}_{S,O,i}} d\theta \\ &= \int \theta(\hat{\pi}) \prod_{i=1}^{K} \theta(\pi_i)^{\alpha_{S,O,i}} d\theta \\ &= P(\hat{\pi} \mid \theta) \cdot P(\theta \mid \boldsymbol{\alpha}_{S,O}) \\ &= P(\theta \mid \hat{\pi}, \boldsymbol{\alpha}_{S,O}) \end{aligned}$$

This conjugacy result means that updating the persona model is very simple—we simply add 1 to $\boldsymbol{\alpha}_{S,O}$ in the position corresponding to the projected persona (as in Equation 6).

In Δ_1, Self updates their model of Other considering all of the personae that Other *might have* been projecting by evoking τ—propagating uncertainty about the projected persona to the update function. In Δ_2, Self assumes that Other is using projecting the maximum likelihood (given τ) persona, $\hat{\pi}$, and updates the posterior accordingly. It would be interesting to compare Δ_2 to a second-order model that is uncertain about the projected persona. Unfortunately, the Dirichlet distribution is not conjugate over the likelihood, $P(\tau \mid \theta)$, meaning that the traditional Bayesian posterior, $P(\theta \mid \tau, \boldsymbol{\alpha}_{S,O})$, is not itself Dirichlet, but rather a mixture of Dirichlet distributions.

Nevertheless, the second-order model performs better than the first-order model in preliminary signaling games simulations. After ten exchanges, a second-order listener's model of the speaker is closer to the speaker's actual persona distribution than that of a first-order listener. Furthermore, as discussed in the following section, similar probabilistic models of the *category adjustment effect* make a

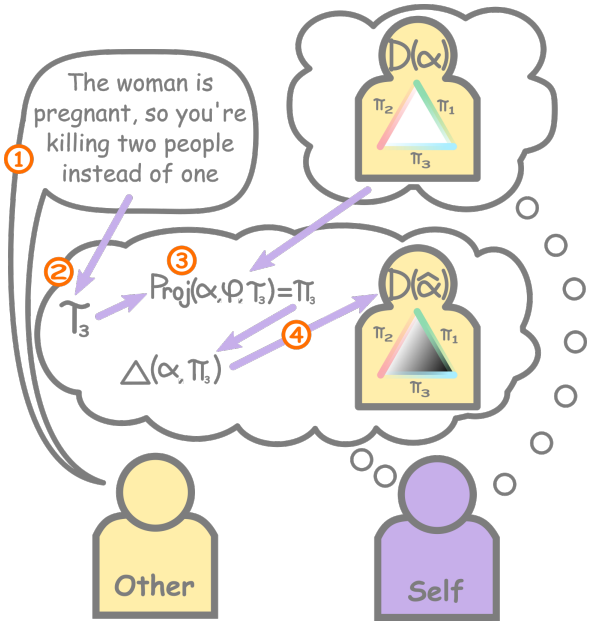


Figure 2: Using the same Ψ, Π, and φ_π's as Figure 1, Self updates their second-order representation of Other as follows: (1) Other utters 1-51. (2) Self interprets 1-51 as evoking τ_3. (3) Self interprets τ_3 as projecting π_3, according to Equation 5. (4) Self updates their prior according to Equation 6.

similar assumption.

4 The category adjustment effect

The category adjustment effect is a phenomenon in which the perception of a stimulus is biased in the direction of the centre of the category in which it falls. Category effects are, for example, an explanation for why phonetic differences are easier to detect when they cross phoneme boundaries (Liberman et al., 1967; Feldman et al., 2009).

The category adjustment model (Huttenlocher et al., 2000), describes the category adjustment effect in explicitly Bayesian terms, with the category acting as a prior distribution over stimuli. Cibelli et al. (2016), use this model to test a version of the Sapir-Whorf hypothesis. In a series of experiments they show that the semantics of colour terms have an effect on colour perception. For example, when asked to recall the colour of a displayed colour swatch, speakers were biased towards the mean of the colour category in which the swatch fell.

Eckert (2008) defines the meaning of a linguistic variable, its *indexical field*, as the "constellation of ideologically related meanings" that arises in virtue of the variable's relationship with one or more personae. Viewed through the lens of category adjustment, the social interpretation of a linguistic variable is mediated by the social categories (personae) associated with it. In our model, idealogical relatedness is represented by the conditional distribution of topoi given a persona (φ from §3). This distribution corresponds to the prior from the category adjustment model. This framework suggests two empirical questions for future work.

The first question concerns the propagation of uncertainty about the projected persona. Is it the case that, as in the original category adjustment model, the interpretation of a social signal is mediated by a *single nearest category* (the projected persona from §3.2), or does it take into account all of the personae that the speaker *might be* projecting (as in §3.1)?

Second, is there a category adjustment effect on the listener's judgment of *which* topos is being evoked? Since we are focused on updates to the listener's model of the speaker's persona, we don't model uncertainty about the evoked topos, but a given argument may have multiple possible warrants. It seems reasonable to assume that the listener would take their persona model of the speaker and associations between personae and topoi into account when judging which topos was evoked.

5 Information state update

In order to use the above technique to account for social meaning dynamics in interaction, we integrate our model with an information state update account of dialogue, an approach successfully used to model various dialogue phenomena (Larsson and Traum, 2000; Ginzburg, 2012). We see this as pointing to a general method for incorporating previous work on social meaning, for example, Burnett (2017), into an account of incremental updates of social meaning in a idealogical context. This continues the work of Breitholtz and Cooper (2019).

To represent the evolving information states of agents involved in interaction, we use dialogue gameboards (Lewis, 1979; Ginzburg, 1994; Larsson, 2002; Ginzburg, 2012). In order

to account for coordination phenomena in dialogue, such as misunderstandings and clarifications, it is important that the information state of the participants are modelled as separate gameboards, representing each agent's view of the conversational game currently being played. The gameboards are split into two fields, one for information that the speaker takes to be private, one field for information that he or she takes to be shared in the dialogue. On our account dialogue participants are represented twice on the DGB. In Figure 3 we see that the shared information about the participants is just referential. The information about perceived personae of the dialogue participants can be found in the private-field of the DGB, where the labels 'other' and 'self' are associated with the corresponding individuals in the shared field. The superscripted up arrow indicates that the path points to an object three levels up in the record type.

As an interaction progresses the DGBs of the participants evolve in accordance with update rules. In Figure 4 we represent the update rule '$f_{\text{UpdatePersonae}}$' which is a function which takes an information state and an utterance event and returns a type for the updated information state. This function is used in the action rule 'UpdatePersonae' given in Figure 5. This action rule has three conditions. The first one requires that the agent's, S, current information state, $s_{i,S}$, is judged by S to be of some type, T. The second condition requires that T is a subtype of the type required for r in '$f_{\text{UpdatePersonae}}$'. The third condition requires that the current utterance, u^*, is of the type required for u in '$f_{\text{UpdatePersonae}}$'. If these conditions are fulfilled S is licensed or "afforded" (indicated by the wavy line) to make a judgement about S's updated information state, $s_{i+1,S}$, namely that it of the type which S judged the current information state to be of asymmetrically merged, indicated by $\boxed{\wedge}$ with the result of applying the update function to the current information state and the current utterance. The operation of asymmetric merge on record types in TTR corresponds to priority unification in feature based systems. It will preserve all the information in both types except that if the two types have different information on a given path then the information from the second type will be in the result but not that from the first type. (See Cooper and Ginzburg, 2015 and Cooper, in prep for more details.)

These definitions rely on two types which depend on the set of topoi, Ψ, which are currently under consideration. The first type is *Persona*(Ψ). A witness for this type is a distribution over Ψ. (In a more complete treatment this would just be one of a number of components that make up a persona.) That is,

> f : *Persona*(Ψ) iff f is a function with domain Ψ and range in $[0, 1]$ such that $\sum_{t \in \Psi} f(t) = 1$

The second type we use is *PersConcFunc*(Ψ), the type of Persona Concentration Functions for Ψ. This is defined as

$$(\textit{Persona}(\Psi) \rightarrow \textit{Real}_{(0,\infty+)})$$

That is, *PersConcFunc*(Ψ) is the type of functions from distributions over Ψ to positive real numbers greater than 0.

6 Conclusion

In this paper we present a probabilistic model that accounts for the social meaning of topoi. We suggest that, as in the case of colour perception, the interpretation of social signals is subject to a category adjustment effect induced by social categories, or personae. Finally, we incorporate this model into an integrated account of linguistic interaction. We do this by defining a TTR update rule which is referenced in an action rule showing how speakers change their model of their interlocutor based on social signalling.

We see three major avenues for future work stemming from the basic model presented here. First, systems of social meaning are not monolithic or static—we should account for variation and change in the available personae, topoi, and the associations between the two. Second, this model could be used in a game-theoretic analysis of argumentation. Based on the persona that a speaker projects, which topoi should an interlocutor use to warrant their arguments? Finally, as discussed in §4, how does the listener's model of the speaker persona affect which topos they interpret as warranting the speaker's argument?

$$\left[\begin{array}{l}\text{private:}\left[\begin{array}{l}\text{topoi:set}(\textit{Topos})\\ \text{participants:}\left[\begin{array}{l}\text{other:}\left[\begin{array}{l}\text{x=}\Uparrow^3\text{shared.participants.O:}\textit{Ind}\\ \text{pcf:}\textit{PersConcFun}(\Uparrow^2\text{topoi})\end{array}\right]\\ \text{self:}\left[\begin{array}{l}\text{x=}\Uparrow^3\text{shared.participants.S:}\textit{Ind}\\ \text{pcf:}\textit{PersConcFun}(\Uparrow^2\text{topoi})\end{array}\right]\end{array}\right]\end{array}\right]\\ \text{shared:}\left[\begin{array}{l}\text{topoi:}\left[\begin{array}{l}\text{prev:}\textit{RecType}\\ \text{curr:}\left[\begin{array}{l}\text{topos:}\textit{Topos}\\ \text{speaker:}\textit{Ind}\end{array}\right]\end{array}\right]\\ \text{participants:}\left[\begin{array}{l}\text{O:}\textit{Ind}\\ \text{S:}\textit{Ind}\end{array}\right]\end{array}\right]\end{array}\right]$$

Figure 3: Representation of participants on the DGB

$$\lambda r\text{:}\left[\text{private:}\left[\begin{array}{l}\text{topoi:set}(\textit{Topos})\\ \text{participants:}\left[\text{other:}\left[\begin{array}{l}\text{x:}\textit{Ind}\\ \text{pcf:}\textit{PersConcFunc}(\Uparrow^2\text{topoi})\end{array}\right]\right]\end{array}\right]\right].$$
$$\lambda u\text{:}\left[\text{s-event:}\left[\begin{array}{l}\text{sp=}r\text{.private.other.x:}\textit{Ind}\\ \text{topos:}\textit{Topos}\\ \text{proj-pers=proj(topos, s-event.sp):}\textit{Topos}\end{array}\right]\right].$$
$$\left[\text{private:}\left[\begin{array}{l}\text{topoi=}r\text{.private.topoi:set}(\textit{Topos})\\ \text{participants:}\left[\text{other:}\left[\text{pcf=}\Delta_2(r\text{.private.participants.other.pcf, }u\text{.proj-pers):}\textit{PersConcFunc}(\Uparrow^2\text{topoi})\right]\right]\end{array}\right]\right]$$

Figure 4: $\mathrm{f}_{\textsc{UpdatePersonae}}$

$$s_{i,S} :_S T$$

$$T \sqsubseteq \left[\text{private:}\left[\begin{array}{l}\text{topoi:set}(\textit{Topos})\\ \text{participants:}\left[\text{other:}\left[\begin{array}{l}\text{x:}\textit{Ind}\\ \text{pcf:}\textit{PersConcFunc}(\Uparrow^2\text{topoi})\end{array}\right]\right]\end{array}\right]\right]$$

$$u^* :_S \left[\text{s-event:}\left[\begin{array}{l}\text{sp=}s_{i,S}\text{.private.other.x:}\textit{Ind}\\ \text{topos:}\textit{Topos}\\ \text{proj-pers=proj(topos, s-event.sp):}\textit{Topos}\end{array}\right]\right]$$

$$\overline{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}$$

$$s_{i+1,S} :_S T \boxed{\wedge} \mathrm{f}_{\textsc{UpdatePersonae}}(s_{i,S})(u^*)$$

Figure 5: UpdatePersonae: Updating personae on the DGB according to the second-order model

7 Acknowledgements

This work was partly funded by Riksbankens Jubileumsfond for the Advancement of the Humanities and Social Sciences, Project P16-0805:1 *Dialogical Reasoning in Patients with Schizophrenia*.

References

Jean-Claude Anscombre et al. 1995. Théorie des topoï. *Hermès*, 15:185–198.

Jon Barwise and John Perry. 1983. *Situations and Attitudes*. MIT Press.

Ellen Breitholtz. 2014. Reasoning with topoi - towards a rhetorical approach to non-monotonicity. In *Proceedings of the 50:th anniversery convention of the AISB*, pages 190–198. AISB.

Ellen Breitholtz and Robin Cooper. 2019. Integrating personae in a TTR account of interaction. Presented at Integrating Approaches to Social Meaning.

Heather Burnett. 2017. Signalling games, sociolinguistic variation and the construction of style. *Linguistics and Philosophy*, pages 1–32.

Emily Cibelli, Yang Xu, Joseph L. Austerweil, Thomas L. Griffiths, and Terry Regier. 2016. The Sapir-Whorf Hypothesis and Probabilistic Inference: Evidence from the Domain of Color. *PLOS ONE*, 11(7):e0158725.

Robin Cooper. in prep. From perception to communication: An analysis of meaning and action using a theory of types with records (TTR). Draft of book chapters available from `https://sites.google.com/site/typetheorywithrecords/drafts`.

Robin Cooper and Jonathan Ginzburg. 2015. Type theory with records for natural language semantics. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*, second edition, pages 375–407. Wiley-Blackwell.

Oswald Ducrot. 1980. *Les Échelles Argumentatives*. Les Éditions de Minuit.

Penelope Eckert. 1989. *Jocks and burnouts: Social categories and identity in the high school*. Teachers college press.

Penelope Eckert. 2000. *Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. Number 27 in Language in Society. Blackwell Publishers, Malden, Mass.

Penelope Eckert. 2008. Variation and the indexical field. *Journal of Sociolinguistics*, 12(4):453–476.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of variation. *Annual Review of Anthropology*, 41:87–100.

Naomi H. Feldman, Thomas L. Griffiths, and James L. Morgan. 2009. The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4):752–782.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998.

Jonathan Ginzburg. 1994. An update semantics for dialogue. In *Proceedings of the Tilburg International Workshop on Computational Semantics*, Tilburg, The Netherlands.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press.

Richard A Hudson. 1996. *Sociolinguistics*. Cambridge university press.

J. Huttenlocher, L. V. Hedges, and J. L. Vevea. 2000. Why do categories affect stimulus judgment? *Journal of Experimental Psychology. General*, 129(2):220–241.

William Labov. 1963. The Social Motivation of a Sound Change. *WORD*, 19(3):273–309.

Staffan Larsson. 2002. *Issue-Based Dialogue Management*. PhD Thesis, University of Gothenburg, Gothenburg, Sweden.

Staffan Larsson and David Traum. 2000. Information state and dialogue management in trindi dialogue move engine tool kit. *Natural Language Engineering*, 6:323–240.

Mary Lavelle, Patrick GT Healey, and Rosemarie McCabe. 2012. Is nonverbal communication disrupted in interactions involving patients with schizophrenia? *Schizophrenia bulletin*, 39(5):1150–1158.

David Lewis. 1970. General Semantics. *Synthese*, 22(1/2):18–67.

David Lewis. 1979. Scorekeeping in a Language Game. In Rainer Bäuerle, Urs Egli, and Arnim von Stechow, editors, *Semantics from Different Points of View*, Springer Series in Language and Communication, pages 172–187. Springer, Berlin, Heidelberg.

A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. 1967. Perception of the speech code. *Psychological Review*, 74(6):431–461.

Dogwhistles as Identity-based Interpretative Variation

Quentin Dénigot and **Heather Burnett**
Laboratoire de Linguistique Formelle
5, rue Thomas Mann
F-75205 Paris Cedex 13
qdenigot@linguist.univ-paris-diderot.fr
heather.susan.burnett@gmail.com

Abstract

The following paper presents a formal model for the description of *dogwhistles*. Dogwhistles are a class of expressions often used in political discourse that aim at being interpreted in different ways by listeners of different communities. The model presented here describes this phenomenon using a variation on the Social Meaning Games framework that uses probability distributions over possible interpretation functions.

1 Introduction

Pragmatics has underlined the importance of context in determining the meaning of utterances, and Gricean pragmatics in particular has established a normative framework for the successful transmission of a message between two cooperating agents (Grice, 1975). The insights into human communication that are Grice's conversational maxims have led to formal implementations since Lewis 1969. Most notably, the maxim of quality is the basis for the emergence of scalar implicatures in the Rational Speech Act (RSA) framework (Frank and Goodman, 2012). Grice's maxims, much like Lewis' signalling games, only seek to describe situations where language is used for the sole goal of transmitting accurate information from one speaker to a listener. This is in part what is meant by "cooperation": both sides share the same goal of having the information properly transmitted (whether it be by choosing the right message for the speaker, or choosing the right interpretation for the listener).

This vision, however, only describes a subset of language. It is reasonable, for example, to think that the information content of a linguistic utterance is not limited to the content of the message itself, but that the way in which the message is articulated, either in terms of pronunciation or choice of words, can convey information about the speaker themself. This is what sociolinguists call *social meaning* (Eckert, 2008, 2012): the part of a linguistic signal that conveys information about the person producing the signal rather than the content of the signal itself. It has been shown that intuitions about the social meaning contained in certain accents, for example, has an influence on the reception of a message by the listeners, leading to systematic interpretations of signals that could be at odds with the message conveyed by the content of the message (Acton, 2020). The traditional approach is limited in its scope in the sense that it fails to account for the existence of at least two sources of what could be called "information" in any given linguistic utterance: message content and social meaning.

Works on Social Meaning Games (SMG) (Burnett, 2017, 2019) fill this gap by offering a framework based on game theory (like Lewis' works and like many formal approaches to pragmatics, including RSA) which treats socially significant linguistic variation as another source of meaning. This leads to a variation on signaling games in which the *personae* signaled by the speaker and retrieved by the listener have to match in order to maximize both players' utilities. Crucially, we are talking of *personae*, not *social identity*, because we have to account for cases where the speaker is trying to convey a specific set of traits about themself to the listener for a given goal; they are trying to communicate how they want to be seen in this situation. Here, we are reaching a point where the maxim of quality is, to some extent, flouted, or at least not as relevant.

Examples used for the illustration of SMGs in Burnett 2019 are often political in nature. Political discourse (whether in debates or speeches) is a great field of inquiry for these phenomena because they involve speakers that are publicly known and for whom we can usually access several discourses, including discourses in many different contexts. SMGs can give us an intuitive

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 17–25

view of how social meaning is conveyed, which is key in understanding political discourse, but they fail to account for situations described as *dogwhistle politics*. The term *dogwhistle* refers to a class of expressions often used in political discourse; the goal in using them is to convey two different messages to two different communities. SMGs do not take into account the fact that the audience of a political discourse might be ideologically heterogeneous, leading to differing interpretations of a given message according to prior beliefs and social background. The goal of this work is to define what form situations of dogwhistling might take and to give a formal model describing the contexts in which they are more likely to be used.

2 *Dogwhistles* and *dogwhistle politics*

Dogwhistle politics is generally defined as sending a message to an audience in such a way that a subset of the audience will understand the message differently from the rest of the audience. In more political terms, it is a "way of sending a message to certain potential supporters in such a way as to make it inaudible to others whom it might alienate or deniable for still others who would find any explicit appeal along those lines offensive" (Goodin and Saward, 2005).

To what extent are such practices indeed noticed and what effects do they *actually* have on public opinion? There is a compelling literature on the subject, showing notably that phrases like "*inner cities*" can be responsible for the fact that discussions of nonracial policies can be biased by racial thinking in White voters (Hurwitz and Peffley, 2005) while having a different effect on Black voters (White, 2007). Likewise, it has been shown that the use of religious discourse can also have a significant impact on both opinions and voting intentions for Evangelical voters (Calfano and Djupe, 2009; Albertson, 2015). The effects of dogwhistle speech are backed by empirical evidence and these effects are congruent with the effects that are intuitively attached to the practice: dogwhistle speech reinforces the support of core supporters while being ignored by moderates, in situations where explicit reference to religion or race has negative effects on moderates.

As far as the intentional use of such terms is concerned, we can mention Kuo 2006, who clearly acknowledges it:

> "We threw in a few obscure turns of phrase known clearly to any evangelical, yet unlikely to be noticed by anyone else [. . .]"

The topic, however, has barely been discussed in the linguistics literature. Several theories exist regarding how and why dogwhistles actually work and only recently (starting with Stanley 2015) have these efforts focused on analyzing the language *per se* and trying to give a linguistically consistent description of the phenomenon.

A first approach consists in saying that dogwhistle words have an *explicit* meaning and an *implicit* meaning. This is the approach favored by Mendelberg 2001; Stanley 2015; Henderson and McCready 2019b and Saul 2018. One way of thinking about this (Mendelberg, 2001) is *ambiguity*, each word would have several meanings, for example one racial and the other nonracial, and the use of that term would trigger (or not) one or both of the interpretations in the audience. This makes intuitive sense, but it has important problems, one of them being that the ambiguity that takes place here does not appear to be symmetrical. Khoo 2017 uses the counterexample of the ambiguous word "*funny*" in English, which can either mean "*humorous*" or "*strange*", and remarks that (1) poses no problem.

(1) Smith is a funny man who is not humorous.

Compare with a sentence like (2), which sounds very uncanny.

(2) #Smith is an inner-city pastor who is from, works and lives, in the suburbs.

If the word "*inner-city*" was indeed ambiguous between a racial and a nonracial meaning, one should be able to cancel out one of the two meanings, but it appears that the nonracial meaning is not cancellable, whereas (3) does not seem to cause any weirdness in terms of interpretation.

(3) Smith is an inner-city pastor who is not African American.

If the word "*inner-city*" was properly ambiguous, one would call upon either one of its meanings while disregarding, or even cancelling the other, and this does not appear to be the case.

Stanley 2015 proposes an approach relying on the concepts of *at-issue* and *not-at-issue* contents. The idea here is that dogwhistle words would not be ambiguous *per se*, but that through con-

ventional use, they have acquired a secondary, *not-at-issue* meaning. The problem with this approach, however, is underlined in Henderson and McCready 2019b and Khoo 2017: conventional meanings are generally thought to be non-cancellable, which makes the crucial deniability part of dogwhistles impossible. Compare, for example, with slurs, where the added conventional meaning that gives the listener information about the speaker's attitude towards the community they are referring to is not cancellable (examples from (Henderson and McCready, 2019b)), compare (4) with (5), where "welfare" is thought to dogwhistle a negative attitude towards social programs:

(4) A: Angela Merkel is a kraut!
B: What do you have against Germans?
A: #I don't have anything against Germans. Why do you think I might?

(5) A: Donald is on welfare.
B: What do you have against social programs?
A: I don't have anything against social programs. Why do you think I might?

That deniability is a key point of dogwhistles that differentiates them from slurs or other lexical items imbued with added conventional meaning.

3 Formal model

There have been very few attempts at sketching out a formal representation of dogwhistles and their use, and we argue that any attempt at doing so should present a solution that satisfies the following properties of the phenomenon: dogwhistles are cases of INTERPRETATIVE VARIABILITY, where different listeners should assign different interpretations to a speaker's single utterance. Dogwhistles are most common in situations of POLITICAL CONFLICT between conversational participants (Goodin and Saward, 2005; Saul, 2018; Stanley, 2015). Furthermore, interpretative variability is IDENTITY-BASED: listeners who attribute a religious identity, or *persona* (Eckert, 2008; Agha, 2003), similar to theirs to the speaker will be more likely to interpret the dogwhistle in the religious way than those who believe the speaker holds no specific religious beliefs (Albertson, 2015). Since racist interpretations are conditioned on but not determined by identity, use of a dogwhistle often provides some PLAUSIBLE DENIABILITY to the speaker, which can be useful to them to satisfy the political requirements of a diverse audience. This deniability is important because a SAVVY OPPONENT, someone who does not share the speaker's political ideology but who understands the racist way the dogwhistle can be used, can call the speaker out for this use (Stanley, 2015; Saul, 2018). Finally, as observed by Khoo 2017, whether or not an expression will display identity-based interpretative variability depends on its SPECIFIC FORM: expressions that are truth-conditionally equivalent to *inner city*, such as *city center*, are not semantically variable in the same way.

3.1 Previous approaches

Because of the strategic aspect of dogwhistling, authors such as Henderson and McCready 2019b,a and Asher and Paul 2018a have found game-theoretic pragmatics to be useful for solving the puzzles described above. Henderson & McCready propose a framework in which a speaker, S, sends a dogwhistle message m_D to a listener, L. L has particular beliefs about the persona of the speaker, and they update their beliefs about the world by taking into account L's hypothesized persona and the m_D's literal meaning. To account for SPECIFIC FORM, Henderson & McCready (2019b:6) use axiom schemata (6) which are triggered by the form of the dogwhistle. In the case of *inner city*, S's use of this message (and this message only) triggers the proposition "All neighborhoods at the center of the city are urban African American" in the mind of L, which then allows L to infer that *African American neighborhood* is S's intended meaning. They say (p.8), "The following axiom (6) states that, given that a speaker S with a persona π uses the dogwhistle *inner city*, and given that the hearer believes that inner city neighborhoods are all African American, then normally the speaker intends the inference from his phrasing to this enriched meaning to be made".

(6) $$\begin{aligned} & Use(S, \pi, [inner\_city]) && \wedge \\ & Bel(L, \forall x(inner\_city(x) && \rightarrow \\ & urban\_AA\_neighborhood(x)) && > \\ & Intend(S, Bel(L, urban\_AA\_neighborhood(x))) \end{aligned}$$

Although this innovative framework provides a game-theoretic foundation for identity based interpretative variability, we argue it could be improved. For one thing, axioms such as (6) are required for each dogwhistle, even though pat-

terns of speaker/listener behavior are exactly what game-theoretic systems aim to derive. The account for SAVVY OPPONENT is also not clear: according to (Henderson and McCready, 2019b), the listener beliefs mentioned in (6) are necessary for the dogwhistled content to be inferred; however, politically informed non-racist listeners can detect dogwhistles without them.

3.2 Dogwhistle games

We therefore propose to modify the system presented in Henderson and McCready 2019b to arrive at one which can account for SAVVY OPPONENTS and in which we can prove statements similar to (6) as theorems. Our proposal also builds on Asher and Paul 2018b,a, who highlight the importance of POLITICAL CONFLICT in dogwhistles and use a special *Jury* player to determine conversational success. Rather than invoking an abstract *Jury*, we will have dogwhistles arise from political conflict between the conversational participants themselves.

A *dogwhistle game* G_{DW} is a tuple:
$G_{DW} = \langle \{S, L^i, L^j\}, W, M, \text{PERS}, \text{INT}, \text{I-LEX}, Pr_\pi(\cdot), Pr_w(\cdot), \mu_S, U_S \rangle$
where

- S, L^i, L^j are the speaker and two listeners.
- W is a set of worlds w.
- M is a set of messages m.
- PERS is a set of personae π.
- INT is a set of interpretation functions $[\![\cdot]\!]$.
- I-LEX : PERS $\rightarrow \Delta$INT is a function from personae to probability distributions over interpretation functions.
- $Pr_\pi(\cdot)$ is a probability distribution over personae.
- $Pr_w(\cdot)$ is a probability distribution over worlds.
- μ_S is S's preference function from worlds to N of the form $\mu_S(w)$ where w stands for a message interpretation.
- U_S, a utility function from $M \times W$ to R of standard RSA form.

As is the case in Henderson and McCready 2019b, we have a set of words $w \in W$, a set of messages $m \in M$, and a set of personae $\pi \in$ PERS. We differ from their model in that we have a **set** of interpretation functions: $[\![\cdot]\!] \in$ INT and a lexical interpretation function, I-LEX, mapping personae to probability distributions over INT. The idea is that a speaker's persona will be informative for their form-meaning associations. Given that listeners are rarely certain about the state of the world or even S's political identity, we will represent this uncertainty as prior probability distributions over worlds ($Pr_w(\cdot)$) and personae ($Pr_\pi(\cdot)$).

Following Asher and Paul 2018a, we will allow considerations other than communication to influence S's actions. As in standard SMGs (Burnett, 2019), we have a preference function μ; in our case however, it is applied to preferred worlds for the speaker. The idea behind this is that in dogwhistling situations, we can assume the speaker might not respect the maxim of quality. The goal of the speaker is not to communicate the "*actual state of the world*" but to ensure the support of the audience (by conveying a state of the world that satisfies their beliefs). It is important to point out that the goal of the speaker is, to some extent, to deceive the audience: we are *not* in a cooperative situation, although our listeners will largely assume that we are. In other words, we are in an asymmetrical context.

We conceptualize the interaction situation as parallel to a signaling game for the listeners: they are trying to figure out S's message. Correspondingly, L^i and L^j's interpretation process will closely follow the *Rational Speech Act* model (Frank and Goodman, 2012). In our situation, however, the speaker is duplicitous, and we will represent this duplicity by the use of a preference function μ^*, that takes as input ordered pairs of worlds corresponding to each listener's possible interpretation. Similarly, this duplicitous speaker has their own U^*_S utility function that also takes ordered pairs of worlds as input.

3.3 Determining the listeners' interpretations

The listeners in this model are almost identical to standard RSA listeners, in that they also infer their interpretations from what a speaker faced with a literal listener would say. In standard RSA, speakers and listeners can reason about each other's rea-

soning, leading to an interpretation of messages that relies on their *literal meaning*, but is not necessarily determined by it.

There is one thing in our model that is added to the listeners: it is assumed that they have priors over the possible personae of the speaker and that they can derive different interpretation functions from these priors. A key point of the model presented here is that there exist different possible interpretations for a given message. This will be illustrated with an example in section 4.

From these two priors, using the I-LEX function, listeners can associate a probability distribution over interpretation functions dependent on the priors over personae that they have. The intuition behind this is that listeners assume that a speaker displays a certain persona, and they assume that people belonging to the group that the speaker appears to belong to speak in a certain way. The strength of these assumptions depends on the speaker.

The probability $P(\llbracket\cdot\rrbracket)$ that a certain interpretation function will be used is computed as follows :

(7) For all $\llbracket\cdot\rrbracket, P(\llbracket\cdot\rrbracket) = \sum_{\pi \in \text{PERS}} Pr(\pi) * \text{I-LEX}(\pi, \llbracket\cdot\rrbracket)$

Then each message can be interpreted using one interpretation function or another:

(8) $Pr(w|\llbracket m\rrbracket) = \frac{Pr(\{w\} \cap \llbracket m\rrbracket)}{Pr(\llbracket m\rrbracket)}$

And finally, the meaning of any given message, taking into account all the ways it could have been meant, uses both values, giving us the *literal listener*:

(9) $P_{L_0}(w|m) = \sum_{\llbracket\cdot\rrbracket \in \text{INT}} P(\llbracket\cdot\rrbracket) * Pr(w|\llbracket m\rrbracket)$

The subsequent steps are similar to standard RSA in that the speaker computes the utility of each message, and using this utility score for a given message, we can have probability distribution over the different messages the speaker can send, given what they want to convey. Where we differ from standard RSA is in the use of the μ function giving the preference of the speaker over *possible interpretations*. What we assume is that listeners have a bias towards thinking that speakers think like them and therefore have similar preferences.

The utility of the speaker is computed as follows, where C is a cost function:

(10) $U_S(m, w) = log(P_{L_0}(w|m)) - C(m)$

The probability distribution over their possible messages is computed as follows, where α is a temperature parameter governing how much variability the system allows:

(11) $P_S(m|w) \propto exp(\alpha U_S(m, w)) * exp(\alpha' \mu(w))$

Adapting from (Burnett, 2019), we can infer from the μ function and the P_S values a probability distribution over the possible messages to have an idea of the speaker's behavior as it is envisioned by all speakers using the following two formulae:

(12) $P_W(w; \mu) = \frac{exp(\alpha' * \mu(w))}{\sum_{w' \in W} exp(\alpha' \mu(w'))}$

(13) $\mathcal{P}_S(m) = \sum_w P_W(w; \mu) * P_S(m|w)$

Finally, given those assumptions about the speaker, pragmatic listeners L_1 will try to interpret the meaning of what was just said by S using:

(14) $P_{L_1}(w|m) \propto Pr(w) * P_S(m|w)$

3.4 The duplicitous speaker

This would be a classic RSA model involving one speaker and one listener. But the case of dogwhistles, it is assumed that there are several listeners in the crowd. More specifically, there are at least two different listeners with different opinions, and the goal of the speaker is to satisfy them both.

We will distinguish several instances of the speaker in this model. Some of these instances are the speakers we described in the previous section. We will call them the *honest speakers*; they are speakers that standard listeners assume they are talking to. But the context in which dogwhistles appear call for another kind of speaker, that we will call *duplicitous*. The duplicitous speaker differs from the honest speakers in one major way: they always consider pairs of worlds when computing any of the aforementioned probabilities. Therefore, instead of a utility function $U_S : M \times W \rightarrow R$, they have a utility function U_S^* defined as follows, where the indices i and j serve to differentiate the two different listeners:

(15) $U_S^*(m, \langle w, w' \rangle) = log(P_{L_0}^i(w|m)) + log(P_{L_0}^j(w'|m)) - C(m)$

Similarly, the preference function μ becomes μ^*

and takes ordered pairs of worlds as inputs. It is important that the pairs of worlds are ordered, because the duplicitous speaker seeks to treat the two different listeners differently.

3.5 The savvy listener

We mentioned previously that a satisfactory model for dogwhistles should take into account the possibility of a *savvy listener*, i.e. a listener that can see through what the speaker is saying and infer that there are in fact several messages being communicated here.

In this model, the savvy listener is equivalent to a listener that would act like the duplicitous speaker does; in other words, the savvy listener also takes into account the fact that there are various listeners with various beliefs and preferences, and assumes that the speaker is going to try and take advantage of that fact. Therefore, they assume that the speaker uses a function μ^* and a utility function U_S^* and computes the intended meaning using this supplementary data, using (16):

(16) $P^{Savvy}(\langle w, w' \rangle | m) \propto P_W(\langle w, w' \rangle; \mu^*) * P_{S_{Dup}}(m | \langle w, w' \rangle)$

4 '*Inner cities*' example

By way of illustration, we will focus on the often-used example of *"inner cities"*. Following (Henderson and McCready, 2019a,b), let S be Representative Paul Ryan trying to discuss issues in urban areas while also trying to gather the support of his more right-wing voters. Let L^i be a stand-in for a more conservative voter and L^j a less conservative voter. More specifically, L^j refuses to speak openly of race in that context and Ryan would lose their support if he did, whereas L^i would be more likely to offer their support to Ryan if he did take into account race in his discourse.

This is the perfect situation for a dogwhistle. Here is the original statement by Ryan:

(17) We have got this tailspin of culture, in our *inner cities* in particular, of men not working and just generations of men not even thinking about working or learning the value and the culture of work.

Let M be the set of expressions available to S. To simplify, we will distinguish between three possible messages: {"inner cities", "city centers", "African-American neighborhoods"}. These differ by being more or less open references to race, with "city centers" ignoring race completely and "African-American neighborhoods" putting it front and center. "Inner cities" is our dogwhistle term. Because of those properties, we will label these messages as follows:

$M = \{m_D, m_{\neg R}, m_R\}$

To this we add the set W of worlds, which distinguishes between worlds where the issue is about race and worlds where it is not:

$W = \{w_R, w_{\neg R}\}$

Let PERS contain the possible personae of "racist conservative" and "non-racist conservative" :

PERS $= \{\pi_i, \pi_j\}$

Finally, we also set the following :

- INT $= \{[\![\cdot]\!]_i, [\![\cdot]\!]_j\}$

  There are two interpretation functions, one corresponding to each of the personae we are considering here.

- I-LEX$(\pi_\rho, [\![\cdot]\!]_\rho) = 1$

  For simplicity, we assume that for any persona π_ρ, our listeners will associate it invariably with the corresponding interpretation function, meaning that they assume that people displaying persona π_ρ will always mean $[\![m]\!]_\rho$.

- $Pr_w^{L_{1/2}}(w_{R/\neg R}) = 0.5$

  The probability distribution over worlds is uniform, meaning that people have no priors regarding what is going to be said.

- $Pr_\pi^{L^i}(\pi_i) = 0.6 = Pr_\pi^{L^j}(\pi_j)$
  $Pr_\pi^{L^i}(\pi_j) = 0.4 = Pr_\pi^{L^j}(\pi_i)$

  Each listener assumes that the speaker is a bit more likely to display one persona over the other. In this scenario, we can interpret this as L^i recognizing themselves more in persona π_i and therefore having a bias towards thinking that S is more likely to be displaying that same persona, and symmetrically for L_j.

- For simplicity of computation, we assume that messages are *costless* and we set temperature parameters α are set to 1.

See Table 1 for the result of applying each interpretation function $[\![\cdot]\!]$ to each message m.

$Pr(w\|[\![\cdot]\!]_i)$	w_R	$w_{\neg R}$
m_R	1	0
$m_{\neg R}$	0	1
m_D	1	0
$Pr(w\|[\![\cdot]\!]_j)$	w_R	$w_{\neg R}$
m_R	1	0
$m_{\neg R}$	0	1
m_D	0	1

Table 1: Interpretation of each message according to different interpretation functions.

The important value we want to have here are the probabilities ascribed to the interpretation of each message $P_{L_{i/j}}(w|m)$ for each listener, as well as the $P_S(m|w)$ score for each message of what the speaker is expected to say given that they wish to communicate w. After the relevant computations, applying the μ functions in table 2, we have the numbers in table 3 and table 4.

	w_R	$w_{\neg R}$
μ^i	2	1
μ^j	0	2

Table 2: μ functions as envisioned by honest listeners. L^j assumes that the speaker does not want to convey race-based interpretations (because they themselves despise them).

$P_{L_1^i}(w\|m)$	w_R	$w_{\neg R}$
m_R	1	0
$m_{\neg R}$	0	1
m_D	≈ 0.567	≈ 0.432
$P_{L_1^j}(w\|m)$	w_R	$w_{\neg R}$
m_R	1	0
$m_{\neg R}$	0	1
m_D	≈ 0.254	≈ 0.745

Table 3: Interpretations for honest pragmatic listeners of each message depending on their priors.

What we can see with these numbers is that honest listeners believe that honest speakers would be more likely to use m_R if they wish to convey w_R, but that using m_D is not seen as impossible. Similarly for $m_{\neg R}$ and $w_{\neg R}$. What it also shows, however, is that if they hear m_D, they are more likely to interpret it according to the interpretation function that they deemed more probable, given their prior beliefs about the speaker.

$P_S^i(m\|w)$	w_R	$w_{\neg R}$
m_R	0.625	0
$m_{\neg R}$	0	≈ 0.714
m_D	0.375	≈ 0.286
$P_S^j(m\|w)$	w_R	$w_{\neg R}$
m_R	≈ 0.714	0
$m_{\neg R}$	0	≈ 0.770
m_D	≈ 0.286	≈ 0.230

Table 4: Speaker probabilities for an honest speaker for each listener.

We can use (12) and (13) to have a better idea of the behavior of the speaker. In this case, we obtain that $\mathcal{P}_S^i(m_D) \approx 0.351$ and $\mathcal{P}_S^j \approx 0.271$, so the use of m_D will be somewhat unexpected, but still more or less in keeping with the idea of such a speaker.

We now consider the duplicitous speaker S_{Dup}, who uses the literal listener values along with the μ^* function presented in table 6. The duplicitous speaker uses these in conjunction with (15), giving us the values in table 5. Using (12) and (13) again to have a better picture of how such a speaker could be predicted to act, we find that $\mathcal{P}_{S_{Dup}}(m_D) \approx 0.752$. Such a speaker is much more likely to use a dogwhistle!

$P_{S_{Dup}}(m\|\langle w, w'\rangle)$	m_R	$m_{\neg R}$	m_D
$\langle w_R, w_R\rangle$	$\approx$0.806	0	$\approx$0.194
$\langle w_R, w_{\neg R}\rangle$	0	0	1
$\langle w_{R\neg R}, w_R\rangle$	0	0	1
$\langle w_{R\neg R}, w_{\neg R}\rangle$	0	$\approx$0.806	$\approx$0.194

Table 5: Speaker probabilities for a duplicitous speaker following the μ^* function in table 6

		w^j	
μ^*		w_R	$w_{\neg R}$
w^i	w_R	0	2
	$w_{\neg R}$	0	1

Table 6: $\mu^*(\langle w^i, w^j\rangle)$ function used by the duplicitous speaker, their main objective is not to be seen as racist by L^j.

A savvy listener L^{Savvy} in this framework is simply a listener who assumes the duplicity of the speaker and bases their interpretation of the speaker message using $P_{S_{Dup}}$ instead of P_S. To compute the intention of the speaker, L^{Savvy} uses the P_W values used at the previous step by the duplicitous listener, following (16) leading to the numbers in table 7. The savvy listener interprets that when the dogwhistle is used there is a very high chance that the speaker is trying to appeal to audiences with opposing points of view!

$P_{L^{Savvy}}(< w, w' > \mid m)$	m_R	$m_{\neg R}$	m_D
$\langle w_R, w_R \rangle$	1	0	$\approx$0.021
$\langle w_R, w_{\neg R} \rangle$	0	0	$\approx$0.811
$\langle w_{R\neg R}, w_R \rangle$	0	0	$\approx$0.109
$\langle w_{R\neg R}, w_{\neg R} \rangle$	0	1	$\approx$0.058

Table 7: Interpretations for savvy listener of each message according to their priors about the speaker.

We argue that our model accounts for the main properties of dogwhistles in the following ways:

- INTERPRETATIVE VARIABILITY: the listeners do not assign the exact same interpretations to dogwhistles. In our example, L^i thinks it is just a bit more likely that m_D conveys a racial meaning rather than no racial meaning at all, and L^j has the opposite view.

- POLITICAL CONFLICT: the use of m_D only presents interest if there is a situation of political conflict, reflected in the duplicitous speaker preferences μ^*. In our example, L^j understanding w_R is what the speaker desires least; whereas, they want for L^i to understand w_R. In fact it can be shown that if we do not have political conflict in this sense, then the game reduces to a signaling game and the utility of using an ambiguous message like m_D is always lower than that of the other two m.

- IDENTITY-BASED: the system used here is identity-based given the fact that priors over the persona of the speaker have an influence on the interpretation function that will be favored (in our simple example, it fully determines it).

- PLAUSIBLE DENIABILITY: As long as members of the audience acknowledge that others might be from different social groups and use different interpretation functions, the meaning of the dogwhistle is never completely clear.

- SAVVY LISTENER: The savvy listener in our model is an individual who assumes that the speaker is being duplicitous and that they have motives beyond communicating information about the world. Otherwise, they use the same mechanisms as other speakers.

5 Conclusion

We have constructed a model that allows us to describe the processes behind the use of dogwhistles by using mechanisms that are already widely used in pragmatics to describe scalar implicatures and social meaning interpretation. We have assumed that a group of more "naïve" listeners use regular RSA-style computations to infer the probable meaning of a dogwhistles utterance while still taking into account the fact that the words used could have different meanings for other audiences. Duplicitous speakers willing to convey two different messages to audiences with different preferences and biases will use dogwhistles to do so and it is likely that they will be understood in the way they intended given what they were assuming of the crowd.

Meanwhile, savvy listeners assume that speakers are indeed duplicitous and conclude that the most likely interpretation after hearing a dogwhistle term is that the speaker is trying to appeal to two different audiences.

Following a similar path and adding iterations in the reasoning, we could easily model other roles in the dogwhistling game, including for example speakers who use dogwhistles in order to specifically trigger savvy listener reactions and then defend themselves from accusations of duplicity by calling out savvy listeners for *ad hominem/appeal to motive* positions.

We think that the framework used here could be generalized to other cases of identity-based interpretation, including cases outside the realm of political discourse, where the meaning intended by speakers is sometimes vague enough to trigger various interpretations from listeners.

References

Eric Acton. 2020. Pragmatics and the third wave. *Social Meaning and Linguistic Variation: Theorizing the Third Wave*.

Asif Agha. 2003. The social life of cultural value. *Language & communication*, 23(3-4):231–273.

Bethany L Albertson. 2015. Dog-whistle politics: Multivocal communication and religious appeals. *Political Behavior*, 37(1):3–26.

Nicholas Asher and Soumya Paul. 2018a. Bias in semantic and discourse interpretation. *arXiv preprint arXiv:1806.11322*.

Nicholas Asher and Soumya Paul. 2018b. Strategic conversations under imperfect information: epistemic message exchange games. *Journal of Logic, Language and Information*, 27(4):343–385.

Heather Burnett. 2017. Sociolinguistic interaction and identity construction: The view from game-theoretic pragmatics. *Journal of Sociolinguistics*, 21(2):238–271.

Heather Burnett. 2019. Signalling games, sociolinguistic variation and the construction of style. *Linguist and Philos*, 42:419–450.

Brian Robert Calfano and Paul A Djupe. 2009. God talk: Religious cues and electoral support. *Political Research Quarterly*, 62(2):329–339.

Penelope Eckert. 2008. Variation and the indexical field 1. *Journal of sociolinguistics*, 12(4):453–476.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41:87–100.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Robert E. Goodin and Michael Saward. 2005. Dog whistles and democratic mandates. *The Political Quarterly*, 76(4):471–476.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Robert Henderson and Elin McCready. 2019a. Dogwhistles and the at-issue/non-at-issue distinction. In *Secondary Content*, pages 222–245. Brill.

Robert Henderson and Elin McCready. 2019b. How dogwhistles work. In *JSAI International Symposium on Artificial Intelligence*, pages 231–240. Springer.

Jon Hurwitz and Mark Peffley. 2005. Playing the race card in the post–willie horton era: The impact of racialized code words on support for punitive crime policy. *Public Opinion Quarterly*, 69(1):99–112.

Justin Khoo. 2017. Code words in political discourse. *philosophical topics*, 45(2):33–64.

David Kuo. 2006. *Tempting faith: An inside story of political seduction*. Simon and Schuster.

David K. Lewis. 1969. *Convention: A Philosophical Study*. Wiley-Blackwell.

Tali Mendelberg. 2001. The race card: Campaign strategy. *Implicit Messages, and the*.

Jennifer Saul. 2018. Dogwhistles, political manipulation, and philosophy of language. *New Work on Speech Acts*, pages 360–383.

Jason Stanley. 2015. *How propaganda works*. Princeton University Press.

Ismail K White. 2007. When race matters and when it doesn't: Racial group differences in response to racial cues. *American Political Science Review*, 101(2):339–354.

Conditional answers and the role of probabilistic epistemic representations

Jos Tellings
Utrecht University / Utrecht Institute of Linguistics
Trans 10, 3512 JK Utrecht
the Netherlands
j.l.tellings@uu.nl

Abstract

Conditional utterances can be used in discourse as answers to regular, non-conditional questions in situations of partial knowledge of the answerer. I claim that the probabilities that interlocutors assign to each other's possible epistemic states are a measure of the relevance of conditional answers. A second criterion that makes a conditional answer 'if p, then q' relevant has to do with the dependency between p and q that is conveyed in the statement. A conditional answer counts as relevant when this dependency leads the question asker to shift from a decision problem about q to an alternative, easier, decision problem about p.

1 Introduction

1.1 Conditionals as answers

The study of conditional sentences (*if-then* sentences) constitutes a vast area within formal semantics. A lot of this work considers the internal structure of conditionals: how do tense, aspect, and modality contribute to the combination of pragmatic and semantic effects observed for conditional utterances? (see e.g. von Fintel, 2011 for an overview). The use of conditionals in conversation is relatively less well studied. Yet, corpus research shows that conditionals are a very common utterance type (Ferguson, 2001; for some preliminary corpus data from the Europarl corpus, see Tellings, 2020). A prominent view in semantic theories of discourse interaction is to model conversational contexts as a structured collection of (possibly implicit) information requests, and answers to these requests. These requests are called *questions under discussion* (QUDs, Roberts, 1996/2012), or *issues* in the framework of Inquisitive Semantics (Ciardelli et al., 2018). Hence, a natural start of a theoretically and computationally robust study of conditionals in discourse is to ask how conditional utterances behave as **answers to questions**.

Work on conditionals as answers to questions is surprisingly scarce, given that both question-answer models of discourse, and formal theories of the meaning of conditional sentences, have been major themes in the literature on semantics and pragmatics. I argue that the most important empirical setting to study is when conditionals answer regular, non-conditional questions. Indeed, we observe that conditional utterances make good answers to all types of non-conditional questions, including polar questions (1a), alternative questions (1b) and *wh*-questions (1c):

(1) a. A: Will John come to the party?
B: If he finishes his work, he will.

b. A: Do you want coffee or tea?
B: If it is freshly made, I would like coffee.

c. A: What will John cook for dinner?
B: If he managed to buy parmesan cheese, he will make pasta.

Here, the conditional form of the answer was not driven by the form of the question, but newly introduced by B on the basis of conditional knowledge he has.[1] What is worth noting about the cases in (1), is that B's answers are not the maximally informative *congruent* answers that are typically studied in semantic theories of questions (congruent answers would be 'yes'/'no' to a polar question, 'coffee' or 'tea' in (1b), and a constituent answer such as 'pasta' or 'steak' in (1c)). This is because the conversations in (1) crucially involve *partial knowledge*. In (1a), B does not know for

[1] Throughout, I will use A(lice) and female pronouns to refer to the question asker, and B(ob) and male pronouns to refer to the answerer.

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 26–33

a fact whether John will come, so the pragmatically most informative answers 'Yes' and 'No' are ruled out. However, he is not completely ignorant about the situation, he has *conditional knowledge*: B knows that if John finishes his work, he will come.

From these observations, my central research question emerges: in what situations do speakers give a conditional answer, and how do speakers choose a particular conditional answer as their optimal response in comparison to other conditional and non-conditional answer options?

These questions can be made more explicit by introducing some theoretical concepts on the relevance of answers. In game-theoretic models of Gricean pragmatics (see Benz and Stevens, 2018 for an overview), the relevance of an answer with respect to a given question has been formalized by modeling the question as a decision problem, and assigning potential answers a **utility value**. Different variants of game-theoretic pragmatics (e.g. older utility-based accounts (van Rooij, 2004), optimal answer models (Benz and van Rooij, 2007), and rational speech act (RSA) models (Goodman and Frank, 2016)) use the notion of utility in different ways, but the common view is that B chooses an answer that maximizes utility with respect to the decision problem A is trying to resolve.

Framed in these terms, the aim is to assign a utility value to conditional statements. By considering some specific examples, I will propose two probability measures that are proportional to the utility of a conditional answer. The first, discussed in section 2, is an epistemic condition related to the probabilities assigned to possible epistemic states of the interlocutors. The second, discussed in section 3, has to do with the dependency between the propositions expressed in *if*-clause and main clause. There we will also see that the simple approach of applying an existing utility-based framework to a material conditional proposition of the form '$p \supset q$' does not work.

Before all this, we will look at conditional answers from a somewhat different domain, conditional perfection (in §1.2), and show that two approaches to conditional answers in earlier literature do not meet the desiderata that I have outlined above (in §1.3).

1.2 Conditional perfection

An additional reason to study conditional answers comes from the pragmatic phenomenon of **conditional perfection**, or the strengthening of conditionals to biconditionals. The following datum from van Canegem-Ardijns and van Belle (2008) is an example:

(2) If you pay your contribution, you may participate in the barbecue.
implicature: if you don't pay, you may not participate

Conditional perfection is widely studied in the pragmatics literature (Geis and Zwicky, 1971; de Cornulier, 1983; von Fintel, 2001; van Canegem-Ardijns, 2010; among others), and various different mechanisms for deriving the implicature have been proposed (see e.g. van der Auwera, 1997 for an overview). A major question in the work on conditional perfection is why the implicature arises in some cases, such as (2), but not in others, such as (3) (from von Fintel, 2001):

(3) If this cactus grows native to Idaho, then it is not an *Astrophytum*.
$\not\rightarrow$ If this cactus doesn't grow native to Idaho, it is an *Astrophytum*.

In more recent work on conditional perfection, it has been proposed that perfection occurs when a conditional is interpreted as an exhaustive answer to the question under discussion (Herburger, 2015, cf. von Fintel, 2001; see Cariani and Rips, 2017 for an experimental approach to this idea).

(4) If you work hard you will succeed.
Exhaustification: ⟨. . . and only if you work hard you will succeed⟩
(Herburger, 2015)

In unrelated work on exhaustive answers, it has been proposed that whether an answer is interpreted as exhaustive or not (*mention-all* or *mention-some*) depends on "human concerns" underlying the asking of the question, which can again be modeled in game-theoretic pragmatical models in terms of the decision problem the speaker is trying to solve (van Rooij, 2004).

Hence, in order to understand conditional perfection better, we need to understand when conditionals are interpreted as mention-some answers,

and how conditional answers correspond to the speakers' interests. Therefore, a study of the utility of conditional answers with respect to the interlocutors' interests can contribute to the understanding of conditional perfection.

1.3 Earlier work

In order to further appreciate the approach to conditionals taken here – as answers to regular questions –, it is worth briefly reviewing earlier work on conditional answers. I will mention two lines of work, which however in my opinion are not representative of the wider problem of conditional sentences in discourse that I address here.

The first is work on conditional utterances as answers to conditional questions, as in example (5) from Isaacs and Rawlins (2008, 276):

(5) A: If Alfonso comes to the party, will Joanna leave?
B: If he comes, Joanna will leave.

The reason that these types of question-answer combinations are not very insightful for my purposes, is that here the conditional answer merely mimics the conditional form of the question. Hence, B will have had no independent grounds to choose a conditional form for his answer.

In the same vein, Ippolito (2013) proposes that counterfactual conditionals are answers to conditional questions under discussion (CQUDs), as the following example taken from her paper illustrates:

(6) [CQUD: If the weather had been fine, would Jones be wearing his hat?]
If the weather had been fine, Jones would be wearing his hat.

This illustrates my point that studying conditionals as answers leads to a better overall understanding of conditional statements in general, because these CQUDs are generally implicit, and Ippolito's work is not part of understanding (counterfactual) conditionals in discourse. However, the same point about form parallelism in question and answer can be made for (6). Moreover, Ippolito (2013) does not take into consideration that the *if*-clause and main clause of a conditional can have different information structural statuses, depending on how they are used in a dialogue context. In fact, both *if*-clause and main clause may be in focus (see e.g. Farr, 2011; Tellings, 2016, §4.4).

The second line of work I want to mention here is Hesse et al. (2018), because it is methodologically closer to what I aim to do (cf. also Stevens et al., 2016), but studies a different kind of conditional expression, namely *speech act conditionals*/SACs (also known as *biscuit conditionals*). They give the following example, illustrating 'positive', 'negative', and 'alternative' speech act conditionals as answers to a polar question:

(7) A: Is there a restaurant close to the apartment?
   a. B: If you enjoy eating out, there is an Italian restaurant in the street. [PSAC]
   b. B: If you enjoy eating out, there is an Italian restaurant in the neighboring quarter. [NSAC]
   c. B: If you enjoy eating out, there is an Italian restaurant as well as a food court nearby. [ASAC]

Hesse et al. provide a model based on this specific example of a client asking a real estate agent questions about an apartment that predicts when a SAC is generated. The model does not, however, address the issue of why the answer is expressed by a SAC, and not by some other linguistic construction. For example, the response to the question in (7) could also be expressed as in (8B):

(8) A: Is there a restaurant close to the apartment?
B: Ah, you like eating out? Yes, there is an Italian restaurant in the street.

Hesse et al. state that the choice between SACs in (7) and answers such as (8B) "depends on discourse-dependent and stylistic reasons" (p. 103), but do not elaborate.

In conclusion, both the work on conditional questions, and the work on speech act conditional answers, take a rather limited view on conditional answers, and does not take into account the general relationship between the two.

2 An epistemic licensing condition

I will start by looking at some specific situations in which conditional answers are licensed.

First, let me set out the boundary conditions as introduced in the previous section. I assume that A asks a question $?q$, and B answers 'if p, (then)

q', written as '$p \rightarrow q$'. It follows that $\neg K_A ?q$ (this is the standard assumption of speaker ignorance: A does not ask a question if she knows the answer to it already). As pointed out above, conditional answers are uttered in situations of partial knowledge, so I moreover assume that $\neg K_B ?q$ (otherwise, B would have given a complete answer 'yes' or 'no'). Finally, I assume that $\neg K_B ?p$ (this is a standard presumption on conditional utterances; if B knows whether p, he wouldn't utter a conditional sentence with 'if p'). These conditions can be seen in the following example of a licensed conditional answer:

(9) [Alice calls to the IT help desk]
A: Did I install my printer correctly?
B: If there is a printer icon on the desktop, you installed it correctly.

The three conditions introduced above are satisfied, because A doesn't know whether she installed her printer correctly, B doesn't know either (he is at a different place), and B doesn't know whether there is a printer icon on A's desktop (B doesn't have access to A's screen).

I claim that the felicity of B's answer in (9) has to do with the fact that B knows that A knows, or can easily verify, whether there is a printer icon on her desktop. More generally, I claim that the utility of a conditional answer depends on the probability that B assigns to that A knows about the antecedent p, more formally on $P_B(K_A ?p)$:[2] the higher this value is, the more useful the conditional answer. The specific context in (9) represents the 'extreme' case in which $P_B(K_A ?p) = 1$ (or sufficiently close to 1 by contextual standards), i.e. $K_B K_A ?p$. Here $K_B K_A ?p$ is a reasonable assumption, given that B knows that A has access to her screen and can easily verify the truth value of p.

In this epistemic setting, B entertains two possible candidates for A's epistemic state: one in which $K_A p$, and one in which $K_A \neg p$.[3] In the former case ($K_A p$), '$p \rightarrow q$' is a highly useful answer, because by modus ponens, A can conclude that q, and solve her decision problem $?q$. In the case that $K_A \neg p$, the answer '$p \rightarrow q$' is useful when it undergoes conditional perfection. In that case, A concludes that $\neg q$ by modus ponens. This links to the problem of characterizing the distribution of conditional perfection described in §1.2 above. This suggests the possibility of a second process that leads to conditional perfection: not only the exhaustive interpretation of a conditional answer (recall (4) above), but also a type of backwards reasoning on the part of A of the steps just described. Informally, A gets the perfection implicature, because she reasons that B's conditional answer is only relevant in the $\neg q$-state when perfection happens. Various details of this proposal about conditional perfection need to be worked out, which I leave aside for now.

Observe that a close variant of the conversation in (9) exists, in which instead of 'if p', an intermediate question $?p$ is uttered (B_1 below):

(10) A: Did I install my printer correctly?
B_1: Is there a printer icon on the desktop?
a. A: Yes.
B_2: Then you installed it correctly.
b. A: No.
B_2: Then there is a problem.[4]

This shows that there is a connection between conditionalizing an answer with 'if p' (in (9)), and asking an intermediate question $?p$ (in (10)). One of the default pragmatic conditions for a speaker A asking a question $?q$ to a speaker B, is ADDRESSEE COMPETENCE. This refers to the condition that A thinks that B is likely to know the answer, or in other words that $P_B(K_A ?q)$ is high. If you want to know what the French word for 'rhubarb' is, you better ask somebody who knows French, rather than somebody who doesn't know French. Of course, there are exceptions to this default rule (e.g. in exam questions), but in general, this assumption is often satisfied: either trivially in addressee-directed questions such as 'How are you?', 'What did you do yesterday?', etc., or in the case of more factual questions as in (9). Literature on the pragmatics of questions has mostly ignored the condition of ADDRESSEE COMPETENCE, pre-

[2] This mixture of probability and epistemic logic is formalized in models such as van Benthem et al. (2009). The '$?p$' notation comes from inquisitive epistemic logic (Ciardelli and Roelofsen, 2015), in which the equivalence $K_A ?p \leftrightarrow (K_A p \vee K_A \neg p)$ comes out as a logical validity, not an abbreviation.

[3] The relation between the epistemic representation '$K_A ?p$' and the two candidates for epistemic states of A that B entertains, is formally present in inquisitive epistemic logic (see fn. 2), since $K_A ?p$ abbreviates $K_A \{p, \neg p\}$. The same holds for other types of questions.

[4] Observe that in both (10a) and (10b), B_2 requires some sort of anaphoric expression 'then' or 'in that case'.

sumably because selecting which person you will ask your question to seems to be an issue that falls outside of linguistics proper. However, the data presented here show that the ADDRESSEE COMPETENCE condition ('$P_B(K_A?p)$ is high') is in fact linguistically relevant, because it doubles as a licensing condition for conditional answers with 'if p' in the given epistemic situation.

The context in (9), in which $P_B(K_A?p)$ takes its maximal value of 1, is perhaps a somewhat uncommon situation. It is instructive to consider cases toward the other end of the scale, where $P_B(K_A?p)$ is low, or indeed 0 (i.e. B knows for a fact that A does not know whether p). Two cases illustrating this epistemic situation are given below.

(11) A: Did it rain yesterday?
B: #If the atmospheric pressure was no higher than 1020 mBar and the squall line progression halted over Western Massachusetts, it did.

(12) [epistemic situation: B knows that A is unaware of John's work situation]
A: Will John come to the party? (=(1a))
B: If he finishes his work, he will.

Example (11) illustrates a situation in which a conditional answer is pragmatically odd (even though it may be true). Here B knows that A does not know whether the proposition in the *if*-clause is true, nor does A have an easy way to verify its truth or falsity (unless A is a professional meteorologist). Example (12), on the other hand, contains a conditional answer that may be licensed in the same epistemic situation without problem: imagine that B knows for a fact that A has no knowledge about John's work situation. This does not render the conditional answer unacceptable, despite the fact that the epistemic condition proposed in this section has not been satisfied.

I will come back to the difference between (11) and (12) in section 3, but to complete the line of argumentation, consider the following. That (12) is indeed licensed in the situation given above, and is different from the earlier example in (9), can be tested by using the connection between (9) and (10) as a diagnostic: for (12), changing the *if*-clause into an intermediate question does not work, see (13).

(13) [epistemic situation: B knows that A is unaware of John's work situation]
A: Will John come to the party?
#B_1: Did he finish his work?
⋮

In the epistemic context just sketched, ADDRESSEE COMPETENCE is violated for B_1.[5] The fact that (12) is nonetheless an acceptable conditional answer shows that in addition to the epistemic condition proposed in this section, there is a second way in which conditional answers can be licensed. This second condition, however, does not license intermediate questions.

The broader conclusion of this section is that the discussed data offer insights into the sort of information that interlocutors keep track of in the course of conversation. In a theory of discourse dynamics such as Farkas and Bruce (2010), interlocutors keep track of each other's discourse commitments. I have shown that this should also include the representations of each other's epistemic states, here represented by probability distributions over their epistemic commitments. Speakers have prior representations about each other's epistemic states, including certain and presumed knowledge about other speakers' knowledge. These representations get updated over the course of the conversation, as new information is provided and new issues are raised. Interlocutors keep track of what other speakers know, and reason about this probabilistically. This is an aspect of probability in meaning that, to the best of my knowledge, has not been addressed in earlier literature.

3 Conditional dependency as relevance

I claim that when the epistemic licensing condition from section 2 is not fulfilled, as in the setting of example (12), a conditional answer can still be relevant, namely when the conditional dependency between p and q that is conveyed by the conditional utterance is relevant information.

The problem in formalizing this dependency is that the belief in a conditional statement 'if p, (then) q' is not simply equal to the conditional probability of q given p, as Lewis (1976) famously showed. This is also the reason why many game-theoretic pragmatic accounts, which are based on

[5] A similar observation can be made for (11): the intermediate question 'Was the atmospheric pressure [...] Western Massachusetts?' is unacceptable.

probabilities of utterances, cannot be straightforwardly applied to conditional answers.

Van Rooij and Schulz (2019) propose an assertability condition for conditional utterances that takes these problems into account:

(14) Van Rooij and Schulz's (2019) assertability condition for conditionals '$p \rightarrow q$':

$$\Delta^* P_p^q := \frac{P(q|p) - P(q|\neg p)}{1 - P(q|\neg p)} \quad \text{should be high.}$$

This condition incorporates the idea that a conditional statement should convey a dependency between p and q. For example, it rules out cases such as:

(15) #If it is sunny today, Jan Ullrich won the Tour de France in 1997.

(15) comes out as true by virtue of its consequent being true, but the conditional is not assertable because there is no (causal) dependency between *if*-clause and main clause (see van Rooij and Schulz, 2019 for details on the link between conditional probabilities and causality).

Van Rooij and Schulz use their condition (14) as a criterion for asserting conditionals, but I claim it is also used in the other direction: updating one's belief state upon hearing a conditional answer.

Utility of a conditional Can this general notion of relevance – conditional dependency as relevant information – be expressed in terms of the notion of utility?

In utility-based frameworks, one starts with a utility function $U(a, w)$ that assigns a utility value to an action a in world w. Then, the notion of the expected utility of an action a given a proposition f is introduced (Benz and van Rooij, 2007):

$$\mathrm{EU}(a|f) = \sum_w P(w|f) \cdot U(a, w).$$

The next step is to define a notion of utility value of a message. This is done in different ways by different authors, but below is one proposal (Benz and van Rooij, 2007, 67):

$$\mathrm{UV}(f) = \max_i \mathrm{EU}(a_i|f).$$

I argue that one cannot simply represent the utility of a conditional answer as $\mathrm{UV}(p \supset q)$, i.e. by computing the utility value of the material conditional. I will identify three problems with this approach.

The first problem is that the truth conditions of the material conditional do not include the notion of conditional dependency as in (14) (the difference between truth conditions and assertability conditions is made clear by examples such as (15), see van Rooij and Schulz, 2019 for further discussion).

A second issue is that the conditional answers we have seen so far do not look like conditional propositions, but rather like conditional speech acts (or *conditional assertions*, see Stalnaker, 2011).[6] This intuition is supported by the observation that the consequent of conditional answers need not be a full proposition, but can take the form of a fragment answer, or in the case of the polar question (16a), 'yes' or 'no':

(16) a. A: Will John come to the party?
B: If he finishes his work, yes.

b. A: Do you want coffee or tea?
B: If it is freshly made, coffee.

c. A: What will John cook for dinner?
B: If he managed to buy parmesan cheese, pasta.

In each case, the consequent fragment answer (underlined in (16)) is, by itself, a valid answer to A's question as a whole. This suggests that the function of the *if*-clause is to conditionalize the answering speech act, rather than forming a conditional proposition $p \supset q$.

Finally, a more general problem with the decision problem approach is that the utility function and the set of actions is usually considered to be a fixed part of the model (a typical definition of a decision problem is a triple $\langle (\Omega, P), \mathcal{A}, U \rangle$ in which $\mathcal{A}$ is the set of actions, and U the utility function; Benz and van Rooij, 2007, 66). For example, take the very familiar example (17), with a toy model for a corresponding utility function.

[6] There may be some terminological confusion here with the notion of speech act conditionals / SACs that was mentioned above in relation to the work of Hesse et al. (2018). A SAC (maybe better called *biscuit conditional*) is a specific type of conditional sentence in which the consequent is factual, and does not depend on the truth of the *if*-clause. A conditional speech act / conditional assertion relates to the more general position that uttering a conditional sentence is not a single speech act that expresses a conditional proposition, but rather a combination of two speech acts. See Stalnaker (2011) for further discussion, cf. Ebert et al. (2014).

(17) Where can I buy an Italian newspaper?

P	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
W	w_1 : only @X	w_2 : only @Y	w_3 : @X&Y
a_X	6	0	4
a_Y	0	6	4

Here the set of actions $\mathcal{A} = \{a_X, a_Y\}$ (a_X: go to place X; a_Y: go to place Y) is fixed as part of the model. This is an unnatural assumption, because the asker of (17) most likely didn't have places X and Y in mind when she asked the question (otherwise something like 'Should I go to X or Y to buy an Italian newspaper?' would be more natural).

This problem also appears when a conditional answers a non-conditional question. For a decision problem representing a polar question $?q$, Alice will only distinguish between q-worlds and non-q-worlds for her utility function. She is not aware that the utility of her actions depends on p. In other words, Alice's U-function will not distinguish between a world w_1 in which p and $\neg q$ are true, and a world w_2 in which $\neg p$ and $\neg q$ are true. However, the conditional answer does distinguish between such worlds. In fact, the dependency between p and q is what is conveyed by the answer, and makes it relevant.

Switching decision problems Instead of the above-mentioned approach of applying existing utility-based frameworks to conditional propositions, I argue that conditional answers can be used to indicate that speaker A's original decision problem $?q$ can be reduced to a different decision problem $?p$ that is easier to resolve.

In (12), Alice's original decision problem was whether John came to party or not. Bob does not have full information to directly resolve A's decision problem, but does have conditional knowledge about a dependency between John's work and his coming to the party. By giving the conditional answer, he conveys to Alice that there is an alternative way to resolve her problem, namely by finding out about the progress of John's work.

The difference between (11) and (12) from section 2 can now be understood: whereas (12) allows to shift to an alternative decision problem that is (potentially) easier to resolve, finding out about the meteorological facts in B's answer in (11) is more difficult than the resolving the original problem of whether it rained. In other words, the answer in (11) invites A to shift to an alternative decision problem that is more difficult than the original one, rendering it uncooperative.

Hence, the study of conditional answers argues for a dynamic turn in utility-based pragmatics of question-answer pairs: answers can lead to updating decision problems (and utility functions alongside) in the course of the conversation. As far as I know, such a dynamic take on utilities has not been proposed before. A way to formalize this idea is work in progress, by employing a probabilistic dynamic semantics (Yalcin, 2012), and combining probabilistic belief update with the utility function.

4 Conclusion

I have outlined various reasons for investigating conditional utterances from the perspective of answers to questions: understanding the use of conditionals in discourse, their information-structural properties, and the distribution of conditional perfection. Then I outlined some specific examples in which conditional answers are licensed, and argued that interlocutors have representations of each other's epistemic states. They update, and reason probabilistically about these representations in the course of the conversation. The utility of a conditional answer is measured by $P_B(K_A?p)$, but there is a second way in which the information conveyed by a conditional utterance counts as relevant, over and above the epistemic condition. Learning about the conditional dependency between p and q is relevant for A in the process of resolving her decision problem '$?q$', when this dependency allows her to switch from the original problem '$?q$' to an alternative decision problem '$?p$' that is easier to resolve.

References

Johan van der Auwera. 1997. Pragmatics in the last quarter century: The case of conditional perfection. *Journal of Pragmatics*, 27:261–274.

Johan van Benthem, Jelle Gerbrandy, and Barteld Kooi. 2009. Dynamic update with probabilities. *Studia Logica*, 93(1):67–96.

Anton Benz and Robert van Rooij. 2007. Optimal assertions, and what they implicate. A uniform game theoretic approach. *Topoi*, 26:63–78.

Anton Benz and Jon Stevens. 2018. Game-theoretic approaches to pragmatics. *Annual Review of Linguistics*, 4:173–191.

Ingrid van Canegem-Ardijns. 2010. The indefeasibility of the inference that if not-A, then not-C. *Journal of Pragmatics*, 42(1):1–15.

Ingrid van Canegem-Ardijns and William van Belle. 2008. Conditionals and types of conditional perfection. *Journal of Pragmatics*, 40:349–376.

Fabrizio Cariani and Lance J. Rips. 2017. Experimenting with (conditional) perfection. Ms.

Ivano Ciardelli, Jeroen Groenendijk, and Floris Roelofsen. 2018. *Inquisitive semantics*. Oxford University Press.

Ivano A. Ciardelli and Floris Roelofsen. 2015. Inquisitive dynamic epistemic logic. *Synthese*, 192(6):1643–1687.

Benoît de Cornulier. 1983. *If* and the presumption of exhaustivity. *Journal of Pragmatics*, 7:247–249.

Christian Ebert, Cornelia Ebert, and Stefan Hinterwimmer. 2014. A unified analysis of conditionals as topics. *Linguistics & Philosophy*, 37(5):353–408.

Donka Farkas and Kim Bruce. 2010. On reacting to assertions and polar questions. *Journal of Semantics*, 27:81–118.

Marie-Christine Farr. 2011. Focus Influences the Presence of Conditional Perfection: Experimental Evidence. In Ingo Reich, Eva Horch, and Dennis Pauly, editors, *Proceedings of Sinn & Bedeuting 15*, pages 225–239. Universaar – Saarland University Press, Saarbrücken.

Gibson Ferguson. 2001. If you pop over there: a corpus-based study of conditionals in medical discourse. *English for Specific Purposes*, 20(1):61–82.

Kai von Fintel. 2001. Conditional strengthening: a case study in implicature. Ms., MIT.

Kai von Fintel. 2011. Conditionals. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Semantics. An International Handbook of Natural Language Meaning. Volume 2*, pages 1515–1538. De Gruyter.

Michael Geis and Arnold Zwicky. 1971. On invited inferences. *Linguistic Inquiry*, 2(4):561–566.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Elena Herburger. 2015. Conditional Perfection: the truth and the whole truth. In Sarah D'Antonio Mary Moroney and Carol Rose Little, editors, *Proceedings of SALT 25*, pages 615–635. LSA.

Christoph Hesse, Maryam Mohammadi, Maurice Langner, Judith Fischer, Anton Benz, and Ralf Klabunde. 2018. Communicating an understanding of intention: Speech act conditionals and modified numerals in a Q/A system. In Laurent Prévot, Magalie Ochs, and Benoît Favre, editors, *Proceedings of SemDIAL 2018 (AixDial)*, pages 103–111.

Michela Ippolito. 2013. Counterfactuals and Conditional Questions under Discussion. In Todd Snider, editor, *Proceedings of SALT 23*, pages 194–211. LSA.

James Isaacs and Kyle Rawlins. 2008. Conditional Questions. *Journal of Semantics*, 25:269–319.

David Lewis. 1976. Probabilities of conditionals and conditional probabilities. *The Philosophical Review*, 85(3):297–315.

Craige Roberts. 1996/2012. Information Structure: Towards an Integrated Formal Theory of Pragmatics. *Semantics & Pragmatics*, 5:6:1–69. Published version of a 1996 manuscript with the same title.

Robert van Rooij. 2004. Utility of Mention-Some Questions. *Research on Language and Computation*, 2:401–416.

Robert van Rooij and Katrin Schulz. 2019. Conditionals, causality and conditional probability. *Journal of Logic, Language and Information*, 28:55–71.

Robert Stalnaker. 2011. Conditional propositions and conditional assertions. In Andy Egan and Brian Weatherson, editors, *Epistemic Modality*, pages 227–248. Oxford University Press, Oxford.

Jon Scott Stevens, Anton Benz, Sebastian Reuße, and Ralf Klabunde. 2016. Pragmatic question answering: A game-theoretic approach. *Data & Knowledge Engineering*, 106:52–69.

Jos Tellings. 2016. *Counterfactuality in discourse*. Ph.D. thesis, UCLA.

Jos Tellings. 2020. Translation mining in the domain of conditionals: first results. Presentation at *Computational Linguistics in the Netherlands* (CLIN) 30.

Seth Yalcin. 2012. Context probabilism. In Maria Aloni, Vadim Kimmelman, Floris Roelofsen, Galit W. Sassoon, Katrin Schulz, and Matthijs Westera, editors, *Logic, language and meaning. 18th Amsterdam Colloquium*, pages 12–21. Springer.

Linguistic interpretation as inference under argument system uncertainty: the case of epistemic *must*

Brandon Waldon
Stanford University Department of Linguistics, 450 Jane Stanford Way
Stanford, CA 94305 USA
bwaldon@stanford.edu

Abstract

Modern semantic analyses of epistemic language (incl. the modal *must*) can be characterized by the 'credence assumption': speakers have full certainty regarding the propositions that structure their epistemic states. Intuitively, however: a) speakers have graded, rather than categorical, commitment to these propositions, which are often never fully and explicitly articulated; b) listeners have higher-order uncertainty about this speaker uncertainty; c) *must* ϕ is used to communicate speaker commitment to some conclusion ϕ and to indicate speaker commitment to the premises that condition the conclusion. I explore the consequences of relaxing the credence assumption by extending the argument system semantic framework first proposed by Stone (1994) to a Bayesian probabilistic framework of modeling pragmatic interpretation (Goodman and Frank, 2016).[1]

1 Introduction

Natural language contains a variety of means for expressing one's epistemic state. The best-studied of these in the semantics literature are the epistemic modal auxiliaries *must* and *may/might*:

(1) a. Ann: *Where is Peter?*
b. Mary: *He* {*may/might/must*} *be in his office.*

There is broad agreement in the literature that Mary's response in (1b) is comprised in part by a conclusion - *Peter is in his office* - the 'prejacent' over which the modal takes semantic scope. Additionally, the consensus is that the epistemic modal expresses a connection between the prejacent and a set of salient premises - most commonly, things that are known and/or assumed by the speaker and/or her interlocutor. Roberts (2019) notes that the details beyond these points of agreement are matters of debate; in particular, theoreticians disagree over the following two questions:

1. How do we specify the premises - the body of information, assumptions, or other contextually-supplied propositions which condition a modalized statement?

2. In what way are the premises related to the conclusion ϕ encoded as the prejacent of a modalized statement?

A well-discussed desideratum of a successful theory of epistemic modality is that it should provide an understanding of the perceived weakness of *must*. An observation going back to Karttunen (1972) is that modalized statements of the form *must* ϕ appear to mark weak speaker commitment to the prejacent compared to the unmodalized counterpart, 'bare' ϕ. The observation stems from consideration of contexts such as (2):

(2) (In the context of direct observation of rain):
a. # It must be raining outside.
b. It is raining outside.

Answers to Roberts' original questions fall into three categories. Restricted quantificational accounts (Kratzer, 1991) posit that the set of premises includes propositions known to a contextually-salient individual (most often the speaker) or group of individuals (containing the speaker and her interlocutor), as well as a contextually-specified set of assumptions which further restrict the space of epistemic possibilities. *Must/might* ϕ quantify universally and existentially over this space, respectively: *must* expresses that the conclusion is true at all possible worlds where the known and assumed propositions are true; *might* expresses that the conclu-

[1]I gratefully acknowledge Cleo Condoravdi, Judith Degen, Atticus Geiger, Daniel Lassiter, Christopher Potts, three PaM reviewers, and Stanford's Construction of Meaning workshop for valuable feedback. All errors are mine.

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 34–40

sion is true at at least one of those worlds. Perceived weakness of *must* is accounted for on this analysis because *must* ϕ - unlike bare ϕ - allows for the possibility that $\neg\phi$ is true in worlds compatible with the known propositions (but incompatible with the assumed ones). In contrast, unrestricted quantificational accounts (von Fintel and Gillies, 2010) posit that *must* quantifies over a space of epistemic possibilities that is unconstrained by contextually-salient assumptions. On this approach, *must* ϕ is incompatible with $\neg\phi$, and *must*'s infelicity in (2) is a consequence of a violation of independently-stipulated felicity conditions.[2] Finally, probabilistic accounts (Swanson, 2006; Lassiter, 2016) vary in their commitments regarding how to specify the premises but consider *must/might* to be operators which take as their input the premises and output a statement about the likelihood of the truth of the prejacent.

All of the above approaches can be characterized by what I call the credence assumption - that is, that however we specify the premises and their relation to the conclusion, speakers have full certainty about the premises upon which they can rely for the purposes of inference in a given context. This assumption is desirable from the standpoint of analytic simplicity; moreover, it provides a way of analyzing modal disagreement, as in (3):

(3) a. Ann: *It might be raining outside.*

b. Mary: *No, it cannot be raining outside!*

On the credence assumption, Ann makes a statement regarding the possibility of rain given a set of known and/or assumed propositions. Mary assesses this statement and disagrees: she has a different (yet also deterministic) understanding of the premises operable in this discourse.[3]

Intuitively, however, speakers' epistemic states are much more complex than the credence assumption allows. Namely, these states involve graded, rather than deterministic, commitments to propositions that are often never fully and explicitly articulated by speakers who produce statements of the form *must/might* ϕ. Listeners in turn have uncertainty about the premises which their interlocutors deem to be relevant for the purposes of inference and deliberation; indeed, *must/might* ϕ is informative not just because it conveys speaker's epistemic commitment to the prejacent but also because it conveys something about the speaker's underlying knowledge and assumptions about the world, and about how she is likely to use available information in the future.

I develop a quantitative framework of modeling interpretation of *must* that relaxes the credence assumption. In doing so, I offer a formal means of representing how these constructions can be informative with respect to speaker commitment to the conclusion as well as with respect to the premises that the speaker believes are operable in context. My point of departure is the argument system semantic framework of Stone (1994), followed by a probabilistic enrichment of that framework rooted in a Bayesian understanding of linguistic inference (Goodman and Frank, 2016). On this approach, communication proceeds between agents who are uncertain about what premises can (and should) be relied upon for the purposes of present and future inference and deliberation. Interlocutors align this uncertainty in part via communicative exchange. Finally, I consider implications of this approach for our understanding of conversational dynamics and the common ground.

2 Argument system semantics

Stone (1994) observes that in contexts such as (1), the *must*-variant of Mary's response is infelicitous if the context does not make clear (to Ann) the basis on which Mary's conclusion is made: Mary's conclusion about Peter may come from the fact that Mary has ruled out all possible other places Peter could be, or perhaps it is 3pm on a Tuesday and Peter is always in his office at that time. If Ann cannot recover Mary's argument in support of the conclusion, then *must* is infelicitous in (1). On Stone (1994)'s semantics, *must* ϕ is true iff a (possibly defeasible) argument A - made somehow salient in the context - justifies concluding ϕ given an argument system $\mathcal{K}$. I recapitulate the relevant details of Stone's analysis below.[4]

2.1 Formal preliminaries

Let $\mathcal{K}$ be an argument system, comprised of a set of established propositions K (ground formulae

[2] von Fintel and Gillies (2010), for example, contend that *must* ϕ presupposes that ϕ has not yet been settled in context.

[3] Or perhaps Mary agrees with Ann on the premises but disagrees regarding their relationship to the conclusion.

[4] I focus on Stone's argument system semantics because his formalism provides a way of verifying the relationship between a conclusion and the premises that condition it. This is crucial for my analysis, which captures how listeners infer speaker beliefs about premises having only observed conclusions asserted by the speaker. But similar results could be achievable with other semantic 'backends'.

K_C and logical rules of inference K_N) and a set of defeasible inferential rules Δ. Arguments for ground formulae are defined as follows:[5]

Definition 1: A set *T* of instantiations of elements of Δ is an ARGUMENT for *h* ($\langle T, h\rangle_{\mathcal{K}}$) iff: (1) $K \cup T \vdash h$; (2) $K \cup T \nvdash \bot$; and (3) for no $T' \subset T, K \cup T' \vdash h$.

The first clause of Definition 1 specifies that an argument for a ground formula (comprised of elements of Δ), added to K, entails the formula; the second specifies that the argument must be consistent with K; the third specifies that the argument must be minimal. Stone also introduces the notion of a sub-argument: an argument which can be computed from the premises of another argument:

Definition 2: $\langle S, j\rangle_{\mathcal{K}}$ is a SUBARGUMENT of $\langle T, h\rangle_{\mathcal{K}}$ if and only if $S \subseteq T$.

Stone emphasizes that subarguments of $\langle T, h\rangle_{\mathcal{K}}$ need not play a role in concluding h. Rather, the set of subarguments for h include all arguments which can be generated from T given argument system $\mathcal{K}$. This means that counterarguments to $\langle T, h\rangle_{\mathcal{K}}$ can do so by targeting not only subarguments necessary to conclude h from T but any inference generated from T given $\mathcal{K}$.

Definition 3: $\langle T_1, h_1\rangle_{\mathcal{K}}$ COUNTERARGUES $\langle T_2, h_2\rangle_{\mathcal{K}}$ at $\langle T, h\rangle_{\mathcal{K}}$ if and only if $\langle T, h\rangle_{\mathcal{K}}$ is a subargument of $\langle T_2, h_2\rangle_{\mathcal{K}}$ and $K \cup \{h, h_1\} \vdash \bot$.

A counterargument defeats an argument if the former is more specific - if it "takes more particulars of the context into consideration" (1994: 6).

Definition 4: $\langle T_1, h_1\rangle_{\mathcal{K}}$ is more SPECIFIC than $\langle T_2, h_2\rangle_{\mathcal{K}}$ if and only if: (1) for all ground formulae *e*, if $K_N \cup \{e\} \cup T_1 \vdash h_1$ but $K_N \cup \{e\} \nvdash h_1$, then $K_N \cup \{e\} \cup T_2 \vdash h_2$; and (2) there is some ground e such that $K_N \cup \{e\} \cup T_2 \vdash h_2$, $K_N \cup \{e\} \cup T_1 \nvdash h_1$, and $K_N \cup \{e\} \nvdash h_2$.

Definition 5: $\langle T_1, h_1\rangle_{\mathcal{K}}$ DEFEATS $\langle T_2, h_2\rangle_{\mathcal{K}}$ if $\langle T_1, h_1\rangle_{\mathcal{K}}$ counterargues $\langle T_2, h_2\rangle_{\mathcal{K}}$ at $\langle T, h\rangle_{\mathcal{K}}$ and $\langle T_1, h_1\rangle_{\mathcal{K}}$ is more specific than $\langle T, h\rangle_{\mathcal{K}}$

The first clause of Definition 4 states that the conclusion of the less specific argument h_2 must be entailed by the argument system coupled with the argument's defeasible premises T_2, provided the argument system is one in which the more specific argument's conclusion h_1 only follows with the addition of its premises T_1. The second clause states that there must be some argument system which entails the conclusion of the less specific argument h_2 on the basis of the premises T_2 but is inconsistent with the conclusion of the more specific argument h_1 on the basis of premises T_1.

An argument system justifies an argument "whenever [the argument] has no counterarguments which are not themselves defeated" (1994: 6). To formalize this, Stone introduces the concepts of supporting arguments and interfering arguments, defined inductively to capture the fact that for an argument to be justified it must not be defeated at any level of sub-argumentation.

Definition 6: All arguments are level 0 supporting and interfering arguments.

- An argument $\langle T_1, h_1\rangle_{\mathcal{K}}$ is a level $(n+1)$ supporting argument if and only if no level *n* interfering argument counters it at any of its subarguments.
- An argument $\langle T_1, h_1\rangle_{\mathcal{K}}$ is a level $(n+1)$ interfering argument if there is no level n interfering argument which defeats it.

Definition 7: An argument $\langle T, h\rangle_{\mathcal{K}}$ JUSTIFIES h in $\mathcal{K}$ if and only if there is some m such that for all $n \geq m$, $\langle T, h\rangle_{\mathcal{K}}$ is a level n supporting argument. h is justified in $\mathcal{K}$ if some $\langle T, h\rangle_{\mathcal{K}}$ justifies it in $\mathcal{K}$.

2.2 Illustration

First, let $\mathcal{K}_0$ be an argument system consisting of ground formulae K_{0_C}, logical rules K_{0_N}, and defeasible rules Δ_0.[6] Assume K_{0_N} contains a forward chain inferential rule, and Δ_0 consists of the following defeasible rules about matches and heat:

A: `match(x) ∧ strike(x) > lit(x)`
B: `match(x) ∧ strike(x) ∧ wet(x) > ¬lit(x)`
C: `lit(x) > hot(x)`

Let K_{0_C} contain two ground formulae `match(m1)` and `strike(m1)`. By Definition 1, we generate two arguments, A_1 and A_2:

A_1: $\langle\{\text{A}\}, \texttt{lit(m1)}\rangle_{\mathcal{K}_0}$
A_2: $\langle\{\text{A}, \text{C}\}, \texttt{hot(m1)}\rangle_{\mathcal{K}_0}$

Now, consider Stone's semantics for *must*:

(4) *Must* ϕ is true in $\mathcal{K}$ iff $\mathcal{K} \vDash \langle A, \phi\rangle_{\mathcal{K}}$

That is, *Must* ϕ is true if a contextually-salient argument *A* justifies concluding ϕ in a given argument system. Given $\mathcal{K}_0$, (5) is true:

(5) The match must have lit.

[5] All definitions below can be found in Stone (1994): p. 6.

[6] This example is based largely on one from Stone (1994).

Note that the semantics of *must* requires (in the case of 5) that the argument (A_1) is contextually-salient; otherwise, *must* is undefined. Assuming this condition is met, we can verify the truth of (5) by considering, by Definition 7, whether concluding `lit(m1)` from A_1 is justified in $\mathcal{K}_0$. It is: the only arguments that could interfere would have as their conclusion $\neg$`lit(m1)`, but these arguments cannot be generated from $\mathcal{K}_0$ because `wet(x)` is not in K_{C_0} (and Rule B cannot be invoked).

Now consider a second argument system $\mathcal{K}_1$, which differs minimally from $\mathcal{K}_0$ in that `wet(m1)` *is* an additional ground formula. Thus, A_3 is generated in addition to A_1 and A_2:

A_3: $\langle \{\text{B}\}, \neg\texttt{lit(m1)} \rangle_{\mathcal{K}_1}$

In $\mathcal{K}_1$, (5) is false. Note first that the only argument from which `lit(m1)` can be concluded is A_1. Thus, as in $\mathcal{K}_0$, (5) can only be true if A_1 is justified. It is not in $\mathcal{K}_1$: A_3 defeats A_1 because the former is more specific than the latter.

2.3 Interim discussion

Stone's system provides a straightforward account of *must*'s perceived weakness, if we can assume that direct observation of h adds h to the set of ground formulae by default. Consider Definition 1: an argument for h in $\mathcal{K}$ must have as its premises the minimal set of defeasible rules of inference which - coupled with the set of established ground formulae and the logical rules of inference in $\mathcal{K}$ - entails h. If h is already in the set of ground formulae, then the minimal set of required premises is empty. The prediction is that there is no argument A that can meet the definedness conditions of *must* ϕ if ϕ is already established in $\mathcal{K}$.

Argument systems and speaker epistemic states are assumed to be one and the same on this analysis, meaning that this analysis can be characterized by the credence assumption. We might imagine one speaker whose epistemic state is akin to argument system $\mathcal{K}_0$, and another whose state is akin to $\mathcal{K}_1$. On this analysis, it is clear why these two agents might disagree over (5): the statement is true given the former argument system and false given the latter. Below, I explore the properties of an extension that relaxes the credence assumption.

3 Probabilistic argument system semantics

In the context of Stone's analysis, relaxing the credence assumption amounts to revising our assumptions regarding the speaker's relationship to $\mathcal{K}$. I define a space of possible argument systems Z, which I allow to vary according to their ground formulae and defeasible rules of inference. I assume that speakers are uncertain as to what precise argument system is the relevant one for the purposes of inference and decision making in context. That is, there may be some uncertainty as to whether particular ground formulae can be taken to be true at the world of evaluation, or there may be uncertainty as to whether certain defeasible rules of inference may (or should) be employed in a particular context. I assume that Z is specified such that the ground formulae that the speaker considers likely to be true are in many (but not perhaps not all) of the elements of Z; the same is assumed modulo the defeasible rules of inference.[7]

We can then define speaker commitment to a proposition ϕ on the basis of some argument A as the likelihood that $\langle A, \phi \rangle$ justifies ϕ given possible argument systems in Z. *Must* ϕ, then, is a comment on this likelihood value: if the likelihood exceeds a certain contextually-specified threshold, then the statement is true:

(6) *Must* ϕ is true in Z iff $\text{P}\big(\langle A, \phi \rangle\big)_Z > \theta$,
where $\text{P}\big(\langle A, \phi \rangle\big)_Z = \frac{\sum_{\mathcal{K} \in Z} \mathcal{K} \vdash \langle A, \phi \rangle_{\mathcal{K}}}{|Z|}$

Speakers produce *must* ϕ to convey their degree of belief that ϕ is a valid conclusion on the basis of an argument A, given their argument system uncertainty. But importantly, the precise nature of this argument system uncertainty - the precise value of Z - is not transparent to the listener: the listener has prior beliefs about possible values of Z that are updated according to the conclusions that a speaker draws (and argumentation that she employs to draw those conclusions) in a particular context. Observation of *must* ϕ, then, allows the listener to update her uncertainty about the speaker's Z distribution.

The speaker's production of *must* ϕ is determined by a utility function of utterances given intended meanings that balances informativity against production cost. Following Goodman and Frank (2016), the model of a pragmatic speaker S_1 is defined partly in reference to a literal L_0 listener whose interpretations are a function of utterances' literal truth/falsity given possible intended

[7] That is, on this analysis, every individual argument system is assumed to have uniform probability. Alternatively, as a reviewer suggests, one could suppose that some elements of Z are more likely than others a priori (and that the truth conditions of *must* ϕ are sensitive to this non-uniform prior).

meanings. The space of possible meanings that the speaker could try to convey are possible valuations of the speaker's Z distribution (from which - by 6 - the speaker's commitment to ϕ can be computed).

The literal L_0 listener, then, can be modeled as a conditional probability distribution over possible valuations of Z given observation of some utterance u and a contextually-supplied value for the probability threshold θ. Following Lassiter and Goodman (2013), who model interpretation of gradable adjectives using a threshold-based semantics, I assume that this θ variable is passed from L_0 and eventually estimated by the pragmatic L_1 listener (defined below) from a prior distribution over values of θ.

$$P_{L_0}(Z|u,\theta) = P_{L_0}(Z|[\![u]\!]^\theta = 1) \times P(Z)$$

The pragmatic speaker selects utterances to convey intended meanings according to their contextual informativeness for L_0 as well as the cost of utterance production. Below, α is a speaker optimality parameter, and C is a cost function defined for all possible utterance choices: all else equal, the greater $C(u)$, the lower the probability that S_1 selects u to convey a particular message.

$$P_{S_1}(u|Z,\theta) \propto \exp(\alpha \times \log(P_{L_0}(Z|u,\theta)) - C(u))$$

The pragmatic listener L_1's interpretations of utterances are a function of expected behavior of S_1, as well as prior expectations about the likelihood of different possible meanings and prior expectations about the threshold value θ:

$$P_{L_1}(Z,\theta|u) \propto P_{S_1}(u|Z,\theta) \times P(Z,\theta)$$

Thus, interpretation of *must* is a joint inference about the state of the world (speaker beliefs regarding the justifiability of concluding ϕ) and the value of a semantic threshold variable θ. These speaker beliefs - $\mathrm{P}\big(\langle A, \phi\rangle\big)_Z$ - can be computed given values the argument variable supplied categorically by the context (A) and one additional variable (Z) which is inferred under uncertainty.

3.1 Illustration

For this illustration, I assume that the listener has uniform prior beliefs over the threshold value θ and that she considers two possible utterance production choices: *must* - whose truth conditions are as in (6) - and a trivially true null message.[8] The listener assumes that the speaker has full certainty about the following features of the argument system: the ground formulae (consisting of propositions `match(m1)`, `strike(m1)`, and `wet(m1)`), the logical rules of inference (including a forward chain operation), and a subset of the defeasible rules of inference (i.e. the listener assumes that Rule A features in every candidate argument system considered by the speaker). However, there are two other inferential rules - Rules B, and C from above - the status of which the speaker is uncertain: elements of Z may individually feature one, both, or neither of these rules.

For this illustration, assume that the listener has observed the speaker utter (5). Intuitively, this utterance conveys a high degree of speaker commitment to the prejacent (`lit(m1)`), but it should also convey something to the listener about the speaker's argument system uncertainty: since it is established that the speaker recognizes that `wet(m1)`, in uttering (5) the speaker has signalled that she finds it unlikely that Rule B is a relevant premise in this context.

The pragmatic listener must infer the value of Z under uncertainty; that is, she will not know the precise proportion of elements of Z that contain inferential Rules B and/or C (or neither). In other words, let β be the speaker's degree of belief that Rule B is in the contextually-relevant argument system; and let γ stand in for speaker beliefs about Rule C. The pragmatic listener updates her beliefs about the values of β and γ by observing the speaker's utterance production choices in context, in addition to inferring the value of the threshold θ. For this illustration, I assume uniform prior beliefs over values for β, γ, and θ and make the simplyfing assumption that $C(\textit{must})$ is equal to 1 while the null message has zero cost.[9] I arbitrarily set the optimality parameter α to 4.

In the computational implementation of this example, 10,000 samples are drawn from $P_{L_1}(\beta, \gamma, \theta|\text{must}(\texttt{lit(m1)}))$ using Markov Chain Monte Carlo sampling, with the assumption that the contextually-salient argument A is A_1.[10] Marginal posterior distributions over values for the inferred parameters are presented in Figure

[8] I follow Lassiter and Goodman (2013) in introducing this null utterance choice, which is an implementational necessity in the absence of utterance alternatives. Adding plausible linguistic alternatives to the model - including *might* ϕ and bare ϕ - does not drastically alter the patterns presented here.

[9] The prior distributions over values for β, γ, and θ are discrete distributions with uniform probability mass on 11 evenly-spaced values on the interval $[0, 1]$.

[10] The implementation was programmed using WebPPL (Goodman and Stuhlmüller, 2014). Code is available at https://github.com/bwaldon/probmust.

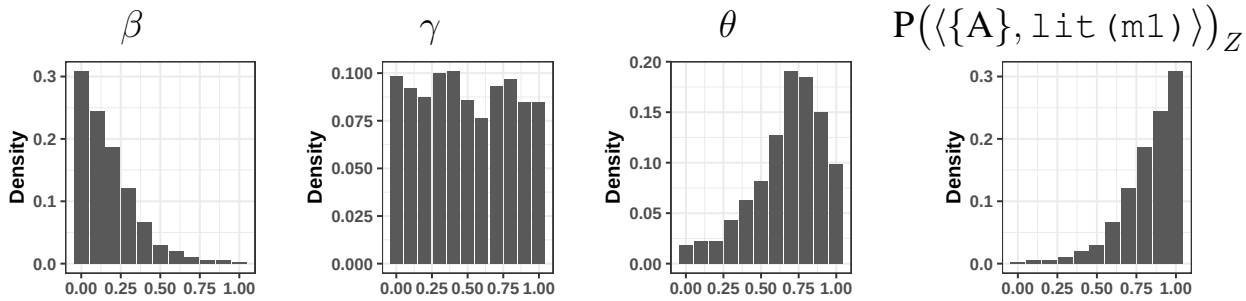


Figure 1: Marginal posterior distributions over values of β, listener beliefs about the speaker's expectation that Rule B is in the contextually-relevant argument system; γ, listener beliefs about the speaker's expectation that Rule C is in this state; θ, the threshold for *must*, and speaker commitment to `lit(m1)` on the basis of inferential rule A, calculated by approximating a posterior distribution over values of Z from posterior values of β and γ. Degree of speaker commitment is anti-correlated with β and not correlated with γ.

1. As a sanity check, we see that the posterior over values of γ is effectively uniform. This is exactly what is to be expected, as the inclusion of Rule C in the argument system has no bearing on the justifiability of concluding `lit(m1)`; thus, the speaker's production of (5) is not informative for the listener regarding the status of Rule C. However, the posterior over values of β suggests that the listener has learned something regarding the speaker's beliefs about Rule B.

Recall that the presence of this inferential rule in the argument system has the consequence that $\langle[\texttt{match(x)} \wedge \texttt{strike(x)} > \texttt{lit(x)}], \texttt{lit(m1)}\rangle$ is not justified (given our assumptions about the ground formulae and possible rules of inference from above). But *must*-`lit(m1)` was asserted on the basis of argument $\texttt{match(x)} \wedge \texttt{strike(x)} > \texttt{lit(x)}$; thus, after observing the speaker produce (5), the listener considers it relatively unlikely that the speaker expects Rule B - $\texttt{match(x)} \wedge \texttt{strike(x)} \wedge \texttt{wet(x)} > \neg\texttt{lit(x)}$ - to be a relevant inferential rule.

4 Discussion and conclusion

On this picture, disagreement functions differently than on analyses characterized by the credence assumption. On that assumption, we could understand disagreements over *must* ϕ as stemming from interlocutors' differences regarding their (deterministic) beliefs about the status of the premises or regarding the relationship of the premises to ϕ. The probabilistic enrichment explored here makes the story slightly more complicated. Consider the illustration above: a listener who hears a speaker utter (5) in this context is likely to disagree with that speaker, if the listener's own uncertainty involves high expectation that Rule B is relevant for the purposes of inference (and hence the listener has relatively low commitment to the prejacent, *the match is lit*). But what is the source of the disagreement? It cannot be that the listener knows definitively that she and the speaker have drastically different expectations regarding what inferential premises can be relied on in this context. However, the speaker's production of (5) is highly suggestive of such a difference: it is quite likely that the speaker puts relatively little weight in the chance that wet matches will light, even when struck. As a consequence, it is quite likely that the speaker has a high degree of belief that the match is lit. Disagreement, then, is triggered by the listener being fairly certain that her argument system uncertainty - her internal Z distribution - is substantially different from her interlocutor's.[11]

This suggests a way of understanding context and communicative exchange that complements the conventional "common ground" approach of Stalnaker (2002), whereby context records the propositions that interlocutors accept ('treat as true'), and communicative exchange involves proposals to update this common ground via addition of new propositions. In particular, my analysis suggests a means of formally modeling another layer of the context concerned with the uncertainty that interlocutors bring to bear on propositions not necessarily treated as categorically true. Epistemic linguistic constructions (e.g. *must*) facilitate coordination of this uncertainty between interlocutors, by communicating properties of this uncertainty from a particular epistemic vantage point.

[11] A more precise understanding of modal disagreement in this framework - for example, how do we quantify the conditions giving rise to disagreement? - is left to future work.

References

Kai von Fintel and Anthony S. Gillies. 2010. Must . . . stay . . . strong! *Natural Language Semantics*, 18(4):351–383.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Noah D Goodman and Andreas Stuhlmüller. 2014. The Design and Implementation of Probabilistic Programming Languages. http://dippl.org. Accessed: 2020-6-17.

Lauri Karttunen. 1972. Possible and must. In J. Kimball, editor, *Syntax and Semantics*, volume 1, pages 1–20. Academic Press, New York.

Angelika Kratzer. 1991. Modality. In Arnim von Stechow and Dieter Wunderlich, editors, *Semantics: An International Handbook of Contemporary Research*, volume 7, pages 639–650. de Gruyter, Berlin.

Daniel Lassiter. 2016. Must, knowledge, and (in)directness. *Natural Language Semantics*, 24(2):117–163.

Daniel Lassiter and Noah D Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Proceedings of Semantics and Linguistic Theory*, volume 23, pages 587–610.

Craige Roberts. 2019. The character of epistemic modality: Evidential indexicals. *Ms., Ohio State University*.

Robert Stalnaker. 2002. Common ground. *Linguistics and Philosophy*, 25(5/6):701–721.

Matthew Stone. 1994. The reference argument of epistemic must. In *Proceedings of the International Workshop on Computational Semantics*, volume 1, pages 181–190.

Eric Swanson. 2006. *Interactions with context*. Ph.D. thesis, Massachusetts Institute of Technology.

Linguists Who Use Probabilistic Models Love Them: Quantification in Functional Distributional Semantics

Guy Emerson
Department of Computer Science and Technology
University of Cambridge
gete2@cam.ac.uk

Abstract

Functional Distributional Semantics provides a computationally tractable framework for learning truth-conditional semantics from a corpus. Previous work in this framework has provided a probabilistic version of first-order logic, recasting quantification as Bayesian inference. In this paper, I show how the previous formulation gives trivial truth values when a precise quantifier is used with vague predicates. I propose an improved account, avoiding this problem by treating a vague predicate as a distribution over precise predicates. I connect this account to recent work in the Rational Speech Acts framework on modelling generic quantification, and I extend this to modelling donkey sentences. Finally, I explain how the generic quantifier can be both pragmatically complex and yet computationally simpler than precise quantifiers.

1 Introduction

Model-theoretic semantics defines meaning in terms of *truth*, relative to *model structures*. In the simplest case, a model structure consists of a set of *individuals* (also called *entities*). The meaning of a content word is a *predicate*, formalised as a *truth-conditional function* which maps individuals to *truth values* (either *truth* or *falsehood*). Because of this precisely defined notion of truth, model theory naturally supports logic, and has become a prominent approach to formal semantics. For detailed expositions, see: Cann (1993); Allan (2001); Kamp and Reyle (2013).

Mainstream approaches to distributional semantics represent the meaning of a word as a vector (for example: Turney and Pantel, 2010; Mikolov et al., 2013; for an overview, see: Emerson, 2020b). In contrast, Functional Distributional Semantics represents the meaning of a word as a truth-conditional function (Emerson and Copestake, 2016; Emerson, 2018). It is therefore a promising framework for automatically learning truth-conditional semantics from large datasets.

In previous work (Emerson and Copestake, 2017b, §3.5, henceforth E&C), I sketched how this approach can be extended with a probabilistic version of first-order logic, where quantifiers are interpreted in terms of conditional probabilities. I summarise this approach in §2 and §3.

There are four main contributions of this paper. In §4.1, I first point out a problem with my previous approach. Quantifiers like *every* and *some* are treated as precise, but predicates are vague. This leads to trivial truth values, with *every* trivially false, and *some* trivially true.

Secondly, I show in §4.2–4.4 how this problem can fixed by treating a vague predicate as a distribution over precise predicates.

Thirdly, in §5 I look at vague quantifiers and generic sentences, which present a challenge for classical (non-probabilistic) theories. I build on Tessler and Goodman (2019)'s account of generics using Rational Speech Acts, a Bayesian approach to pragmatics (Frank and Goodman, 2012). I show how generic quantification is computationally simpler than classical quantification, consistent with evidence that generics are a "default" mode of processing (for example: Leslie, 2008; Gelman et al., 2015).

Finally, I show in §6 how this probabilistic approach can provide an account of donkey sentences, another challenge for classical theories. In particular, I consider generic donkey sentences, which are doubly challenging, and which provide counter-examples to the claim that donkey pronouns are associated with universal quantifiers.

Taking the above together, in this paper I show how a probabilistic first-order logic can be associated with a neural network model for distributional semantics, in a way that sheds light on longstanding problems in formal semantics.

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 41–52

2 Generalised Quantifiers

Partee (2012) recounts how quantifiers have played an important role in the development of model-theoretic semantics, seeing a major breakthrough with Montague (1973)'s work, and culminating in the theory of *generalised quantifiers* (Barwise and Cooper, 1981; Van Benthem, 1984).

Ultimately, model theory requires quantifiers to give truth values to propositions. An example of a logical proposition is given in Fig. 1, with a quantifier for each logical variable. This also assumes a neo-Davidsonian approach to event semantics (Davidson, 1967; Parsons, 1990).

Equivalently, we can represent a logical proposition as a *scope tree*, as in Fig. 2. The truth of the scope tree can be calculated by working bottom-up through the tree. The leaves of the tree are logical expressions with free variables. They can be assigned truth values if each variable is fixed as an individual in the model structure. To assign a truth value to the whole proposition, we work up through the tree, quantifying the variables one at at time. Once we reach the root, all variables have been quantified, and we are left with a truth value.

Each quantifier is a non-terminal node with two children – its *restriction* (on the left) and its *body* (on the right). It quantifies exactly one variable, called its *bound variable*. Each node also has *free variables*. For each leaf, its free variables are exactly the variables appearing in the logical expression. For each quantifier, its free variables are the union of the free variables of its restriction and body, minus its own bound variable. For a well-formed scope tree, the root has no free variables. Each node in the tree defines a truth value, given a fixed value for each free variable.

The truth value for a quantifier node is defined based on its restriction and body. Given values for the quantifier's free variables, the restriction and body only depend on the quantifier's bound variable. The restriction and body therefore each define a set of individuals in the model structure – the individuals for which the restriction is true, and the individuals for which the body is true. We can write these as $\mathcal{R}(v)$ and $\mathcal{B}(v)$, respectively, where v denotes the values of all free variables.

Generalised quantifier theory says that a quantifier's truth value only depends on two quantities: the cardinality of the restriction $|\mathcal{R}(v)|$, and the cardinality of the intersection of the restriction and body $|\mathcal{R}(v) \cap \mathcal{B}(v)|$. Table 1 gives examples.

$$\forall x\ \mathit{picture}(x) \rightarrow$$
$$\exists z \exists y\ \mathit{tell}(y) \wedge \mathit{story}(z) \wedge \textsc{arg}1(y, x) \wedge \textsc{arg}2(y, z)$$

Figure 1: A first-order logical proposition, representing the most likely reading of *Every picture tells a story*. Scope ambiguity is not discussed in this paper.

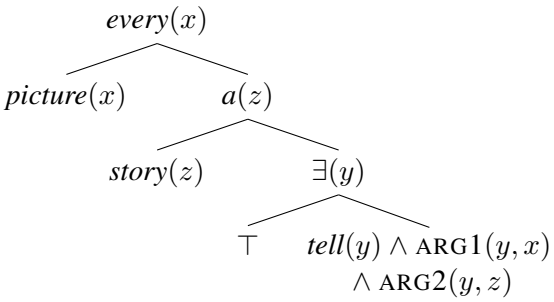


Figure 2: A scope tree, equivalent to Fig. 1 above. Each non-terminal node is a quantifier, with its bound variable in brackets. Its left child is its restriction, and its right child its body.

Quantifier	Condition
some	$\lvert\mathcal{R}(v) \cap \mathcal{B}(v)\rvert > 0$
every	$\lvert\mathcal{R}(v) \cap \mathcal{B}(v)\rvert = \lvert\mathcal{R}(v)\rvert$
no	$\lvert\mathcal{R}(v) \cap \mathcal{B}(v)\rvert = 0$
most	$\lvert\mathcal{R}(v) \cap \mathcal{B}(v)\rvert > \frac{1}{2}\lvert\mathcal{R}(v)\rvert$

Table 1: Classical truth conditions for precise quantifiers, in generalised quantifier theory.

3 Generalised Quantifiers in Functional Distributional Semantics

Functional Distributional Semantics defines a probabilistic graphical model for distributional semantics. Importantly (from the point of view of formal semantics), this graphical model incorporates a probabilistic version of model theory.

This is illustrated in Fig. 3. The top row defines a distribution over situations, each situation being an event with two participants.[1] This generalises a model structure comprising a *set* of situations, as in classical situation semantics (Barwise and Perry, 1983). Each individual is represented by a *pixie*, a point in a high-dimensional space, which represents the features of the individual. Two individuals could be represented by the same pixie, and the space of pixies can be seen as a conceptual space in the sense of Gärdenfors (2000, 2014).

[1] For situations with different structures (multiple events or different numbers of participants), we can define a family of such graphical models. Structuring the graphical model in terms of semantic roles makes the simplifying assumption that situation structure is isomorphic to a semantic dependency graph such as DMRS (Copestake et al., 2005; Copestake, 2009). In the general case, the assumption fails. For example, the ARG3 of *sell* corresponds to the ARG1 of *buy*.

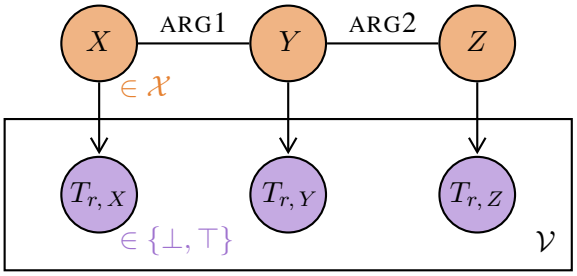


Figure 3: Probabilistic model theory, as formalised in Functional Distributional Semantics. Each node is a random variable. The plate (box in bottom row) denotes repetition of nodes.
Top row: pixie-valued random variables X, Y, Z together represent a situation composed of three individuals. They are jointly distributed according to the semantic roles ARG1 and ARG2. Their joint distribution can be seen as a probabilistic model structure.
Bottom row: each predicate r in the vocabulary $\mathcal{V}$ has a probabilistic truth-conditional function, which can be applied to each individual. This gives a truth-valued random variable for each individual for each predicate.

The bottom row of the graphical model defines a distribution over truth values, so that each predicate has some probability of being true of each individual. Each predicate can therefore be seen as a probabilistic truth-conditional function.

In this paper, I will not discuss learning such a model (for an up-to-date approach, see: Emerson, 2020a). Instead, the focus is on how we can manipulate a trained model, to move from single predicates to complex propositions.

In previous work (E&C), I sketched an account of quantification. The idea is to follow generalised quantifier theory, but with a truth-valued random variable for each node in the scope tree. Similarly to the classical case, the distributions for these nodes are defined bottom-up through the tree.

In the classical theory, we only need to know the cardinalities $|\mathcal{R}(v)|$ and $|\mathcal{R}(v) \cap \mathcal{B}(v)|$. In fact, all the conditions in Table 1 can be expressed in terms of the ratio $\frac{|\mathcal{R}(v) \cap \mathcal{B}(v)|}{|\mathcal{R}(v)|}$. It therefore makes sense to consider the conditional probability $\mathbb{P}(b \mid r, v)$, because this uses the same ratio, as shown in (1).[2]

$$\mathbb{P}(b \mid r, v) = \frac{\mathbb{P}(r, b \mid v)}{\mathbb{P}(r \mid v)} \tag{1}$$

More precisely, B and R are truth-valued random variables for the body and restriction, and V is a tuple-of-pixies-valued random variable, with one pixie for each free variable. Intuitively, the truth of a quantified expression depends on how likely B is to be true, given that R is true.[3]

Quantifier	Condition
some	$\mathbb{P}(b \mid r, v) > 0$
every	$\mathbb{P}(b \mid r, v) = 1$
no	$\mathbb{P}(b \mid r, v) = 0$
most	$\mathbb{P}(b \mid r, v) > \frac{1}{2}$

Table 2: Truth conditions for precise quantifiers, in terms of the conditional probability of the body given the restriction (and given all free variables). These conditions mirror Table 1.

Truth conditions for quantifiers can be defined in terms of $\mathbb{P}(b \mid r, v)$, as shown in Table 2. For these precise quantifiers, the truth value is deterministic – if the condition in Table 2 holds, the quantifier's random variable Q has probability 1 of being true, otherwise it has probability 0. However, taking a probabilistic approach means that we can naturally model vague quantifiers like *few* and *many*. I did not give further details on this point in E&C, but I will expand on this in §5.

4 Quantification with Vague Predicates

Truth-conditional functions that give probabilities strictly between 0 and 1 are motivated for both practical and theoretical reasons. Practically, such a function can be implemented as a feedforward neural network with a final sigmoid unit (as used by E&C), whose output is never exactly 0 or 1. Theoretically, using intermediate probabilities of truth allows a natural account of vagueness (Goodman and Lassiter, 2015; Sutton, 2015, 2017).

However, as we will see in the following subsection, intermediate probabilities pose a problem for E&C's account of quantification.

4.1 Trivial Truth Values

Combining the conditions in Table 2 with vague predicates causes a problem, which can be illustrated with a simple example. Consider a model structure containing only a single individual, and consider only the single predicate *red*, which is true of this individual with probability p. Now consider the sentences (1) and (2).

[2] I use uppercase for random variables, lowercase for values. I abbreviate $\mathbb{P}(X = x)$ as $\mathbb{P}(x)$, and $\mathbb{P}(T = \top)$ as $\mathbb{P}(t)$. For example, $\mathbb{P}(b \mid r, v)$ means $\mathbb{P}(B = \top \mid R = \top, V = v)$.

[3] This would not seem to cover so-called *cardinal quantifiers* like *one* and *two*. Under Link (1983)'s lattice-theoretic approach, a model structure contains plural individuals, so numbers can be treated as normal predicates like adjectives.

(1) Everything is red.
(2) Something is red.

The body of each quantifier is simply the predicate *red*. For simplicity, we can assume that *everything* and *something* put no constraints on their restrictions. We need to calculate $\mathbb{P}(b \mid r, v)$. There are no free variables, and R is always true, so this is simply $\mathbb{P}(b)$. Because there is only one individual, this is simply the probability p.

This means that (1) is true iff $p = 1$, and (2) is false iff $p = 0$. However, we have seen above how predicates will never be true with probability exactly 0 or exactly 1. This means (1) is always false, and (2) is always true, even though we have assumed nothing about the individual!

4.2 Distributions over Precise Predicates

To avoid the problem in §4.1, we must only combine precise quantifiers with precise predicates (i.e. classical truth-conditional functions). To do this, we can view a vague predicate not as defining a probability of truth for each individual, but as defining a distribution over precise predicates. This induces a distribution for the quantifier.

Consider the example in §4.1. With probability p, *red* is a precise predicate that is true of the individual. In this case, both (1) and (2) are true. With probability $1-p$, *red* is a precise predicate that is false of the individual. In this case, both (1) and (2) are false. Combining these cases, both (1) and (2) are true with probability p, which has avoided trivial truth values.

Formalising a vague predicate as a distribution over precise predicates was also argued for by Lassiter (2011). It can be seen as an improved version of supervaluationism (Fine, 1975; Kamp, 1975; Keefe, 2000, chapter 7), which avoids the problem of higher-order vagueness, as shown by Lassiter.

4.3 Probabilistic Scope Trees

To generalise the account in §4.2 to arbitrary scope trees (see §4.4) and vague quantifiers (see §5), it is helpful to introduce a graphical notation for *probabilistic scope trees*, illustrated in Fig. 4. This makes the E&C account easier to visualise. The improved proposal in this paper modifies how the distribution for each truth value node is defined.

For a classical scope tree, the truth of a quantifier node depends on its free variables, and is defined in terms of the extensions of its restriction and body, in a way that removes the bound variable. For a probabilistic scope tree, the distribution for a quantifier node is conditionally dependent on its free variables, and is defined in terms of the distributions for its restriction and body, marginalising out the bound variable. The distributions at the leaves of the tree are defined by predicates, inducing a distribution for each quantifier node as we work up through the tree.

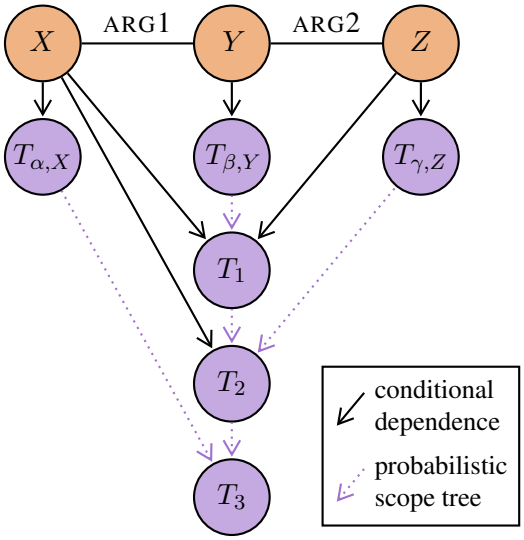


Figure 4: A probabilistic scope tree. T_1, T_2, T_3 correspond to non-terminal nodes in Fig. 2, going up through the tree. One random variable is marginalised out at a time, until T_3 is no longer dependent on any variables.

Fig. 4 corresponds to Fig. 2, if we set α, β, γ to be *picture*, *tell*, *story*. The distributions for $T_{\alpha,X}$, $T_{\beta,Y}$, $T_{\gamma,Z}$ are determined by the predicates. We have three quantifier nodes in the classical scope tree, and hence three additional truth value nodes in the probabilistic scope tree. We first define a distribution for T_1, which represents the $\exists(y)$ quantifier, and which depends on its free variables X and Z. It is true if, for situations involving the fixed pixies x and z, there is *nonzero* probability that they are the ARG1 and ARG2 of a telling-event pixie. Next, we define a distribution for T_2, which represents the $a(z)$ quantifier, and depends on the free variable X. It is true if, for situations involving the fixed pixie x and story pixie z, there is *nonzero* probability that T_1 is true. Finally, we define a distribution for T_3, which represents the *every*(x) quantifier, and has no free variables. It is true if, for situations involving a picture pixie X, we are *certain* that T_2 is true.

4.4 Probabilistic Scope Trees with Vague Predicates as Distributions

In this section I show how to define the quantifier nodes in §4.3 so that they are nontrivial.

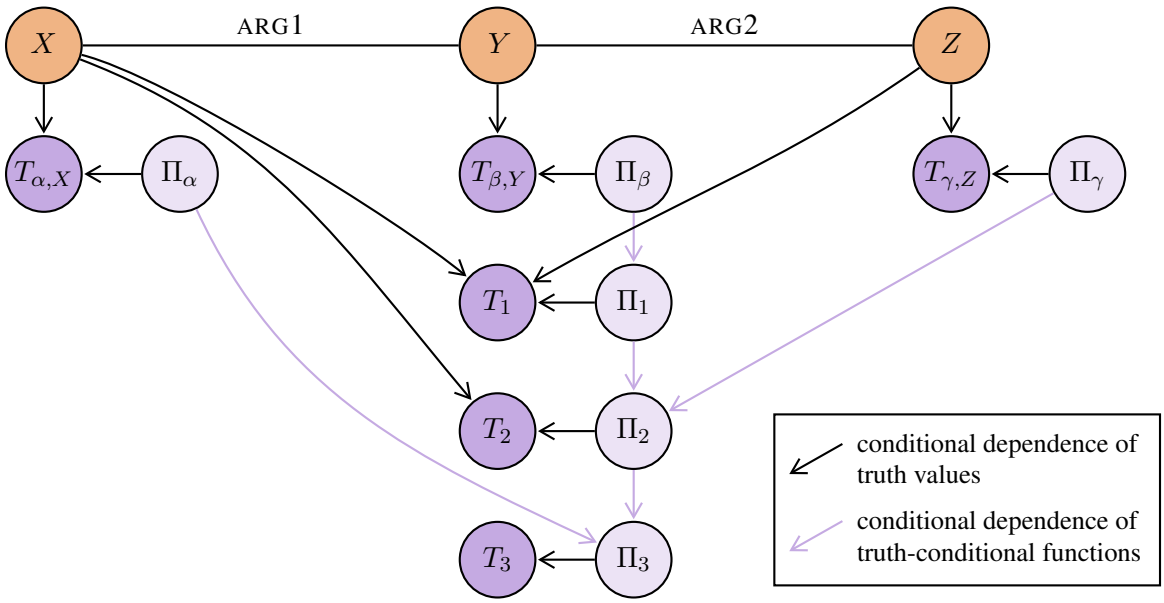


Figure 5: The probabilistic scope tree in Fig. 4, explicitly showing random variables over precise functions.

To explicitly represent each vague truth-conditional function as a random variable over precise functions, we need to add a function node for each truth value node in the graphical model. For example, this transforms Fig. 4 into Fig. 5.

For a truth value node T that is a leaf of the scope tree (the second row of Fig. 5), the distribution $\mathbb{P}(t)$ over truth values follows the description in §4.2. A precise predicate $\pi : \mathcal{X} \to \{\top, \bot\}$ maps pixies to truth values. Given π and a pixie x, the distribution for T is deterministic: $T = \pi(x)$ with probability 1. A distribution Π over precise predicates π defines a vague predicate ψ, by marginalising out this distribution:[4]

$$\mathbb{P}(t \mid x) = \psi(x) = \mathbb{E}_\pi[\pi(x)] \tag{2}$$

More generally, a truth value node Q is dependent on its free variables V. We can represent this in terms of a precise function $\pi : \mathcal{X}^n \to \{\top, \bot\}$, where n is the number of free variables. Given values v for the free variables, a distribution Π over precise predicates π defines a vague predicate ψ, by marginalising out this distribution:

$$\mathbb{P}(q \mid v) = \psi(v) = \mathbb{E}_\pi[\pi(v)] \tag{3}$$

What remains to be shown is that the E&C account of quantification (in §3) can be adapted so that a quantifier's distribution Π_Q over precise functions π_Q can be defined in terms of its restriction function π_R and body function π_B. This can be seen as probabilistic semantic composition: the aim is to combine two truth-conditional functions to produce a distribution over truth-conditional functions. This is illustrated by the nodes Π_1, Π_2, Π_3 in Fig. 5, which are conditionally dependent on other function nodes (indicated by the purple edges), forming a probabilistic scope tree.

Expanding (3) so it is dependent on the restriction and body functions, we have (4). The aim is now to re-write the distribution for Q, using an adapted version of E&C, in order to derive π_Q in terms of π_R and π_B. As explained in §3, the E&C account defines Q using the conditional probability $\mathbb{P}(b \mid r, v)$. More precisely, $\mathbb{P}(q \mid v) = f_Q(\mathbb{P}(b \mid r, v))$ for some f_Q, such as those defined by Table 2. With vague functions now considered as distributions over precise functions, the conditional probability must be amended to $\mathbb{P}(b \mid r, v, \pi_R, \pi_B)$, as in (5), given precise functions π_R and π_B for the restriction and body. This can be re-written as a ratio of probabilities (corresponding to the classical sets), summing over possible values for the bound variable(s) U, as in (6). We can factorise out the distribution for U, according to the conditional dependence structure (illustrated in Fig. 5), as in (7). Finally, we can express R and B in terms of the functions π_R and π_B, and write the sum as an expectation, as in (8). Note that π_R and π_B take $u \cup v$ as an argument – by definition of a scope tree, if we combine a quantifier's bound and free variables, we get the free variables of its restriction and body. I have written $u \cup v$ rather than $\{u\} \cup v$, to leave open the possi-

[4] I write expectations with a subscript to indicate the random variable being marginalised out. To write the expectation in (2) explicitly as a sum: $\mathbb{E}_\pi[\pi(x)] = \sum_\pi [\pi(x)\mathbb{P}(\pi)]$.

bility that the quantifier has more than one bound variable, which will be relevant in §5.

$$\mathbb{P}(q \mid v, \pi_R, \pi_B) = \mathbb{E}_{\pi_Q \mid \pi_R, \pi_B}[\pi_Q(v)] \quad (4)$$

$$= f_Q\left(\mathbb{P}(b \mid r, v, \pi_R, \pi_B)\right) \quad (5)$$

$$= f_Q\left(\frac{\sum_u \mathbb{P}(b, r, u \mid v, \pi_R, \pi_B)}{\sum_u \mathbb{P}(r, u \mid v, \pi_R, \pi_B)}\right) \quad (6)$$

$$= f_Q\left(\frac{\sum_u \mathbb{P}(u \mid v)\,\mathbb{P}(r, b \mid u, v, \pi_R, \pi_B)}{\sum_u \mathbb{P}(u \mid v)\,\mathbb{P}(r \mid u, v, \pi_R)}\right) \quad (7)$$

$$= f_Q\left(\frac{\mathbb{E}_{u|v}\left[\pi_R(u \cup v)\,\pi_B(u \cup v)\right]}{\mathbb{E}_{u|v}\left[\pi_R(u \cup v)\right]}\right) \quad (8)$$

(8) gives a probability of truth, hence a vague function. Viewing it as a distribution over precise functions (as in §4.2), we finally have a definition of π_Q in terms of π_R and π_B. Concretely, π_Q returns truth iff (8) is above a threshold. A uniform distribution over thresholds in $[0, 1]$ gives a distribution over such functions.

Abbreviating the notation, we can write (9). A quantifier's truth-conditional function depends on the restriction and body functions, marginalising out the bound variable. The ratio of expectations mirrors the classical ratio of cardinalities.

$$\pi_Q \sim f_Q\left(\frac{\mathbb{E}_u[\pi_R \pi_B]}{\mathbb{E}_u[\pi_R]}\right) \quad (9)$$

We can now recursively define functions for quantifier nodes, given functions in the leaves. We can therefore see Fig. 4 as an abbreviated notation for Fig. 5. The dotted edges do not indicate conditional dependence of *truth values*, but conditional dependence of *truth-conditional functions*.

5 Vague Quantifiers and Generics

While *some*, *every*, *no*, and *most* can be given precise truth conditions, other natural language quantifiers are vague. In particular, we can consider the terms *few* and *many*.[5]

Under a classical account (for example: Barwise and Cooper, 1981), *many* means that $\mathcal{R}(v) \cap \mathcal{B}(v)$ is large compared to $\mathcal{R}(v)$, but how large is underspecified; similarly, *few* means this ratio is small. The underspecification of a proportion can naturally be represented as a distribution. So, we can define the meaning of a vague generalised quantifier to be a function from $\mathbb{P}(b \mid r, v)$ to a probability of truth, as illustrated in Fig. 6.

[5] Partee (1988) surveys work suggesting that *few* and *many* are ambiguous between a vague cardinal reading and a vague proportional reading. As mentioned in Footnote 3, we can treat cardinals as predicates rather than quantifiers.

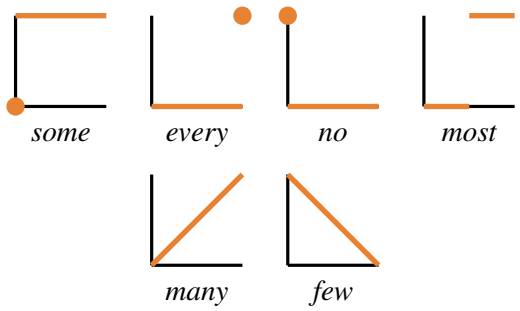


Figure 6: Probabilities of truth for various quantifiers. Each x-axis is $\mathbb{P}(b \mid r, v, \pi_R, \pi_B)$, and each y-axis is $\mathbb{P}(q \mid v, \pi_R, \pi_B)$, plotting the function f_Q in orange. All axes range from 0 to 1. Quantifiers in the bottom row are vague, requiring intermediate probabilities.

A particularly challenging case of natural language quantification involves *generic* sentences, such as: *dogs bark*, *ducks lay eggs*, and *mosquitoes carry malaria*. Generics are ubiquitous in natural language, but they are challenging for classical models, because the truth conditions seem to depend heavily on lexical semantics and on the context of use (for discussion, see: Carlson, 1977; Carlson and Pelletier, 1995; Leslie, 2008).

While it is tempting to treat generic quantification as underspecification of a precise quantifier (for example: Herbelot, 2010; Herbelot and Copestake, 2011), this is at odds with evidence that generics are easier for children to acquire than precise quantifiers (Hollander et al., 2002; Leslie, 2008; Gelman et al., 2015), and also easier for adults to process (Khemlani et al., 2007).

In contrast, Tessler and Goodman (2019) analyse generic sentences as being semantically simple, with the complexity coming down to pragmatic inference. They use Rational Speech Acts (RSA), a Bayesian approach to pragmatics (Frank and Goodman, 2012; Goodman and Frank, 2016). In this framework, literal truth is separated from pragmatic meaning. Communication is viewed as a game where a listener has a prior belief about a situation, and a speaker wants to update the listener's belief. Given a truth-conditional function, a *literal listener* updates their belief by conditioning on truth, ruling out situations for which the function returns false. A *pragmatic speaker* who observes a situation can choose an utterance which is informative for a literal listener – in particular, the utterance which maximises a literal listener's posterior probability for the observed situation. A *pragmatic listener* can update their belief by conditioning on a pragmatic speaker's utterance.

Tessler and Goodman's insight is that this inference of *pragmatic* meanings can account for the behaviour of generic sentences. The literal meaning of a generic can be simple (it is more likely to be true as the proportion increases), but the pragmatic meaning can have a rich dependence on the world knowledge encoded in the prior over situations. For example, *Mosquitoes carry malaria* does not mean that all mosquitoes do (in fact, many do not) but it can be informative for the listener: as most animals never carry malaria, even a small proportion is pragmatically relevant.

Building on this, we could model the generic quantifier by setting f_Q as the identity function (the same as *many* in Fig. 6). From (8), the probability of truth is then as shown in (10). However, marginalising out Π_R and Π_B is computationally expensive, as it requires summing over all possible functions. We can approximate this by reversing the order of the expectations, and so marginalising out Π_R and Π_B before U, as shown in (11), where ψ_R and ψ_B are vague functions. Evaluating a vague function is computationally simple.

$$\mathbb{E}_{\pi_R,\pi_B}\left[\frac{\mathbb{E}_{u|v}\left[\pi_R(u \cup v)\,\pi_B(u \cup v)\right]}{\mathbb{E}_{u|v}\left[\pi_R(u \cup v)\right]}\right] \tag{10}$$

$$\approx \frac{\mathbb{E}_{u|v}\left[\psi_R(u \cup v)\,\psi_B(u \cup v)\right]}{\mathbb{E}_{u|v}\left[\psi_R(u \cup v)\right]} \tag{11}$$

Abbreviating this, similarly to (9), we can write:

$$\psi_Q = \frac{\mathbb{E}_u\left[\psi_R \psi_B\right]}{\mathbb{E}_u\left[\psi_R\right]} \tag{12}$$

For precise quantifiers, using vague functions gives trivial truth values (discussed in §4.1), but for generics, (10) and (11) give similar probabilities of truth. To put it another way, a vague quantifier doesn't need precise functions. Modelling generics with (10) was driven by the intuition that generics are vague but semantically simple. The alternative in (11) is even simpler, because we only need to calculate $\mathbb{E}_{u|v}$ once in total, rather than once for each possible π_R and π_B. This would make generics computationally simpler than other quantifiers, consistent with the evidence that they are easier to acquire and to process.

In fact, (11) takes us back to E&C's conditional probability, as shown in (13).

$$\psi_Q(v) = \frac{\sum_u \mathbb{P}(u \mid v)\,\mathbb{P}(r \mid u, v)\,\mathbb{P}(b \mid u, v)}{\sum_u \mathbb{P}(u \mid v)\,\mathbb{P}(r \mid u, v)} = \mathbb{P}(b \mid r, v) \tag{13}$$

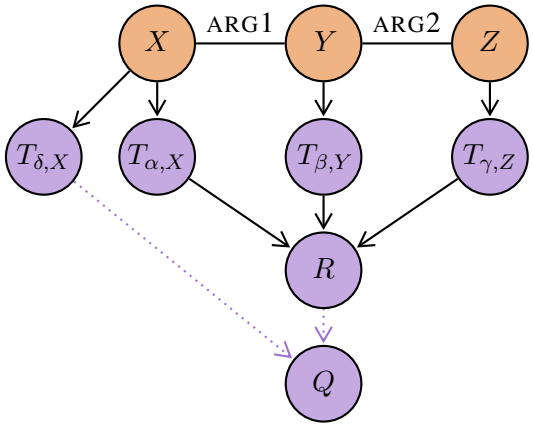


Figure 7: Emerson and Copestake (2017a)'s logical inference, re-analysed as generic quantification. R is the restriction, the logical conjunction of $T_{\alpha,X}$, $T_{\beta,Y}$, and $T_{\gamma,Z}$, while $T_{\delta,X}$ is the body. Generic quantification gives $\mathbb{P}(q) = \mathbb{P}(t_{\delta,X} \mid t_{\alpha,X}, t_{\beta,Y}, t_{\gamma,Z})$, marginalising out all three bound variables (X, Y, and Z).

This means the logical inference proposed by Emerson and Copestake (2017a) can in fact be seen as generic quantification. This is illustrated in Fig. 7, which corresponds to a sentence like *Rooms that have stoves are kitchens*, if α, β, γ, δ are set to *room*, *have*, *stove*, *kitchen*.[6]

Not only does this approach to quantification deal with both precise and vague quantifiers in a uniform way, it can also explain why generics are easier to process than precise quantifiers.

6 Donkey Sentences

An example of a donkey sentence is shown in (3). They are challenging for classical semantic theories, because naive composition, shown in (4), leaves a variable (y) outside the scope of its quantifier (Geach, 1962). The tempting solution in (5) requires a universal quantifier for an indefinite (*a donkey*), which would be non-compositional.[7]

(3) Every farmer who owns a donkey feeds it.

(4) $\forall x\big[\big(\mathit{farmer}(x) \wedge \exists y[\mathit{donkey}(y) \wedge \mathit{own}(x,y)]\big) \rightarrow \mathit{feed}(x,y)\big]$

(5) $\forall x \forall y\big[\big(\mathit{farmer}(x) \wedge \mathit{donkey}(y) \wedge \mathit{own}(x,y)\big) \rightarrow \mathit{feed}(x,y)\big]$

Kanazawa (1994), Brasoveanu (2008), and King and Lewis (2016) discuss how donkey sentences seem to admit multiple readings, which vary in the strength of their truth conditions, and which depend on both lexical semantics and the

[6] An example from RELPRON (Rimell et al., 2016).
[7] For simplicity, (4) and (5) suppress event variables.

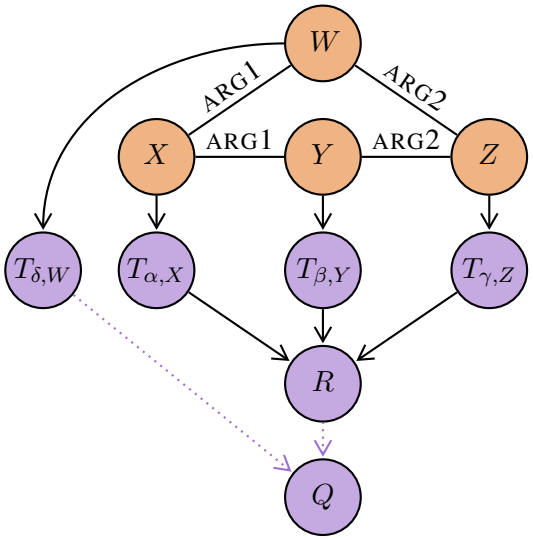


Figure 8: Analysis of a generic donkey sentence, using generic quantification. The quantifier node Q has restriction R (a logical conjunction) and body $T_{\delta,W}$.

context of use. This kind of dependence is exactly what Tessler and Goodman (2019) explained using RSA, so I will apply the same tools here.

As discussed in §5, generics are more basic than classical quantifiers, so I first consider generic donkey sentences, as illustrated in (6)–(8). An analysis of (3) is given in Appendix A.

(6) Farmers who own donkeys feed them.

(7) Linguists who use probabilistic models love them.

(8) Mosquitoes which bite birds infect them with malaria.

Example (8) shows it is inappropriate to use a universal quantifier: not all mosquitoes carry malaria, and not all bitten birds are infected (even if bitten by a malaria-carrying mosquito). However, this sentence still communicates that malaria is spread between birds by mosquitoes. This relies on pragmatic inference, from prior knowledge that most animals cannot spread malaria.

Despite the challenge for classical theories, generic donkey sentences can be straightforwardly handled by my proposed probabilistic approach. An example is shown in Fig. 8, which corresponds to (6), if α, β, γ, δ are set to *farmer*, *own*, *donkey*, *feed*. Intuitively, the more likely it is that a farmer owning a donkey implies the farmer feeding the donkey, the more likely it is for the sentence to be true. Given world knowledge and a discourse context, this can lead to a sharp threshold for being uttered, using RSA's pragmatic inference.

7 Related Work

Functional Distributional Semantics is related to other probabilistic semantic approaches. Goodman and Lassiter (2015) and Bernardy et al. (2018, 2019) represent meaning as a probabilistic program. This paper brings Functional Distributional Semantics closer to their work, because a probabilistic scope tree can be seen as a probabilistic program. An important practical difference is that Functional Distributional Semantics represents all predicates in the same way (as functions of pixies), allowing a model to be trained on corpus data.

Probabilistic TTR (Cooper, 2005; Cooper et al., 2015) also represents meaning as a probabilistic truth-conditional function. However, in this paper I have provided an alternative compositional semantics, in order to deal with vague quantifiers and generics. In principle, my proposal could be incorporated into a probabilistic TTR appproach. Furthermore, although Cooper et al. (2015) discuss learning, they assume a richer input than available in distributional semantics.

Some hybrid distributional-logical systems exist (for example: Lewis and Steedman, 2013; Grefenstette, 2013; Herbelot and Vecchi, 2015; Beltagy et al., 2016), but these do not discuss challenging cases like generics and donkey sentences.

Explaining the multiple readings of donkey sentences using pragmatic inference has been proposed using non-probabilistic tools (for example: Champollion, 2016; Champollion et al., 2019). I have provided a concrete computational method to calculate such inferences, in the same way that Tessler and Goodman (2019) have provided a concrete account of generics.

8 Conclusion

In this paper, I have presented a compositional semantics for both precise and vague quantifiers, in the probabilistic framework of Functional Distributional Semantics. I have re-interpreted previous work in this framework as performing generic quantification, building on the approach of Tessler and Goodman (2019). I have shown how generic quantification is computationally simpler than classical quantification, consistent with evidence that generics are a "default" mode of processing. Finally, I have presented examples of generic donkey sentences, which are doubly challenging for classical theories, but straightforward under my proposed approach.

Acknowledgements

This paper builds on chapter 7 of my PhD thesis (Emerson, 2018), and I would like to thank my PhD supervisor Ann Copestake, for her support, advice, and suggestions. I would also like to thank the anonymous reviewers, for pointing out areas that were unclear, and suggesting additional areas for discussion.

I am supported by a Research Fellowship at Gonville & Caius College, Cambridge.

References

Keith Allan. 2001. *Natural language semantics*. Blackwell.

Jon Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4(2):159–219.

Jon Barwise and John Perry. 1983. *Situations and Attitudes*. Massachusetts Institute of Technology (MIT) Press.

Islam Beltagy, Stephen Roller, Pengxiang Cheng, Katrin Erk, and Raymond J. Mooney. 2016. Representing meaning with a combination of logical and distributional models. *Computational Linguistics*, 42(4):763–808.

Jean-Philippe Bernardy, Rasmus Blanck, Stergios Chatzikyriakidis, and Shalom Lappin. 2018. A compositional Bayesian semantics for natural language. In *Proceedings of the 1st International Workshop on Language, Cognition and Computational Models*, pages 1–10.

Jean-Philippe Bernardy, Rasmus Blanck, Stergios Chatzikyriakidis, Shalom Lappin, and Aleksandre Maskharashvili. 2019. Bayesian Inference Semantics: A modelling system and a test suite. In *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 263–272.

Adrian Brasoveanu. 2008. Donkey pluralities: plural information states versus non-atomic individuals. *Linguistics and Philosophy*, 31(2):129–209.

Ronnie Cann. 1993. *Formal semantics: an introduction*. Cambridge University Press.

Gregory N. Carlson. 1977. *Reference to kinds in English*. Ph.D. thesis, University of Massachusetts at Amherst.

Gregory N. Carlson and Francis Jeffry Pelletier, editors. 1995. *The generic book*. University of Chicago Press.

Lucas Champollion. 2016. Homogeneity in donkey sentences. In *Proceedings of the 26th Semantics and Linguistic Theory Conference (SALT 26)*, pages 684–704.

Lucas Champollion, Dylan Bumford, and Robert Henderson. 2019. Donkeys under discussion. *Semantics and Pragmatics*, 12(1):1–45.

Robin Cooper. 2005. Austinian truth, attitudes and type theory. *Research on Language and Computation*, 3(2-3):333–362.

Robin Cooper, Simon Dobnik, Staffan Larsson, and Shalom Lappin. 2015. Probabilistic type theory and natural language semantics. *Linguistic Issues in Language Technology (LiLT)*, 10.

Ann Copestake. 2009. Slacker semantics: Why superficiality, dependency and avoidance of commitment can be the right way to go. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1–9.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal Recursion Semantics: An introduction. *Research on Language and Computation*, 3(2-3):281–332.

Donald Davidson. 1967. The logical form of action sentences. In Nicholas Rescher, editor, *The Logic of Decision and Action*, chapter 3, pages 81–95. University of Pittsburgh Press. Reprinted in: Davidson (1980/2001), *Essays on Actions and Events*, Oxford University Press.

Guy Emerson. 2018. *Functional Distributional Semantics: Learning Linguistically Informed Representations from a Precisely Annotated Corpus*. Ph.D. thesis, University of Cambridge.

Guy Emerson. 2020a. Autoencoding pixies: Amortised variational inference with graph convolutions for Functional Distributional Semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Guy Emerson. 2020b. What are the goals of distributional semantics? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Guy Emerson and Ann Copestake. 2016. Functional Distributional Semantics. In *Proceedings of the 1st Workshop on Representation Learning for NLP (RepL4NLP)*, pages 40–52. Association for Computational Linguistics.

Guy Emerson and Ann Copestake. 2017a. Variational inference for logical inference. In *Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML)*, pages 53–62. Centre for Linguistic Theory and Studies in Probability (CLASP).

Guy Emerson and Ann Copestake. 2017b. Semantic composition via probabilistic model theory. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*, pages 62–77. Association for Computational Linguistics.

Kit Fine. 1975. Vagueness, truth and logic. *Synthese*, 30(3-4):265–300.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Peter Gärdenfors. 2000. *Conceptual spaces: The geometry of thought*. Massachusetts Institute of Technology (MIT) Press.

Peter Gärdenfors. 2014. *Geometry of meaning: Semantics based on conceptual spaces*. Massachusetts Institute of Technology (MIT) Press.

Peter Thomas Geach. 1962. *Reference and generality: An examination of some medieval and modern theories*. Cornell University Press.

Susan A. Gelman, Sarah-Jane Leslie, Alexandra M. Was, and Christina M. Koch. 2015. Children's interpretations of general quantifiers, specific quantifiers and generics. *Language, Cognition and Neuroscience*, 30(4):448–461.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.

Noah D. Goodman and Daniel Lassiter. 2015. Probabilistic semantics and pragmatics: Uncertainty in language and thought. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*, 2nd edition, chapter 21, pages 655–686. Wiley.

Edward Grefenstette. 2013. Towards a formal distributional semantics: Simulating logical calculi with tensors. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 1–10. Association for Computational Linguistics.

Aurélie Herbelot. 2010. *Underspecified quantification*. Ph.D. thesis, University of Cambridge.

Aurélie Herbelot and Ann Copestake. 2011. Formalising and specifying underquantification. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, pages 165–174. Association for Computational Linguistics.

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 20th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 22–32. Association for Computational Linguistics.

Michelle A. Hollander, Susan A. Gelman, and Jon Star. 2002. Children's interpretation of generic noun phrases. *Developmental Psychology*, 38(6):883–394.

Hans A. W. Kamp. 1975. Two theories about adjectives. In Edward L. Keenan, editor, *Formal Semantics of Natural Language*, chapter 9, pages 123–155. Reprinted in: Stephen Davis and Brendan S. Gillon, editors (2004), *Semantics: A Reader*, chapter 26, pages 541–562; Kamp (2013), *Meaning and the Dynamics of Interpretation: Selected Papers of Hans Kamp*, pages 225–261.

Hans A. W. Kamp and Uwe Reyle. 2013. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, volume 42 of *Studies in Linguistics and Philosophy*. Springer.

Makoto Kanazawa. 1994. Weak vs. strong readings of donkey sentences and monotonicity inference in a dynamic setting. *Linguistics and Philosophy*, 17(2):109–158.

Rosanna Keefe. 2000. *Theories of Vagueness*. Cambridge Studies in Philosophy. Cambridge University Press.

Sangeet Khemlani, Sarah-Jane Leslie, Sam Glucksberg, and Paula Rubio Fernandez. 2007. Do ducks lay eggs? how people interpret generic assertions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, pages 395–400.

Jeffrey C. King and Karen S. Lewis. 2016. Anaphora. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, fall 2016 edition. Metaphysics Research Lab, Stanford University.

Daniel Lassiter. 2011. Vagueness as probabilistic linguistic knowledge. In Rick Nouwen, Robert van Rooij, Uli Sauerland, and Hans-Christian Schmitz, editors, *Vagueness in Communication: Revised Selected Papers from the 2009 International Workshop on Vagueness in Communication*, chapter 8, pages 127–150. Springer.

Sarah-Jane Leslie. 2008. Generics: Cognition and acquisition. *The Philosophical Review*, 117(1):1–47.

Mike Lewis and Mark Steedman. 2013. Combined distributional and logical semantics. *Transactions of the Association for Computational Linguistics (TACL)*, 1:179–192.

Godehard Link. 1983. The logical analysis of plurals and mass terms: A lattice-theoretical approach. In Rainer Bäuerle, Christoph Schwarze, and Arnim von Stechow, editors, *Meaning, Use and the Interpretation of Language*, chapter 18, pages 303–323. Walter de Gruyter. Reprinted in: Paul Portner and Barbara H. Partee, editors (2002), *Formal semantics: The essential readings*, chapter 4, pages 127–146.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations (ICLR), Workshop Track*.

Richard Montague. 1973. The proper treatment of quantification in ordinary English. In K. Jaakko J. Hintikka, Julius M. E. Moravcsik, and Patrick Suppes, editors, *Approaches to Natural Language*, number 49 in Synthese Library, chapter 10, pages 221–242. Kluwer Academic Publishers. Reprinted in: Paul Portner and Barbara H. Partee, editors (2002), *Formal semantics: The essential readings*, chapter 1, pages 17–34.

Terence Parsons. 1990. *Events in the Semantics of English: A Study in Subatomic Semantics*. Current Studies in Linguistics. Massachusetts Institute of Technology (MIT) Press.

Barbara H. Partee. 1988. Many quantifiers. In *Proceedings of the Eastern States Conference on Linguistics (ESCOL)*, pages 383–402. Ohio State University. Reprinted in: Partee (2004), *Compositionality in Formal Semantics*, pages 241–258, Blackwell.

Barbara H. Partee. 2012. The starring role of quantifiers in the history of formal semantics. In *The Logica Yearbook 2012*, pages 113–136. College Publications.

Laura Rimell, Jean Maillard, Tamara Polajnar, and Stephen Clark. 2016. RELPRON: A relative clause evaluation dataset for compositional distributional semantics. *Computational Linguistics*, 42(4):661–701.

Peter R. Sutton. 2015. Towards a probabilistic semantics for vague adjectives. In Henk Zeevat and Hans-Christian Schmitz, editors, *Bayesian Natural Language Semantics and Pragmatics*, chapter 10, pages 221–246. Springer.

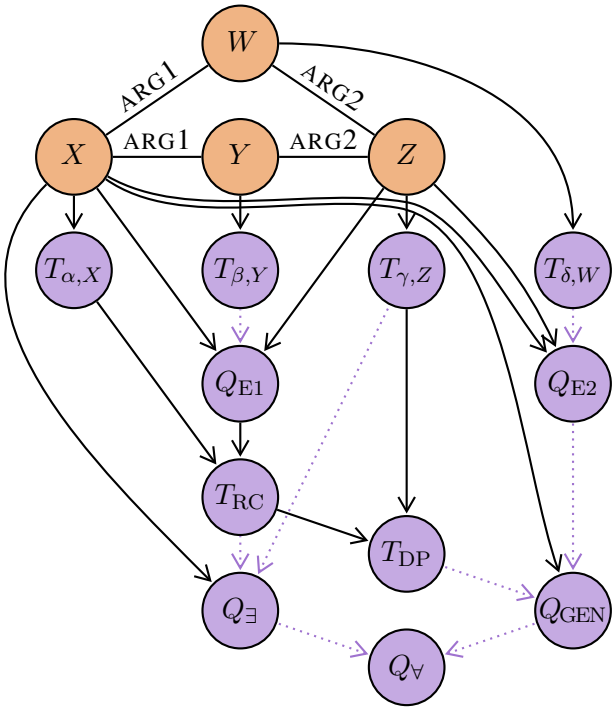
Peter R. Sutton. 2017. Probabilistic approaches to vagueness and semantic competency. *Erkenntnis*.

Michael Henry Tessler. 2018. *Communicating generalizations: Probability, vagueness, and context*. Ph.D. thesis, Stanford University.

Michael Henry Tessler and Noah D. Goodman. 2019. The language of generalization. *Psychological Review*, 126(3):395–436.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Johan Van Benthem. 1984. Questions about quantifiers. *The Journal of Symbolic Logic*, 49(2):443–466.



A Classical Donkey Sentences

In this analysis of a classical donkey sentence, the donkey pronoun is associated with a generic quantifier, while all other quantifiers are precise. The generic quantifier allows the range of readings associated with donkey sentences.

The above figure corresponds to example (3), if α, β, γ, δ are set to *farmer*, *own*, *donkey*, *feed*. Intuitively, this analysis says that, if all farmers who own at least one donkey feed at least a proportion p of their donkeys, then this sentence is true with probability p.

The probability of truth gradually increases with the proportion p. Given world knowledge and a discourse context, this can lead to a sharp threshold proportion above which it is uttered, using pragmatic inference in the RSA framework. If distinguishing small proportions is pragmatically relevant, the *weak* reading becomes preferred. If distinguishing large proportions is pragmatically relevant, the *strong* reading becomes preferred.

I will now go over all nodes in the graph. Firstly, the distributions for $T_{\alpha,X}$, $T_{\beta,Y}$, $T_{\gamma,Z}$, $T_{\delta,W}$ are determined by the predicates.

The remaining truth value nodes are labelled for convenience. T_{RC} and T_{DP} are logical conjunctions (for the relative clause and donkey pronoun, respectively). The remaining five nodes are quantifier nodes, each quantifying one variable.

Note that Z is quantified twice (by $Q_{\exists}$ and Q_{GEN}). This would be surprising in a classical logic, but is not a problem here – marginalising out a random variable means that the quantifier node is not dependent on that variable, but the random variable is still part of the joint distribution, so it can be referred to by other nodes. Because of this double quantification, the scope "tree" is actually a scope DAG (directed acyclic graph).

Q_{E1} and Q_{E2} marginalise out the event variables, respectively Y and W, with trivially true restrictions and bodies $T_{\beta,Y}$ and $T_{\delta,W}$, leaving free variables X and Z. They can be treated like *some* in Fig. 6. For given pixies x and z, Q_{E1} is true if x owns z; Q_{E2} is true if x feeds z.

$Q_{\exists}$ marginalises out Z, with T_{RC} as restriction and Q_{E1} as body, leaving the free variable X. It can be treated like *some* in Fig. 6. For a given x, it is true if x is a farmer and there is a donkey z such that Q_{E1} is true.

Q_{GEN} also marginalises out Z, with T_{DP} as restriction and Q_{E2} as body, leaving the free variable X. It uses the generic quantifier, as in (11). For a given x, it considers donkeys z for which Q_{RC} is true; the probability of truth is the proportion of such z for which Q_{E2} is true (out of donkeys owned by farmer x, the proportion fed by x).

Finally, $Q_{\forall}$ marginalises out X, with $T_{\exists}$ as restriction and Q_{GEN} as body, leaving no free variables. It is treated as in Fig. 6. It is true if, whenever $T_{\exists}$ is true of x, Q_{GEN} is true of x, considering Q_{GEN} as a distribution over precise functions.

Fast visual grounding in interaction: bringing few-shot learning with neural networks to an interactive robot

José Miguel Cano Santín[1] **Simon Dobnik**[1,2] **Mehdi Ghanimifard**[1,2]
[1]Department of Philosophy, Linguistics and Theory of Science (FLoV)
[2]Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
[1]jmcs990@gmail.com [2]{simon.dobnik,mehdi.ghanimifard}@gu.se

Abstract

The major shortcomings of using neural networks with situated agents are that in incremental interaction very few learning examples are available and that their visual sensory representations are quite different from image caption datasets. In this work we adapt and evaluate a few-shot learning approach, Matching Networks (Vinyals et al., 2016), to conversational strategies of a robot interacting with a human tutor in order to efficiently learn to categorise objects that are presented to it and also investigate to what degree transfer learning from pre-trained models on images from different contexts can improve its performance. We discuss the implications of such learning on the nature of semantic representations the system has learned.

1 Introduction

Robots need to ground real world objects and entities that they see with their cameras to natural language that they hear or generate when interacting with a human tutor. There are four properties that distinguish this kind of machine learning from the approaches that are based on a corpus: (i) agents have to make reliable classifications already after being exposed to a few examples; (ii) they may utilise background knowledge; (iii) learning is not done in large batches of examples but in small increments as the interaction unfolds; and (iv) the mechanisms of interaction are used to control the rate of learning.

Matching Networks (Vinyals et al., 2016) is a neural network algorithm designed for one-shot and few-shot learning of a classifier in a standard dataset. For object classification, their principal advantage is their capability to learn objects from few labelled instances rapidly. This property makes them a possible candidate for modelling meaning representations for robot interactions. The learning algorithm for Matching Networks fits in the supervised learning paradigm. Its central idea is influenced from the meta-learning paradigm with memory-augmentation (Santoro et al., 2016) which has not been evaluated for interactive scenarios. In this work, we implement Matching Networks and evaluate its performance in critical interactive scenarios which includes both offline and online learning with a simulated interactive agent.[1]

The contributions of this paper are as follows: (i) we create our own implementation of the Matching Network model from (Vinyals et al., 2016) and (ii) integrate it with interactive strategies of a situated agent; (iii) we test the performance of such a simulated agent (a) on baseline recognition of objects presented to the agent, (b) where the support set of image categories is from another context, and (c) where new object categories must be learned; (iv) we discuss implications of interactive learning using this model for representations of meaning and semantics of natural language and outline several future experiments in which the model could be exploited and improved.

2 Matching Networks

Network Architecture Matching Networks are composed of four sub-modules: (1) convolutional networks $\texttt{ConvNet}_\theta(\cdot)$ to extract basic visual features from an image, (2) support set embeddings $\texttt{g}_\theta(\cdot)$ to encode visual features of the few-shots of labelled instances in memory, (3) target embeddings $\texttt{f}_\theta(\cdot)$ to encode visual features of a new image, (4) the matching layer $\texttt{att}(\cdot,\cdot)$, which compares the target representation with support embeddings in a memory. The output of the matching layer can be interpreted as attention on categories in the support set.

As shown in Figure 1, the Matching Networks

[1]Additional performance metrics of our implementation on Omniglot (Lake et al., 2015) and ImageNet (Russakovsky et al., 2015) are reported in Appendix.

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 53–61

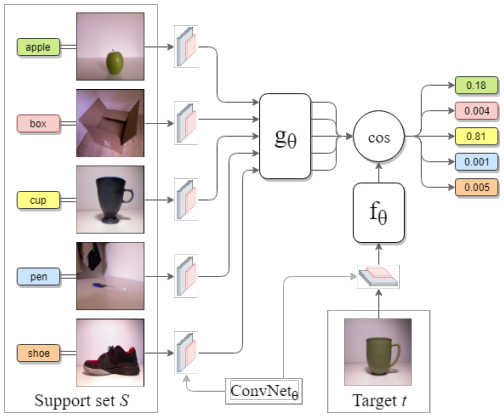


Figure 1: A diagram demonstrating a one-shot learning task with five labelled categories. First, five labelled images of the support set (S) are embedded in memory. Then, an unlabelled target image (t) is also encoded. The matching layer computes the cosine similarity between the embeddings of t and the embeddings of categories in the support set and aggregates attention scores over the categories in the output.

compare an unlabelled target image (t) with the labelled images of a support set (S) to determine how similar t is to different categories in S. After encoding images of the support set S with $\texttt{ConvNet}_\theta$, the support set embedding module ($\texttt{g}_\theta$) is responsible to implicitly learn the interdependency between the visual features and categories by creating an embedding space for images where their similarity must correspond to their shared category. After encoding the target image (t) with $\texttt{ConvNet}_\theta$, the target embedding module ($\texttt{f}_\theta$) is expected to learn the projection of convolutional visual features onto the embedding space of categories constructed by ($\texttt{g}_\theta$). After obtaining the output from the embedding functions, the matching layer computes the similarities between the target embedding and the support set embeddings as if it attends on the categories of similar images in the support set:

$$\hat{y}_t = \sum_{i=1}^{n} \texttt{att}(x_i, t) \cdot y_i$$

where, the support set $S = \{(x_i, y_i)\}_{i=1}^{n}$ consists of n instances of labelled images x_i and their one-shot encoded labels y_i. $\hat{y}_t$ represents the predicted distribution of categories $P(\cdot|t, S)$ of t, conditioned on given S.

Training Algorithm Notably, the labelled images in the support set are the augmented memory of the neural networks. The training objective is to minimise the error in categorisation by learning prototype representations of the images of each category in the support set. In principle, components of the model can be trained in an end-to-end fashion. With categorical cross entropy loss, in each supervised learning step, the predicted log-likelihood of the correct labels is back-propagated to all parameters of the modules:

$$loss(y_t, \hat{y}_t) = -\log(y_t \cdot \hat{y}_t)$$

Implementation We have implemented a variation of Matching Networks in TensorFlow (Abadi et al., 2015) with a small number of parameters. In our implementation, `g` and `f` are one-layer feed forward networks with `ReLU` activation and L2 normalisation, `ConvNet` is VGG16 (Simonyan and Zisserman, 2014) and the matching layer is the cosine similarity of two vectors with `softmax` normalisation with no trainable parameters.

$$\texttt{ConvNet}_{\theta_0}(x) = \text{VGG16}_{\theta_0}(x)$$

$$\texttt{g}_{\{w_1,b_1\}}(x) = \frac{\texttt{ReLU}(w_1 \cdot x + b_1)}{||\texttt{ReLU}(w_1 \cdot x + b_1)||_2}$$

$$\texttt{f}_{\{w_2,b_2\}}(x) = \frac{\texttt{ReLU}(w_2 \cdot x + b_2)}{||\texttt{ReLU}(w_2 \cdot x + b_2)||_2}$$

$$\texttt{att}(x_i, t) = \frac{e^{(\texttt{f}(t) \cdot \texttt{g}(x_i))}}{\sum_{j=1}^{n} e^{(\texttt{f}(t) \cdot \texttt{g}(x_j))}}$$

In the normalisation of `f` and `g` in `att`, their dot product is equivalent to their cosine similarity. The code of this implementation is openly available.[2]

Differences with (Vinyals et al., 2016) Unlike (Vinyals et al., 2016), in the implementation of `g` and `f` we only process one image at a time instead of using bigger networks to process the support set in one go with BiLSTM (for `g`) and attLSTM (for `f`) (Vinyals et al., 2016, p. 3). While the incentive behind using LSTMs (Hochreiter and Schmidhuber, 1997) is their ability to encode dependencies between items, they also encode sequential dependencies which are not beneficial for sets of labelled images in S. This was why we do not use LSTMs, despite its promising results in (Vinyals et al., 2016).

Our goal is to use the model in an interactive scenario where the support set grows incrementally. Since the implemented design requires retraining the parameters, smaller models are preferable. Therefore, we chose the simplest solution

[2] https://github.com/jcanosan/Interactive-robot-with-neural-networks

that is sufficient to give reliable results. We simply concatenate the images in the support set vector to retrain the re-parameterised embeddings with a random order of categories.

Transfer of pre-trained visual features We aim to exploit the background knowledge that is transferred from off-the-shelf pre-trained object recognition models trained on images that humans took with their cameras in a variety of settings. VGG16 consist of stacks of convolutional layers with a small receptive field of 3×3, followed by max-pooling layers (see Simonyan and Zisserman, 2014, p. 2). Instead of training the `ConvNet` layers, we use the output of the pre-trained model on the ImageNet dataset (Russakovsky et al., 2015) after the final convolution layer followed by global average pooling as a `ConvNet` module to extract the visual features. We then investigate domain adaptation measures such as fine-tuning on images and experiment with adding new images to the support set as described in Section 4.

We expect that photographic images taken by humans contain different distributions of objects, backgrounds and attention or focus on objects from those that our agent can capture with its camera. Yosinski et al. (2014) point out that the benefit of using pre-trained networks decreases when the task or the data employed in the pre-training stage is very different from the target task. However, they found that "features transferred from distant tasks are better than random weights" when initialising a task and transferred features help improve performance "even after substantial fine-tuning on a new task, which could be a generally useful technique for improving deep neural network performance" (see Yosinski et al., 2014, p. 8).

3 Interactive grounding of visual objects

Robotic systems have to learn constantly new knowledge about their dynamic environment, for example when they encounter new objects. One of the major differences between learning from datasets and learning interactively is that the system must be able to use the knowledge that it has learned very quickly, after seeing only a couple of examples. Therefore, in this respect few-shot learning is ideally suited for this domain. However, an interacting agent with a human tutor can exploit mechanisms of human interaction (*interaction strategies*) to learn faster, for example by being able to control how information is presented to the human or by requesting new information that it is missing (Skočaj et al., 2010).

Our situated agent setup is based on the KILLE framework (Dobnik and de Graaf, 2017). It consists of a stationary Microsoft's Kinect v1 RGB-D sensor connected to an Ubuntu 16.04 system. The sensor is supported by the *Freenect* drivers[3] integrated in the Robot Operating System (ROS) framework (Quigley et al., 2009). Our system could use any RGB camera supported by ROS. Its intended scenario of usage is grounding of (small) objects on a table top that a human tutor presents to it.

Several interaction strategies can be defined in this domain (Skočaj et al., 2009; Dobnik and de Graaf, 2017). Because in this work we explore the integration of the few-shot learning architecture with a focus on data sparseness and its contextual dependence, we investigate two strategies to learn objects. Firstly, the human tutor can present the object to the agent and describe what it is (e.g. *This is an apple*). The image of the object is saved to the dataset. Then, depending on the number of examples that the system has seen of this object category, it either waits to see more examples (e.g. *Please, show me more examples of apple*), *learns a new category* (e.g. *I am learning apple*) if it has five images of this category, or *updates* its knowledge (e.g. *I am updating my systems on apple*) if it has seen more than five images.

The second interaction strategy allows the robot to understand and respond to questions about objects that are presented to it (e.g. *What is this?*). The robot first attempts to classify the object presented and includes the highest scoring label in its answer and the certainty of the guess based on the score of the label (e.g. *This is an apple*; *I think this is an apple*; *I am not sure. Is this an apple?*; *I don't know what is this. Please, tell me.*). The user can then provide feedback and affirm the guess or tell the system the correct label.

The application of the Matching Networks has two stages: training and prediction. When training, we build the support set S with a few (k) images of each labelled class (n). Then, we build a target set t from the images from S to train our model. Hence, we use S as both the support set and the target images in the training phase. When a user asks the system to identify an object, the same support set for training is used (S) but the

[3] https://openkinect.org/

target t is the new input taken with the camera. The target must belong to one of the categories of S.

The online training of the matching network is performed every time the robot needs to *learn a new category* or *update* its knowledge of one of the existing categories. To *learn a new category*, the system collects five images of this category and adds the images as a new category of objects to the support set S which now has an extra category. To *update* the knowledge of an existing category, the system includes in the support set S the new image taken by the camera and trains the model with the new S. This update of S is performed (i) when a user presents an object to the robot and (ii) when the robot incorporates feedback from the user after incorrectly classifying an object or after clarifying a category of an object that it was not sufficiently confident in.

With few images and categories the training is fast and the learning is reliable for our robot interaction. However, as S grows, so does the training time. As it is expected that the system would need to constantly learn new objects and categories, this issue will have to be addressed in the future. Having to wait for a long time for retraining the model every time we change the size of the support set is not so good from the perspective of user experience and scalability of the system. One strategy we could take is to implement an additional module of a short-term memory for the current objects and wait with the learning updates until when the robot is not interacting with a user and is resting.

In order to measure the success of learning under different conditions we automatised the testing procedure. To test each configuration with identical data, we created a Small Objects daTAset (SOTA), a collection of 400 images with equal distribution of images over 20 categories.[4] Each image depicts a single small everyday object (such as *apple*, *pen* or *shoe*), which was placed in front of the robot's camera while interacting with our system in real time.

4 Matching networks in interaction

4.1 Baseline performance on SOTA

In this first (baseline) experiment we simulate the object recognition with the Matching Networks outside of the interactive scenario on the collected images of the SOTA dataset.

[4]The dataset is available in our GitHub repository.

		1-shot	5-shot	10-shot
5 labels	Accuracy	66.0%	90.0%	94.0%
	Encoding	1.12s	1.63s	2.15s
	Training	1.43s	3.57s	7.27s
20 labels	Accuracy	41.5%	71.0%	86.5%
	Encoding	1.41s	1.93s	2.39s
	Training	3.26	12.15s	25.99s

Table 1: SOTA evaluation. The table shows the accuracy of the model evaluated on the same dataset and the time taken for encoding images and training.

The Matching Networks are trained as outlined in Section 3: we build a support set (S) with n categories or labels (5 or 20) and k images per label (1, 5 or 10-shot). During training, the images from S are also used as target images, making the number of training instances of $k \times n$. For instance, when training a model with 20 labels (n) and 5-shot (k), there are 100 training instances of target images. During evaluation, the images that we use as targets are the remaining 10 images per each category from SOTA that have not been used to train the model and so they are unknown to the system.

Results in Table 1 show that the model performs much better as more images are added to each label. However, increasing the number of shots also increases the training time. Although 1-shot per category would be ideal in terms of time performance, the results indicate that we need more images per category to achieve good recognition performance. For balancing both training time and performance, five images per category seems to be the optimal setting.

In this experiment we have not applied any selection on the support dataset. That is, we have taken images from SOTA at random to build S with them (the selection is saved so we can reproduce the same S). Different selections of the support set might improve or worsen the performance of the system. The selection of the categories could be derived by another process modelling how humans categorise and discriminate objects in particular contexts.

4.2 Support set from another context

We argued that it is very useful for a robot to be able to use knowledge from another context. The objective of the experiment is to investigate whether Matching Networks can be benefit from transfer learning on images taken in different contexts with different cameras by human observers,

		5 labels	20 labels
S = SOTA-ext	**1-shot**	71.1%	39.4%
t = SOTA-ext	**5-shot**	91.1%	60.0%
S = SOTA-ext	**1-shot**	75.6%	53.3%
t = SOTA	**5-shot**	86.7%	73.9%

Table 2: Transferring knowledge from other domains. The table compares the accuracy of the model trained using the SOTA-external dataset as both a support and a target set and the model trained using SOTA-external as a support and the SOTA as a target.

which can be trained offline rather than in an example-by-example basis. We have created a collection of images from online datasets (ImageNet) and web search which we call SOTA-external. These images depict a single object, each pertaining to one of the 20 categories in SOTA. There are 100 images in total, 5 per label. Unfortunately, due to copyright issues, SOTA-external cannot be published.

We have devised two training strategies. In the *first scenario* we build a support set (S) with SOTA-external and use S as both support and target (t) for training. In the *second scenario* the S is again the SOTA-external, but the t are the images from SOTA. During evaluation of both, the S is always SOTA-external, which represents the knowledge that the robot already has; and the t are the images from the SOTA dataset, which represent the objects that our robot is trying to recognise. The objective behind the two scenarios is to test if it is sufficient to train our model using data from another context, or is it better to train the system wit target from our context and use external images as support.

As shown in Table 2, the performance on five categories is variable: it is better to use external target with 5-shot learning but SOTA target with 1-shot learning. This suggests that for external targets during training, more shots are required possibly because such targets are more variable. However, with more categories it is clear that using images from our robot context (SOTA) as t during training is better than using images from SOTA-external. Despite the fact that there is a visible gap in performance, the system still performs acceptably when using only SOTA-external. Compared to the baseline in Table 1, SOTA-external with SOTA targets still improves the performance. Therefore, transferred learning from another context helps, possibly because external images introduce more variation in visual features which allow better discrimination of objects. Hence, if there is a scenario with a new room in a house (e.g. kitchen), the robot could have pre-learned categories for common objects found there (e.g. plate, cup, etc.). Then, we can improve this knowledge with new images and labels from the robot camera. Collections of images from external resources may also be useful to learn new categories.

4.3 New categories of objects

The objective of this experiment is to test how the system learns new categories that it encounters in its dynamic environment. How many images are needed for the robot to recognise a new category and what are the implications of this for modelling the interaction strategies of the robot. We simulate the learning process of each label from SOTA by building a support set S which contains 19 labels with five images each, which represent the categories that the robot already knows, plus the remaining label in SOTA which represents the newly learned category. We incrementally increase the number of images in this category from 1 to 5 which gives us five models: 1-shot to 5-shot. The images in S are also used as targets for training. In the evaluation phase we take the same S and measure the recognition accuracy of the newly learned category by using the remaining 15 images of the same label in SOTA as target images.

The results are shown in Figure 2. We can see that four to five images are necessary for most of the labels. Some labels are clearly easier to learn than others. For instance, *bottle*, *box*, *clothespin* and *marker* do not reach more than 40% accuracy, while *apple*, *book* and *stuffed toy* do not need more than three images to achieve 80% accuracy. The plot also shows some irregular fluctuations in accuracy when adding new images to some labels. For instance, the accuracy of *boot* achieves 73.3% in 3-shot, but then drops again to 46.7%. It appears that some images confuse the model, for instance images with different objects or depicting an object from an unusual perspective. The irregularities could be due to the object background.

5 Discussion and conclusions

We have explored how a neural network algorithm for one-shot learning, Matching Networks (Vinyals et al., 2016), can be used in an interactive scenario between a robot agent and a human tu-

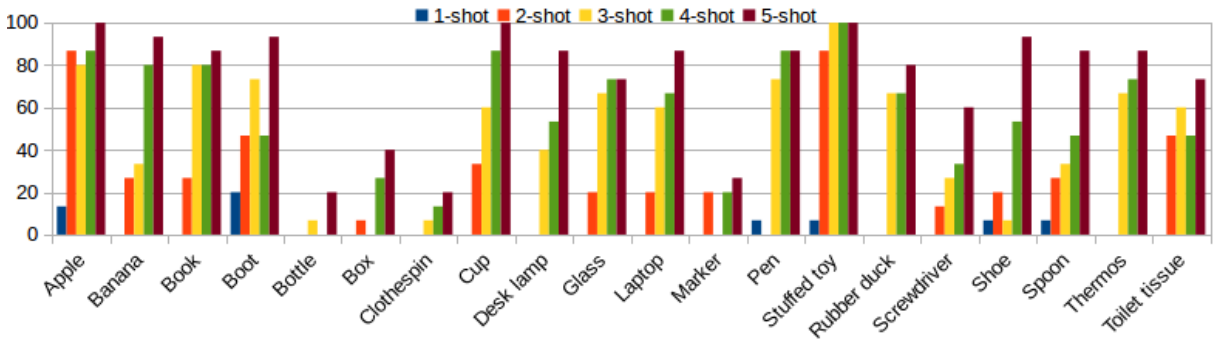


Figure 2: Results on learning new labels. The k-shot learned label is specified under the x axis and each of the bars represent the accuracy of the classification from 1-shot (left) to 5-shot (right).

tor. Our evaluation demonstrates that fast learning with neural networks is possible for this task and it is effective with a few categories.

The Matching Networks do not learn how to directly ground words describing objects in visual features (Roy, 2002; Dobnik, 2009). Instead, it relies on a small support set, contextually introduced objects that it stores, and then learns how to discriminate their categories from one another. This is a particularly appropriate meaning representation for situated agents: (i) the system is able to express gradient beliefs: the outputs of the network are probabilities of the target belonging to each category; (ii) contextual priming information can be integrated by controlling the categories of objects in the support set which demonstrates a potential for modelling top-down attention of an agent (Lavie et al., 2004; Dobnik and Kelleher, 2016); (iii) it models the Saussurean notion of lexical meaning, namely that language is a system of differences without positive terms (Saussure et al., 1983); (iv) as a result the system is very robust and discrimination of categories can already be achieved with a small number of examples.

The system could be extended in several ways, which will be the focus of our future work. First, additional experiments will involve a more distant type of knowledge transfer, to offline pre-train the Matching Networks also on datasets of images with large number of categories and then fine-tune the pre-trained model in the local domain. Another extension is to have a process in place to control the size of the support set over time and transfer the matching knowledge to the memory as required. The update procedure that trains the model from scratch every time affects the user experience negatively, especially when the support set becomes large. Additional procedures could be added to deal with this issue: (1) a training process scheduled when the user is not interacting; (2) a training process of a new model during the interaction when this is still using the older model; (3) implementing methods in the framework of Continual Learning to prevent catastrophic forgetting when updating our models (Greco et al., 2019; Hayes et al., 2019). We will also investigate the influence of using different support sets in terms of variation of objects within a category, variation of objects sampled from different views and selection of object categories as referred to in the current conversation. New interactive strategies with the robot will be investigated and implemented as a way to make the interactive learning of our framework more efficient, for example to *unlearn* efficiently incorrectly learned labels (Skočaj et al., 2009; Dobnik and de Graaf, 2017).

As another direction of future work, we noted that a uniform image background in our experiments negatively influences both performance and scalability. This could be countered by automatic localisation and segmentation of relevant regions of images either with bounding boxes around objects (Girshick et al., 2013; Anderson et al., 2018) or using soft-attention over regions of the image (Xu et al., 2015; Lu et al., 2016) to remove the background of the objects.

Finally, further studies of this framework should go beyond learning of visual grounding of objects. One direction of our future work is to learn grounding of relations including spatial relations with few-shot learning and matching knowledge. Another direction is to consider linguistic and distributional knowledge that could be transferred from cross-modal resources (Lazaridou et al., 2014) as an external matching knowledge.

Acknowledgements

We are grateful to the anonymous reviewers for their helpful comments on our earlier draft. The research of Dobnik and Ghanimifard reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014- 39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, United Kingdom.

Simon Dobnik and Erik de Graaf. 2017. Kille: a framework for situated agents for learning language through interaction. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 162–171, Gothenburg, Sweden. Association for Computational Linguistics.

Simon Dobnik and John D. Kelleher. 2016. A model for attention-driven judgements in Type Theory with Records. In *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue*, volume 20, pages 25–34, New Brunswick, NJ USA.

Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv*, arXiv:1910.02509 [cs.LG].

Claudio Greco, Barbara Plank, Raquel Fernández, and Raffaella Bernardi. 2019. Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3601–3605, Florence, Italy. Association for Computational Linguistics.

Tyler L. Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. 2019. REMIND your neural network to prevent catastrophic forgetting. *arXiv*, arXiv:1910.02509 [cs.LG].

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Nilli Lavie, Aleksandra Hirst, Jan W de Fockert, and Essi Viding. 2004. Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, 133(3):339–354.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414, Baltimore, Maryland. Association for Computational Linguistics.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016. Knowing when to look: adaptive attention via a visual sentinel for image captioning. *arXiv*, arXiv:1612.01887 [cs.CV].

Morgan Quigley, Ken Conley, Brian P. Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y. Ng. 2009. ROS: an open-source robot operating system. In *ICRA Workshop on Open Source Software*.

Sachin Ravi and Hugo Larochelle. 2017. Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Deb Roy. 2002. Learning visually-grounded words and syntax for a scene description task. *Computer speech and language*, 16(3):353–385.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks.

In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1842–1850, New York, New York, USA. PMLR.

Ferdinand de Saussure, Charles Bally, Albert Sechehaye, Albert Riedlinger, and Roy Harris. 1983. *Course in general linguistics*. Duckworth, London.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv*, arXiv:1409.1556 [cs.CV].

Danijel Skočaj, Miroslav Janiček, Matej Kristan, Geert-Jan M. Kruijff, Aleš Leonardis, Pierre Lison, Alen Vrečko, and Michael Zillich. 2010. A basic cognitive system for interactive continuous learning of visual concepts. In *ICRA 2010 workshop ICAIR - Interactive Communication for Autonomous Intelligent Robots*, pages 30–36, Anchorage, AK, USA.

Danijel Skočaj, Matej Kristan, and Aleš Leonardis. 2009. Formalization of different learning strategies in a continuous learning framework. In *Proceedings of the Ninth International Conference on Epigenetic Robotics; Modeling Cognitive Development in Robotic Systems*, pages 153–160. Lund University Cognitive Studies.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. *arXiv*, arXiv:1606.04080 [cs.LG].

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? *arXiv*, arXiv:1411.1792 [cs.LG].

A Appendix

Matching Networks (ours)			
5 labels	**1-shot**	**5-shot**	**10-shot**
Accuracy	96.0%	99.0%	99.0%
Encoding	0.94s	0.86s	0.98s
Training	1.40s	3.77s	6.76s
20 labels	**1-shot**	**5-shot**	**10-shot**
Accuracy	71.0%	93.0%	98.0%
Encoding	0.83s	0.81s	0.82s
Training	3.25	11.81s	22.59s
Vinyals et al. (2016)			
5 labels	**1-shot**	**5-shot**	**10-shot**
Accuracy	98.1%	98.9%	
20 labels	**1-shot**	**5-shot**	**10-shot**
Accuracy	93.8%	98.7%	

Table 3: Validation of our Matching Networks on the Omniglot dataset in comparison to the figures cited in (Vinyals et al., 2016, p. 5) for their model. Encoding is the number of seconds that took to encode the support set images with VGG16. Training is the number of seconds to train the Matching Networks.

Matching Networks (ours)			
5 labels	**1-shot**	**5-shot**	**10-shot**
Accuracy	75.8%	89.8%	98.8%
Encoding	1.12s	1.63s	2.15s
Training	1.43s	3.57s	7.27s
20 labels	**1-shot**	**5-shot**	**10-shot**
Accuracy	52.5%	74.2%	82.6%
Encoding	1.41s	1.93s	2.39s
Training	3.26	12.15s	25.99s
Vinyals et al. (2016)			
5 labels	**1-shot**	**5-shot**	**10-shot**
Accuracy	46.6%	60%	

Table 4: Validation on miniImageNet. Encoding is the number of seconds that took to encode the support set images with VGG16. Training is the number of seconds to train the Matching Networks. Results are compared to the results cited in (Vinyals et al., 2016, p. 7) for their model.

A.1 System validation

Tables 3 and 4 show the results of validation of our Matching Networks and interactive strategies on the standard datasets and in comparison to the implementation in (Vinyals et al., 2016). To this end, we simulated the learning process as we do in Section 4.1 but with two standard offline datasets.

In Table 3 we use the Omniglot dataset (Lake et al., 2015), which consists of 1623 grey-scale images that represent characters from 50 different alphabets. Each of the categories in this dataset contains 20 images of the same character hand-drawn by different people.

In Table 4 we use miniImageNet, a sub-set of ImageNet, containing 60,000 images distributed equally over 100 categories (600 per category). This makes this dataset more suitable for "rapid prototyping and experimentation" than the full dataset (see Vinyals et al., 2016, p. 6). Since the categories used in (Vinyals et al., 2016) were not released with their dataset, our splits are the ones proposed in (Ravi and Larochelle, 2017). The 100 categories are divided into three splits: 64 for training (*train* split), 16 for validation (*val*) and 20 for testing (*test*).

Discrete and Probabilistic Classifier-based Semantics

Staffan Larsson
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg, Sweden
sl@ling.gu.se

Abstract

We present a formal semantics (a version of Type Theory with Records) which places classifiers of perceptual information at the core of semantics. Using this framework, we present an account of the interpretation and classification of utterances referring to perceptually available situations (such as visual scenes). The account improves on previous work by clarifying the role of classifiers in a hybrid semantics combining statistical/neural classifiers with logical/inferential aspects of meaning. The account covers both discrete and probabilistic classification, thereby enabling learning, vagueness and other non-discrete linguistic phenomena.

1 Introduction

Marconi (1997) distinguishes inferential and referential meaning. *Inferential* word meanings enable inferences from uses of the word. Such meanings are sometimes referred to as "high level" or "symbolic", and are typically modelled in formal semantics. *Referential* meaning, on the other hand, allows speakers to identify objects and situations referred to. Referential meaning is sometimes referred to as "low-level" or "subsymbolic". Our working hypothesis is that referential meaning can be modelled using classifiers that output formal representations (Larsson, 2011, 2015), thus connecting "high level" formal representations to "low level" perceptual information. This is a way of addressing the *symbol grounding problem* put forward by (Harnad, 1990) in a way that is compatible with formal semantics.

We also want a framework where meanings can be learned from interactions, and where dialogue participants can coordinate on meanings (Larsson and Myrendal, 2017). To enable this, intensions need to be represented independently of extensions as structured objects which can be modified (updated), and include classifiers of perceptual data.

In formal semantics in the Montague tradition (Montague, 1974), the meaning of a word such as "dog" is taken to be its extension, i.e. the set of all dogs in the world (or in a possible world). This type of semantic theory does not represent intensions independently of extensions, which makes it less well suited for modelling aspects of referential meaning using classifiers. For example, modelling a classifier extensionally (as, say, a set of ordered pairs of inputs and outputs) seems to require some external classifier procedure to produce these sets. Taken as a model of natural language meaning, this suggests an counter-intuitive and unrealistic account of how humans encounter new situations and classify them. Furthermore, such a theory would exclude classification and classification learning (which we take to be part of the acquisition of word meanings) from semantic theory proper, when we in fact believe that it is central to semantics. For these reasons and others, traditional Montagovian semantics does not seem to us to be a satisfactory framework for classifier-based semantics. However, we do believe that it is crucial that the accumulated insights from work in formal semantics over the last 50 decades are integrated with the ideas put forward in this paper.

2 Background

We are developing a formal *judgement-based semantics* where notions such as perception, classification, judgement, learning and dialogue coordination play a central role (Cooper, 2005; Larsson and Cooper, 2009; Larsson, 2011; Dobnik et al., 2011; Cooper, 2012; Dobnik and Cooper, 2013; Cooper et al., 2015). A key idea introduced in Larsson (2011) and Larsson (2015) is the modelling of referential meanings as classifiers of real-

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 62–68

valued (perceptual) data, and training these classifiers in interaction with the world and other agents.

There is a growing body of work in computational and formal semantics which is in line with the approach taken here (Kennington and Schlangen, 2015; Andreas et al., 2016; Schlangen et al., 2016; Ghanimifard and Dobnik, 2017; Shore and Skantze, 2018). We propose a way of connecting this line of work to formal semantics, to enable combining it with the successes of formal semantics (compositionality, quantification, etc.).

Using a Type Theory with Records (Cooper et al., 2014), Larsson (2015) presents a formal semantics for perception, using classifiers to model the relation between perception and linguistic utterances. This paper substantially improves on the formal machinery used in Larsson (2015) and incorporates insights from the implemented version of TTR (Cooper, 2019) as well as related work on visual question answering (Utescher, 2019).

3 TTR: A brief introduction

We will be formulating our account in a Type Theory with Records (TTR). We can here only give a brief and partial introduction to TTR; see also Cooper (2005) and Cooper (2012). To begin with, $s : T$ is a judgment that some s is of type T. To make explicit who is making this judgment, the of-type relation may be subscripted with an agent A, as in $:_A T$. One *basic type* in TTR is Ind, the type of an individual; another basic type is $\mathbb{R}$, the type of real numbers. Given that T_1 and T_2 are types, $T_1 \rightarrow T_2$ is a *functional type* whose domain is objects of type T_1 and whose range is objects of type T_2.

Next, we introduce *records* and *record types*. If $a_1 : T_1, a_2 : T_2(a_1), \ldots, a_n : T_n(a_1, a_2, \ldots, a_{n-1})$, where $T(a_1, \ldots, a_n)$ represents a type T which depends on the objects $a_1, \ldots, a_n$, the record to the left in Figure 1 is of the record type to the right.

In Figure 1, $\ell_1, \ldots \ell_n$ are *labels* which can be used elsewhere to refer to the values associated with them. A sample record and record type is shown in Figure 2.

Types constructed with predicates may be *dependent*. This is represented by the fact that arguments to the predicate may be represented by labels used on the left of the ':' elsewhere in the record type. In Figure 2, the type of c_{man} is dependent on ref (as is c_{run}).

If r is a record and ℓ is a label in r, we can use a *path* $r.\ell$ to refer to the value of ℓ in r. Similarly, if T is a record type and ℓ is a label in T, $T.\ell$ refers to the type of ℓ in T. Records (and record types) can be nested, so that the value of a label is itself a record (or record type). As can be seen in Figure 2, types can be constructed from predicates, e.g., "run" or "man". Such types are called *ptypes* and correspond roughly to propositions in first order logic. Given a set of predicates and a set of possible arguments, the set of possible ptypes is $\mathbf{PType}$, thus allowing for polymorphic predicates. The arity of a ptype P is a set of tuple of types $Arity(P)$. For example $Arity(\text{run}) = \{\langle Ind \rangle\}$.

A fundamental type-theoretical intuition is that something of a ptype T is whatever it is that counts as a proof of T. One way of putting this is that "propositions are types of proofs". In Figure 2, we simply use prf(T) as a placeholder for proofs of T; below, we will show how low-level perceptual input can be included in proofs.[1]

4 The left-or-right game

As an illustration, we follow Larsson (2015) in using a simple dialogue game called the left-or-right (LoR) game. In this game, one agent places objects on a square surface, and the other agent classifies these objects as being to the right or not. In first language acquisition, training of perceptual meanings typically takes place in situations where the referent is in the shared focus of attention and thus perceivable to the dialogue participants. We assume that our DPs (dialogue participants) are able to establish a shared focus of attention. A (simple) sensor collects some information (sensor input) from the environment and emits a real-valued vector. The sensor is assumed to be oriented towards the object in shared focus of attention.

5 Classifiers and TTR

Again following Larsson (2015), we formalise the notion of a simple perceptron classifier and provide its TTR type. The input to the classifier func-

[1] Note that TTR is not proof-theoretic like may other type theories. TTR proofs are more like *witnesses* in situation semantics (Barwise and Perry, 1983) or the *proof objects* in intuitionistic type theory (Martin-Löf and Sambin, 1984). For instance, there are no canonical proofs in TTR; there can be several non-equivalent proofs of the same ptype. This is related to the fact that types in TTR are intensional, i.e., there can be several different types with the same extension. Also, there is no notion of a proof method in TTR.

$$\begin{bmatrix} \ell_1 & = & a_1 \\ \ell_2 & = & a_2 \\ \dots & & \\ \ell_n & = & a_n \\ \dots & & \end{bmatrix} : \begin{bmatrix} \ell_1 & : & T_1 \\ \ell_2 & : & T_2(l_1) \\ \dots & & \\ \ell_n & : & T_n(\ell_1, l_2, \dots, l_{n-1}) \end{bmatrix}$$

Figure 1: Schema of record and record type

$$\begin{bmatrix} \text{ref} & = & \text{obj}_{123} \\ \text{c}_{\text{man}} & = & \text{prf}(\text{man}(\text{obj}_{123})) \\ \text{c}_{\text{run}} & = & \text{prf}(\text{run}(\text{obj}_{123})) \end{bmatrix} : \begin{bmatrix} \text{ref} & : & \text{Ind} \\ \text{c}_{\text{man}} & : & \text{man}(\text{ref}) \\ \text{c}_{\text{run}} & : & \text{run}(\text{ref}) \end{bmatrix}$$

Figure 2: Sample record and record type

tion π_{right} is (1) a parameter record specifying a weight vector w (a vector of real numbers) and a threshold t (a real number) and (2) a situation record specifying an object in the focus of attention, foo, and a sensor reading sr (a vector of real numbers). Whereas a (non probabilistic) classifier normally gives a Boolean output (corresponding to whether the neuron triggers or not), we want as output a ptype (or the negation thereof). The argument of the ptype predicate (right) is the object in the shared focus of attention, i.e. the value of the field foo in the situation record.

(1) $\pi_{right} : \begin{bmatrix} \text{w}:\mathbb{R}^+ \\ \text{t} :\mathbb{R} \end{bmatrix} \rightarrow \begin{bmatrix} \text{foo}:\textit{Ind} \\ \text{sr} :\mathbb{R}^+ \end{bmatrix} \rightarrow \textit{Type}$

such that if

- par:$\begin{bmatrix} \text{w} & : & \mathbb{R}^+ \\ \text{t} & : & \mathbb{R} \end{bmatrix}$ and
- $r : \begin{bmatrix} \text{foo} & : & \textit{Ind} \\ \text{sr} & : & \mathbb{R}^+ \end{bmatrix}$,

then $\pi_{right}(par, r)$ =

$$\begin{cases} \text{right}(r.\text{foo}) & \text{if } r.sr \cdot par.\text{w} > par.\text{t} \\ \neg\, \text{right}(r.\text{foo}) & \text{otherwise} \end{cases}$$

Note that the function itself is defined outside TTR. This allows any classifier to used with TTR, no matter how it is implemented. Classifiers can also be non-binary, as shown here for a fruit classifier FC:

(2) $\pi_{fruit} : \mathbb{R}^+ \rightarrow \begin{bmatrix} \text{foo} & : & \textit{Ind} \\ \text{img} & : & \textit{Image} \end{bmatrix} \rightarrow$ *Type*

such that if

- par:$\mathbb{R}^+$ and
- $r : \begin{bmatrix} \text{foo} & : & \textit{Ind} \\ \text{img} & : & \textit{Image} \end{bmatrix}$,

then $\pi_{fruit}(par, r)$ =

$$\begin{cases} \text{apple}(r.\text{foo}) & \text{if FC}(r.\text{img}, par)\text{=Apple} \\ \text{orange}(r.\text{foo}) & \text{if FC}(r.\text{img}, par)\text{=Orange} \\ \text{pear}(r.\text{foo}) & \text{if FC}(r.\text{img}, par)\text{=Pear} \\ \dots & \\ \neg\, \text{fruit}(r.\text{foo}) & \text{otherwise} \end{cases}$$

6 Putting classification at the core

In this section, we present a version of TTR which explicitly puts classifiers at the core of what it is to understand natural language in relation to a perceived situation. This version replaces that of (Larsson, 2015) types and gives a clearer and more perspicuous account of how judgement and classification are related.

6.1 Meanings for predicates

We start by accounting for predicate meanings in TTR. Several types of expressions in natural language (nouns, verbs, adjectives) can be modelled semantically using predicates. We will represent the (perceptual) meaning of predicates as records containing four fields:

- Classifier parameters (params): a (possibly empty) record containing classifier parameters (e.g. weight vectors)
- Background meaning (bg): a record type representing assumptions about the context of utterance (presuppositions)
- Interpretation function (interp), taking a situation of type bg and providing a ptype encoding a contextual interpretation of an utterance in the context of that situation

- Classification function (clfr) that can be used to make a judgement as to whether an (interpreted) utterance correctly describes a situation

Accrdingly, we define the type *Mng* of a meaning entry as follows:

(3) $Mng = \begin{bmatrix} \text{params} & : & Rec \\ \text{bg} & : & RecType \\ \text{intrp} & : & \text{bg} \rightarrow Type \\ \text{clfr} & : & \text{bg} \rightarrow Type \end{bmatrix}$

Predicate meanings are defined for a predicate with a certain arity. It is convenient to have a looking function outputting the meaning of the predicate used in a given ptype. We define such a function *Pred* as follows (where $P(a_1, \ldots, a_n)$ is a ptype, $P(a_1, \ldots, a_n) \in \mathbf{PType}$):

(4) $Pred(P(a_1, \ldots, a_n)) = P_{\langle T_1, \ldots, T_n \rangle}$

where

- $\langle T_1, \ldots, T_n \rangle \in Arity(P)$
- $a_1 : T_1, \ldots, a_n : T_n$

For example, we get:

(5) $Pred(\text{right}(obj_{45})) = \text{right}_{\langle Ind \rangle}$

Next, we define a function *PredMng* for looking up the meaning of a predicate, whose domain is $\{P_A \mid P \in \mathbf{Pred}, A \in Arity(P)\}$ and whose range is in $\{r \mid r : Mng\}$. For example,

(6) $PredMng(\text{right}_{\langle Ind \rangle}) =$

$$\begin{bmatrix} \text{params} & = & \begin{bmatrix} \text{w} = [0.800 \quad 0.010] \\ \text{t} = 0.090 \end{bmatrix} \\ \text{bg} & = & \begin{bmatrix} \text{sr}_{\text{pos}} & : & \mathbb{R}^+ \\ \text{foo} & : & Ind \end{bmatrix} \\ \text{intrp} & = & \lambda r : \text{bg} \cdot \text{right}(r.\text{foo}) \\ \text{clfr} & = & \lambda r : \text{bg} \cdot \pi_{\text{right}}(\text{params}, r) \end{bmatrix}$$

We also define the interpretation of "right":

(7) $[\![\text{right}]\!] = PredMng(\text{right}_{\langle Ind \rangle}).\text{intrp}$

Finally, we define

(8) $Clfr(T) = PredMng(Pred(T)).\text{clfr}$

For example,

(9) $Clfr(\text{right}(obj_{45})) =$

$$\lambda r : \begin{bmatrix} \text{sr}_{\text{pos}} : \mathbb{R}^+ \\ \text{foo} : Ind \end{bmatrix} \cdot \pi_{\text{right}}\left(\begin{bmatrix} \text{w} = [0.800 \quad 0.010] \\ \text{t} = 0.090 \end{bmatrix}, r \right)$$

6.2 Classification and witness conditions

We now get to the crux of how to put classifiers at the heart of our semantics. According to (Cooper, in progress), for $T \in \mathbf{PType}$,

(10) $s : T$ iff $s \in F(T)$

where $F(T)$ is the *witness cache*, for type T – a set of situations (in the case of ptypes) previously judged to be of type T. The witness cache for a type and an agent can represent the history of judgements made by that agent with respect to the type.

We modify this definition to include witness conditions along the lines of PyTTR (Cooper, 2019) defined with respect to the classifier associated with the predicate of the ptype:

(11) $s : T$ iff $Clfr(T)(s) = T$ or $s \in F(T)$

This definition puts classifiers at the core of TTR. New judgements are made using the *Clfr* function. Previous judgements can be stored in the witness cache for T.

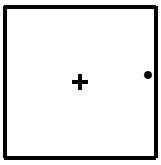
One issue that arises is in what to do first: apply the classifier, or check the witness cache? We do not take a stand on this issue here, but we note that checking the witness chache first makes sense provided it can be assumed to be up to date. Given that classifiers can be continuously trained on new instances, previous judgements may no longer be valid (in the sense that if they were made using the retrained classifier, the results would be different). Guaranteeing the validity of the witness cache would require that any changes in the classifier(s) related to a type T result in purging or re-evaluating the history of potentially affected judgements stored in the witness cache.

7 Putting the model to work

In this section, we show an illustrative example of how the framework above might be put to work in the context of the LoR game, when contextually interpreting utterances and when deciding whether they describe the situation correctly.

7.1 Interpretation

Assume that an agent A places an object on the surface and says "That one is to the right", or just "Right".

(12) 

Agent B watches and gets a position sensor reading $\begin{bmatrix}0.900 & 0.100\end{bmatrix}$ which is part of B's take on the current situation (s_1):

(13) $s_1 = \begin{bmatrix} \text{sr}_{\text{pos}} & = & \begin{bmatrix}0.900 & 0.100\end{bmatrix} \\ \text{foo} & = & obj_{45} \end{bmatrix}$

B now interprets A's utterance in the context the situation s_1 by computing $[\![\text{right}]\!](s_1)$, which gives the result $[\![\text{right}]\!](s_1) = \text{right}(obj_{45})$. How does this happen? Recall that $[\![\text{right}]\!] = \mathit{PredMng}(\text{right}_{\langle Ind\rangle})$.intrp, which means that

(14) $[\![\text{right}]\!](s_1) =$

$(\mathit{PredMng}(\text{right}_{\langle Ind\rangle}).\text{intrp})(s_1) =$

$(\lambda r : \begin{bmatrix}\text{sr}_{\text{pos}}:\mathbb{R}^+ \\ \text{foo} :\mathit{Ind}\end{bmatrix} \cdot \text{right}(r.\text{foo}))($
$\begin{bmatrix}\text{sr}_{\text{pos}}=\begin{bmatrix}0.900 & 0.100\end{bmatrix} \\ \text{foo} =obj_{45}\end{bmatrix}) =$

$\text{right}(obj_{45})$

7.2 Classification

Next, B decides if A's utterance correctly describes (her take on) the situation, i.e. if

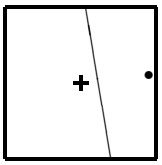
(15) $s_1 : [\![\text{right}]\!](s_1)$, i.e., if $s_1 : \text{right}(obj_{45})$

For T=right(obj_{45}), we get

(16) s:right(obj_{45}) iff
($\mathit{PredMng}(\text{right}_{\langle \text{Ind}\rangle}).\text{clfr})(s)$= right($obj_{45}$) or
$s \in F(\text{right}(obj_{45}))$

In Figure 3, we show how this is checked for for s_1.

The result is that ($\mathit{PredMng}(\text{right}_{\langle \text{Ind}\rangle}).\text{clfr})(s)$= right($obj_{45}$). Hence, $s_1 : \text{right}(obj_{45})$ and (equivalently) $s_1 : [\![\text{right}]\!](s_1)$. Consequently, this round of the LoR game plays out thus:

(17) 

A: "right"
B: "okay"

8 Vagueness and Probabilistic TTR

In Fernández and Larsson (2014), we formulate a Bayesian noisy threshold classifier for vague concepts such as "tall". The classifier is trained on previous observations of tall entities, and is sensitive to the kind of entity (skyscraper, human, basketball player, ...). Instead of a binary judgement, the classifier returns an probability distribution over ptypes. This account connects to the probabilistic extension of TTR (Cooper et al., 2014, 2015).

Adapting from Fernández and Larsson (2014) to our current framework, the meaning of the vague predicate "tall" could be formalised thus:

(18) $\text{PredMng}(\text{tall}_{\langle Ind\rangle})$
$= \begin{bmatrix} \text{bg} = \begin{bmatrix} \text{c} : \text{Type} \\ \text{x} : \text{c} \\ \text{h} : \mathbb{R} \end{bmatrix} \\ \text{params}= \begin{bmatrix} \mu = \mu_{tall} \\ \sigma = \sigma_{tall} \end{bmatrix} \\ \text{intrp} = \lambda r{:}\text{bg}\cdot\text{tall}(r.\text{x}) \\ \text{clfr} = \lambda r{:}\text{bg}\cdot\kappa_{tall}(\sigma(r.\text{bg.c}),\ \mu(r.\text{bg.c}),\ r.\text{h}) \end{bmatrix}$

$\kappa_{tall} : (\mathbb{R}, \mathbb{R}, \text{bg}) \rightarrow [0, 1]$

We are here employing a noisy *probabilistic threshold* (cf. Lassiter (2011)) – a normal random variable, represented by the parameters of its Gaussian distribution, the mean μ and the standard deviation σ (the noise width). Note that the probabilistic threshold depend on the semantic class of the individual being classified:

(19) μ_{tall}:*Type* $\rightarrow \mathbb{R}$

(20) σ_{tall}:*Type* $\rightarrow \mathbb{R}$

Interpretation works exactly as in the non-probabilistic case. Regarding classification, the probabilistic version of (11) above (ignoring the witness cache for the moment) is simply:

(21) $\text{p}(s : T) = \mathit{Clfr}(T)(s)$

Since the output of the clfr function is now a probability, so is the result of classification.

(22) p(s:tall(sally))$\in [0, 1]$

9 Conclusion

We presented a version of Type Theory with Records which places classifiers at the core of semantics. Using this framework, we present an account of the interpretation and classification of utterances referring to perceptually available information (such as a visual scene). The account improves on previous work by clarifying the role of

$$
\begin{aligned}
&(\mathit{PredMng}(\text{right}_{\langle \text{Ind} \rangle}).\text{clfr})(s_1) \\
&= (\lambda r : \begin{bmatrix} \text{sr}_{\text{pos}}{:}\mathbb{R}^+ \\ \text{foo}{:}\mathit{Ind} \end{bmatrix} \cdot \pi_{\text{right}}(\begin{bmatrix} \text{w} = \begin{bmatrix} 0.800 & 0.010 \end{bmatrix} \\ \text{t} = 0.090 \end{bmatrix}, r)(\begin{bmatrix} \text{sr}_{\text{pos}} = \begin{bmatrix} 0.900 & 0.100 \end{bmatrix} \\ \text{foo} = \mathit{obj}_{45} \end{bmatrix}) \\
&= \pi_{\text{right}}(\begin{bmatrix} \text{w} = \begin{bmatrix} 0.800 & 0.010 \end{bmatrix} \\ \text{t} = 0.090 \end{bmatrix}, \begin{bmatrix} \text{sr}_{\text{pos}} = \begin{bmatrix} 0.900 & 0.100 \end{bmatrix} \\ \text{foo} = \mathit{obj}_{45} \end{bmatrix}) \\
&= \begin{cases} \text{right}(\mathit{obj}_{45}) & \text{if } \begin{bmatrix} 0.900 & 0.100 \end{bmatrix} \cdot \begin{bmatrix} 0.800 & 0.010 \end{bmatrix} > 0.090 \\ \neg\, \text{right}(\mathit{obj}_{45}) & \text{otherwise} \end{cases} \\
&= \text{right}(\mathit{obj}_{45})
\end{aligned}
$$

Figure 3: Example classification derivation

classifiers in a hybrid semantics combining statistical/neural classifiers with logical/inferential aspects of meaning. The account covers both discrete and probabilistic classification, thereby enabling learning, vagueness and other non-discrete linguistic phenomena.

This account is intended as a starting point for a comprehensive account of semantics encompassing both referential and inferential meaning. Issues to explore include e.g. how referential meanings are coordinated between DPs, and how compositionality works for referential meaning (Larsson, 2017).

Acknowledgements

This work was supported by grant 2014-39 from the Swedish Research Council (VR) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Learning to compose neural networks for question answering. *arXiv preprint arXiv:1601.01705*.

J. Barwise and J. Perry. 1983. *Situations and Attitudes*. The MIT Press.

Robin Cooper. 2005. Austinian truth, attitudes and type theory. *Research on Language and Computation*, 3:333–362.

Robin Cooper. 2012. Type theory and semantics in flux. In Ruth Kempson, Nicholas Asher, and Tim Fernando, editors, *Handbook of the Philosophy of Science*, volume 14: Philosophy of Linguistics. Elsevier BV. General editors: Dov M. Gabbay, Paul Thagard and John Woods.

Robin Cooper. 2019. Types as learnable cognitive resources in pyttr. In Cleo Condoravdi and Tracy Holloway King, editors, *Tokens of Meaning: Papers in Honor of Lauri Karttunen*, pages 569–586. CSLI Publications.

Robin Cooper. in progress. *Type theory and language - From perception to linguistic communication*.

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2014. A probabilistic rich type theory for semantic interpretation. In *Proceedings of the EACL Workshop on Type Theory and Natural Language Semantics (TTNLS)*.

Robin Cooper, Simon Dobnik, Staffan Larsson, and Shalom Lappin. 2015. Probabilistic type theory and natural language semantics. *LiLT (Linguistic Issues in Language Technology)*, 10.

Simon Dobnik and Robin Cooper. 2013. Spatial descriptions in type theory with records. In *Proceedings of IWCS 2013 Workshop on Computational Models of Spatial Language Interpretation and Generation (CoSLI-3)*, pages 1–6, Potsdam, Germany. Association for Computational Linguistics.

Simon Dobnik, Staffan Larsson, and Robin Cooper. 2011. Toward perceptually grounded formal semantics. In *Proceedings of the Workshop on Integrating Language and Vision at NIPS 2011*, Sierra Nevada, Spain. Neural Information Processing Systems Foundation (NIPS).

Raquel Fernández and Staffan Larsson. 2014. Vagueness and learning: A type-theoretic approach. In *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics (∗SEM 2014)*, Dublin, Ireland. The *SEM 2014 Organizing Committee.

Mehdi Ghanimifard and Simon Dobnik. 2017. Learning to compose spatial relations with grounded neural language models. In *IWCS 2017-12th International Conference on Computational Semantics-Long papers*.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1990):335–346.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the Conference for the Association for Computational Linguistics (ACL)*, pages 292–301.

Staffan Larsson. 2011. The ttr perceptron: Dynamic perceptual meanings and semantic coordination. In *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2011)*, Los Angeles (USA). Institute for Creative Technologies.

Staffan Larsson. 2015. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 25(2):335–369. Published online 2013-12-18.

Staffan Larsson. 2017. Compositionality for perceptual classification. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.

Staffan Larsson and Robin Cooper. 2009. Towards a formal view of corrective feedback. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition, EACL*, pages 1–9, Athens, Greece. Association for Computational Linguistics.

Staffan Larsson and Jenny Myrendal. 2017. Dialogue acts and updates for semantic coordination. In *Proc. SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*, pages 52–59.

Dan Lassiter. 2011. Vagueness as probabilistic linguistic knowledge. In R. Nowen, R. van Rooij, U. Sauerland, and H. C. Schmitz, editors, *Vagueness in Communication*. Springer.

Diego Marconi. 1997. *Lexical competence*. MIT press.

P. Martin-Löf and G. Sambin. 1984. *Intuitionistic type theory*. Studies in proof theory. Naples: Bibliopolis.

Richard Montague. 1974. *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New Haven. Ed. and with an introduction by Richmond H. Thomason.

David Schlangen, Sina Zarriess, and Casey Kennington. 2016. Resolving References to Objects in Photographs using the Words-As-Classifiers Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Todd Shore and Gabriel Skantze. 2018. Using lexical alignment and referring ability to address data sparsity in situated dialog reference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2288–2297.

Ronja Utescher. 2019. Visual ttr-modelling visual question answering in type theory with records. In *Proceedings of the 13th International Conference on Computational Semantics-Student Papers*, pages 9–14.

Social Meaning in Repeated Interactions

Robert Henderson
Department of Linguistics
University of Arizona
rhenderson@arizona.edu

Elin McCready
Department of English
Aoyama Gakuin University
mccready@cl.aoyama.ac.jp

Abstract

Judgements about communicative agents evolve over the course of interactions both in how individuals are judged for testimonial reliability and for (ideological) trustworthiness. This paper combines a theory of social meaning and persona with a theory of reliability within a game-theoretic view of communication, giving a formal model involving interactional histories, repeated game models and ways of evaluating social meaning and trustworthiness.

1 Overview

Social meaning has been a topic of much recent attention in computational linguistics and in semantics and pragmatics (Yoon et al., 2016; Burnett, 2018; McCready, 2019). One reason for this has been the need to address and identify bad actors in online speech through automatic means. To this end, there has been significant research in the computational linguistics and artificial intelligence communities in this domain. Within semantics and pragmatics, the motivation has been to identify and understand the kinds of meanings carried by expressions with socially significant content, and to find ways of formally modeling their effects on discourse, norms of behavior, and non-linguistic structures.

One active area of research has been political and hate speech. In many cases, though of course not all (for example slurs: see e.g. Camp 2013; Davis and McCready 2018 for work on this topic), it is difficult to determine what counts as hate speech, or what aspects of speech have political overtones. Prominent in this area is the phenomenon of *dogwhistles*, expressions which have the dual function of signaling a speaker's (usually objectionable or controversial) political stance to a set of savvy interpreters with the requisite background to catch the coded message, while appearing to those not in the know as carrying only more innocuous meanings. Work in this area is the starting point for the present paper.

Henderson and McCready (2018, 2017) present a theory of dogwhistles set in an extension of the game-theoretic framework proposed by Burnett (2018). The theory will be detailed in §2, but in essence involves a game in which utilities depend on recognition of the persona the speaker means to express, where a persona is understood as a kind of social role or stance on certain socially relevant issues (cf. Jaffe 2009). Henderson and McCready (2019) extend this work to an account of trust in communication, which they take to contrast with the notion of reliability in McCready 2015, which takes testimonial reliability to be determined by communicational histories and initial judgements about the likelihood that a source is reliable; this theory is outlined in §3. The basic idea of Henderson and McCready 2019 is to ground a notion of trust on social meaning: since social meanings and personas can signal shared values and goals, it is sensible to trust someone on that basis regardless of the degree to which one finds them reliable in the sense of truth-tracking in communicative behavior.

The main goal of this paper is to combine these two views into one coherent one. Judgements about communicative agents evolve over the course of interactions both in how individuals are judged for testimonial reliability and for (ideological) trustworthiness. A formal model of this necessitates combining the insights of McCready 2015 on histories and repeated game models and those of Henderson and McCready 2019 on ideology and trust. This paper proposes an extension of McCready 2015 which takes social meaning into account, and how social presentation can change over time; this extension is presented in §4, after

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 69–72

which the paper concludes with some future directions in §5.

2 Social Meaning and Dogwhistles

This section briefly describes the theory of dogwhistles given by (Henderson and McCready, 2018). Dogwhistles are prevalent in political speech, and also of course used elsewhere; they serve to show the ideologies and social or political stances and views of the speaker in a way which is both deniable and accessible only to those aware of the coded language they utilize. Further, the meanings they convey are not obviously part of any of the traditional categories of semantic and pragmatic meaning: at-issue content, presupposition, conversational implicature and so on. Henderson and McCready (2018) pursue an analysis which ties dogwhistles directly to the expression of social meaning, and claim they fall into a new kind of category of meaning.

Within sociolinguistics, the category of *indexical meanings* has been used for decades (e.g. Eckert 2008; Silverstein 2003). Such meanings are tied to (for example) phonological or stylistic features and express aspects of the speaker's identity; as such, their efficacy is contingent on recognition by the interpreter of the kinds of identity associated with the feature. Burnett (2018) provides a game-theoretic model for such features using a modified version of standard signaling games involving *personas*, roughly definable as social presentations, which are quite various and cover traits such as social features such as friendliness/professionalism and political ideologies. In her model, utilities depend on hearer recovery of the speaker's presented persona and the way in which hearers assign value, positive or negative, to that persona.

Henderson and McCready (2018) extend this model to provide an analysis of dogwhistles. The basic idea is that the coded message which savvy listeners retrieve from dogwhistles is available as a result of recognizing the speaker's ideological presentation as modeled in the form of a persona. Thus Burnett's model must be extended to allow interpreters to vary in the degree to which they associate particular messages with personas. Utilities are then calculated according to (2), which combines the value of the social meaning of the message (1), which depends on the affective values of the range of personas consistent with the message and likelihood of recovering each persona from the message, with the value assigned to its truth-conditional content, positive only in case the hearer arrives at the true state of affairs on the basis of the message. The two aspects of meaning are weighted with values δ and γ which reflect the relative importance assigned to social and truth-conditional meaning respectively.

(1) $U_S^{Soc}(m, L) = \sum_{p \in [m]} ln(Pr(p|m)) + \nu_S(p)Pr(p|m) + \nu_L(p)Pr(p|m)$

Speaker strategies σ are functions from pairs of states and personas to messages; listener strategies ρ are functions from messages to such pairs. Let $\rho(\sigma(p, t)) = (p', t')$. Then

(2) $US(m, L) = US_{Soc}(m, L) + EU(m, L)$, where $EU(m, L) = \sum_{t \in T} \mathbf{Pr}'(t) \times U(t, m, L)$, where $U(t, m, L) > 0$ if $t \in \rho(m)$ and else $= 0$ (cf. van Rooij 2008).

This view will be combined in §4 with the view of McCready (2015) on reliability, which we turn to next.

3 Reliability

McCready (2015) presents a model of how epistemic agents can make judgements about the reliability of an individual's testimony. Reliability here refers exclusively to the degree to which the individual's utterances can be expected to accurately convey information about the world, so reliability corresponds to the probability with which the individual's testimony conveys the truth. According to this work, such judgements come from two sources: initial impressions of an individual's reliability based on experience and world knowledge, and learning about reliability from interactions with that individual.

The first aspect comes into play when making initial judgements about an agent's reliability. Many have observed that such judgements are conditioned on aspects of presentation – e.g. clothing, grooming, context, and various properties like age, race, gender, and physical form which, when used as bases for judging reliability, often lead to pernicious results (Fricker, 2007) – together with stereotypical judgements about how such properties correlate with truth-telling and reliability (see McCready and Winterstein 2019). In the present paper, we are more concerned with the second aspect: the way in which agent interaction

influences subsequent judgements about reliability.

Here, the basic model is frequentist. Testimonial interaction with an agent produces a *history* consisting of a record of that agent's utterances and the way in which they track truth, modeled in terms of records of their actions in a repeated game; simplifying slightly, each action a performed by agent i in each iteration of a game g is entered into the record as $a_i = \langle \varphi, \tau \rangle$, where φ is the content of the utterance and τ indicates its truth or lack thereof; so τ is selected from $\{T, F, ?\}$, for 'true', 'false', and 'indeterminate/unknown' respectively. The value ? is selected when the content either cannot be verified to be true or false at the present time or if it is unclear whether it has a truth-value at all, as in utterances containing only nontruthconditional content or more controversial cases such as sentences expressing subjective judgements ('Life is beautiful.'). Records then have the form $Hist_g = \langle a_1, \ldots, a_n \rangle$, for a game g with n repetitions.

In this setting, the degree of reliability assigned to an agent R_a is defined as, where $t_a = \Sigma_{i \in 1,\ldots,n} val(2(a_i)) = T$ (where '2' is a projection function picking out the second element of the tuple) and $f_a = \Sigma_{i \in 1,\ldots,n} val(2(a_i)) = F$,

$$R_a =_{def} \frac{t_a}{t_a + f_a}.$$

This simple treatment can be made more sophisticated in various ways (e.g. by weighting more recent interactions over older elements in the history, by introducing awareness, or by introducing other ways to deal with ?-valued elements), but it is sufficient for our purposes to note that all such modifications will still be restricted to judgements about truth-tracking and leave out social meaning entirely.

4 Trust and Reputation

But social meaning is important for decisions about trust. Henderson and McCready (2019) combine the ideas of Henderson and McCready (2018) and McCready (2015) to help understand how communicative agents who are obviously unreliable in a truth-conditional sense can still be trusted; Donald Trump is the obvious example here. According to their proposal, trust is not strictly dependent on truth, but rather can involve ideology.

A lacuna in the proposals of Burnett (2018) and Henderson and McCready (2018) is the way in which hearer values are assigned to personas. One way to valuate personas is to compare them to your own: the more similar, the higher the value assigned. Henderson and McCready (2019) motivate this view via ideological personas: the closer an ideology is to one's own, the more one likes it, since it expresses a similar political stance. It then becomes possible to judge an individual unreliable in the sense of §3 – in that their statements don't consistently track the truth – but still trust them, in the sense that one takes them to have similar goals and thus judges them to act in a way consistent with one's interests. The idea then is that if an agent has a similar enough persona to oneself they can be *trusted*, without precisely being *believed*.

But this idea is not fully formalized, because the only model of discourse-level reliability available is that of McCready (2015), which only covers truth-tracking. Henderson and McCready (2019) observe this point but do not modify the model so that it is capable of handling the full range of facts. The goal of this section is to extend that model to account for a notion of trust.

Burnett (2020) provides a model of personas set within vector spaces of the same sort used to ground formal models of cognitive lexical semantics. On this view, ideological structures have the form $\langle D, sim, \text{PERS}, \mu \rangle$, where $\langle D, sim \rangle$is a $|D|$-dimensional vector space and sim a similarity function on points in such spaces; PERS is a set of points which correspond to personas in this ideological space. μ is a function partitioning personas into positively and negatively valued ones.

In this model, it is easy to see how to incorporate a notion of trust: once the persona expressed by the signaler is extracted by the interpreter, sim is used to compare the personas of signaler and interpreter, yielding a value in the real-numbered interval $[0, 1]$. Given a sufficiently high degree of similarity, the interpreter will be justified (in terms of closeness of interests) in trusting the signaler, in the same way as which reliability was handled by McCready (2015).

To extend this model to discourse-level phenomena and thereby make the actions of agents across the lifespan of testimonial interaction genuinely dependent on both social meaning and reliability, we now integrate this view with the histories of McCready (2015). Game iterations are

now of the form $\langle \varphi, \tau, \pi \rangle$, where φ and τ are as before and $\pi \in$ PERS. Now (3) indicates the degree of trust assigned by the interpreter to the signaler a in the initial state: this is just the degree of similarity between the persona π_1 expressed by a in their first interaction, ie. the first game iteration. (4) indicates how trust is assigned as the interaction continues, simply by averaging the trust assigned before the current iteration with the similarity of the interpreter's and the agent's currently expressed personas.

(3) $\quad trust_a^1 = sim(\pi_1, P)$

(4) $\quad trust_a^{i+1} = \frac{sim(\pi_i, P) + trust_a^i}{2}$

This system is extremely simple and gives a high degree of importance to the latest interaction of the two agents; this is easy to modify, but we find it intuitive to let the latest interaction of agents be highly determinative of how they judge trustworthiness via social aspects of persona and ideological communication.

5 Conclusions and Directions

This paper has integrated the model of testimonial reliability of McCready (2015) with the model of trust of Henderson and McCready (2019) via a notion of persona similarity in vector spaces. This integration is successful and brings together notions of reliability in terms of truth-telling and reliability in terms of common interests and ideological similarity, on the assumption that the latter is to be understood in terms of personas. In future work, we intend to incorporate the valuation function μ and thereby rethink the notion of persona. We think that it is likely that agents judge others not just on the basis of the persona they communicate but also in terms of how they evaluate such personas, ie. their general ideological stance. This requires incorporating valuations into the notion of persona in general, an extension of the model of Burnett (2020). Doing so is the next step in the current project.

Acknowledgments

Thanks to Daisuke Bekki and Heather Burnett for discussion.

References

Heather Burnett. 2018. Signalling games, sociolinguistic variation and the construction of style. *Linguistics and Philosophy*.

Heather Burnett. 2020. A persona-based semantics for slurs. To appear in *Grazer Philosophische Studien*.

Elisabeth Camp. 2013. Slurring perspectives. *Analytic Philosophy*, 54(3):330–349.

Christopher Davis and Elin McCready. 2018. The instability of slurs. To appear in *Grazer Philosophische Studien*.

Penelope Eckert. 2008. Variation and the indexical field. *Journal of sociolinguistics*, 12(4):453–476.

Miranda Fricker. 2007. *Epistemic Injustice*. Oxford University Press.

Robert Henderson and Elin McCready. 2017. How dogwhistles work. In *Proceedings of LENLS 14*. JSAI.

Robert Henderson and Elin McCready. 2018. Dogwhistles and the at-issue/not-at-issue distinction. In Daniel Gutzmann and Katherine Turgay, editors, *Secondary Content*, pages 191–210. Brill.

Robert Henderson and Elin McCready. 2019. Dogwhistles, trust and ideology. In *Proceedings of the 22nd Amsterdam Colloquium*, pages 152–160. ILLC.

Alexandra Jaffe. 2009. Introduction: the sociolinguistics of stance. In Alexandra Jaffe, editor, *Stance: Sociolinguistic Perspectives*, pages 3–28. Oxford University Press.

Elin McCready. 2015. *Reliability in Pragmatics*. Oxford University Press.

Elin McCready. 2019. *The Semantics and Pragmatics of Honorification: Register and Social Meaning*. Oxford University Press.

Elin McCready and Grégoire Winterstein. 2019. Testing epistemic injustice. *Investigationes Linguisticae*, 41:86–104.

Michael Silverstein. 2003. Indexical order and the dialectics of social life. *Language and Communication*, 23:193–229.

Robert van Rooij. 2008. Game theory and quantity implicatures. *Journal of Economic Methodology*, pages 261–274.

Erica J Yoon, Michael Henry Tessler, Noah D Goodman, and Michael C Frank. 2016. Talking with tact: Polite language as a balance between kindness and informativity. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Towards functional, agent-based models of *dogwhistle* communication

Robert Henderson
Department of Linguistics University of Arizona
rhenderson@email.arizona.edu

Elin McCready
Department of English
Aoyama Gakuin University
mccready@cl.aoyama.ac.jp

Abstract

Henderson and McCready 2017, 2018, 2019 build a novel theory of so-called 'dogwhistle' communication by extending the *social meaning games* of Burnett 2017. This work reports on an ongoing project to build systems to model the evolution of dogwhistle communication in a population based on probability monads (Erwig and Kollmansberger, 2006; Kidd, 2007). The ultimate results will be useful not just for dogwhistles, but modeling the diffusion and evolution of social meaning in populations in general. The initial results presented here is a computational implementation of Henderson and McCready 2018, which will serve as the basis for models with multiple speakers and repeated interactions.

1 Introduction

It is the 2016 US presidential election and Jill Stein is in a predicament. She is doing a Reddit AMA[1] and has just been asked about vaccines. We assume she believes her base is uniformly anti-corporate, but also contains a passionate anti-vax minority that hold a position others in her party don't like. She knows that her anti-corporate bona fides are solid, but the question wouldn't be coming up unless there was some uncertainty about her stance on vaccines. This is the perfect occasion for a dogwhistle. She says:

> By the same token, being "tested" and "reviewed" by agencies tied to **big pharma** and the **chemical industry** is also problematic.

Phrases like 'big pharma' and 'chemical industry' could be read as generic anti-corporate speak, but people familiar with anti-vax discourse know that these phrases are a staple of that genre. By using phrases like this, the general population of listeners, who are unfamiliar with anti-vax discourse, might assume that Stein is being anti-corporate, while anti-vaxers, who are obviously familiar with this discourse, might assume that Stein is one of them because she speaks like one of them. That is, Stein can use a dogwhistle to signal allegiance to ingroup members, while signaling a different, more palatable allegiance to naive outgroup members. Crucially, this signal is plausibly deniable. When outgroup members who were savvy about anti-vax discourse called out Stein for using dogwhistles, she could fall back on statements she had made asserting the efficacy of vaccines.

This is an isolated example, but dogwhistles have been well studied in the political science (Albertson, 2015; Hurwitz and Peffley, 2005; Mendelberg, 2001; White, 2007) and advertising literature (Kanner, 2000; Palmer, 2000). In particular, Albertson 2015 shows that religious dogwhistles are in fact effective in signaling religious affiliation to ingroup prospective voters in ways that non-religious voters who would othwerise disapprove of religious appeals in politics are unable to detect. The linguistic work on dogwhistles is sparse, but Henderson and McCready 2017, 2018, 2019 build a novel theory by extending the *social meaning games* of Burnett 2017, which itself builds of of work in game-theoretic pragmatics, in particular, Bayesian Rational Speech Act theory (e.g., Goodman and Frank 2016; Franke and Jäger 2016; Franke and Degen 2016).

Having an account of dogwhistles, especially one that connects with social meaning and pragmatic reasoning more broadly, is an advance, but the proposal in Henderson and McCready 2017, 2018, 2019 makes no attempt to study the dynamics of dogwhistles in a population. As we have seen though, even the rather schematic Stein ex-

[1] An AMA ('Ask Me Anything') is an online forum for free discussion hosted by Reddit.

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 73–77

ample, the structure of the population is critical. The Stein example imagines a scenario with three groups—ingroup, naive outgroup, and savvy outgroup. Dogwhistles should evolve under the following conditions: (i) ingroup members, in virtue of speaking with each other, should develop linguistic variants that occur at a lower rate than outgroup members, (ii) most outgroup members (naive) should be unaware of these linguistic variants that signal ingroup members, though some savvy outgroup members may be away of ingroup language, and (iii) group membership is punished by outgroup members, but rewarded by ingroup members. In this scenario, and in a specific communication event, speakers could choose one of these linguistic variants if the structure of the audience (proportion of ingroup / savvy outgroup / naive outgroup) is such that it will lead to a positive payoff.

Understanding how dogwhistles arise, are used in particular speech situations, and then fall out of use clearly calls for some kind of agent-based modeling. The fact that Henderson and McCready do not do so is due to lack of tooling. There are currently no existing, off-the-shelf resources for computationally modeling populations of agents playing social meaning games, or even the simpler games discussed in the RSA literature. This paper will present ongoing efforts to develop agent-based models of dogwhistle communication in a population based on probability monads (Erwig and Kollmansberger, 2006; Kidd, 2007), especially the implementation of simple RSA models in Bumford and Charlow 2018.

2 Dogwhistles in social meaning games with probability monads

The Haskell[2] type system provides a clean way to lay out social meaning models, that is, the models in which expressions have their social meaning.[3] One of the core ideas of so-called *Third Wave* variationist sociolinguistics (see Eckert 2012 for a review), is that sociolinguistic practice is deeply creative, with speakers, though their linguistic choices constantly creating a place for themselves in social space. Under this view, linguistic variation is the ferment from which speakers construct, entrench, and mutate social identities though their stylistic practices. Eckert (2008) calls this ferment the indexical field, which is made up of opposing features. Speakers, through selection of language variants creatively construct persona which are sets of these features.

In our Stein example, we treat features as types which can be grouped into an indexical field, or list of indexes of opposing types.

(1)
```
data Feature =
 AntiVax | ProVax | ProCorp |
     AntiCorp
type Index = [Feature]
indices =
 [[AntiVax,ProVax], [AntiCorp,
     ProCorp]]
```

A persona is a maximally consistent list of features drawn from the indexical field. The set of all personas is what Burnett (2017) calls the Eckert-Montague Field. We can generate all possible personas, or the EMField, from selecting one Feature by each Index in every way possible.[4]

(2)
```
personae :: [Index] -> EMField
personae p = sequence p
```

We can now introduce messages with social meaning. We assume that messages have their normal truth conditional meaning, but when we turn to social meaning they are not interpreted in worlds, but instead denote sets of Features. In the Stein example, we assume that expressions like 'Big Pharma' is both anti-vax and anti-corporate language, while Stein could have selected some other variant, like 'Corporate Scientists', which would be anti-corporate, but in virtue of invoking science, could be interpreted as pro-vax.

(3)
```
data Message = BigPharma | CorpSci
type Denotation = Message -> [
    Feature]
deno :: Denotation
deno BigPharma = [AntiVax, AntiCorp]
deno CorpSci = [ProVax, AntiCorp]
```

The effect of uttering one of these variants is for the listener to rule out assigning the speaker any persona that is inconsistent with that variant. That is, 'Big Pharma' tells the listener the speaker is definitely not both pro-vax and pro-corporate. Eval implements this logic, which takes a message and a context (some field of possible personas),

[2]We direct the intererested reader who is not familiar with Haskell to Hackage, https://hackage.haskell.org/, which provides documentation of built-in functions used here.

[3]One can see the complete code discussed here with a working example at https://github.com/bkeej/SocialMeaningExp/blob/master/src/RSAsoc.hs

[4]Not all these personas may be active in a community. Following a reviewer's suggestion, we could set the prior that an agent bears such a persona to 0 in a community to model this.

and returns just those personas that overlap with the denotation of the message.

(4)
```
type Lexicon = Message -> EMField ->
  [Persona]
eval :: Lexicon
eval m f = nub $ [i | i <- f,
                      p <- i,
                      p `elem` (deno m)]
```

This completes the implementation of the model theoretic aspects of social meaning in Henderson and McCready 2017, 2018, 2019. The real action takes place as speakers use these expressions and listeners infer their personas in a probabilistic setting. Before defining this, though, note that there are actually different kinds of listeners, and these are meant to react differently to dogwhistles. We have ingroup listeners, as well as two kinds of outgroup listeners, those savvy to ingroup language and those who are naive.

(5)
```
data Group = Ingroup | Naive | Savvy
```

The listener's priors for the speaker's persona, as well as how speakers of different personas tend to speak, will now now be conditioned what group they belong to. The probability monad toolkit as described in Kidd 2007; Erwig and Kollmansberger 2006 and implemented in Bumford and Charlow 2018 is built on a set of monad transformers that enrich monads with probabilistic notions that can be computed in the background (e.g., weights, Bayes' theorem, etc.), separating them from code describing the structure at hand.

For instance, the PerhapsT monad transformer attaches probabilities to each computation in the list monad, while the MaybeT monad transformer allows us to throw out branches of of the computation that fail, which permits an implementation of Bayes' theorem via normalizing probabilities of non-failed branches.

(6)
```
type BBDist = MaybeT DDist
```

We start by setting priors for personas via calls to weighted, which constructs a weighed distribution from a list of weights and values.

(7)
```
personaPrior :: Dist m =>
  Group -> m Persona
personaPrior g =
  weighted [Mass 5 [ProVax, ProCorp
      ]...
```

In principle, priors for the speaker's persona can vary by listener type, but for this example, we assume that all listeners are fairly certain Stein is not ProVax,ProCorp=5%, most likely not AntiVax,ProCorp=%15, but think it is equally likely that she is AntiCorp,AntiVax=40% or AntiCorp,AntiVax=40%. This uncertainty is what makes using a dogwhistle a potentially profitable strategy.

Listeners also have beliefs about the probability that they will hear certain messages. The fact that these beliefs can vary by listener type is what will make a particular linguistic expression a dogwhistle. That is, an Ingroup member on knowing a speaker is AntiVax might expect them to use BigPharma because they are familiar with anti-vax rhetoric (the same for Savvy outgroup members). In contrast, a Naive outgroup member would assign a lower probability, maybe placing more probability on purely AntiCorp speakers using the phrase. We see this in the definition of messagePrior:

(8)
```
messagePrior :: Dist m =>
  Group -> Persona -> m Message
messagePrior Ingroup [AntiVax,
    AntiCorp] =
  weighted [Mass 80 BigPharma,...
messagePrior Naive [AntiVax,AntiCorp
    ] =
  weighted [Mass 15 BigPharma,...
```

Finally, we can define the recursive RSA-style reasoning, following the example in Bumford and Charlow 2018, where the literal speaker produces messages based on their persona and priors on how speakers with that persona speak, while listeners guess the speaker's persona based on their priors and a model of what the literal speaker will do. By providing higher integers we get a tower of back-and-forth, probabilistic reasoning between speakers and listeners. Note the guard condition in the literal speaker. The computation will fail for messages whose denotation is not consistent with the given persona. This triggers a reapportioning of probability mass over the surviving branching by the monad transformer BBDist. The result is that Bayesian reasoning happens in the background, while we preserve a clean presentation in code of the structure of these games, exactly as promised by the probability monads.

(9)
```
speaker :: Int -> Group -> Persona
    ->
Lexicon -> BBDist Message
speaker n g p sem = bayes $ do
  m <- messagePrior g p
  scaleProb m $
  if n <= 0 --lit. speaker
   then guard (p `elem` sem m field)
  else do --lit. listener
   p' <- listener n g m sem
   guard (p' == p)
  return m
```

(10)
```
listener :: Int -> Group -> Message
    ->
 Lexicon -> BDDist Persona
listener n g m sem = bayes $ do
  p <- personaPrior g
  m' <- speaker (n-1) g p sem
  guard (m' == m)
  return p
```

With this recursive reasoning, we can already observe the dogwhistle effect. For instance, assuming the message priors above, on hearing Stein say 'Big Pharma', an Ingroup or Savvy outgroup member assigns a 60% probability that Stein is AntiVax, up from 40%, while the Naive outgroup member only assign a 42% chance, just slightly up from the prior of 40%.

Starting from the speaker's perspective (i.e., literal-speaker vs. literal-listener) makes sense in these sociolinguistic games. Actually, already, before worrying about issues of audience design, we have implemented a probabilistic model so-called 'First Wave' sociolinguistics. That is, speakers are assigned a persona and mechanistically produce variants at a rate given by that speech community—i.e., by messagePrior. We have seen that we can produce the dogwhistle effect even in this First Wave model. As discussed above, Third Wave sociolinguistics is much richer, assuming that speakers (along with their listeners) are constantly collaboratively choosing variants to construct a persona.

3 Adding audiences in the Third Wave

One way to think of the system in its current guise is that it purely models information transfer in the social meaning domain. To get a Third Wave theory, one that can handle richer aspects of dogwhistles in agent-based models, we need to endow speakers and listeners with preferences for personas.[5] This will allow speakers, not just to report their persona, but also to pick messages that allow them to have a persona they like (and the audience likes) in a particular situation.

Once again, we take the speaker's perspective and model the social utility of message and person for a speaker given a listener.

[5]Note, there are aspects of Third Wave theory that we do not model like bricolage—agents convey parts of multiple personas at once, or the fact that the indexical field itself is dynamic, i.e., "fluidity".

(11)
```
vL :: Group -> Persona -> Float
vS :: Persona -> Float

uSoc :: Message -> Persona ->
 Group -> Lexicon -> Float
uSoc m p g l =
 pr + (vL g p * pr) + (vS p * pr)
  where Sum pr = sum $ [x | Mass x (
      Just y)
   <- runMassT (runMaybeT
      (RSAsoc.listener 1 g m eval)),
         y == p]
```

Speakers now pick a message based on its efficacy in informing a listener about a persona (i.e., pr in uSoc) modified by how listeners and the speaker themself will react to listeners assigning them that persona (i.e., vS and vL in uSoc), weighted by the probability the lister will assign that persona.

Treating an audience as a list of listener types, the utility of a message for a speaker is just the sum of the utility calculation for each listener.

(12)
```
Type Audience = [Group]
uSSoc :: Audience -> Message ->
 Persona -> Lexicon -> Float
uSSoc a m p l = sum $
 map (\g -> uSoc m p g l) a
```

We now have a 'Third Wave'-style model of social meaning for dogwhistles, and make good on the promise made in Section 1. That is, if Stein looks out at her audience and thinks there are a large number of Naive outgroup members, it will increase her social utility to use an anti-vax dogwhistles. The reason is the low probability of detection by Naive outgroup members will meaning high negative affective value for the AntiVax persona will be weighted downward. In contrast, if the audience is mostly Savvy outgroup members, it will not be safe to do so.

4 Conclusions

This paper provides a computational implementation of Henderson and McCready 2018 using probability monads as implemented in Bumford and Charlow 2018. To make the implementation we extend the RSA-style games in that work with model-theoretic logic for social meaning, uncertainty for messages based on persona type, different listener types, and a social utility function implementing social costs for audiences with one or more listeners. In this way, we get formal verification of the work in Henderson and McCready 2017, 2018, 2019, as well as firm foundation for future work developing functional, agent-based models of *dogwhistle* communication.

References

Bethany L Albertson. 2015. Dog-whistle politics: Multivocal communication and religious appeals. *Political Behavior*, 37(1):3–26.

Dylan Bumford and Simon Charlow. 2018. Prob-tools.

Heather Burnett. 2017. Sociolinguistic interaction and identity construction: The view from game-theoretic pragmatics. *Linguistics and Philosophy*.

Penelope Eckert. 2008. Variation and the indexical field. *Journal of sociolinguistics*, 12(4):453–476.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41:87–100.

Martin Erwig and Steve Kollmansberger. 2006. Functional pearls: Probabilistic functional programming in Haskell. *Journal of Functional Programming*, 16(1):21–34.

Michael Franke and Judith Degen. 2016. Reasoning in reference games: Individual-vs. population-level probabilistic modeling. *PloS one*, 11(5):e0154854.

Michael Franke and Gerhard Jäger. 2016. Probabilistic pragmatics, or why Bayes??? rule is probably important for pragmatics. *Zeitschrift für sprachwissenschaft*, 35(1):3–44.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Robert Henderson and Elin McCready. 2017. How dogwhistles work. In *Proceedings of LENLS 14*. JSAI.

Robert Henderson and Elin McCready. 2018. Dogwhistles and the at-issue/not-at-issue distinction. In Daniel Gutzmann and Katherine Turgay, editors, *Secondary Content*, pages 191–210. Brill.

Robert Henderson and Elin McCready. 2019. *Signaling without Saying: The Semantics and Pragmatics of Dogwhistles*. To appear from Oxford University Press.

Jon Hurwitz and Mark Peffley. 2005. Playing the race card in the post—Willie Horton era the impact of racialized code words on support for punitive crime policy. *Public Opinion Quarterly*, 69(1):99–112.

Bernice Kanner. 2000. Hide in plain sight. *Working Woman*, 25(3):14.

Eric Kidd. 2007. Build your own probability monads. *Draft paper for Hac 07 in Freiburg*, 7.

Tali Mendelberg. 2001. *The race card: Campaign strategy, implicit messages, and the norm of equality*. Princeton University Press.

Kimberly Palmer. 2000. Gay consumers in the driver's seat: Subaru's new ad campaign is among those signaling to homosexual buyers. *The Washington Post*.

Ismail White. 2007. When race matters and when it doesn't: Racial group differences in response to racial cues. *American Political Science Review*, 101(2):339–354.

Stochastic frames

Annika Schuster
Department of Philosophy
Heinrich Heine University Düsseldorf
annika.schuster@uni-duesseldorf.de

Corina Strößner
Department of Philosophy II
Ruhr-University of Bochum
corina.stroessner@rub.de

Peter R. Sutton
Department of General Linguistics
Heinrich Heine University Düsseldorf
peter.sutton@uni-duesseldorf.de

Henk Zeevat
ILLC
University of Amsterdam
h.w.zeevat@uva.nl

Abstract

In the frame hypothesis (Barsalou, 1992; Löbner, 2014), human concepts are equated with frames, which extend feature lists by a functional structure consisting of attributes and values. For example, a bachelor is represented by the attributes GENDER and MARITAL STATUS and their values 'male' and 'unwed'. This paper makes the point that for many applications of concepts in cognition, including for concepts to be associated with lexemes in natural languages, the right structures to assume are not merely frames but stochastic frames in which attributes are associated with (conditional) probability distributions over values. The paper introduces the idea of stochastic frames and three applications of this idea: vagueness, ambiguity, and typicality.

1 Background: Frames

Frames originated in Minsky (1974) and were further developed in the field of cognitive science by Barsalou (1992). Frames extend feature lists by a functional structure consisting of attributes and values. Petersen (2007) developed a precise formalisation of recursive frames, in which frames are connected directed graphs, with labeled nodes, labeled arrows and a central node. The labels on arrows are interpreted as partial functions over the domain (attributes) and the labels on nodes as classes of elements of the domain (values). A frame F applies to an object x if there is a function f that assigns x to the central node, and that assigns an object in the annotated class to each node and is such that whenever an arrow labeled a_i leads from node q_j to node q_k, $a_i(f(q_j)) = f(q_k)$. A function typ assigns an element from the set of types **TYPE** to every node (see Figure 1). Figure 2 demonstrates how the frame structure applies to a particular individual (the black cat Felix).

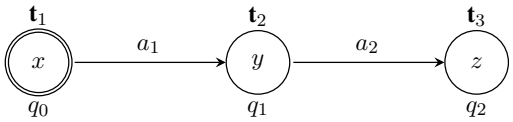


Figure 1: Schema for a frame, F, (a recursive attribute-value structure) with a central node with a value x such that: $a_1, a_2 \in A$, the set of attributes; $q_0, q_1, q_2 \in Q$, the set of nodes; $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3 \in$ **TYPE**. F applies to x iff there is some function f such that $f(q_0) = x : \mathbf{t}_1$, $f(q_1) = y : \mathbf{t}_2$, $f(q_2) = z : \mathbf{t}_3$ where $a_1(f(q_0)) = f(q_1)$ and $a_2(f(q_1)) = f(q_2)$.

Figure 2: A partial frame of a specific black cat Felix, following the schema in Figure 1.

The type of a node contains the semantic information associated with that node. For example, the value for the attribute AGE is necessarily a time, and for humans it is a value between 0 and approximately 120 years. The set **TYPE** is usually accompanied by an ontology (Carpenter, 1992), which includes additional information about the relation between the classes. For example, it determines whether they are exclusive ('red' and 'blue') or one contains the other ('crimson' and 'red').

According to the frame hypothesis, frames are the "single general format of representations" in human cognition (Löbner, 2014, 23). However, this hypothesis requires restrictions on the set of admissible attributes and types to be empirically meaningful. The frame hypothesis can be empirically grounded if we assume that the types and functions are *natural* in human cognition. From

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 78–85

Binder et al. (2016), for example, one can conclude that smell, color, touch, texture, size, weight, sound, and shape are -in combination- vital for the recognition of many concrete natural objects (such as trees and rocks) and animate individuals (such as animals). From this, one can conclude that such attributes are natural and could be some of the attributes on which a natural frame is based.

Let us make a general remark on concepts that also applies to frames. Speaking of concepts, which are usually associated with lexemes, prejudices one into thinking that these concepts are autonomous parts of cognition. However, it seems equally correct to think of lexemes getting a high activation when a certain configuration in semantic memory is activated. In their role as what is expressed by natural language lexemes, concepts (or their frames) are not necessarily more than an isolatable chunk of mental life that does not exist independently of the processes from which it is isolated. The same remark holds for one of the questions we consider in this paper: what are the necessary ingredients of a (stochastic) frame? For many purposes, the explanation can be carried out with a limited notion of a particular frame, based on an abstraction over concrete cases or instances. In our terms, this means, for example, assuming a particular attribute-value frame structure in any given case. For implementation purposes, such limitations are essential, but that does not mean there is a realm of concept-like entities that exhibit these limitations.

2 Stochastic frames

Stochastic frames are the stochastic version of the frames defined in Petersen (2007). This incorporates the probabilistic semantics of Sutton (2015) (also see Cooper et al., 2015) in which the basic idea is that the nodes are associated probability distributions over possible values.

Formally, a (minimally) stochastic frame has a recursive attribute-value structure with a central node with a value x and where each node in the frame has a type from a type hierarchy, just as classical frames do. Where (minimal) stochastic frames differ from classical frames is that the values of attributes need not be categorical (given as particular entities). Instead, they may be probability distributions over entities/values of the relevant type. For example, for a stochastic frame F' that contains an attribute COLOR, the range of this attribute is a probability distribution over entities of the type **Color** (points in the color space). If F' contains an attribute HEIGHT, the range of this attribute is a probability distribution over entities of the type **Height** (values on a measurement scale).

A simple example is given in Figure 3 for a cat, Felix, where the agent does not know what color fur Felix has. This contrasts with the non-stochastic, classical frame in Figure 2 in which a categorical value for Felix's fur color is recorded. Some values in the stochastic frame can be categorical (technically, an assignment of probability 1 to a single value). For example, the value in the stochastic Felix frame for the attribute FUR is assumed to be categorical in this way if the agent knows that Felix has fur. However, in stochastic frames, values of attributes may be distributions over multiple values each with > 0 probability values. For example, a distribution over colors for the value of attribute FUR (such that this distribution may be generated by the agent's experiences of the typical fur colors of cats, see section 3.3).

Figure 3: A partial stochastic frame for a cat Felix where X is a probability distribution over points in the color space.

This minimal conception of a stochastic frame is extended by the addition of constraints. Constraints are encoded as conditional probabilities that capture stochastic relations between the values that different nodes of the frame can obtain (see §2.2).

As a further extension, probabilities can also be embedded at the level of attributes, not just values. For example, the probability that cats have fur (that the cat frame has a FUR attribute), given that an agent may consider fur-less cats to be a possibility.

The minimal and extended characterisations of stochastic frames given above extend the notion of classical frames. Stochastic values (and attributes) allow us to model the uncertainty an agent has about a particular entity or class of entities. Combined with stochastic constraints on relations be-

tween nodes, as we will show, we can furthermore model the knowledge that agents have regarding the distributions of properties of entities of some particular class.

Stochastic frames and classical frames more or less converge when it comes to *ground frames*, i.e., frames for specific individuals in which every attribute has a categorical value. For example, the stochastic frame of a particular instance of 'bachelor' (say John) resembles the corresponding classical frame. (The difference between them is that, technically, the values of attributes in a stochastic ground frame are probability distributions over a single value in a classical frame.)

Where classical and stochastic frames diverge is with respect to uninstantiated frames and frames which are only partially grounded (where some attribute values are not given categorical values). A set of ground frames can be taken as the list of observations on which an uninstantiated stochastic frame or a partially grounded stochastic frame is based, where the list of observations corresponds to the probabilities the frame assigns. This is the case for the partially grounded stochastic frame in figure 3. The frame assigns categorical values to the referent of *Felix* and to the stuff that makes up Felix's fur, but the distribution over colors of fur is based on observed instances of cats and the colors of fur they have. Thinking in terms of ground instances gives a simple transition from thinking in terms of belonging to classes with a given probability to thinking in terms of distributions over values.

2.1 Definitions and prototypes

In its non-stochastic form, the frame hypothesis on concepts fits best with a classical theory of concepts, where concepts are defined by necessary and jointly sufficient conditions of category membership. For example, the concept 'bachelor' is defined as a male, adult person, who is unmarried. The classical view can be traced back to Plato's dialogues and was also fundamental in the early development of formal logic. Frames extend this view by the further demand that they are a quantifier-free conjunction of atoms of the form: x belongs to class C and attribute a_i maps x to y. Formulated in this way, the frame hypothesis seems to rule out any view of concepts in which they do not characterise necessary and sufficient conditions, such as the prototype theory.

Starting with the writings of Wittgenstein (1953) the classical view began to lose credibility. The vagueness of concepts, as well as many other empirical results, can be taken as evidence that the classical view in which concepts provide necessary and sufficient conditions for their application is not on the right track (see Margolis and Laurence, 1999, 27).

Meanwhile, other approaches have been developed and have gained prominence. The most widely discussed one is the prototype theory of concepts, going back to Eleanor Rosch and her collaborators (Rosch, 1973; Rosch and Mervis, 1975; Rosch, 1978; Rosch et al., 1976). It explains the application of a concept to an instance in terms of its similarity to a so-called prototype. This prototype can be understood as a central instance, for example a focal color (Rosch, 1973), but it can also be an idealised representation of the concept (Rosch, 1978). In all variants of prototype theory, a central idea is that concepts are based on an overall similarly of instances rather than on defined features that are common to all instances. Conceptual spaces theory (Gärdenfors, 2000) is also based on similarity, which is understood as an inverse of geometric distance. In this approach, concepts are equated with areas in conceptual spaces and instances with points in these areas. Gärdenfors (2000) emphasizes the relation to the prototype view, according to which an instance is matched to C if it is similar to the the centroid (central point) of the geometric area covered by C. On this understanding, a prototype is a central point in the category. However, if the prototype is not seen as a central point but as a typicality weighted summary of properties one finds in the category, prototypes *are* stochastic frames: they express which properties are likely and in this respect typical.

2.2 Constraints

Constraints were already thought to be an important part of frames in Barsalou (1992), where positive dependencies are marked by a "+" and negative ones by a "-". A good example is the concept of a bird. Birds have different principal modes of locomotion (flying, swimming, and walking), and birds also have different physical features such as webbed feet, or clawed feet. Flying birds with

clawed feet are more typical. However, there are correlations between the swimming, walking, and flying of birds to the feet type. While birds normally have clawed feet, the webbed structure is more expected for birds that swim. In a stochastic frame, the relations between properties of birds are captured as conditional probabilities. Figure 4 shows a partial stochastic frame annotated with such probabilistic constraints. These constraints, allow us not only to model the typical properties of birds, but also to reason about properties of entities on the basis of partial information, for example, that swimming birds have a high probability of having webbed feet.

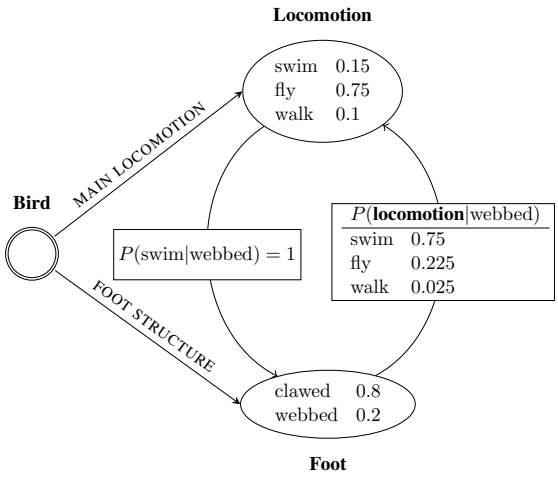


Figure 4: A partial frame for an arbitrary bird with a probabilistic constraint governing the connection between having webbed feet and the main means of locomotion.

The rest of the paper runs through three applications of stochastic frames, superficially, since these applications are covered elsewhere in greater depth. The point of including them here is to make it clear these applications need stochastic frames and that they require very similar characterisations of stochastic frames. This convergence together with the realisation that frames need to be replaced by stochastic frames if the frame hypothesis is accepted as true formed the basis of the authors' cooperation for this paper.

3 Three Applications

3.1 Vague predication

Probabilistic models of vague expressions such as *tall* can capture how an utterance of a sentence such as (1) can reduce the uncertainty hearers have about the way the world is, for example, John's height, as well as about what contextual standards are in play regarding the meaning of *tall* (Lassiter, 2011).

(1) John is tall for a basketball player

In this section, we outline, first, how stochastic frames can incorporate this insight of probabilistic models of vagueness. We then discuss why a frame-based analysis for gradable adjectives is advantageous when dealing with more complex varieties of adjectival modification than the example in (1).

We start with a derivation for *John is tall* (see Figure 5). The subject NP denotes a frame for *John* the central node of which is typed **Person**. This frame includes height information relative to this type (possibly affected by assumptions relating to gender etc.), namely an attribute HEIGHT, the value of which is a probability distribution over heights (we assume for convenience that the unit of measurement is centimetres). On the assumption that no other size information is known about John, and on the assumption that John is a man, this distribution should reflect the sizes of men and so have a mean value around 1.75m, the average height of men in the authors' country of residence (we suppress gender information in Figure 5).

The interpretation of *tall*, $[\![\text{tall}]\!]$, we propose, is a function on the value of a HEIGHT attribute in a frame such that $[\![\text{tall}]\!]$ can compose with any frame that contains a HEIGHT attribute. We propose that $[\![\text{tall}]\!]$ furthermore encodes a function f_{tall} that is applied to the value of this attribute. Where the value of a HEIGHT attribute is represented as a tuple $\langle \mu, \sigma \rangle$ of the mean (μ) and standard deviation (σ) of a Gaussian distribution, the function f_{tall} is such that $f_{tall}(\langle \mu, \sigma \rangle) = \langle n\mu, m\sigma \rangle$ for some positive factor n and some negative factor m. In other words, we propose that $[\![\text{tall}]\!]$ shifts up one's expectations as to average height (relative to the height expectations for the concept to which *tall* applies, and decreases the variance.

To derive *tall for a basketball player*, we propose that $[\![\text{tall}]\!]$ first modifies a basketball player frame (see Figure 5 for a schematic derivation). As with the modification of the **Person** frame (that was instantiated by John) in the previous case, the function f_{tall} applies to the value of the HEIGHT attribute. However, in this case, background knowledge about the heights of basketball players can

be different from the heights of people in general (we tend to know that the former are taller). On the assumption that basketball players are believed to be on average 200cm tall, the effect of applying f_{tall} is to shift this mean upwards and reduce the standard deviation. The effect of this is that the expected height of John, given the information in (1) is drawn from a different distribution over heights than if one were told that *John is tall*, without specifying a comparison class.

Finally, we propose a constraint on the felicitous use of comparison class *for*-PPs, namely that the type for the *for*-PPs (e.g., **BBP** (basketball player)) must be a subtype of the implicit type for the subject NP (e.g., **Person** in *John is tall for a basketball player*). This correctly predicts the oddity of sentences such as *John is tall for a bush*.

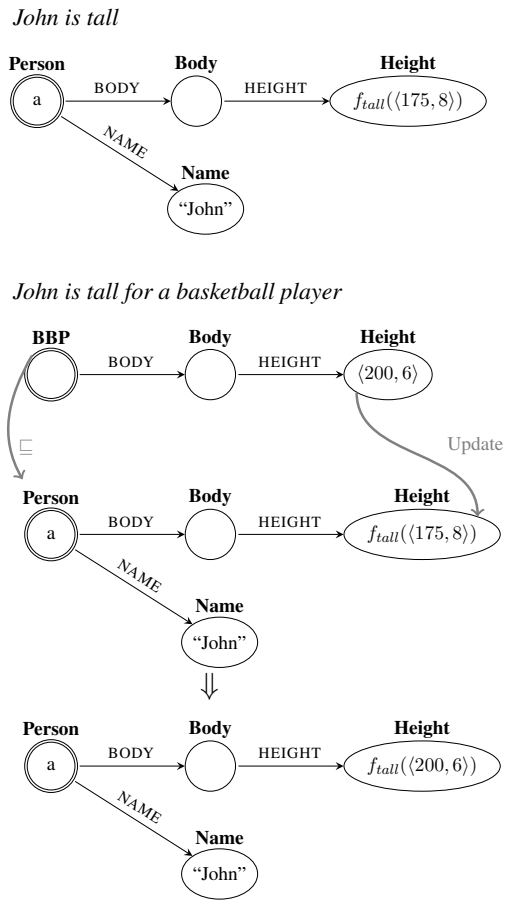


Figure 5: Schema for deriving *John is tall* and *John is tall for a basketball player* in stochastic frames.

The treatment for *tall* sketched above has in common with other probabilistic approaches to vagueness that vague adjectives convey probabilistic information that can be used to infer the probability of some object having observable properties, such as a particular height, given the knowledge and beliefs a hearer has about the way the world is. (See, among others, Sutton, 2015; Lassiter, 2011; Égré, 2017). An advantage of these approaches is that they can be formulated in such a way as to not assume the presence of hidden sharp boundaries in linguistic knowledge (a point above which entities are P and below which, they are not-P (see Sutton, 2018, for further discussion). It is unclear how any of these insights could be modelled within a classical frame.

The advantage of combining a probabilistic approach to vagueness with frame theory (using stochastic frames) is the incorporation of a theory of adjectival modification. For example, *red pen* can be naturally understood as meaning, among other things, a pen that writes in red, a pen that has a red casing, or a pen with a red lid. However, the value *red* must relate to some aspect of the pen (or of an object related to the pen in a specific context). In other words the locus for adjectival modification is underspecified but nonetheless constrained. Frame theory captures this in terms of the number of attributes of the right type there are in the frame. For example, $[\![\text{tall}]\!]$ can apply only to HEIGHT attributes in an NP frame and $[\![\text{red}]\!]$ can apply only to COLOR attributes in an NP frame.

3.2 Ambiguity

This section is a recapitulation of the relevant parts of Zeevat et al. (2015).

Words in natural languages can be studied for their contribution to the truth-conditions of the expressions they are part of in particular contexts. This gives rise to a bewildering number of non-equivalent readings (meanings in use) for high-frequency verbs with a long history: 84 for the verb "fall", 71 for "run" where it is by no means clear that these are all the readings one needs, given that yet further uses are frequently reported. The situation is similar for many nouns and adjectives.

Stochastic frames can accommodate those readings, since they can deal with different possibilities with preferences due to the frequency with which the reading occurs, but mere accommodation is not what one wants. Human language users effortlessly disambiguate in these cases in linear time[1]. This ability is modeled by stochas-

[1]This follows from the empirical work on which the Mar-

tic frames in which the lexical ambiguity is not captured as list of readings, but as one integrated structure, with different preferentially weighted options, once an optimal use is made of the context as an additional resource in a unification like process.

Lexical stochastic frames for verbs will have many nodes that are presupposed from the context. The standard cases are the obligatory arguments of the verbs. Saying that such arguments are obligatory means that it is obligatory for the context to supply the relevant information. But talking of obligatory arguments is just one aspect of the more general case that many of the readings require values from noun phrases or prepositional phrases or from presupposition resolution to the linguistic and non-linguistic context to be possible. Interactions with the context will therefore deliver a particular version of the part of the concept that refers to given material. By unification, this will also change the part of the concept that contains the new information. In addition, the stochastic frame will contain alternatives with different probabilities and in disambiguation, the more probable alternative will be systematically preferred.

As reported in Zeevat et al. (2015), an approach of this kind was hand-tested on the verb "fall" (all readings in David Copperfield (Dickens, 2000)) with full success. The approach uses a logical representation of equivalence classes of stochastic frames (the ones that give the same inequalities). This allows the different users to learn their own probabilities, converging on the same equivalence class under enough exposure to uses of the word. The logical representation using (comparative) probabilistic preferences rather than full-fledged probabilities is human readable and can also be taken as the object that establishes the (near-)unity of the verbal meaning (what all meanings in use have in common) and of the different versions of the meaning of the same word that users learn.

The approach in Zeevat et al. (2015) is an implementation and extension of the approach to lexical ambiguity pioneered by Smolensky (1991) and further developed by Hogeweg (2009), which can be described as: take the maximal amount of content that fits the context. Coercion is part of the mechanism, not a separate process. The model gives a far more detailed picture than just a set of semantic features for lexical meanings. Going for strongest readings is what distinguishes it from approaches such as Asher (2011), which rely on contextual disambiguation and coercion only. Stochastic frames are more conservative than Casasanto and Lupyan (2015) in assuming that observed meanings are stored and serve as a basis for computing meanings in use. Stochastic frames can be learnt and meanings in use can be computed from them by methods that are within the current state of the art.

3.3 Typicality

In section 2.1, we pointed out that stochastic frames fit well to the prototype theory of concepts: understanding prototypes as a weighted sum of property probabilities means to take a stochastic frame *to be* the prototype of the concept. An understanding of prototype concepts as weighted attribute value structures has already been used by Smith et al. (1988) for explaining modifications such as "red apple". Prototype frames extend this by explicitly using probabilistic weights. In this section, we aim to show that stochastic frames can be used to model one of the core phenomena of prototype concepts, namely the existence of typical and atypical category members. For example, apples are typical fruit, while avocados are not. The structures in (2) are partial frames with probability information for fruit, apple and avocado.

(2) [fruit
COLOUR: red 0.3 green 0.1 yellow 0.3 orange 0.2 other 0.1,
TASTE: sweet 0.6 sour 0.3 other 0.1]
[apple
COLOUR: red 0.5 green 0.2 yellow 0.2 orange 0 other 0.1
TASTE: sweet 0.8 sour 0.1 other 0.1]
[avocado
COLOUR: red 0 green 0.7 yellow 0 orange 0 other 0.3,
TASTE: sweet 0 sour 0 other 1]

Probability information can be used to define diagnostic and frequent properties, i.e. attribute values V, such as sweet taste (Schurz, 2012):

(3) A property V is frequent for a class C iff $P(V|C)$ is high
A property V is diagnostic for a class C iff $P(C|V)$ is high

cus parser is based (Marcus, 1980)

The latter is well-known as the notion of *cue validity*. It allows a definition of the diagnosticity of an attribute A in (4), where $V_1, V_2, ..., V_n$ are alternative values of the attribute A:

(4) $diag(A, C) = max(P(C|V_1), P(C|V_2), ..., P(C|V_n))$

The similarity Sim of the probability distributions of properties on one attribute (i.e., the frequency of the values) in a concept C and another concept, for example, a subcategory SC, can be compared in terms of (5):

(5) $Sim(C, SC|A)) = \sum_{i=1}^{n} min(P(V_i|C), P(V_i|SC))$

Sim can be used to express that the probability distribution of COLOR in 'apple' is quite similar to the one for 'fruit' $(0.3 + 0.1 + 0.2. + 0.1 = 0.6)$ but not so similar for 'avocado' and 'fruit' $(0 + 0.1 + 0 + 0 + 0.1 = 0.2)$.

Finally, the typicality of a subcategory is determined as the diagnosticity-weighted average similarity in all contributing attributes:

(6) $typ(C, SC) = \sum_{i=1}^{n} \frac{diag(A_i|C)}{\sum_{i=1}^{n} diag(A_i|C)} Sim(C, SC|A_i)$

With this formula, one can quantify how typical fruit apples or avocados are as an diagnosticity weighted average of similarities in all contributing attributes.

4 Final remarks

The paper presents a notion of a stochastic frame that represents concepts and the linguistic knowledge of agents in terms of attribute-value structures in which values may only occur with some probability. We outlined how probabilistic constraints on stochastic frames facilitate reasoning about probable features (attribute values) in conditions of uncertainty. What comes out of this can be interpreted a formalisation of the prototype theory of concepts in which all other theories of concepts can be understood as special cases. By trivialising the distributions, one obtains the classical view. (Products of) regions in conceptual spaces are obtained by deriving such regions from actual distributions (it becomes hard to see such an account of concepts as properly different from the prototype view).

Taking this common core, we also outlined three areas in which stochastic frames have obvious applications: vague predication, lexical ambiguity, and the typicality of kinds. A shared property of these phenomena is arguably that, in all cases, individuals must reason with complex, multifaceted concepts in conditions of uncertainty, be this uncertainty about the extension of a term (vagueness), uncertainty about the meaning of a term in use (lexical ambiguity), or uncertainty about properties of the instances (the typicality of kinds). For an explanation of all of these cases, we seem to need not only something along the lines of a probabilistic component to drive the reasoning process and model graded or fuzzy phenomena, but also a means of applying this reasoning tool to different aspects or properties of the entities being reasoned about. Representational structures such as frames give us the structure we need in this respect. Stochastic frames, therefore, give us the right combination of conceptual structure and a formal theory of reasoning.

Acknowledgements

This research was funded by the German Research Foundation (DFG) funded project: CRC 991 *The Structure of Representations in Language, Cognition, and Science*, specifically projects C09, D01 and a Mercator Fellowship awarded to Henk Zeevat. We would like to thank audiences at CoST 2019 at HHU Düsseldorf, the workshop on Records, Frames, and Attribute Spaces held at ZAS in Berlin, March 2018, and the Workshop on Uncertainty in Meaning and Representation in Linguistics and Philosophy held in Jelenia Góra, Poland, February, 2018.

References

Nicholas Asher. 2011. *Lexical Meaning in Context: A Web of Words*. Cambridge University Press.

Lawrence. W. Barsalou. 1992. Frames, concepts, and conceptual fields. In E. Kittay and A. Lehrer, editors, *Frames, fields, and contrasts: New essays in semantic and lexical organization*, pages 21–74. Erlbaum.

Jeffrey R Binder, Lisa L Conant, Colin J Humphries, Leonardo Fernandino, Stephen B Simons, Mario Aguilar, and Rutvik H Desai. 2016. Toward a brain-based componential semantic representation. *Cogn Neuropsychol*, 33(3-4):130–174.

Bob Carpenter. 1992. *The logic of typed feature structures*. Cambridge: Cambridge University Press.

Daniel Casasanto and Gary Lupyan. 2015. All concepts are ad hoc concepts. In Eric Margolis and Stephen Laurence, editors, *The Conceptual Mind: New Directions in the Study of Concepts*, pages 543–566. MIT Press.

Robin Cooper, Simon Dobnik, Staffan Larsson, and Shalom Lappin. 2015. Probabilistic type theory and natural language semantics. *LILT*, 10(4).

Charles Dickens. 2000. *David Copperfield*. Modern Library, New York.

Paul Égré. 2017. Vague judgment: A probabilistic account. *Synthese*, 194(10):3837–3865.

Peter Gärdenfors. 2000. *Conceptual Space: The Geometry of Thought*. The MIT Press, Cambridge, MA.

Lotte Hogeweg. 2009. *Word in Process. On the interpretation, acquisition and production of words*. Ph.D. thesis, Radboud University Nijmegen.

Daniel Lassiter. 2011. Vagueness as probabilistic linguistic knowledge. In R. Nouwen, U. Sauerland, H.C. Schmitz, and R. van Rooij, editors, *Vagueness in Communication*. Springer.

Sebastian Löbner. 2014. Evidence for frames from human language. In *Frames and concept types*, pages 23–67. Springer International Publishing.

Mitchell P. Marcus. 1980. *Theory of Syntactic Recognition for Natural Languages*. MIT Press, Cambridge, MA, USA.

Eric Margolis and Stephen Laurence. 1999. Concepts and cognitive science. In Eric Margolis and Stephen Laurence, editors, *Concepts: Core Readings*. MIT Press, Cambridge, MA.

Marvin Minsky. 1974. A framework for representing knowldege. *MIT-AI Laboratory Memo*, 306.

Wiebke Petersen. 2007. Representation of concepts as frames. In J Skilters, editor, *Complex Cognition and Qualitative Science. The Baltic International Yearbook of Cognition, Logic and Communication Vol. 2*, pages 151–170. Springer International Publishing.

Eleanor Rosch. 1973. Natural categories. *Cognitive Psychology*, 4:328–350.

Eleanor Rosch. 1978. Principles of categorization. In *Cognition and Categorization*, pages 27–48. Lawrence Elbaum Associates.

Eleanor Rosch and Carolyn B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4):573–605.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439.

Gerhard Schurz. 2012. Prototypes and their composition from an evolutionary point of view. In W. Hinzen, E. Machery, and Werning M., editors, *The Oxford Handbook of Compositionality*, pages 530–553. Oxford University Press, Oxford, UK.

Edward E. Smith, Daniel N. Osherson, Lance J. Rips, and Margaret Keane. 1988. Combining prototypes: A selective modification model. *Cognitive science*, 12(4):485–527.

Paul Smolensky. 1991. Connectionism, constituency and the language of thought. In M. Loewer and G. Rey, editors, *Meaning in Mind: Fodor and His Critics*, pages 201–227. Blackwell, Oxford.

Peter R. Sutton. 2015. Towards a probabilistic semantics for vague adjectives. In Hans-Christian Schmitz and Henk Zeevat, editors, *Language, Cognition, and Mind*. Springer.

Peter R. Sutton. 2018. Probabilistic approaches to vagueness and semantic competency. *Erkenntnis*, 83(4):711–740.

Ludwig Wittgenstein. 1953. *Philosophische Untersuchungen*. Suhrkamp Verlag, Frankfurt am Main.

Henk Zeevat, Scott Grimm, Lotte Hogeweg, Sander Lestrade, and E. Allyn Smith. 2015. Representing the lexicon: Identifying meaning in use via overspecification. In Kata Balogh and Wiebke Petersen, editors, *Proceedings of Workshop Bridging Formal and Conceptual Semantics (BRIDGE-14)*. Düsseldorf University Press.

A toy distributional model for fuzzy generalised quantifiers

Mehrnoosh Sadrzadeh
Department of Computer Science
University College London
m.sadrzadeh@ucl.ac.uk

Gijs Wijnholds
Utrecht Institute of Linguistics OTS
Utrecht University
g.j.wijnholds@uu.nl

Abstract

Recent work in compositional distributional semantics showed how bialgebras model generalised quantifiers of natural language. That technique requires working with vector space over power sets of bases, and therefore is computationally costly. It is possible to overcome the computational hurdles by working with fuzzy generalised quantifiers. In this paper, we show that the compositional notion of semantics of natural language, guided by a grammar, extends from a binary to a many valued setting and instantiate in it the fuzzy computations. We import vector representations of words and predicates, learnt from large scale compositional distributional semantics, interpret them as fuzzy sets, and analyse their performance on a toy inference dataset.

1 Introduction

The work of [10] showed how one can reason about generalised quantifiers using bialgebras over the category of sets and relations over a fixed powerset object (powerset of a universe of discourse). This provides us with an abstract categorical semantics, which when instantiated to category of sets and relations, one will obtain a truth-theoretic semantics. The abstract setting, however, can also be instantiated to category of vector spaces and linear maps, in which one obtains a compositional distributional semantics, in the style of [6, 9]. The downside is that the resulting vector spaces span over powersets of the usual set of bases and the complexity of reasoning in the setting explodes. It is also not very clear how can one learn the new basis vectors, consisting of sets of vectors, rather than just one vectors. One solution would be to move to a fuzzy setting, as done in [17]. The rationale behind this move is as follows: fuzzy sets have been encoded in the category of sets and many valued relations and the categorical setting of [10] also instantiates to these categories. We demonstrate the details of this construction in the Springer Outstanding Contributions volume in honor of M. Ardeshir. In that paper, we show that the categorical version of fuzzy sets $\mathbf{V}$-$\mathbf{Rel}$ of sets and many valued relations, is compact closed and define over it the necessary bialgebras to encode Zadeh's fuzzy generalised quantifiers.

In this paper, we spare the categorical technicalities, and review the definitions of generalised quantifiers in a compositional relational setting (sets and relations) guided by an elementary generative grammar. Independently, we also review the definitions of fuzzy generalised quantifiers of Zadeh, using the notions of fuzzy sets and possibility distributions. We then explain how fuzzy sets can be modelled by many valued relations and define a compositional semantics for sentences with fuzzy generalised quantifiers in this setting. Finally, we interpret vectors as fuzzy sets and show how the many valued semantic computations can be done over vectorial data. We demonstrate the workings of our model on toy vectors extracted from real data and compute a degree of truth for quantified sentences containing them. In order to ground our semantics, i.e. conclude that these computations are sound, we use the results in a toy inference task and analyse the results.

Finally, although fuzzy concepts are often motivated by vague predicates such as short and tall, fuzzy generalised quantifiers have a large, if not full, overlap with natural language generalised quantifiers. Most of the latter are non-logical and consider words such as "almost, many, most, few", and these are exactly the fuzzy generalised quantifiers that Zadeh deals with. Further, and as we will see in the examples that have been worked out in the paper, one can apply our methodology to deal with logical part of generalised quantifiers, i.e. words such as "all" and "some".

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 86–94

2 Generalised Quantifiers as Relations

We briefly review the theory of generalised quantifiers in natural language as presented in [2]. Consider the fragment of English generated by the following context free grammar.

S $\rightarrow$ NP VP
VP $\rightarrow$ V NP
NP $\rightarrow$ Det N
NP $\rightarrow$ John, Mary, ...
N $\rightarrow$ cat, dog, man, ...
VP $\rightarrow$ sneeze, sleep, ...
V $\rightarrow$ love, kiss, ...
Det $\rightarrow$ some, all, no, most, almost all, several, ...

A model for the language generated by this grammar is a pair $(U, [\![\ ]\!])$, where U is a universal reference set and $[\![\ ]\!]$ is an inductively defined interpretation function. In order to keep the semantic simple, we will not fully follow formal semantics guidelines and shall not treat noun phrases as general quantifiers. Noun phrases and nouns are treated similarly and by sets. The $[\![\ ]\!]$ of terminals is thus defined via the following cases:

1. The interpretation of a determiner d generated by 'Det $\rightarrow d$' is the following map:

$$[\![d]\!] \colon \mathcal{P}(U) \rightarrow \mathcal{PP}(U)$$

It assigns to each $A \subseteq U$, a family of subsets of U. The images of these interpretations are referred to as *generalised quantifiers*. For logical quantifiers, these are:

$$\begin{aligned} [\![\text{some}]\!](A) &= \{X \subseteq U \mid X \cap A \neq \emptyset\} \\ [\![\text{every}]\!](A) &= \{X \subseteq U \mid A \subseteq X\} \\ [\![\text{no}]\!](A) &= \{X \subseteq U \mid A \cap X = \emptyset\} \\ [\![n]\!](A) &= \{X \subseteq U \mid\mid X \cap A \mid = n\} \end{aligned}$$

A similar method is used to define non-logical quantifiers, for example "most A" is defined to be the set of subsets of U that has 'most' elements of A, "few A" is the set of subsets of U that contain 'few' elements of A, and similarly for 'several' and 'many'.

Generalising the two cases above, provides us with the following definition for semantics $[\![d]\!](A)$ of any generalised quantifier d:

$$\{X \subseteq U \mid \quad X \text{ has } d \text{ elements of } A\}$$

2. The interpretation of a terminal $y \in \{np, n, vp\}$ generated by either of the rules 'NP $\rightarrow$ np, N $\rightarrow$ n, VP $\rightarrow$ vp' is $[\![y]\!] \subseteq U$. That is, noun phrases, nouns and verb phrases are interpreted as subsets of the reference set.

3. The interpretation of a terminal y generated by the rule V $\rightarrow$ y is $[\![y]\!] \subseteq U \times U$. That is, verbs are interpreted as binary relations over the reference set.

The semantics of $[\![\ ]\!]$ on non-terminals is defined according to the following cases:

1. The interpretation of expressions generated by the rule 'NP $\rightarrow$ Det N' is:

$$[\![\text{Det N}]\!] = [\![d]\!]([\![n]\!])$$

where $X \in [\![d]\!]([\![n]\!])$ iff $X \cap [\![n]\!] \in [\![d]\!]([\![n]\!])$, for Det $\rightarrow d$ and N $\rightarrow n$. This condition is often referred to as *conservativity* or the *living on* property of generalised quantifiers.

2. The interpretations of expressions generated by other rules are as usual:

$$\begin{aligned} [\![\text{V NP}]\!] &= [\![v]\!]([\![np]\!]) \\ [\![\text{NP VP}]\!] &= [\![vp]\!]([\![np]\!]) \end{aligned}$$

Here, for $R \subseteq U \times U$ and $A \subseteq U$, by $R(A)$ we mean the forward image of R on A, that is $R(A) = \{y \mid (x, y) \in R, \text{for } x \in A\}$. To keep the notation unified, for R a unary relation $R \subseteq U$, we use the same notation and define $R(A) = \{y \mid y \in R, \text{for } x \in A\}$, i.e. $R \cap A$.

The 'meaning' of a sentence in this setting is its truth value. So we have that a sentence is true iff $[\![\text{NP VP}]\!] \neq \emptyset$ and false otherwise. For the cases of quantified sentences considered in this paper, i.e. sentences with quantified subject and object phrases, a truth value is defined as follows:

1. A sentence of the form 'Det N VP' is *true* iff $[\![\text{Det N VP}]\!] = [\![vp]\!] \cap [\![n]\!] \in [\![\text{Det N}]\!]$ and *false* otherwise.

2. A sentence of the form 'NP V Det N' is *true* iff $[\![\text{NP V Det N}]\!] = [\![n]\!] \cap [\![v]\!]([\![np]\!]) \in [\![\text{Det N}]\!]$ and *false* otherwise.

For example, the sentence 'some cats slept' with a quantifier at the subject phrase is true iff $[\![\text{slept}]\!] \cap$

$[\![\text{cats}]\!] \in [\![\text{some cats}]\!]$, that is, whenever the set of things that sleep and are cats is a non-empty set. Similarly, a sentence with a quantified phrase at its object position, for instance, 'Cats like some rats' is true iff $[\![\text{rats}]\!] \cap [\![\text{likes}]\!]([\![\text{cats}]\!]) \in [\![\text{some rats}]\!]$, that is, whenever, the set of things that are liked by cats and are rats is a non-empty set. Similarly, the sentence 'Cats liked three rats' is true iff the set of things that are liked by cats and are rats has three elements in it.

3 Zadeh's Fuzzy Generalised Quantifiers

In this section we review definitions of fuzzy sets and quantifiers, as done by Zadeh [17]. A fuzzy set is a set whose elements have a corresponding weight associated to them. For a set A, the weight μ_i of element u_i is interpreted as the degree of membership of u_i in A. The fuzzy set A with n elements is represented symbolically by a sum:

$$A = \mu_1 u_1 + \mu_2 u_2 + \cdots + \mu_n u_n$$

The cardinality of a fuzzy set is defined via the notion of *sigma-count*, defined below:

$$\Sigma Count(A) = \Sigma_{i=1}^{n} \mu_i$$

Terms whose degrees of membership fall below a certain threshold, may be omitted from the sum. This is to avoid a situation where a large number of terms with low degrees become equivalent to a small number of terms with high degrees.

The quantified sentences Zadeh considers are built from two basic forms: "There are Q A's" and "Q A's are B's". Each of these propositions induces a possibility distribution. Zadeh provides the following insights for the analysis of these quantified propositions. "There are Q A's" implies that the probability of event A is a fuzzy probability equal to Q. "Q A's are B's" implies that the conditional probability of event B given event A is a fuzzy probability which is equal to Q. Most statements involving fuzzy probabilities may be replaced by semantically equivalent propositions involving fuzzy quantifiers and this is the statement we work with in this paper.

The fuzzy semantics of a proposition p is interpreted as "the degree of truth of p", or the possibility of p, where possibility is treated in an elementary way, i.e. a function from a set to the unit interval. In order to compute this, one translates p into a *possibility assignment equation*, denoted by $\Pi_{(X_1,\cdots,X_n)} = F$, where F is a fuzzy subset of the universe of discourse U and $\Pi_{(X_1,\cdots,X_n)}$ is the joint possibility distribution over (explicit or implicit) variables $X_1, \cdots, X_n$ of p. For instance, the proposition "Vickie is tall" is translated as:

$$\Pi_{Height(Vickie)} = TALL$$

Here, $TALL$ is a fuzzy subset, $Height(Vickie)$ is a variable implicit in "Vickie is tall", and $\Pi_{Height(Vickie)}$ is the possibility distribution of this variable. Following Zadeh, we use the $=$ sign, but are aware that this makes the reading awkward. The reader is encouraged to treat (as we did) $=$ as an informal assignment. The above possibility assignment equation implies that

$$Poss\{Height(Vickie) = u\} = \mu_{TALL}(u)$$

where $Poss\{X = u\}$ the possibility that X is u, for u a specified value. The above thus states that "the possibility that height of Vickie is u is equal to $\mu_{TALL}(u)$, that is, is the grade of membership of u in the fuzzy set $TALL$. Quantified sentences are translated in a similar way. For instance, "Vickie has several credit cards", is translated to the following:

$$\Pi_{Count(CreditCards(Vickie))} = SEVERAL$$

Suppose that 4 is compatible with the meaning of "several" with degree 0.8, then the above implies that, for instance, the possibility that Vickie has 4 credit cards is

$$Poss\{Count(CreditCards(Vickie)) = 4\} = 0.8$$

In order to analyse sentences of the general forms "There are Q A's" and 'Q A's are B's", Zadeh assumes that they are semantically equivalent to the following:

$$\begin{aligned} \text{There are } Q\ A\text{'s} &\rightsquigarrow \Sigma Count(A) \text{ is } Q \\ Q\ A\text{'s are } B\text{'s} &\rightsquigarrow Proportion(B|A) \text{ is } Q \end{aligned}$$

Here, $Poportion(B|A)$ is the proportion of elements of B that are in A, aka the relative cardinality of B in A, formally defined as follows:

$$\Pi_{Proportion(B|A)} := \frac{\Sigma Count(A \cap B)}{\Sigma Count(A)}$$

Both $Proportion(B|A)$ and $\Sigma Count(A)$ may be fuzzy or non-fuzzy counts. Zadeh then formalises

the above counts as possibility assignment equations as follows

$$\begin{aligned} \Sigma Count(A) \text{ is } Q &\rightsquigarrow \Pi_{\Sigma Count(A)} = Q \\ Proportion(B|A) \text{ is } Q &\rightsquigarrow \Pi_{Proportion(B|A)} = Q \end{aligned}$$

In the spirit of truth-conditional semantics, the weight of each of the elements of the set can be interpreted as the degree of truth of the proposition denoted by the element. This weight is $Q(\Sigma Count(A))$ for sentences of the form "There are Q A's" and $Q(Proportion(B|A))$ for sentences of the form "Q A's are B's".

Writing $\mu_A(u)$ for the degree of membership of u in the fuzzy set A, we define the intersection of two fuzzy sets A and B as

$$A \cap B = \Sigma_i \ \min(\mu_A(u_i), \mu_B(u_i))\ u_i$$

where i is understood to range over all the elements in A and B (when an element is in A but not in B it will still be represented in A with a degree of 0). A similar version without the Σ is used to define it for the non-fuzzy case.

Example. Let's say we have a universe

$$U = \{u_1, u_2, u_3, u_4, u_5\}$$

and fuzzy sets KP for "kind people" and BP for "big men", defined as follows:

$$\begin{aligned} KP &= 0.5u_1 + 0.8u_2 + 0.2u_3 + 0.6u_4 \\ BM &= 0.8u_1 + 0.3u_2 + 0.1u_3 + 0.9u_4 + 1u_5 \end{aligned}$$

The quantified sentence "Most big men are kind", is translated to the following possibility assignment equation $\Pi_{Proportion(KP|BM)} = MOST$. The intersection of KP and BM is computed as follows:

$$KP \cap BM = 0.5u_1 + 0.3u_2 + 0.1u_3 + 0.6u_4$$

The proportion of big men that are kind, i.e. $Proportion(KP|BM)$, is computed as follows:

$$\frac{\Sigma Count(BM \cap KP)}{\Sigma Count(BM)} = \frac{1.5}{3.1}$$

Suppose that proportions between 0.6 and 0.7 are compatible with the meaning of $MOST$ with degree 0.75. Then, since $\frac{1.5}{3.1} = 0.48$, the degree of truth of our sentence is below 0.75. For the crisp quantifier ALL, the sentence "All big men are kind" is, since only the proportion 1 is compatible with the meaning of ALL with degree 1, which is not the case here.

Possibility distributions can be learnt, e.g. Zadeh develops a test-score procedure by sampling from a database of related data.

4 Fuzzy Generalised Quantifiers as Many Valued Relations

A *many-valued relation* between two sets A and B is denoted by $R : A \nrightarrow B$ and is a function $R : A \times B \to \mathbf{V}$, where $\mathbf{V}$ is a commutative quantale of values, usually the unit interval $[0, 1]$. This function is viewed as a $\mathbf{V}$-valued matrix. We compose two relations $R : A \nrightarrow B$ and $S : B \nrightarrow C$ to get a relation $S \circ R : A \nrightarrow C$ such that

$$(S \circ R)(a, c) = \bigvee_{b \in B} (R(a, b) \bullet S(b, c))$$

holds in $\mathbf{V}$. Here, $\bullet$ and $\bigvee$ are operations on the numbers in the quantale $\mathbf{V}$. When $\mathbf{V}$ is the real interval $[0, 1]$ with operations $\min$ and $\max$, the composition of two $\mathbf{V}$-relations becomes as follows. Given two $\mathbf{V}$-relations $R : A \nrightarrow B$ and $S : B \nrightarrow C$ (so two functions $R : A \times B \to [0, 1]$ and $S : B \times C \to [0, 1]$), the composite $S \circ R : A \nrightarrow C$ is given by

$$(S \circ R)(a, c) = \max_{b \in B} \min(R(a, b), S(b, c)).$$

We refer to sets and many valued relations on them as $\mathbf{V}$-$\mathbf{Rel}$.

A non-fuzzy generalised quantifier d is interpreted as a relation $[\![d]\!]$ over the power set of the universe of discourse $P(U)$, where it relates a subset $A \subseteq U$ to subsets $B \subseteq U$, based on the cardinalities of A and B. The fuzzy version of this quantifier is interpreted as a many valued relation over $P(U)$, where, in fuzzy set notation, it relates A to subsets $u_i \subseteq U$ and assigns to each such subset a degree of membership μ_i. The result is a fuzzy set whose weights come from a possibility distribution over the relative cardinalities of A and u_i's. In Zadeh's notation:

$$[\![d]\!](Proportion(u_i|A)) = \mu_i \qquad (1)$$

For $\mathbf{V} = [0, 1]$ and given a fuzzy generalised quantifier for which we have $\Pi_{Proportion(B|A)} = \overline{[\![d]\!]}$, we define its $\mathbf{V}$-$\mathbf{Rel}$ encoding to be the many valued relation $\overline{[\![d]\!]} : P(U) \nrightarrow P(U)$, with values

Table 1: Sentences of our Toy Dataset and their Annotated Degrees of Entailment

Entry 1	Entry 2	Deg.
people strike	group attacks	4.31
notice advertises	sign announces	5.37
clarify rule	explain process	5.00
recommend development	suggest improvement	5.37
people clarify rule	group explain process	5.00
corporation recommend development	firm suggest improvement	5.375
office arrange task	staff organize work	5.50
editor threatens	application predicts	1.12
progress reduces	development replaces	1.22
confirm number	approve performance	1.81
editor threatens man	application predicts number	1.12
man recall time	firm cancel term	1.62

coming from the possibility distribution of $\overline{[\![d]\!]}$, defined as follows:

$$\overline{[\![d]\!]}(A,B) = \mu_i, \text{ for } \mu_i = [\![d]\!](Proportion(B|A))$$

In order to obtain a many valued relation in $\mathbf{V\text{-}Rel}$, we need a numerical value assigned to subsets A and B of universe. This number is nothing but the weight of $[\![d]\!](Proportion(B|A))$.

The semantics of a sentence of our grammar extends from sets and relations to sets and many valued relations. We define a $\mathbf{V\text{-}Rel}$ model to be the tuple $(\mathbf{V\text{-}Rel}, P(U), \overline{[\![\ ]\!]})$ over a universe of discourse U. In this model, the language constructions are interpreted as follows:

1. A terminal x of either category N,NP, or VP is interpreted as a many valued relation whose value is the degree to which a subset A of the universe is $[\![x]\!]$. This is the relative sigma count of the subset A in $[\![x]\!]$, that is:

$$\star\overline{[\![x]\!]}A := Proportion(A|[\![x]\!])$$

2. A terminal x of category V is interpreted as a many valued relation whose value is the degree to which its image on a subset A of universe is a subset B of the universe, that is the relative sigma count of B in $[\![x]\!](A)$:

$$\star\overline{[\![x]\!]}(A,B) = Proportion(B|[\![x]\!](A))$$

where $[\![x]\!](A)$ is the application of $[\![x]\!]$ to A, resulting in a set $\Sigma_{i=1}^{n}\mu_i b$ where the subscripts of the μ's vary over elements of fuzzy sets A and $[\![v]\!]$, so we have

$$\max_{a_i} \min(\mu_A(a_i), \mu_{[\![v]\!]}(a_i, b_i))$$

Here, μ_A and $\mu_{[\![v]\!]}$ are degrees of memberships of elements of fuzzy sets A and $[\![v]\!]$, respectively.

Using the above interpretation, a quantified sentence s gets a a degree of truth $r \in [0,1]$ as its semantics iff $\overline{[\![s]\!]} = r$ in $(\mathbf{V\text{-}Rel}, \mathcal{P}(U), \overline{[\![\ ]\!]})$. Using this definition, we compute the semantics of the sentence "several cats sleep" with a fuzzy quantifier at subject position becomes as follows:

$$\max_{(A,B)} \min\left(\star\overline{[\![cats]\!]}A, \star\overline{[\![sleep]\!]}B, A\overline{[\![several]\!]}A\cap B\right)$$

This formula will get a maximal value for $A = [\![cats]\!], B = [\![sleep]\!]$ and when assuming that $\Pi_{Proportion(A\cap B|A)} =$ several, in which case the value of semantics becomes as follows:

$$[\![several]\!]\left[\frac{\Sigma Count([\![cats]\!] \cap [\![sleep]\!])}{\Sigma Count([\![cats]\!])}\right]$$

To compute this concretely, suppose that the fuzzy sets $[\![cats]\!]$ and $[\![sleep]\!]$ are defined as follows:

$$[\![cats]\!] = 0.2c_1 + 0.3c_2 + 0.8c_3$$
$$[\![sleep]\!] = 0.5c_1 + 0.4c_2 + 0.4c_3$$

Then the value for "several cats sleep" will be

$$[\![several]\!]\left[\frac{\Sigma Count(0.2c_1 + 0.3c_2 + 0.4c_3)}{0.2c_1 + 0.3c_2 + 0.8c_3}\right]$$
$$= [\![several]\!]\left[\frac{0.9}{1.3}\right]$$

Suppose that the possibility distribution [\![several]\!] will map low values to low values and very high values to low values, but intermediate values would be mapped to a high number as they still represent "several". Thus the proportion $\frac{9}{13}$, which is a high number, will evaluate to a high number. Thus the many valued relation of this statement will be high (a number close to 1). For examples of possibility distributions of some other fuzzy quantifiers, see [17].

Table 2: Degrees of Truth of the Non Quantified Sentences of the Toy Data Set

Entry 1	Entry 2	Deg.
people strike	group attacks	[0.63, 0.42]
notice advertises	sign announces	[0.57, 0.6]
clarify rule	explain process	[0.54, 0.47]
recommend development	suggest improvement	[0.69, 0.59]
people clarify rule	group explain process	[0.50, 0.34]
corporation recommend development	firm suggest improvement	[0.60, 0.49]
office arrange task	staff organize work	[0.61, 0.62]
editor threatens	application predicts	[0.36, 0.49]
progress reduces	development replaces	[0.69, 0.61]
confirm number	approve performance	[0.47, 0.67]
editor threatens man	application predicts number	[0.42, 0.36]
man recall time	firm cancel term	[0.65, 0.11]

The semantics of a sentence "Mice eat several plants" with a fuzzy quantifier at object place is computed as follows. Suppose we have fuzzy sets

$$\begin{aligned}\llbracket mice \rrbracket &= 0.7c_1 + 0.6c_2 + 0.2c_3\\ \llbracket eat \rrbracket &= 0.5(c_1,c_1) + 0.8(c_1,c_3) + 0.2(c_2,c_1)\\ &\quad + 0.3(c_2,c_3) + 0.9(c_3,c_3)\\ \llbracket plants \rrbracket &= 0.2c_1 + 0.3c_2 + 0.6c_3\end{aligned}$$

Then the semantics we get is

$$\llbracket several \rrbracket\Big[\frac{\Sigma Count(\llbracket eat \rrbracket(\llbracket mice \rrbracket) \cap \llbracket plants \rrbracket)}{\Sigma Count(\llbracket plants \rrbracket)}\Big]$$

The application of the verb to its subject gives

$$\llbracket eat \rrbracket(\llbracket mice \rrbracket) = 0.5c_1 + 0.7c_3$$

As a result, the whole expression now evaluates to

$$\begin{aligned}&\llbracket several \rrbracket\Big[\frac{\Sigma Count(0.2c_1 + 0.6c_3)}{\Sigma Count(0.2c_1 + 0.3c_2 + 0.6c_3)}\Big]\\ &\quad = \llbracket several \rrbracket\Big[\frac{0.8}{1.1}\Big]\end{aligned}$$

This gives another relatively high value for the many valued semantics of this sentence, as $Proportion(\llbracket eat \rrbracket(\llbracket mice \rrbracket)|\llbracket plants \rrbracket)$ certainly indicates a case of "several" mice eating plants.

Finally, for the case where we have fuzzy quantifiers at both subject and object places, e.g. in the sentence "Several mice eat most plants", a semantics is computed as follows. Given that the fuzzy sets representing mice and plants are as before and taking the same fuzzy relation for ⟦eat⟧, we compute the meaning of this sentence. Suppose further that ⟦most⟧ is a possibility distribution that assigns the value 0 to numbers below 0.5, and gradually increasing the value for numbers from 0.5 to 1. In this case, first, we compute the application of the quantifiers to their respective noun phrases:

$$\begin{aligned}&\llbracket several \rrbracket[\llbracket mice \rrbracket] =\\ &\arg\max_B \Big(\llbracket several \rrbracket\Big[\frac{\Sigma Count(\llbracket mice \rrbracket \cap B)}{\Sigma Count(\llbracket mice \rrbracket)}\Big]\Big)\end{aligned}$$

If we assume that "several" has the highest value for 0.4, then it would for instance assign to the set $0.4\llbracket mice \rrbracket$ the value $\Sigma_i 0.4\mu_i u_i$ for $\mu_i u_i$ in $\llbracket mice \rrbracket$. The second application gives

$$\begin{aligned}&\llbracket most \rrbracket[\llbracket plants \rrbracket] =\\ &\arg\max_A \Big(\llbracket most \rrbracket\Big[\frac{A \cap \llbracket plants \rrbracket}{\Sigma Count(\llbracket plants \rrbracket)}\Big]\Big)\end{aligned}$$

This will set $A = \llbracket plants \rrbracket$, given that 1 has the highest probability of being "most".

The value of the whole sentence will be the verb applied to the quantified subject and object, hence we obtain

$$\begin{aligned}&\llbracket eat \rrbracket\Big[\llbracket several \rrbracket[\llbracket mice \rrbracket], \llbracket most \rrbracket[\llbracket plants \rrbracket]\Big]\\ =&\llbracket eat \rrbracket\Big[0.4\llbracket mice \rrbracket, \llbracket plants \rrbracket\Big]\\ =&\max_{a,b} \min(\mu_{0.4\llbracket mice \rrbracket}(a), \mu_{\llbracket eat \rrbracket}(a,b), \mu_{\llbracket plants \rrbracket}(b))\\ =&\max\Big(\min(0.28, 0.5, 0.2), \min(0.28, 0.8, 0.6),\\ &\min(0.24, 0.2, 0.2), \min(0.24, 0.3, 0.6),\\ &\min(0.08, 0.9, 0.6)\Big)\\ =&\max(0.2, 0.28, 0.2, 0.24, 0.08) = 0.28\end{aligned}$$

That is, the extent to which several mice eat most plants is 28%.

5 From Many Valued Relations to Vectors Spaces and Linear Maps

By transferring our natural language semantics of quantified sentences from sets and relations to

Table 3: Degrees of Truth of the Quantified Sentences of the Toy Data Set

Entry 1	Entry 2	Deg.
all people strike	several groups attacks	[0.8, 0.8]
all noticesadvertise	many signs announce	[0.8, 0.8]
clarify several rules	explain some processes	[0.8, 0.8]
recommend many developments	suggest several improvements	[0.8, 0.8]
all people clarify rule	several groups explain process	[0.2, 0.8]
all corporations recommend development	many firms suggest improvement	[0.8, 0.2]
several offices arrange task	some staff organize work	[0.8, 0.8]
few editors threaten	all applications predict	[0.8, 0.2]
many progresses reduce	all developments replace	[0.8, 0.2]
confirm several numbers	approve few performances	[0.8, 0.2]
few editors threaten man	all applications predict number	[0.8, 0.2]
few men recall time	many firms cancel term	[0.2, 0.2]

sets and many valued relations, we generalised our Boolean-valued true-false semantics to a many valued semantics with degrees of truth from the unit interval $[0, 1]$. Transferring sets and many valued relations to vector spaces and linear maps enables us to compute the meaning of our sentences via quantitative reasoning on the statistical data provided in distributional semantics. A distributional vector for a target word w is seen as a fuzzy set whose degrees of membership are the degrees of co-occurrences of w with a set of context words c, or the degrees of contextual relevance of w to c, or other similar readings. In this section we use this interpretation and implement the formulae for obtaining the degrees of truth of fuzzy generalised quantifiers on vectors obtained from the combined UKWac/Wackypedia corpus [16], extracted using normalised co-occurrence counts, to ensure that the vectors indeed represent fuzzy sets. As Zadeh only provides semantics for a set of two atomic quantifiers without invoking grammatical compositionality, we use the many valued semantics of the previous section to compute the semantics of our quantified sentences compositionally.

We start with the sentence entailment dataset of [13] and later combine it with the entailed generalised quantifiers of [1]. The former dataset is a small dataset of 12 pairs of sentences, classified in three bands: high entailment, medium entailment, and low entailment. Each band is annotated with human judgements with a number in the range 1-7, representing the degree of entailments between each pair of sentences. The bands of entailment are decided upon based on the averages of the annotations. For the purpose of this paper, we only work with clear (non-)entailments and thus only present data for the high and low bands. In the high band, both subjects/objects and verbs/verb phrases/intransitive verbs entail each other. These got an average annotation of 4 and above. In the low band are the non-entailing entries, i.e. neither of the subjects/objects or verbs/verb phrases/intransitive verbs entail each other. These got an average annotation of 2 or under. The entries of the dataset and their annotated degrees of entailment are given in table 1.

We implement our fuzzy semantics on the sentences of each entry and obtain degrees of truth. These degrees are given in table 2, with a horizontal line separating the high and low bands of entailment. A list of entailing and non-entailing quantifiers were provided in [1]. The list is a mixture of logical and generalised quantifiers. Entailment for logical quantifiers, e.g. 'all' and 'some', is clear and so we drop these as well as numerical quantifiers such as 'both' which are not fuzzy. The case of generalised quantifiers is the interesting one. Here, the degrees of truth of the non quantified sentences will change after applying the quantifiers to them, as we saw in the examples of the previous section for the quantifier 'several'. Thus we work with the generalised quantifier subset of the data set of [1]. These are as follows:

Entailing	Non-Entailing
all, several	several, all
all, many	many, all
several, some	several, few
many, several	few, all
	few, many

We use the entailing ones to strengthen the entailments of the high band and the non-entailing ones to weaken the entailments of the low band. This provides us with the following dataset, the positive entailments are separated from the negative ones by a bar in the table. Hardcoding the meaning of generalised quantifiers, in the same way as we did in the previous section, provides

,

Table 4: Degrees of Entailment for Sentences Before and After Quantification

Entry 1	Entry 2	Deg.1	Deg.2
people strike	group attacks	-	+
notice advertises	sign announces	+	+
clarify rule	explain process	-	+
recommend development	suggest improvement	-	+
people clarify rule	group explain process	-	+
corporation recommend development	firm suggest improvement	-	-
office arrange task	staff organize work	+	+
editor threatens	application predicts	+	-
progress reduces	development replaces	-	-
confirm number	approve performance	+	-
editor threatens man	application predicts number	-	-
man recall time	firm cancel term	-	+

us with new degrees of truth for each pair of sentences of our dataset, which we give in table 3.

The intuitions for the hard coding of quantifiers are obtained following [17] and are stipulated in the table below:

quantifier	hard coding
all	high $\rightarrow$ 0.8, low $\rightarrow$ 0.2
some	very low $\rightarrow$ 0.2, the rest $\rightarrow$ 0.8
several	low $\rightarrow$ 0.2, high $\rightarrow$ 0.2, interm. $\rightarrow$ 0.8
many	low $\rightarrow$ 0.2, high $\rightarrow$ 0.8, interm. $\rightarrow$ 0.2
few	low $\rightarrow$ 0.8, high $\rightarrow$ 0.2, interm. $\rightarrow$ 0.2

In this table, the quantifier 'all' expectedly sends high numbers to high numbers and low numbers to low numbers, whereas e.g. 'several' maps low values to low values, very high values to low values, but intermediate values to high values. In order to be able to compare the resulting numbers in a uniform way we apply the convention that 'low' is 0.2, 'high' is 0.8 and intermediate is 0.5.

Given a pair of degrees of truth, we now compute an entailment and a degree for it using the definition of fuzzy entailment [17, 8]. A fuzzy proposition p entails another fuzzy proposition q iff q is less specific than p. For p and q two fuzzy sets, this is defined to be the point wise ordering between their possibilistic distributions. After a proportion is computed on the fuzzy sets, the ordering becomes the ordering between their computed degrees of truth, i.e. on $\Pi_{Proportion(B|A)}$ for the proportion of elements of B in A. A degree of entailment is computed from a pair of degrees of truth, by subtracting them. Whenever this number is positive, we mark it with a + sign and whenever it is negative we mark it with a - sign. If the result is 0, it must have come from a case where the degrees of truth of each entry of the pair is the same. Since the ordering is $\leq$ and not strict, these cases stand for as a full entailment and given a + sign. The degrees of entailment of the entries of our dataset thus becomes as in the Deg. 1 column of table 4, the two high and low bands are combined in one table with a bar separating them.

This means that out of the seven cases of positive entailment, only two (the ones marked with a +) are predicted correctly and out of the five cases of negative entailment, only three (the ones marked with -) re predicted correctly. However, after applying the generalised quantifiers to them, the results improve, as shown in the Deg. 2 column of table 4. Here, the + signs increase from two to six, predicting all but one case correctly. The number of the - signs also increase by 1, so the model is predicting that 4 out of the 5 entries do not entail each other correctly.

6 Conclusion and Future Work

We showed how the compositional semantics of the generalised quantifiers of natural language extends from a binary setting to a many valued setting and how this latter can be used to model fuzzy generalised quantifier. In a quest to relate these developments to large scale data learnt from distributional semantics, we interpreted vectors as fuzzy sets and computed degrees of truth for the sentences of a toy dataset. We then extended the dataset with quantifiers and showed how these computations can be used for an inference task.

One essential piece of work is experimenting with the model on main stream inference datasets such as SNLI and Fracas. We mainly chose to work with this small dataset since the details of its high and low bands and the annotations for them where published and we could use them to provide a detailed case by case analysis. A theoretical direction is to use the logic of fuzzy sets, e.g. that of [12], to develop a logic for distributional

data. Quantifiers are known to impose contextual restrictions on their domains, e.g. in the donkey sentences. In previous work [15, 14] authors have shown how compositional distributional semantics deals with these issues. Finding out the added value of working in a fuzzy setting for such examples remains to be worked out.

Finally, we are aware of the tension that exists between the membership values of fuzzy sets and the probabilities that come from the normalised distributional vector representations. In their simplest forms, the latter probabilities are log likelihood estimates resulting from co-occurrence counts of a corpus of text. Fuzzy values, however, are obtained from distribution of the individuals in the domain of vague predicates in a model. Relating these two should be possible by machine learning and alongside recent work on Bayesian inference semantics [5], unified functional distributional models [7], and distributional model theoretic approach [11, 3, 4].

References

[1] Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 23–32, Avignon, France, 2012. Association for Computational Linguistics.

[2] Jon Barwise and Robin Cooper. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4:159–219, 1981.

[3] I. Beltagy, Stephen Roller, Pengxiang Cheng, Katrin Erk, and Raymond J. Mooney. Representing meaning with a combination of logical and distributional models. *Computational Linguistics*, 42:763–808, 2016.

[4] Islam Beltagy, Cuong Chau, Gemma Boleda, Dan Garrette, Katrin Erk, and Raymond Mooney. Montague meets Markov: Deep semantics with probabilistic logical form. In *Second Joint Conference on Lexical and Computational Semantics*, pages 11–21, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics.

[5] Jean-Philippe Bernardy, Rasmus Blanck, Stergios Chatzikyriakidis, Shalom Lappin, and Aleksandre Maskharashvili. Bayesian inference semantics: A modelling system and a test suite. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 263–272, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[6] B. Coecke, M. Sadrzadeh, and S. Clark. Mathematical Foundations for Distributed Compositional Model of Meaning. Lambek Festschrift. *Linguistic Analysis*, 36:345–384, 2010.

[7] Guy Emerson and Ann Copestake. Functional distributional semantics. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 40–52, Berlin, Germany, 2016. Association for Computational Linguistics.

[8] L. Godo and L. Valverde. Entailment and inference in fuzzy logic using fuzzy preorders. In *1992 Proceedings of the EEE International Conference on Fuzzy Systems*, pages 587–594, San Diego, CA, USA, 1992. IEEE.

[9] Edward Grefenstette and Mehrnoosh Sadrzadeh. Concrete models and empirical evaluations for the categorical compositional distributional model of meaning. *Computational Linguistics*, 41:71–118, 2015.

[10] Jules Hedges and Mehrnoosh Sadrzadeh. A generalised quantifier theory of natural language in categorical compositional distributional semantics with bialgebras, 2016.

[11] Aurélie Herbelot and Eva Maria Vecchi. Building a shared world: mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Lisbon, Portugal, 2015. Association for Computational Linguistics.

[12] Vilém Novák. *Fuzzy Sets in Natural Language Processing*, volume 165 of *The Springer International Series in Engineering and Computer Science*, pages 185–200. Springer, 1992.

[13] M. Sadrzadeh, D. Kartsaklis, and E. Balkır. Sentence entailment in compositional distributional semantics. *Ann Math Artif Intell*, 82:189–218, 2018.

[14] Mehrnoosh Sadrzadeh. Quantifier scope in categorical compositional distributional semantics. In *Proceedings of the Workshop on Semantic Spaces at the Intersection of NLP, Physics and Cognitive Science*, volume 221 of *EPTCS*, pages 49–57, 2016.

[15] G. Wijnholds. A proof-theoretic approach to scope ambiguity in compositional vector space models. *Journal of Language Modelling*, 6:261–286, 2018.

[16] Gijs Wijnholds and Mehrnoosh Sadrzadeh. Evaluating composition models for verb phrase elliptical sentence embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

[17] Lotfi A. Zadeh. A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with Applications*, 9(1):149 – 184, 1983.

Generating Lexical Representations of Frames using Lexical Substitution

Saba Anwar
Universität Hamburg
Germany
anwar@informatik.uni-hamburg.de

Artem Shelmanov
Skolkovo Institute of Science and Technology
Russia
a.shelmanov@skoltech.ru

Alexander Panchenko
Skolkovo Institute of Science and Technology
Russia
a.panchenko@skoltech.ru

Chris Biemann
Universität Hamburg
Germany
biemann@informatik.uni-hamburg.de

Abstract

Semantic frames are formal linguistic structures describing situations/actions/events, e.g. *Commercial transfer of goods*. Each frame provides a set of roles corresponding to the situation participants, e.g. *Buyer* and *Goods*, and lexical units (LUs) – words and phrases that can evoke this particular frame in texts, e.g. *Sell*. The scarcity of annotated resources hinders wider adoption of frame semantics across languages and domains. We investigate a simple yet effective method, lexical substitution with word representation models, to automatically expand a small set of frame-annotated sentences with new words for their respective roles and LUs. We evaluate the expansion quality using FrameNet. Contextualized models demonstrate overall superior performance compared to the non-contextualized ones on roles. However, the latter show comparable performance on the task of LU expansion.

1 Introduction

The goal of lexical substitution (McCarthy and Navigli, 2009) is to replace a given target word in its context with meaning-preserving alternatives. In this paper, we show how lexical substitution can be used for semantic frame expansion. A semantic frame is a linguistic structure used to describe the formal meaning of a situation or event (Fillmore, 1982). Semantic frames have witnessed a wide range of applications; such as question answering (Shen and Lapata, 2007; Berant and Liang, 2014; Khashabi et al., 2018), machine translation (Gao and Vogel, 2011; Zhai et al., 2013), and semantic role labelling (Do et al., 2017; Swayamdipta et al., 2018). The impact, however, is limited by the scarce availability of annotated resources. Some publicly available resources are FrameNet (Baker et al., 1998) and PropBank (Palmer et al., 2005), yet for many languages and domains, specialized resources do not exist. Besides, due to the inherent vagueness of frame definitions, the annotation task is challenging and requires semanticists or very complex crowd-sourcing setups (Fossati et al., 2013).

Seed sentence: I hope Patti$_{\text{Helper}}$ can help$_{\text{Assistance}}$ you$_{\text{Benefited_party}}$ soon$_{\text{Time}}$.
Substitutes for Assistance: assist, aid **Substitutes for Helper:** she, I, he, you, we, someone, they, it, lori, hannah, paul, sarah, melanie, pam, riley **Substitutes for Benefited_party:** me, him, folk, her, everyone, people **Substitutes for Time:** tomorrow, now, shortly, sooner, tonight, today, later

Table 1: An example of the induced lexical representation (roles and LUs) of the Assistance FrameNet frame using lexical substitutes from a single seed sentence.

We suggest a different perspective on the problem: expanding the FrameNet resource automatically by using lexical substitution. Given a small set of seed sentences with their frame annotations, we can expand it by substituting the *targets* (words corresponding to lexical units of the respective frame) and *arguments* (words corresponding to roles of the respective frame) of those sentences and aggregating possible substitutions into an induced frame-semantic resource. Table 1 shows one such induced example. For this purpose, we have experimented with state-of-the-art non-contextualized (static) word representation models including neural word embeddings, i.e. fastText (Bojanowski et al., 2017), GloVe (Pennington et al., 2014), and word2vec (Mikolov et al., 2013); and distributional thesaurus, i.e. JoBimText (Bie-

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 95–103

mann and Riedl, 2013); and compared their results with contextualized word representations of the state-of-the-art BERT model (Devlin et al., 2019), which has set a new benchmark performance on many downstream NLP applications. To complete the comparison, we also include the lexical substitution model of Melamud et al. (2015), which uses dependency-based word and context embeddings and produces context-sensitive lexical substitutes.

To generate substitutes, we decompose the problem into two sub-tasks: **Lexical unit expansion**: Given a sentence and its *target* word, the task is to generate frame preserving substitutes for this word. **Frame role expansion**: Given a sentence and an *argument*, the task is to generate meaning-preserving substitutes for this argument.

Contributions of our work are (i) a *method* for inducing frame-semantic resources based on a few frame-annotated sentences using lexical substitution, and (ii) an *evaluation* of various distributional semantic models and lexical substitution methods on the ground truth from FrameNet.

2 Related Work

Approaches to semantic frame parsing with respect to a pre-defined semantic frame resource, such as FrameNet, have received much attention in the literature (Das et al., 2010; Oepen et al., 2016; Yang and Mitchell, 2017; Peng et al., 2018), with SEMAFOR (Das et al., 2014) being a most widely known system to extract complete frame structure including target identification. Some works focus on identifying partial structures such as frame identification (Hartmann et al., 2017; Hermann et al., 2014), role labelling with frame identification (Swayamdipta et al., 2017; Yang and Mitchell, 2017), and simple role labelling (Kshirsagar et al., 2015; Roth and Lapata, 2015; Swayamdipta et al., 2018), which is considered very similar to standard PropBank (Palmer et al., 2005) style semantic role labelling, albeit more challenging because of the high granularity of frame roles. These supervised models rely on a dataset of frame-annotated sentences such as FrameNet. FrameNet-like resources are available only for very few languages and cover only a few domains. In this paper, we venture into the inverse problem, the case where the number of annotations is insufficient, similar to the idea of Pennacchiotti et al. (2008) who investigated the utility of semantic spaces and WordNet-based methods to automatically induce new LUs and reported their results on FrameNet.

Our method is inspired by the recent work of Amrami and Goldberg (2018). They suggest to predict the substitutes vectors for target words using pre-trained ELMo (Peters et al., 2018) and dynamic symmetric patterns, then induced the word senses using clustering. Arefyev et al. (2019) takes the idea of substitute vectors from (Amrami and Goldberg, 2018) for the SemEval 2019 (Qasemi-Zadeh et al., 2019) frame induction task and replaces ELMo with BERT (Devlin et al., 2019) for improved performance. Zhou et al. (2019) show the utility of BERT for the lexical substitution task. Lexical substitution has been used for a range of NLP tasks such as paraphrasing or text simplification, but here, we are employing it, as far as we are aware, for the first time to perform expansion of frame-semantic resources.

3 Inducing Lexical Representations of Frames via Lexical Substitution

We experimented with two groups of lexical substitution methods. The first one use no context: non-contextualized neural word embedding models, i.e. fastText (Bojanowski et al., 2017), GloVe (Pennington et al., 2014), and word2vec (Mikolov et al., 2013), as well as distributional thesaurus based models in the form of JoBimText (Biemann and Riedl, 2013). The second group of methods does use the context: here, we tried contextualized word embedding model BERT (Devlin et al., 2019) and the lexical substitution model of Melamud et al. (2015).

3.1 Static Word Representations

These word representations models are inherently non-contextualized as they learn one representation of a word regardless of its context.

Neural Word Embeddings Neural word embeddings represent words as vectors of continuous numbers, where words with similar meanings are expected to have similar vectors. Thus, to produce substitutes, we extracted the k nearest neighbors using a cosine similarity measure. We use pre-trained embeddings by authors models: fastText trained on the Common Crawl corpus, GloVe trained on Common Crawl corpus with 840 billion words, word2vec trained on Google News. All these models produce 300-dimension vectors.

Distributional Thesaurus (DT) In this approach, word similarities are computed using complex linguistic features such as dependency relations (Lin, 1998). The representations provided by DTs are sparser, but similarity scores based on them can be better. JoBimText (Biemann and Riedl, 2013) is a framework that offers many DTs computed on a range of different corpora. Context features for each word are ranked using the lexicographer's mutual information (LMI) score and used to compute word similarity by feature overlap. We extract the k nearest neighbors for the target word. We use two JoBimText DTs: (i) DT built on Wikipedia with n-grams as contexts and (ii) DT built on a 59G corpus (Wikipedia, Gigaword, ukWaC, and LCC corpora combined) using dependency relations as context.

3.2 Contextualized Models

Static word representations fail to handle polysemic words. This paves the way for context-aware word representation models, which can generate diverse word-probability distributions for a target word based on its context.

Melamud et al. (2015) This simple model uses syntax-based skip-gram embeddings (Levy and Goldberg, 2014) of a word and its context to produce context-sensitive lexical substitutes, where the context of the word is represented by the dependency relations of the word. We use the original word and context embeddings of Melamud et al. (2015), trained on the ukWaC (Ferraresi et al., 2008) corpus. To find dependency relations, we use Stanford Parser (Chen and Manning, 2014) and collapsed the dependencies that include prepositions. Top k substitutes are produced if both the word and its context are present in the model's vocabulary. Melamud et al. (2015) proposed four measures of contextual similarity which rely on cosine similarity between context and target words, of which we report the two best performing on our task (BalAdd and BalMult).

BERT Although BERT was originally trained to restore masked tokens, it can produce a word distribution even without masking the target word. In this case, it will consider both the context and the semantics of the target word, leading to a more accurate probability distribution. For experiments, we choose one of the largest pre-trained models presented in Devlin et al. (2019), which is bert-large-cased (340M parameters) from the PyTorch implementation by Wolf et al. (2019). We produce a substitute word distribution without masking and selected substitutes with top k probabilities.

4 Experimental Setup

4.1 Datasets

We experimented with FrameNet (Baker et al., 1998) version 1.7. It contains around 170k sentences annotated with $1,014$ frames, $7,878$ types of frame roles, and $10,340$ lexical units. Frame roles and LUs can consist of a single token or multiple tokens. For this work, we have only considered a single-token substitution. The datasets for evaluation were derived automatically from FrameNet. To create a gold standard for LU expansion task, for each sentence containing an annotated LU, we consider other LUs of the corresponding semantic frame as ground truth substitutes. We keep only LUs marked as *verbs* in FrameNet. To make a gold standard for the role expansion task, for each of the sentences that contain an annotation of a given frame role, we consider all the single-word annotations from the rest of the corpus marked with the same role and related to the same frame as ground truth substitutes. The final datasets for experiments contain $79,584$ records for lexical unit expansion and $191,252$ records for role expansion (cf. Tables 4 and 5).

4.2 Evaluation Measures

To evaluate the quality of generated substitutes for a given target word, we use precision at k ($p@k$) top substitutes. To evaluate the quality of the entire list of generated substitutes, we use mean average precision at level k ($MAP@k$):

$$AP^i@k = \frac{1}{min(k, R^i)} \sum_{l=1}^{k} r_l^i \cdot p^i@l,$$

where $MAP@k = \frac{1}{N} \sum_{i=1}^{N} AP^i@k$. Here, N is a total number of examples in the dataset; R^i is a number of possible correct answers for an example i; r_l^i equals 1 if the model output at the level l is correct and 0 if not. We present $p@k$ at levels: 1, 5, 10, as well as $MAP@50$. Sometimes, the post-processing procedure leads to the generation of a list of substitutes shorter than k; we consider the absence of a substitute for a position as a wrong answer of a model.

Lexical Unit Expansion Task				
Algorithm	**p@1**	**p@5**	**p@10**	**MAP@50**
GloVe	0.359	0.243	0.195	0.127
fastText	0.374	**0.273**	**0.222**	0.151
word2vec	0.375	0.263	0.212	0.146
DT wiki	0.301	0.199	0.161	0.102
DT 59g	0.339	0.246	0.202	0.136
BalAdd	**0.380**	0.271	0.220	**0.152**
BalMult	0.379	0.270	0.220	0.151
BERT cased	0.378	0.258	0.203	0.136

Table 2: Evaluation of LU expansion.

Frame Role Expansion Task				
Algorithm	**p@1**	**p@5**	**p@10**	**MAP@50**
GloVe	0.301	0.249	0.200	0.069
fastText	0.182	0.134	0.102	0.028
word2vec	0.319	0.224	0.165	0.051
DT wiki	0.336	0.250	0.211	0.079
DT 59G	0.322	0.247	0.200	0.075
BalAdd	0.381	0.288	0.213	0.073
BalMult	0.379	0.282	0.209	0.073
BERT cased	**0.384**	**0.313**	**0.271**	**0.105**

Table 3: Evaluation of frame role expansion.

4.3 Post-processing

In post-processing, we remove numbers, symbols, special tokens from the generated list. There may also be multiple examples of the same word in different forms, especially word embeddings often produce multiple words with a shared root form. Therefore, we lemmatize the generated substitutes using the Pattern library (Smedt and Daelemans, 2012). The duplicates and the target words are dropped. For the lexical unit expansion task, as we just experiment with verbs, we drop the substitutes that cannot be verbs. We used a dictionary of verbs that aggregates verb lists taken from Pattern, WordNet (Miller, 1995), and FreeLing (Padró and Stanilovsky, 2012).

5 Results

5.1 Lexical Units Expansion Task

The results for the LU expansion task are presented in Table 2. The best performance was achieved by the BalAdd measure of Melamud et al. (2015) with $p@1 = 0.380$ and $MAP@50 = 0.152$. The fastText model achieves a comparable performance and even shows slightly better results for $p@5$ and $p@10$. The DTs considered in our experiments perform worse than word2vec, fastText, and models of Melamud et al. (2015). That is expected since the DTs need much larger datasets for training as compared to embedding-based models. Even though BERT performed comparably to fastText and word2vec, it could not outperform them except for $p@1$. However, a close examination of some examples shows that it does make a difference when the target word is polysemic.

Table 4 in the appendix contains example sentences with highlighted target words and top 5 substitutes generated by all models (along with the ground truth FrameNet annotations). The first example presents an LU that is associated with only one frame in FrameNet. Being unambiguous in meaning, all models produced many matching substitutes. The other two examples present an LU with multiple associated frames, which leads to different senses of the LU. All non-contextualized models could not produce any substitute for the *Abandonment* frame except fastText, and failed completely for the *Causation* frame, whereas BERT has successfully generated a sufficient number of matching substitutes for both examples.

5.2 Frame Role Expansion Task

The evaluation results of the methods for the frame roles expansion task are presented in Table 3. In this experiment, the non-contextualized models were outperformed by BERT with a significant margin with $p@1 = 0.384$ and $MAP@50 = 0.105$. The performance of fastText is worst compared to all models, in contrast to the previous experiment. The DTs perform substantially better than neural word embedding models. The better score is achieved by the DT trained on Wikipedia. The models of Melamud et al. (2015) achieve slightly worse results for $p@1$ and $p@5$ than BERT, but significantly lose in terms of $p@10$ and $MAP@50$.

Table 5 in the appendix enlists several substitutes for semantic roles in a hand-labelled seed sentence. The first example demonstrates several valid matching substitutes, because *Vehicle* is the most common sense of "car". Whereas, the other two examples present an argument with multiple roles. Again, BERT was able to distinguish both senses and produced valid substitutes.

6 Conclusion

We presented a simple practical technique for the generation of lexical representations of semantic frames using lexical substitution with several contextualized and static word representation models

demonstrating that a single frame annotated example can be used to bootstrap a fully-fledged lexical representation of the FrameNet-style linguistic structures. Non-contextualized baseline models proved to be strong baselines, but failed to produce good substitutes for polysemic words (same word but different semantic frame), whereas BERT for such cases produced competitive substitutes. A prominent direction for future work is testing the proposed technology for building frame representations of low-resource languages and domains.

Acknowledgements

We thank the anonymous reviewers for valuable feedback and acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG) under the "JOIN-T 2" project (BI 1544/4-2), the German Academic Exchange Service (DAAD) and the Higher Education Commission (HEC), Pakistan. The work of Artem Shelmanov in writing and experiments with BERT model was supported by the Russian Science Foundation, project #20-11-20166 "Cross-lingual Knowledge Base Construction and Maintenance".

References

Asaf Amrami and Yoav Goldberg. 2018. Word sense induction with neural biLM and symmetric patterns. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4867, Brussels, Belgium. Association for Computational Linguistics.

Nikolay Arefyev, Boris Sheludko, Adis Davletov, Dmitry Kharchev, Alex Nevidomsky, and Alexander Panchenko. 2019. Neural GRANNy at SemEval-2019 task 2: A combined approach for better modeling of semantic relationships in semantic frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 31–38, Minneapolis, MN, USA. Association for Computational Linguistics.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Montréal, QC, Canada. Association for Computational Linguistics.

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Baltimore, MD, USA. Association for Computational Linguistics.

Chris Biemann and Martin Riedl. 2013. Text: now in 2D! A framework for lexical expansion with contextual similarity. *J. Language Modelling*, 1(1):55–95.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40:9–56.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 948–956, Los Angeles, CA, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, USA. Association for Computational Linguistics.

Quynh Ngoc Thi Do, Steven Bethard, and Marie-Francine Moens. 2017. Improving implicit semantic role labeling by predicting semantic frame arguments. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 90–99, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4)*.

Charles J. Fillmore. 1982. Frame Semantics. *In The Linguistic Society of Korea, eds. Linguistics in the Morning Calm. Seoul: Hanshin*, pages 111–137.

Marco Fossati, Claudio Giuliano, and Sara Tonelli. 2013. Outsourcing FrameNet to the crowd. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 742–747, Sofia, Bulgaria. Association for Computational Linguistics.

Qin Gao and Stephan Vogel. 2011. Utilizing target-side semantic role labels to assist hierarchical phrase-based machine translation. In *Proceedings of Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 107–115, Portland, OR, USA. Association for Computational Linguistics.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych. 2017. Out-of-domain FrameNet semantic role labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 471–482, Valencia, Spain. Association for Computational Linguistics.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic frame identification with distributed word representations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1448–1458, Baltimore, MD, USA. Association for Computational Linguistics.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2018. Question answering as global reasoning over semantic abstractions. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 1905–1914, New Orleans, LA, USA. Association for the Advancement of Artificial Intelligence.

Meghana Kshirsagar, Sam Thomson, Nathan Schneider, Jaime Carbonell, Noah A. Smith, and Chris Dyer. 2015. Frame-semantic role labeling with heterogeneous annotations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 218–224, Beijing, China. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, MD, USA. Association for Computational Linguistics.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98/COLING '98, pages 768–774, Montreal, QC, Canada. Association for Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

Oren Melamud, Omer Levy, and Ido Dagan. 2015. A simple word embedding model for lexical substitution. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 1–7, Denver, CO, USA. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., Harrahs and Harveys, NV, USA.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajic, Angelina Ivanova, and Zdenka Uresova. 2016. Towards comparability of linguistic graph banks for semantic parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3991–3995, Portorož, Slovenia. ELDA.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2473–2479, Istanbul, Turkey. European Language Resources Association (ELRA).

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Hao Peng, Sam Thomson, Swabha Swayamdipta, and Noah A. Smith. 2018. Learning joint semantic parsers from disjoint data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1492–1502, New Orleans, LA, USA. Association for Computational Linguistics.

Marco Pennacchiotti, Diego De Cao, Roberto Basili, Danilo Croce, and Michael Roth. 2008. Automatic induction of FrameNet lexical units. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 457–465, Honolulu, Hawaii. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long Papers), NAACL-HLT 2018, pages 2227–2237, New Orleans, LA, USA. Association for Computational Linguistics.

Behrang QasemiZadeh, Miriam R. L. Petruck, Regina Stodden, Laura Kallmeyer, and Marie Candito. 2019. SemEval-2019 task 2: Unsupervised lexical frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 16–30, Minneapolis, MN, USA. Association for Computational Linguistics.

Michael Roth and Mirella Lapata. 2015. Context-aware frame-semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:449–460.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 12–21, Prague, Czech Republic. Association for Computational Linguistics.

Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *Journal of Machine Learning Research*, 13(Jun):2063–2067.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold. *arXiv preprint arXiv:1706.09528*.

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith. 2018. Syntactic scaffolds for semantic structures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3772–3782, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Bishan Yang and Tom Mitchell. 2017. A joint sequential and relational model for frame-semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256, Copenhagen, Denmark. Association for Computational Linguistics.

Feifei Zhai, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2013. Handling ambiguities of bilingual predicate-argument structures for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1136, Sofia, Bulgaria. Association for Computational Linguistics.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.

A Examples of Induced Lexical Semantic Frame Representations

This appendix contains additional examples of lexical substitutions of lexical units (LUs) and roles of the semantic frames resource along with the ground truth from FrameNet. Examples of the LU expansions are presented in Table 4 while roles are presented in Table 5.

Frame: Statement
Seed sentence: The report **stated**$_{\text{state}}$, however , that some problems needed to be solved , principally that of lack of encouragement of cadres and individuals to exercise their democratic right of freedom of expression .
GloVe: explain, note, agree, acknowledge, mention **fastText:** note, explain, indicate, reiterate, opine **word2vec:** comment, note, assert, remark, explain **DT wiki:** say, note, claim, comment, suggest **DT 59g:** note, say, claim, comment, think **BalAdd:** indicate, stipulate, assert, reiterate, say **BalMult:** indicate, stipulate, assert, say, aver **BERT:** say, find, conclude, note, declare **FrameNet gold:** proclaim, mention, claim, detail, profess, tell, caution, allow, propose, comment, preach, reaffirm, avow, challenge, recount, reiterate, pronounce, relate, remark, report, say, speak, state, allege, suggest, conjecture, talk, write, contend, venture, declare, add, hazard, pout, announce, exclaim, smirk, address, confirm, explain, assert, gloat, acknowledge, insist, maintain, note, observe, aver, refute, attest, describe
Frame: Abandonment
Sentence: When their changes are completed , and after they have worked up a sweat , ringers often skip off to the local pub , **leaving**$_{\text{leave}}$ worship for others below .
GloVe: return, back, left, rest, stay **fastText:** left, abandon, return, rejoin, exit **word2vec:** left, return, depart, exit, enter **DT wiki:** visit, enter, join, reach, represent **DT 59g:** visit, enter, occupy, beat, represent **BalAdd:** abandon, quit, allow, depart, prefer **BalMult:** abandon, allow, quit, prefer, cause **BERT:** give, abandon, do, let, left **FrameNet gold:** leave, abandon, forget
Frame: Causation
Seed sentence: Older kids , like Tracy and Kerry , **left**$_{\text{leave}}$ homeless after a recent murder - suicide in Indianapolis claimed Mom and Dad.
GloVe: right, back, left, off, rest **fastText:** left, right, return, lurch, move **word2vec:** return, right, depart, limp, go **DT wiki:** left, right, break, curve, rear **DT 59g:** left, right, break, swell, enlarge **BalAdd:** left, gash, vacate, depart, jolt **BalMult:** left, vacate, gash, jolt, depart **BERT:** left, send, raise, make, help **FrameNet gold:** cause, leave, mean, render, wreak, bring, dictate, sway, force, make, precipitate, send, raise, motivate, induce, put, see

Table 4: LU expansion examples. Green highlighting indicates matches with the gold annotations.

Frame: Vehicle
Seed sentence: I noticed the **car**$_{\text{Vehicle}}$ was bouncing up and down as if someone were jumping on it.
GloVe: vehicle, automobile, truck, auto, drive **fastText:** vehicle, automobile, car–and, car.but, car.it **word2vec:** vehicle, suv, minivan, truck, ford_focu **DT wiki:** vehicle, automobile, truck, sedan, bus **DT 59g:** vehicle, truck, automobile, sedan, jeep **BalAdd:** vehicle, bike, minivan, land-rover, horsebox **BalMult:** vehicle, bike, minivan, land-rover, passat **BERT:** thing, convertible, vehicle, sedan, cruiser **FrameNet gold:** helicopter, airplane, ship, vessel, subway, boat, vehicle, stryker, tank, truck, aircraft, bike, bus, car, train, plane, cab, carriage, automobile, buse, ferry, tram, sedan, taxi, tricycle, submarine, yacht, aeroplane, chopper
Frame: Part_orientational
Seed sentence: Repton was an Anglo-Saxon town, on the south **bank**$_{\text{Part}}$ of the River Trent, and was at one time a chief city of the Kingdom of Mercia.
GloVe: draft, financial, credit, lender, loan **fastText:** bank.the, bank.it, bank.thi, bank.so, bank. **word2vec:** draft, lender, banker, depositor, mortgage_lender **DT wiki:** shore, company, draft, lender, embankment **DT 59g:** lender, company, insurer, draft, brokerage **BalAdd:** aib, citibank, hsbc, bundesbank, riksbank **BalMult:** citibank, aib, hsbc, tsb, bundesbank **BERT:** side, shore, river, west, fork **FrameNet gold:** bottom, rear, north, north-south, northwest, west, side, territory, western, end, south, acquifer, back, left, window, top, heart, face, dynasty, tip, front, coast, southern, northernmost, northern, part, eastern, aegean, base, peak, area, portion, island, edge, sliver, strip, region, east, bank, fork, aisle, wall, shore, feet, leg, paw, quarter, wing, femora, half, halve, reach, slope, sea-board, borderland, ring, step, drawer, lip, realm, claw, border, ridge, foot, summit, door, gate, apse, façade, hemisphere, boundary, section, entrance, province, point, apex, corner, axle, page, pocket, seat, stair, underbelly, crest, layer, floor, button, shelf, flank, frontier, peninsula, hill, underside, coastline, spoiler, tailcone, panel, wheel
Frame: Abounding_with
Seed sentence: For their sledging trick, they love a steep, snow covered **bank**$_{\text{Location}}$ and will lie on the top, facing downhill, then tuck up their front paws so that they slide along upon their chests.
GloVe: draft, financial, credit, lender, loan **fastText:** bank.the, bank.it, bank.thi, bank.so, bank. **word2vec:** draft, lender, banker, depositor, mortgage_lender **DT wiki:** shore, company, draft, lender, embankment **DT 59g:** lender, company, insurer, draft, brokerage **BalAdd:** cahoot, citibank, hsbc, tsb, draft **BalMult:** cahoot, draft, citibank, hsbc, natwest **BERT:** slope, hill, ditch, mountain, river **FrameNet gold:** ringer, it, kitchen, hill, equipment, island, street, nut, place, which, plimsoll, paper, bread, roll, egg, scone, tin, salmon, dish, potatoe, kavo, hillside, fiord, sea, pottery, cuff-link, porcelain, bowl, room, somethe, that, pocket, hand, gorget, finger, office, bookshelve, stall, animal, bird, mushroom, olive, folder, fish, pepper, pension, panel, door, donut, stoneware, tile, window, eye, veal, walnut, i, jeep, collection, frame, mirror, everythe, bedroom, barge, easel, desk, arbour, bank, bar, cinema, appearance, raspberry, ful, glass, mug, tankard, river, goblet, pew, skin, ceil, bookcase, figure, face, plaster, wall, wood, buse, fishing-boat, sign, poplar, curtain, promenade, avenue, pasture, land, another, weapon, bottle, ditch, everywhere, meadow, pasta, depression, church, sandbag, sofa, bubble, car, countryside, closet, hallway, pond, train, road, home, accommodation, dwelling, fireplace, floor, roof, corridor, uniform, bed, oak, bath, dump, nylon, chalet, balcony, machinery, reef, overhead, belt, path, roadway, area, courtyard, terrace, entrance, character, liverpool, toenail, shaft, object, neck, fingerboard, they, unit, table, pot, fingernail, moccasin, tray, goldie, peach, inn, ingushetia, sidewalk, mast, nail, floorboard, rail, plywood, launch, cabin-top, toy, she, anglo-saxon

Table 5: Role expansion examples: Green highlighting indicates matches with the gold annotations.

Informativity in Image Captions vs. Referring Expressions

Elizabeth Coppock, Danielle Dionne, Nathanial Graham, Elias Ganem, Shijie Zhao, Shawn Lin, Wenxing Liu and **Derry Wijaya**
Boston University, MA
ecoppock@bu.edu

1 Introduction

At the intersection between computer vision and natural language processing, there has been recent progress on two natural language generation tasks: *Dense Image Captioning* and *Referring Expression Generation* for objects in complex scenes (Farhadi et al., 2010; Karpathy and Fei-Fei, 2014; Vinyals et al., 2014; Krishna et al., 2017; Mao et al., 2016; Vedantam et al., 2017; Cohn-Gordon et al., 2018, 2019). The former aims to provide a caption for a specified object in a complex scene for the benefit of an interlocutor who may not be able to see it, and may form part of a larger Visual Question Answering (VQA) system (Antol et al., 2015; Goyal et al., 2017; Zhang et al., 2016). The latter aims to produce a referring expression that will serve to identify a given object in a scene that the interlocutor can see. The two tasks are designed for different assumptions about the common ground between the interlocutors, and serve very different purposes, although they both associate a linguistic description with an object in a complex scene. Despite these fundamental differences, the distinction between these two tasks is sometimes overlooked (Mao et al., 2016; Cohn-Gordon et al., 2018, 2019). Here, we undertake a side-by-side comparison between image captioning and reference game human datasets and show that they differ systematically with respect to informativity. We hope that an understanding of the systematic differences among these human datasets will ultimately allow them to be leveraged more effectively in the associated engineering tasks.

2 Background and Predictions

As the purpose of using a referring expression is to distinguish one referent from another, without being overly wordy, a naive expectation would be that referring expressions should contain as much information as is necessary to do that, and no more. In other words, descriptive modifiers are expected to be included only if they are *informative* in the sense of helping to narrow down on the set of potential referents. This kind of behavior is predicted by the Rational Speech Act (RSA) framework (Frank and Goodman, 2012): Speakers optimize their choice of expression through a trade-off between accuracy and cost, and listeners use a Bayesian reasoning process to identify a speaker's referent.

In work on *Referring Expression Generation* (REG; see Krahmer and Van Deemter 2012), RSA has not been viewed entirely without skepticism. Gatt et al. (2013) compare RSA to a Probabilistic Referential Overspecification model (PRO). They conclude that RSA is insufficient because it fails to consider overspecification and preference rankings when generating referring expressions. Baumann et al. (2014) conduct production and interpretation studies that question the assumption that speakers aim to minimize production costs. Their findings suggest that speakers may favor overspecification not only to help the listener, but to avoid the additional cognitive effort.

Amendments to RSA have been proposed in order to account for overinforativity. Degen et al. (2019) do so by adjusting the deterministic semantics that exists in the basic framework to continuous (fuzzy) semantics. Cohn-Gordon et al. (2018) leverage the captions from the Visual Genome corpus (Krishna et al., 2017) in order to define a semantics for an RSA-based referring expression generation system. The incremental nature of their system provides an alternative account of overinformativity, one which explains differences between languages with prenominal and postnominal adjectives (Paula Rubio-Fernandez, 2020).

But overinformativity has its limits: There is

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 104–108

still a basic trade-off between accuracy and cost at work in the realm of referring expressions. This basic premise predicts that referring expressions for objects in scenes with multiple objects of the same type will tend to be longer, as more content is necessary in order to distinguish one referent from another.

Captions are not subject to the same pressures. The purpose of a caption is not to distinguish one object from another, but rather to describe what is in the picture. Hence we predict that the number of objects in a scene with the same type should have a significant impact on the length of a referring expression for an object of that type, but either less or no impact on the length of a caption.

As we will show, this prediction is borne out by the data. We find furthermore that captions generally involve indefinite descriptions while referring expressions use definite descriptions, and referring expressions typically make use of more relational vocabulary (e.g. *left*, *closest*) than captions.

3 Approach

The Visual Genome corpus (Krishna et al., 2017) provides a set of captions for objects in complex scenes, called *region descriptions*. We selected a subset of these images in order to construct a dataset of corresponding referring expressions. Our dataset was constructed based on object types (e.g. horse, phone, vase) such that there exist images with one, two, and three objects of that type (e.g. one horse, two horses, and three horses). For each of the types satisfying this condition, we included two images with a SINGLE instance of the type, two with two instances (DOUBLE), and two with three instances of the type (TRIPLE). A total of 198 images were included, comprising 33 sextuples.

We developed an interactive web-based reference game in which a speaker was matched with a listener, and was told to complete the sentence *Draw a box around* _____, for an object in a complex scene designated with a bounding box (see Figure 1). Participants were randomly assigned the role of speaker or listener and communicated through a modified chat window. The listener was instructed to draw a box around the entity indicated by the speaker, and the box drawn by the listener was shown to the speaker as feedback. We filtered out participants who did not attempt to distinguish one object from another in their responses (e.g. referring to one of three teddy bears as 'toy'), and we normalized the responses, taking into account self-corrections and variations in how speakers interpreted the task (e.g. 'the hose |I mean the horse' was normalized to 'the horse', and 'Draw a box around the center horse' was normalized to 'the center horse').

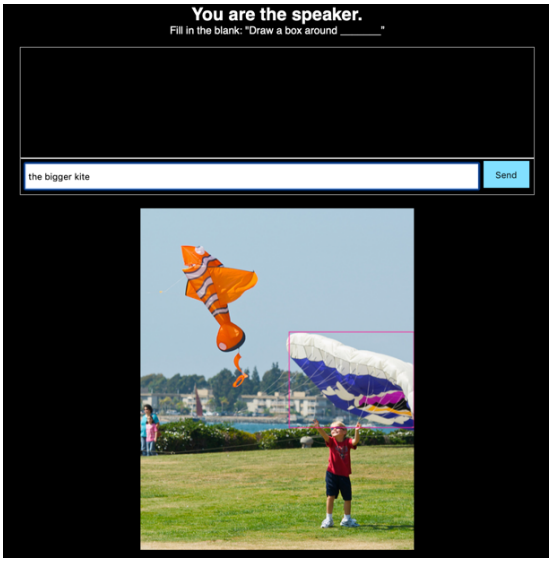


Figure 1: Speaker's point of view in reference game.

Our predictions about length are conditional on whether the referring expressions use a synonym or the same word as the target type; a hyponym would be an alternative strategy to include more specific information. We therefore analyzed the sense relation between the head noun of the description and the target type noun. We used a dependency parser to identify the head noun of the description, and categorized the head noun as a HYPONYM, SYNONYM (or SAME word), or HYPERNYM of the noun corresponding to the target type using WordNet[1].

We then carried out an analysis of the external syntax of these *semi-normalized responses*. Some participants used full definite descriptions, as in *the horse in the middle*, while others left off the initial definite article *horse in the middle*, and others used an even more telegraphic style: *horse in middle*. The variation in style is of interest in its own right, but also makes the descriptions difficult to compare in terms of length. To resolve this, we normalized the responses to make them full noun phrases. We compared the length of the resulting *fully normalized responses*, comparing them to the captions in Visual Genome for the corresponding

[1]https://wordnet.princeton.edu

regions.

4 Results and Discussion

Sample results are shown in Figure 3. In the image with a **single (salient) plane**, over-informative adjectives (*red and white*) are provided to describe the unique salient plane in the image (there is in fact another one in the background), while the referring expression provides just enough information to identify the salient plane (*the plane*). In the image with **three polar bears**, the caption is shorter than the referring expression; the caption simply describes the entity as a polar bear, while the referring expression provides enough information to distinguish the entity from other ones in the scene (the negation of a relational property, *getting licked*). In the image with **two horses**, the caption and the referring expression are of comparable length, but the caption provides non-distinguishing information; the referring expression uses the relational expression *darker* to uniquely identify a referent. In the image with **three planes**, again the caption and the referring expression are of comparable length, and the caption contains enough information to distinguish the referent from the other potential referents in the scene. However, the referring expression uses the relational term *middle*, while the caption describes a non-relational attribute of the object. And of course, the referring expressions use definite articles, while the captions tend to use indefinite articles. These images are representative of the overall set of patterns.

Let us turn now to a quantitative analysis. We note first that the overwhelming majority of the referring expressions we gathered (**94.5%**) were noun phrases headed by the same noun as the target type or a synonym; only 5.5% were a hyponym or a hypernym. We therefore predict for our dataset overall that in images with multiple instances of a given type, referring expressions picking out one of those instances should be longer, in comparison to images with only a single instance of the type.

Of the unique **referring expressions** we gathered, **63% used a definite description**. **Less than 1% used an indefinite description.** The remaining group was predominantly made up of descriptions lacking an initial article, e.g. *horse on (the) left*, with only a handful of exceptions. In contrast, in the corresponding **region descriptions (captions)**, **4.7% used a definite description**, and **39.6% used an indefinite description**. The remaining set were predominantly noun phrases with no initial article (e.g. *large brown bear by a rocky wall*; notice here that the embedded noun phrase is indefinite, however). Perhaps surprisingly, 11.9% of the region descriptions took the form of a sentence, e.g. *Pizza is thin crust* or *The zebra has stripes*. It seems the region description data reflects a range of approaches to the annotation task; this is a source of noise in the data.

We now compare the captions to the referring expressions with respect to length. The results are summarized in Table 1, which shows the mean length of utterances for both region descriptions (captions) and referring expressions, by number of objects of the same type within the image. These results are also visualized in Figure 2, which shows the distribution of lengths (note that the points are jittered, so as to avoid overplotting).

These results support the hypothesis that referring expressions and captions are subject to very different pressures with respect to informativity. Referring expressions include descriptive information for the purpose of distinguishing one referent from another, while captions do not.

Finally, the kind of information that can help to discriminate among referents often consists in relations that instances of the type stand in to each other (e.g. *darker brown*, *in the middle*, *closest*, *on the right*). We defined a *relational modifier* narrowly as a modifier that specifies a characteristic of an object in relation to another instance of the type named by the head noun, excluding gradable size adjectives like *big*. Even on this narrow definition, we find a strong difference between captions and referring expressions, with **captions** exhibiting such modifiers at a rate of **less than one percent**, and **referring expressions** exhibit-

	REF. EXP	CAPTION
SINGLE	2.95 ($sd = 2.2$)	4.45 ($sd = 1.8$)
DOUBLE	4.84 ($sd = 2.9$)	4.46 ($sd = 2.9$)
TRIPLE	5.35 ($sd = 2.7$)	3.45 ($sd = 2.9$)
t	2.1	-0.66
$P(>\|t\|)$	0.039 (*)	0.511 (n.s.)

Table 1: Mean length in words for captions vs. referring expressions, along with t statistics and P-values for OLS-based linear regression models estimating the effect of target type count on length.

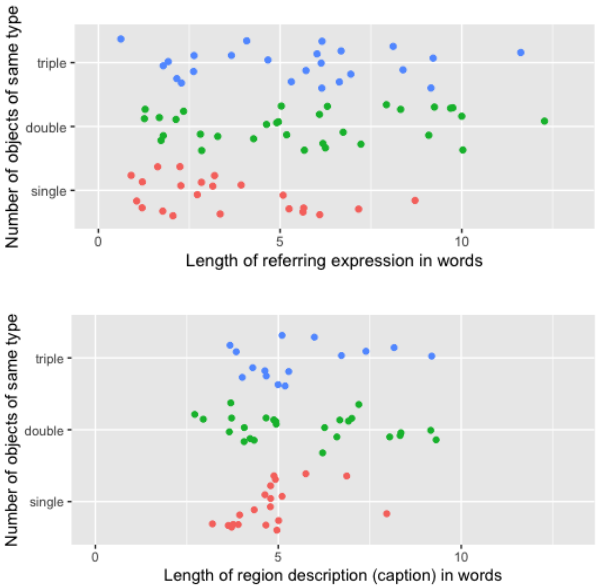


Figure 2: Effect of number of instances of target type on length for referring expressions (top) and captions (bottom) (points jittered).

ing them at a rate of **26.3%**.

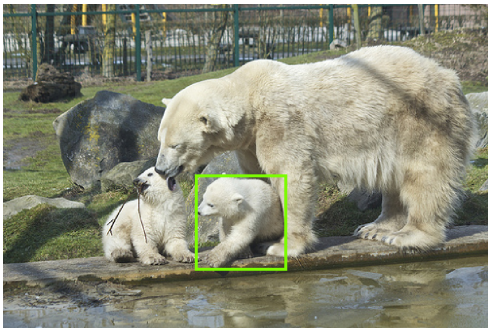
5 Conclusion

This comparison has shown that referring expressions and captions are subject to very different pressures with respect to informativity. When there is only a single instance of a given type (or only one instance that is visually salient), then it suffices to refer to it using 'the [noun]', where '[noun]' identifies the type. A caption, on the other hand, is there to tell someone about the object, so descriptive detail is more likely to be added even when it does not help to identify the referent.

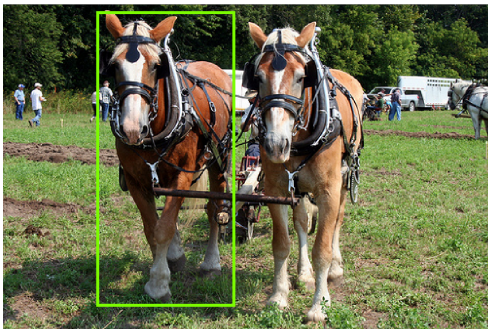
But captions are not systematically longer than referring expressions, either. Descriptive modifiers will be added to a referring expression when they serve the purpose of distinguishing the referent from other ones, i.e., when they are informative. This is why expressions referring to objects of a type that is multiply instantiated within a scene tend to be longer. A caption and the corresponding referring expression may also be equally long, but the kind of information they contain is different: a caption is more likely to contain information that does not help to discriminate among the possible referents. Relational vocabulary is for distinguishing among referents.

We hope that these findings will enable image captioning datasets to be leveraged more effectively in systems for generating expressions that

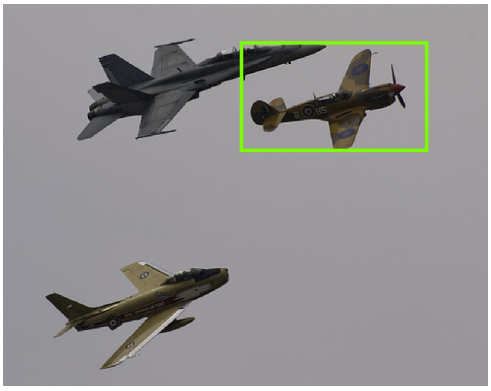
Caption: 'red and white plane'
Ref. Exp.: 'the plane'



Caption: 'a polar bear cub'
Ref. Exp.: 'the bear that's not getting licked'



Caption: 'a brown and white horse'
Ref. Exp.: 'the darker brown horse'



Caption: 'plane with a propeller on the front'
Ref. Exp.: 'the airplane in the middle'

Figure 3: Captions versus referring expressions for selected images.

refer to objects in complex scenes.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Peter Baumann, Brady Clark, and Stefan Kaufmann. 2014. Overspecification and the cost of pragmatic reasoning about referring expressions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 1898–1903.

Reuben Cohn-Gordon, Noah Goodman, and Chris Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, NAACL-HLT*, volume 2, pages 439–443.

Reuben Cohn-Gordon, Noah D. Goodman, and Christopher Potts. 2019. An incremental iterated response model of pragmatics. In *Proceedings of the Society for Computation in Linguistics*, volume 2, pages 81–90.

Judith Degen, Robert X. D. Hawkins, Caroline Graf, Elisa Kreiss, and Noah D. Goodman. 2019. When redundancy is rational: A bayesian approach to 'overinformative' referring expressions. *CoRR*.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision – ECCV 2010*, pages 15–29, Berlin, Heidelberg. Springer Berlin Heidelberg.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998.

Albert Gatt, Roger P. G. van Gompel, Kees van Deemter, and Emiel Krahmer. 2013. Are we bayesian referring expression generators? In *Proceedings of the Cogsci workshop on Production of Referring expressions. Associated with the 35th Annual Conference of the Cognitive Science Society*, Berlin.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Andrej Karpathy and Li Fei-Fei. 2014. Deep visual-semantic alignments for generating image descriptions.

Emiel Krahmer and Kees Van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehensions of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20.

Julian Jara-Ettinger Paula Rubio-Fernandez, Francis Mollica. 2020. Why searching for a blue triangle is different in english than in spanish.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

How does Punctuation Affect Neural Models in Natural Language Inference

Adam Ek **Jean-Philippe Bernardy** **Stergios Chatzikyriakidis**
Centre for Linguistic Theory and Studies in Probability
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{adam.ek,jean-philippe.bernardy,stergios.chatzikyriakidis}@gu.se

Abstract

Natural Language Inference models have reached almost human-level performance but their generalisation capabilities have not been yet fully characterized. In particular, sensitivity to small changes in the data is a current area of investigation. In this paper, we focus on the effect of punctuation on such models. Our findings can be broadly summarized as follows: (1) irrelevant changes in punctuation are correctly ignored by the recent transformer models (BERT) while older RNN-based models were sensitive to them. (2) All models, both transformers and RNN-based models, are incapable of taking into account small relevant changes in the punctuation.

1 Introduction

In recent years models for Natural Language Inference (NLI) have reached almost human-level performance. These models frame inference as a classification problem, whose input is a premise/hypothesis pair. It has been noted that small changes in the pair, can flip the prediction (Glockner et al., 2018). In this paper, we explore the effect of punctuation[1] in neural models in natural language inference.

Small changes in a premise/hypothesis pair are of two kinds. First, the change can be of an irrelevant kind. For example, we can expect that removing a sentence-final stop should not change the relationship between a premise and hypothesis sentence. Second, a textually small change could flip the relationship between hypothesis and premise. For example, adding a negation word is a small textual change that has a lot of semantic content. But it is not only words that can have a large impact on the meaning of a sentence. Commas, for example, may indicate which words belong together and which do not in a list. Ideally, an NLI model should be insensitive to changes of the first kind, but still, properly recognize changes of the second kind.

In this paper, we test both hypotheses for the case of punctuation. Namely:

- (H1) Deep-learning based classifiers are sensitive to irrelevant punctuation.
- (H2) Deep-learning classifiers take relevant punctuation into account correctly.

This work is part of the larger question concerning the ability of NLI models to generalize. There are a number of papers that report several problems of generalizability: Glockner et al. (2018) have shown that several NLI models break considerably easily when, instead of tested on the original SNLI (Bowman et al., 2015) test set, they are tested on a test set which is constructed by taking premises from the training set and creating several hypotheses from them by changing at most one word within the premise. Talman and Chatzikyriakidis (2018) show that NLI models break down when one trains in one dataset, but then test on the test set of a similar dataset (e.g. training on MNLI (Williams et al., 2017) and testing on SNLI). Wang et al. (2019) report problems in generalizability when the two pairs are swapped. The idea is that one should expect the same accuracy for contradiction and neutral when the pairs are swapped (neutral remains neutral, and contradiction remains a contradiction[2]), and a lower accuracy for entailment (given that entailment turns neutral when the pairs are swapped).

[1] The set of punctuation symbols we consider are: ’!"#$%&()*+,-./:;<=>?@[]\^_‘{}|’

[2] Even though one can imagine exotic, non-symmetric definitions of “neutral” and “contradiction”, we are not aware of any system or dataset using such a definition.

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 109–116

2 Datasets and experiments

Our experiments are performed on the Multi-Genre Natural Language Inference (MNLI) corpus (Williams et al., 2017) (and variants thereof, as described below). MNLI consists of 433k human-written sentence pairs labeled with entailment, contradiction and neutral. MNLI contains sentence pairs from ten distinct genres[3] of both written and spoken English. Only five genres are included in the training set. The development and test sets have been divided into matched and mismatched, where the former includes only sentences from the same genres as the training data and the latter include sentences from the remaining genres not present in the training data.

We consider three variants of MNLI:

(*orig*) This variant is the original MNLI with no changes whatsoever.

(p) To obtain this variant we make punctuation consistent throughout examples by adding full stops at the end of each sentence.

($\neg p$) To obtain this variant we remove all non-alphanumeric characters from each sentence. This also remove special characters that are sometimes not classified as punctuation, such as the dollar sign. However, such characters occur so seldom that they have little influence on the results, either way (see Table 1).

Appending a sentence-final stop is in general reasonable, especially for the non-dialogue examples. For the dialogue part of the MNLI dataset, this is unnatural as final stops typically are not expressed in dialogue.

To convey an idea of the amount of data that our transformation impact, we show the raw and relative count[4] of punctuation symbols in Table 1. In total, relative to word-tokens, punctuation symbols account for about 11.5% of the tokens.

SYMBOL	COUNT	%
,	672354	3.544
.	632460	3.334
’	426014	2.246
-	188124	0.992
)	66498	0.351
(	66210	0.349
?	41530	0.219
”	27246	0.144
;	18182	0.096
!	11384	0.060
$	8724	0.046
:	6162	0.033
/	5746	0.030
[	1920	0.010
]	1872	0.010
&	1032	0.005
%	1014	0.005
_	666	0.003
*	186	0.001
@	162	0.001
=	150	0.001
#	114	0.001
+	66	0.0003
‘	24	0.0001
~	12	6.32e-05
\	12	6.32e-05
{	12	6.32e-05

Table 1: Count of punctuation symbols used in the training examples of MNLI.

2.1 Experiments

We perform two sets of experiments:

In the first set, designed to test (H1), we train NLI models for either of the three ($orig$, p, $\neg p$) variants and test on either the p or $\neg p$ variants. Additionally, we train on $orig$ and test on $orig$, as a baseline result.

In the second set, we designed a dataset to test (H2), that is, whether NLI models are able to detect semantically relevant punctuation. This experiment is performed the same way as the first set, but we replace the MNLI test data with our own dataset. The dataset we constructed for this contain a number of problems whose correct label depends on the presence or absence of punctuation. Here are some representative examples (& separates the premise from the hypothesis, label follows in parentheses):

(1) I thank, my mother, Anna, Smith and John & I thank four people (E)

(2) I thank, my mother Anna, Smith and John & I thank two people (C)

[3] face to face conversations, telephone ones, letters, oxford university press publications, etc.

[4] Relative to the number of total tokens in the MNLI dataset

(3) The notion of good, god, is incomprehensible & Good is incomprehensible (E)

(4) The notion of good, god, is incomprehensible & Good god is incomprehensible (C)

The first two examples are cases where the commas are used to denote the conjunction of more than one conjunct. Removing the comma between "my mother" and "Anna" in 2 has a significant effect on counting: what is taken to be two entities in 1, are one in 2. In 3 and 4, we get a different label depending on whether the hypothesis refers to the property "good" (E) or the adjectival modification "good god" (C). The test set consists of 18 examples which can be seen in Table 4.

3 Models

The experiments are performed using three models:

BiLSTM The simplest model is a bidirectional LSTM that encodes the premise and hypothesis, then applies max pooling. The model then concatenates the premise and hypothesis in the standard fashion (Conneau et al., 2017; Talman et al., 2019): $[p; h; p - h; p * h]$ where p is the premise representation and h the hypothesis representation. A three-layer perceptron with leaky ReLU activation between the layers then assigns a class to the example.

HBMP The second model is described by Talman et al. (2019). The model is a three-layer bidirectional LSTM, wherein between the layers a representation is extracted through max pooling. The final representation for each sentence is the concatenation of all intermediate representations $[h_0; h_1, h_2]$. The same representation as with the BiLSTM, $[p; h; p - h; p * h]$ where p and h respectively is the concatenation of all intermediate representations, is then passed to a three-layer perceptron with leaky ReLU activation and dropout.

BERT Our third model is a transformer model, BERT (Devlin et al., 2018). We use the BERT base model from the transformer library (Wolf et al., 2019). To train BERT we use a three layer perceptron with Leaky ReLU activations on top of the BERT model and fine-tune. The BERT model process the premise and hypothesis in parallel and there is no need to explicitly combine them as with the previous models. For the classification of a sentence pair, we use the CLS token generated by BERT that contain a summary of the sentences.

4 Experimental setup

For each architecture (BERT, HBMP, and BiLSTM) we perform experiments by training four models, two trained and validated on the dataset with punctuation and two models trained and validated on the dataset without punctuation. To asses the effect of our data augmentation we test the model on the other dataset, i.e. a model trained and validated without punctuation is tested on the dataset with punctuation. We measure the performance in terms of accuracy.

For HBMP and the BiLSTM models we use the default hyperparameters reported by Talman et al. (2019) with GloVe (Pennington et al., 2014) word embeddings[5]. The BERT model is fine-tuned with the default model hyperparameters. We use the Adam optimizer with a learning rate of 0.00002 and a batch size of 24.

5 Results

5.1 First experiment set

The results from the first experiment are shown below in Table 2. The experiment shows the accuracy for the models trained on the MNLI variations with and without punctuation and their accuracy on all variations.

The results indicate that when the RNN-based models are tested on the same dataset as it is trained on, the results are similar to that of the original model. However, when we test on the opposite dataset the performance drops drastically (about 30 percentage points). We see that the drop in accuracy is about the same for both the matched and mismatched test set. In contrast to the RNN-based models, the transformer model only shows a slight difference in accuracy when presented with test data different from its training data.

5.2 Second experiment set

Full results from the second experiment can be found in Table 4, a subset of the examples can be found in Table 3. The experiment shows the predictions by the HBMP and BERT models trained with and without punctuation on our hand-crafted dataset.

[5]Trained on 840 billion tokens.

MODEL	TEST	MA	MM
BiLSTM$_{orig}$		.724	.723
BiLSTM$_{p}$	p	.723	.724
BiLSTM$_{p}$	$\neg p$	.428	.414
BiLSTM$_{\neg p}$	$\neg p$	.714	.727
BiLSTM$_{\neg p}$	p	.424	.430
HBMP$_{orig}$		.729	.733
HBMP$_{p}$	p	.728	.729
HBMP$_{p}$	$\neg p$	.430	.408
HBMP$_{\neg p}$	$\neg p$	.729	.732
HBMP$_{\neg p}$	p	.436	.427
BERT$_{orig}$		.833	.839
BERT$_{p}$	p	.835	.837
BERT$_{p}$	$\neg p$	.816	.822
BERT$_{\neg p}$	$\neg p$	.819	.820
BERT$_{\neg p}$	p	.830	.833

Table 2: The effect on punctuation on all three models in terms of accuracy of the MNLI dataset. MA indicate the matched and MM the mismatched test split. $original$ is trained on the unaugmented data, p models trained with punctuation and $\neg p$ models trained without punctuation

5.3 Experiment one analysis

The experiment shows that the BLSTM and HBMP models trained with punctuation drops significantly in accuracy when tested on data without punctuation. This indicates that when removing punctuation the model changes its prediction incorrectly. Most of the removed punctuation does not change the meaning, rather some information irrelevant the the relationship between the two sentences (such as sentence-final stop).

Inspecting the output of the HBMP model we can see that in many cases, removing a sentence-final stop flips the models' prediction. In example (5) and (6), both the model trained on punctuation and the one without fail to predict that the final stop does not add any meaning.

(5) P = not yourself .
H = only you . (C)
HBMP$_{p}$ = C
HBMP$_{\neg p}$ = E
BERT$_{p}$ = N
BERT$_{\neg p}$ = N

(6) P = not yourself
H = only you (C)
HBMP$_{p}$ = E
HBMP$_{\neg p}$ = C
BERT$_{p}$ =N
BERT$_{\neg p}$ = N

In examples (7) and (8)[6], the sentence-final stop has been removed, as well as a comma. In such a case, the comma does not add any meaning but acts as a separator of clauses. The removal or addition of this comma flips the prediction of the models. This shows that irrelevant changes both involving commas and sentence-final stops can flip the model's prediction without any semantic motivation.

(7) P = so they set about clearing the land for agriculture , setting fire to massive tracts of forest .
H = as a result , the land was devastated by erosion . (N)
HBMP$_{p}$ = N
HBMP$_{\neg p}$ = C
BERT$_{p}$ = N
BERT$_{\neg p}$ = N

(8) P = so they set about clearing the land for agriculture setting fire to massive tracts of forest
H = as a result the land was devastated by erosion (N)
HBMP$_{p}$ = C
HBMP$_{\neg p}$ = N
BERT$_{p}$ =E
BERT$_{\neg p}$ = C

BERT assigns the neutral class regardless of punctuation in examples (5) to (7), indicating that the choice of punctuation in training and test does not impact its decision. For example (8) there is no punctuation in the premise and hypothesis, but the different BERT models assign two different classes, *entailment* by the model trained on punctuation and *contradiction* by the model trained without punctuation.

A possible explanation for why the accuracy of BERT does not behave similarly to that of the LSTM based models in Table 3 is that the pretraining of BERT allows the model to better ignore variations in the input. However, the HBMP

[6]For clarity, the premise is indicated by P and the hypothesis by H.

n	Premise	Hypothesis	Gold	Pred	Model
0	I thank, my mother, Anna, Smith and John	I thank four people	E	N	HBMP$_{\neg p}$
1	I thank, my mother, Anna Smith and John	I thank three people	E	N	HBMP$_{\neg p}$
8	I hear John says 'come here'	I hear John speaking	E	E	HBMP$_{\neg p}$
9	I hear 'John says come here'	I hear John speaking	C	N	HBMP$_{\neg p}$
14	No, god is good	God is good	E	E	HBMP$_{\neg p}$
15	No god is good	There is no good god	E	E	HBMP$_{\neg p}$
16	No, god is good	There is no good god	C	E	HBMP$_{\neg p}$
17	No god is good	God is good	C	C	HBMP$_{\neg p}$
0	I thank, my mother, Anna, Smith and John	I thank four people	E	E	HBMP$_{p}$
1	I thank, my mother, Anna Smith and John	I thank three people	E	E	HBMP$_{p}$
8	I hear John says 'come here'	I hear John speaking	E	E	HBMP$_{p}$
9	I hear 'John says come here'	I hear John speaking	C	E	HBMP$_{p}$
14	No, god is good	God is good	E	E	HBMP$_{p}$
15	No god is good	There is no good god	E	E	HBMP$_{p}$
16	No, god is good	There is no good god	C	E	HBMP$_{p}$
17	No god is good	God is good	C	C	HBMP$_{p}$
0	I thank, my mother, Anna, Smith and John	I thank four people	E	E	BERT$_{\neg p}$
1	I thank, my mother, Anna Smith and John	I thank three people	E	E	BERT$_{\neg p}$
8	I hear John says 'come here'	I hear John speaking	E	C	BERT$_{\neg p}$
9	I hear 'John says come here'	I hear John speaking	C	E	BERT$_{\neg p}$
14	No, god is good	God is good	E	E	BERT$_{\neg p}$
15	No god is good	There is no good god	E	E	BERT$_{\neg p}$
16	No, god is good	There is no good god	C	E	BERT$_{\neg p}$
17	No god is good	God is good	C	E	BERT$_{\neg p}$
0	I thank, my mother, Anna, Smith and John	I thank four people	E	E	BERT$_{p}$
1	I thank, my mother, Anna Smith and John	I thank three people	E	E	BERT$_{p}$
8	I hear John says 'come here'	I hear John speaking	E	C	BERT$_{p}$
9	I hear 'John says come here'	I hear John speaking	C	E	BERT$_{p}$
14	No, god is good	God is good	E	E	BERT$_{p}$
15	No god is good	There is no good god	E	E	BERT$_{p}$
16	No, god is good	There is no good god	C	E	BERT$_{p}$
17	No god is good	God is good	C	E	BERT$_{p}$

Table 3: Results on a subset of the examples in our constructed dataset. E is entailment, N is neutral and C is contradiction. The model column indicate which HBMP model configuration was used (trained with punctuation p, or without $\neg p$).

model also uses pre-trained information in the form of GLoVE vectors, yet we do not see HBMP handling the discrepancy between the training and the test well. Albeit the pre-training of GLoVE and BERT are different, in the essence they are the same. Both model the meaning of words based on their surroundings in the neural architecture. Thus, the relevant difference between the models relevant to the absence or presence of punctuation is whether the model use self-attention or an LSTM to create representations of sentences. From this, we pose a tentative hypothesis that self-attention more easily learn to ignore irrelevant input tokens for a task than the LSTM. However, to confirm this we need to perform more expensive experiments.

5.4 Experiment two analysis

None of the models perform very well for this dataset. The HBMP$_p$ model has an accuracy of 61.1% while the HBMP$_{\neg p}$ has an accuracy of 48.8%. The BERT$_p$ model has an accuracy of 44.4% while the BERT$_{\neg p}$ has an accuracy of 48.8%.

For example, both models are tricked by comma removal in (2). An interesting case involves cases where the comma is removed from "No, god" turning it into a negative quantifier "no god". The models are tricked when asked to infer "There is no good god" from "No, god is good" (they predict E instead of C). Another example where the models are tricked by comma removal is when listing items. In the example "I thank, my mother, Anna Smith and John" there are three entities being thanked. The comma placement indicates that "Anna Smith" is one person, and not two. Only HBMPp fails to predicts that "I thank three people" is an entailment for this example. The quotation examples are also challenging. Both systems are tricked when they are asked to judge whether "I hear John speaking" follows: a) from "I hear John says 'come here' ", and b) "I hear 'John says come here' ". Both HBMP models correctly predict a) but fail on b). However, they give a

different wrong label, (N) for HBMP$_{\neg p}$ and (E) for HBMP$_p$. For BERT, both the model trained on p and $\neg p$ make the same predictions, further supporting our hypothesis that bert does not take meaningful punctuation into account, even when trained with punctuation.

6 Conclusions

The conclusions of this paper can be summarized as follows:

Only BERT is robust to irrelevant changes in punctuation (H1 is validated for BERT). The other models see a significant drop in performance when for any mismatch of the presence of punctuation between training and testing sets. However, the presence or absence of the full stop at the end of a sentence has little effect.

This statement rests on the observation that punctuation is generally semantically insignificant in MNLI. This fact has not been tested using a model but rather relies on manual inspection of the data.

We have evidence that no model is capable of taking into account cases where punctuation is meaningful. At this stage of our research, this evidence does not rely on a large body of data. This result is not surprising because of the above observation (namely, there is not enough meaningful punctuation in the training set). Yet, we use pre-trained embeddings (BERT) which have been trained on very large dataset, and it could not be ruled out *a priori* that such embeddings did not contain information related to the meaning of punctuation.

As a general remark, it seems to us useful, if not necessary, to extend the present datasets for NLI to include examples where punctuation is actually meaningful. In general, this is part of a discussion of extending current datasets to include cases of inference where more fined-grained phenomena are taken into consideration Chatzikyriakidis et al. (2017); Bernardy and Chatzikyriakidis (2019, 2020). This also connects with the generalization capabilities of NLI models that were briefly brought up in the introduction. However, the goal should not only be to create many diverse datasets that can get very fine-grained for numerous syntactic phenomena. What we further need are models that will have the ability to generalize well to new data after they have been trained on datasets that represent a much more diverse and rich picture of NLI, and are not prone to similar problems as these have been reported in the literature (Glockner et al., 2018; Talman and Chatzikyriakidis, 2018; Wang et al., 2019; Poliak et al., 2018).

7 Future work

In future work, we plan to continue pursuing the question of model generalizability by investigating how neural models for natural language inference can be adapted to take into account fine-grained semantic phenomena. More specifically, how can models be adapted to learn what constitutes a meaningful part of a sentence, in terms of semantics, and what is not meaningful. We can notice that the phenomena of punctuation is primarily ”syntactic sugar”, by constructing a sentence in a certain way syntactically (by inserting or removing punctuation). To exploit this we plan to incorporate syntactic representations of sentences.

Acknowledgments

The research reported in this paper was supported by grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability (CLASP) in the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg.

References

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. 2019. What kind of natural language inference are nlp systems learning: Is this enough? In *13th International Conference on Agents and Artificial Intelligence*.

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. 2020. Improving the precision of natural textual entailment problem datasets. In *Proceeding of LREC 2020*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Stergios Chatzikyriakidis, Robin Cooper, Simon Dobnik, and Staffan Larsson. 2017. An overview of natural language inference data collection: The way forward? In *Proceedings of the Computing Natural Language Inference Workshop*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from

natural language inference data. *arXiv preprint arXiv:1705.02364*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. *arXiv preprint arXiv:1805.02266*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

Aarne Talman and Stergios Chatzikyriakidis. 2018. Testing the generalization power of neural network models across nli benchmarks. *arXiv preprint arXiv:1810.09774*.

Aarne Talman, Anssi Yli-Jyrä, and Jörg Tiedemann. 2019. Sentence embeddings in nli with iterative refinement encoders. *Natural Language Engineering*, 25(4):467–482.

Haohan Wang, Da Sun, and Eric P Xing. 2019. What if we simply swap the two text fragments? a straightforward yet effective way to test the robustness of methods to confounding signals in nature language inference tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7136–7143.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Appendix: Dataset

We present the full dataset we developed below in Table 4 along with the HBMP and MNLI models prediction on the dataset.

n	Premise	Hypothesis	Gold	Pred	Model
0	I thank, my mother, Anna, Smith and John	I thank four people	E	N	HBMP$\neg p$
1	I thank, my mother, Anna Smith and John	I thank three people	E	N	HBMP$\neg p$
2	I thank, my mother Anna, Smith and John	I thank two people	C	E	HBMP$\neg p$
3	I thank, my mother Anna Smith and John	I thank three people	C	E	HBMP$\neg p$
4	I thank, my mother Anna, Smith and John	I thank more than two people	E	N	HBMP$\neg p$
5	I thank my mother Anna, Smith and John	My mother is called Anna Smith	N	N	HBMP$\neg p$
6	I thank my mother, Anna Smith and John	My mother is called Anna Smith	N	N	HBMP$\neg p$
7	I thank my mother Anna Smith and John	My mother is called Anna Smith	E	E	HBMP$\neg p$
8	I hear John says 'come here'	I hear John speaking	E	E	HBMP$\neg p$
9	I hear 'John says come here'	I hear John speaking	C	N	HBMP$\neg p$
10	The notion of good, god, is incomprehensible	Good is incomprehensible	E	N	HBMP$\neg p$
11	The notion of good god is incomprehensible	Good god is incomprehensible	E	E	HBMP$\neg p$
12	The notion of good, god, is incomprehensible	Good god is incomprehensible	C	E	HBMP$\neg p$
13	The notion of good god is incomprehensible	Good is incomprehensible	N	C	HBMP$\neg p$
14	No, god is good	God is good	E	E	HBMP$\neg p$
15	No god is good	There is no good god	E	E	HBMP$\neg p$
16	No, god is good	There is no good god	C	E	HBMP$\neg p$
17	No god is good	God is good	C	C	HBMP$\neg p$
0	I thank, my mother, Anna, Smith and John	I thank four people	E	E	HBMPp
1	I thank, my mother, Anna Smith and John	I thank three people	E	E	HBMPp
2	I thank, my mother Anna, Smith and John	I thank two people	C	E	HBMPp
3	I thank, my mother Anna Smith and John	I thank three people	C	E	HBMPp
4	I thank, my mother Anna, Smith and John	I thank more than two people	E	N	HBMPp
5	I thank my mother Anna, Smith and John	My mother is called Anna Smith	N	N	HBMPp
6	I thank my mother, Anna Smith and John	My mother is called Anna Smith	N	N	HBMPp
7	I thank my mother Anna Smith and John	My mother is called Anna Smith	E	E	HBMPp
8	I hear John says 'come here'	I hear John speaking	E	E	HBMPp
9	I hear 'John says come here'	I hear John speaking	C	E	HBMPp
10	The notion of good, god, is incomprehensible	Good is incomprehensible	E	E	HBMPp
11	The notion of good god is incomprehensible	Good god is incomprehensible	E	E	HBMPp
12	The notion of good, god, is incomprehensible	Good god is incomprehensible	C	E	HBMPp
13	The notion of good god is incomprehensible	Good is incomprehensible	N	C	HBMPp
14	No, god is good	God is good	E	E	HBMPp
15	No god is good	There is no good god	E	E	HBMPp
16	No, god is good	There is no good god	C	E	HBMPp
17	No god is good	God is good	C	C	HBMPp
0	I thank, my mother, Anna, Smith and John	I thank four people	E	E	BERT$\neg p$
1	I thank, my mother, Anna Smith and John	I thank three people	E	E	BERT$\neg p$
2	I thank, my mother Anna, Smith and John	I thank two people	C	C	BERT$\neg p$
3	I thank, my mother Anna Smith and John	I thank three people	C	E	BERT$\neg p$
4	I thank, my mother Anna, Smith and John	I thank more than two people	E	C	BERT$\neg p$
5	I thank my mother Anna, Smith and John	My mother is called Anna Smith	N	E	BERT$\neg p$
6	I thank my mother, Anna Smith and John	My mother is called Anna Smith	N	E	BERT$\neg p$
7	I thank my mother Anna Smith and John	My mother is called Anna Smith	E	E	BERT$\neg p$
8	I hear John says 'come here'	I hear John speaking	E	E	BERT$\neg p$
9	I hear 'John says come here'	I hear John speaking	C	E	BERT$\neg p$
10	The notion of good, god, is incomprehensible	Good is incomprehensible	E	E	BERT$\neg p$
11	The notion of good god is incomprehensible	Good god is incomprehensible	E	E	BERT$\neg p$
12	The notion of good, god, is incomprehensible	Good god is incomprehensible	C	E	BERT$\neg p$
13	The notion of good god is incomprehensible	Good is incomprehensible	N	E	BERT$\neg p$
14	No, god is good	God is good	E	E	BERT$\neg p$
15	No god is good	There is no good god	E	E	BERT$\neg p$
16	No, god is good	There is no good god	C	E	BERT$\neg p$
17	No god is good	God is good	C	E	BERT$\neg p$
0	I thank, my mother, Anna, Smith and John	I thank four people	E	E	BERTp
1	I thank, my mother, Anna Smith and John	I thank three people	E	E	BERTp
2	I thank, my mother Anna, Smith and John	I thank two people	C	C	BERTp
3	I thank, my mother Anna Smith and John	I thank three people	C	E	BERTp
4	I thank, my mother Anna, Smith and John	I thank more than two people	E	C	BERTp
5	I thank my mother Anna, Smith and John	My mother is called Anna Smith	N	E	BERTp
6	I thank my mother, Anna Smith and John	My mother is called Anna Smith	N	E	BERTp
7	I thank my mother Anna Smith and John	My mother is called Anna Smith	E	E	BERTp
8	I hear John says 'come here'	I hear John speaking	E	E	BERTp
9	I hear 'John says come here'	I hear John speaking	C	E	BERTp
10	The notion of good, god, is incomprehensible	Good is incomprehensible	E	E	BERTp
11	The notion of good god is incomprehensible	Good god is incomprehensible	E	E	BERTp
12	The notion of good, god, is incomprehensible	Good god is incomprehensible	C	E	BERTp
13	The notion of good god is incomprehensible	Good is incomprehensible	N	E	BERTp
14	No, god is good	God is good	E	E	BERTp
15	No god is good	There is no good god	E	E	BERTp
16	No, god is good	There is no good god	C	E	BERTp
17	No god is good	God is good	C	E	BERTp

Table 4: Constructed dataset. E is entailment, N is neutral and C is contradiction. The Model column indicate which model was used (trained with punctuation p, or without $\neg p$).

Building a Swedish Question-Answering Model

Hannes von Essen
Chalmers University of Technology
hannes.von.essen@gmail.com

Daniel Hesslow
Chalmers University of Technology
daniel.hesslow@gmail.com

Abstract

High quality datasets for question answering exist in a few languages, but far from all. Producing such datasets for new languages requires extensive manual labour. In this work we look at different methods for using existing datasets to train question-answering models in languages lacking such datasets.

We show that machine translation followed by cross-lingual projection is a viable way to create a full question-answering dataset in a new language. We introduce new methods both for bitext alignment, using optimal transport, and for direct cross-lingual projection, utilizing multilingual BERT.

We show that our methods produce good Swedish question-answering models without any manual work.

Finally, we apply our proposed methods on Spanish and evaluate it on the XQuAD and MLQA benchmarks where we achieve new state-of-the-art values of 80.4 F1 and 62.9 Exact Match (EM) points on the Spanish XQuAD corpus and 70.8 F1 and 53.0 EM on the Spanish MLQA corpus, showing that the technique is readily applicable to other languages.

1 Introduction

The application of supervised machine learning approaches on reading comprehension tasks such as question answering has traditionally been unsuccessful due to a lack of large-scale datasets for training and the lack of powerful enough language models (Hermann et al., 2015). In recent years however, important steps have been made in both areas with the introduction of large-scale datasets (SQuAD (Rajpurkar et al., 2016), SELQA (Jurczyk, Zhai, and Choi, 2016)) and the paradigm-shifting language model BERT.

While this has enabled impressive results for English question answering, there is still a lack of such large-scale datasets in other languages such as Swedish. We therefore explore whether it is possible to train a model for Swedish question-answering using the existing English dataset.

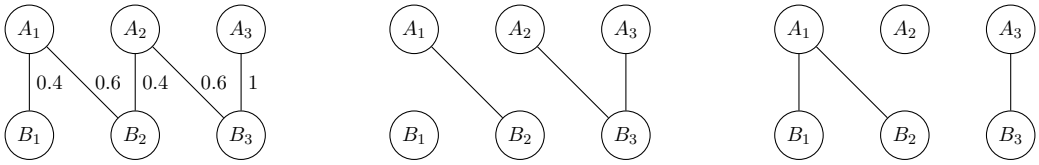
Our two main approaches are to (1) fine-tune a multilingual BERT model on the English SQuAD (Stanford Question Answering Dataset) and see how well it generalizes to Swedish, i.e. doing zero-shot learning, and to (2) machine-translate the English dataset into Swedish and fine-tune a Swedish BERT model on it. We also evaluate various combinations of the two. As SQuAD is based on retrieving the answer from a text, the main challenge with the translation to Swedish consists in determining where the beginning and end positions of the answers are in the translated text, i.e. *projecting* the answer span onto the translated sentence. This is difficult as the translation may change the order of the words (e.g. ”The 1973 oil crisis” $\rightarrow$ ”Oljekrisen 1973”), and the translation also changes depending on the context around it.

We experiment with two approaches to the projection problem: one optimal transport-based solution that creates a word-alignment mapping between the English and the Swedish sentence, and one deep learning-based solution that uses multilingual BERT to find the position of the answer given translations of the English answer and a few words surrounding it, with varying amounts of surrounding words included; something we will refer to as a *translation pyramid*.

We evaluate our model for Swedish QA on our machine-translated versions of the SQuAD dev set, but also apply the method on Spanish to evaluate it on two human-made benchmarks, where we establish a new state-of-the-art for Spanish QA systems. We make both the Swedish and Spanish datasets produced by our method freely available[1].

[1] https://github.com/Vottivott/building-a-swedish-qa-model

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 117–127



(a) The translation probabilities between the source and the target text.

(b) The alignment produced by maximizing the probability from A to B.

(c) The alignment produced by maximizing the probability from B to A

Figure 1: A simple example illustrating the difficulty of word-alignment, a reasonable solution is $A_i \overset{\text{Aligned with}}{\longleftrightarrow} B_i$

2 Related work

(Lee et al., 2018) trained a Korean question-answering model using a machine-translated SQuAD in conjunction with a small manually annotated "seed" dataset of Korean question-answer pairs, which is used to predict the translation certainty to remove poor translations from the dataset. Their method for determining the answer span in the translated text is based on modifying the text to be translated by adding citation marks around the answer span. We found this to be unreliable (sometimes citation marks would be removed or shifted in the translation) and to compromise the translation quality. We therefore translate the original text as is, and instead propose a method for reliably finding the translated answer span afterwards.

(Carrino, Costa-jussà, and Fonollosa, 2019) used an automatic method to translate SQuAD into Spanish. They use a custom-trained neural machine translation model to translate the dataset and then create word alignments using the methods described in (Östling and Tiedemann, 2016) to find the answer span in the translated text. This method is similar to our optimal transport method but uses a different way to find the word alignments.

3 Problem definition

The extractive question answering task we want to solve is defined as follows. Given a context paragraph c (for the SQuAD dataset, these are paragraphs from Wikipedia articles) and a question q, we want to extract the answer from the context paragraph c. This consists in predicting the start and the end position of the answer in c, so that our predicted answer becomes $c_{start:end}$.

The SQuAD dataset essentially consists of tuples of $(c, q, start, end)$ and the translation of the dataset requires not only retrieving the text translations c' and q', but also finding the corresponding ground-truth $start'$ and end' such that the answer is given by $c'_{start':end'}$. This is what we refer to as the *projection* problem.

4 Method

4.1 Projection method I: Optimal Transport Based Word Alignment

We introduce a novel optimal transport-based word alignment method. While optimal transport has been used in NLP to describe the distance between different texts (eg. Word Movers Distance, (Kusner et al., 2015)) we are instead interested in the transport plans which roughly describes which words correspond to which.

Optimal transport is used to account for the fact that given word-wise translation probabilities between the source and the target text, the translation of one word is affected by the translation of other words. See figure 1 for an example. The discrete Wasserstein optimal transport problem, (Villani, 2008), is defined as

$$\begin{aligned} &\underset{\gamma}{\operatorname{argmin}} < \gamma, M >_F \\ s.t. \quad &\gamma \mathbb{1} = p, \quad \gamma^T \mathbb{1} = q, \quad \gamma \geq 0 \end{aligned}$$

where M is the cost matrix, in the case of word alignment it is the negative log of the pairwise translation probabilities. p is the mass for each word in the source and q is the mass for each word in the target, and γ is a matrix describing a transport plan which conserves the mass.

Additionally any good word alignment method must take the position of the source words and target words into account when finding the alignments. Traditionally this has been done by biasing the translation probabilities towards the diagonal, see for example (Dyer, Chahuneau, and Smith, 2013). However to achieve this we instead look at the discrete Gromov-Wasserstein optimal transport problem, (Mémoli, 2011),

$$\underset{\gamma}{\operatorname{argmin}} \sum_{i,i',j,j'} \left\|d_{i,i'} - \bar{d}_{j,j'}\right\|_2^2 * \gamma_{i,j} * \gamma_{i',j'}$$
$$s.t. \quad \gamma \mathbb{1} = p, \quad \gamma^T \mathbb{1} = q, \quad \gamma \geq 0$$

Here d and $\bar{d}$ describe the pairwise distances between the words in the source and the target sentences respectively. Note that the distance functions may not only take into account the index of the word in the sentence but could for example also be designed such that the distance is larger between words that are in different sentences than between words that are in the same sentence. Intuitively this optimization problem tries to find the transport plan which maintains the pairwise distances before and after the translation (Vayer et al., 2018a). Finally, both of these optimization problems can be combined into the so-called fused Gromov-Wasserstein optimal transport problem first introduced in (Vayer et al., 2018b),

$$\underset{\gamma}{\operatorname{argmin}} (1-\alpha) * < \gamma, M >_F +$$
$$\alpha * \sum_{i,i',j,j'} \left\|d_{i,i'} - \bar{d}_{j,j'}\right\|_2^2 * \gamma_{i,j} * \gamma_{i',j'}$$
$$s.t. \quad \gamma \mathbb{1} = p, \quad \gamma^T \mathbb{1} = q, \quad \gamma \geq 0$$

(Vayer et al., 2018b) also propose an optimization method based on conditional gradients to find a local minimum, we use the open source implementation from (Flamary and Courty, 2017). It is worth noting that to be able to use this method successfully one must be able to estimate the pairwise translation probabilities, the mass of each word as well as the distances within the sentences. While finding sufficiently good approximations for the distances within a sentence is easy it is more difficult to find the masses and translation probabilities for all words. We use the cosine similarity of the supervised multilingual word embeddings provided in (Conneau et al., 2017) as M and set the mass of p and q to be uniform.

4.2 Projection method II: Span Projection using Cross-Lingual BERT

Instead of calculating an alignment between all words in the sentences and extracting the translated answer span from it, a different approach is to try to find the projection of the span directly, without any intermediate step. We propose a method that trains on a cross-lingual (mixed English/Swedish) version of the task we want to solve, and then counts on the generalization abilities of multilingual BERT to be able to apply it on the fully Swedish end task.

More specifically, when training, the task is to predict a span of words in an English sentence given Swedish translations of the span and its surrounding words, while in the application of the method the task is performed on a *Swedish* sentence given the same Swedish translations of the span and its surrounding words. This allows us to use the SQuAD dataset itself for training.

The input to the model is designed to both give information about the span and its surroundings, in order to ensure that the correct position is selected when there are multiple occurrences of the answer text. We use Swedish translations of the following: the answer span, the range from two words before to two words after the answer span, and the range from five words before to five words after the answer span, as illustrated in Figure 2. These are sent as input to the projection model, separated by [SEP] tokens, along with the full English (for training) or Swedish (for the end application) sentence. The output heads are identical to the ones used for the SQuAD task itself, allowing us to reuse most of the code from the SQuAD training.

An additional benefit of the multiple translations is increased robustness due to the variation in the translations it creates. With multiple variations of the translation, there is a greater chance that at least one of them will be more similar to the translated paragraph. For example, there is often ambiguity in whether titles of movies and other works should be translated or left in their original language. With the multiple translations, there is a greater chance that one of them will match the language used in the translated paragraph.

One potential benefit of training using SQuAD rather than a more general dataset could be that it priors the network towards predicting answer-like spans, which could help the network make more plausible guesses in unclear situations.

While we used the answer spans from SQuAD for both the training and the application, it would be possible to generate more training data by using random spans from the text instead of just the answer spans, although this would reduce the potential benefit discussed in the last paragraph.

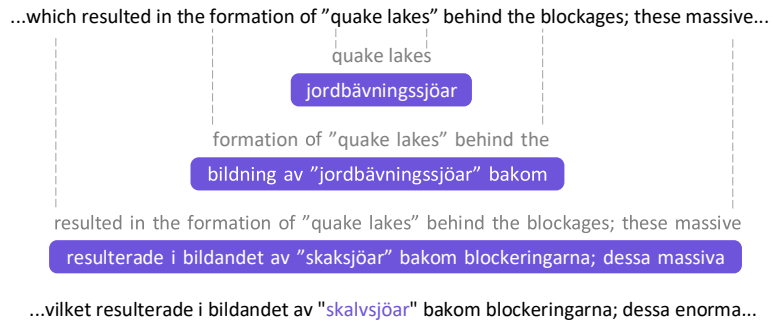


Figure 2: Illustration of the translation pyramid used to give contextual information to the span projection model. All three translations are fed into the network, separated by [SEP] tokens. Note that the answer, *quake lakes*, is translated differently in the expansion levels as either *jordbävningssjöar* or *skaksjöar*, and that both are different from the sentence translation where it is translated as *skalvsjöar*. This illustrates one reason why surrounding words can be important for extra information, and also shows the variation between the different translations, which can make the projection system more robust.

4.3 SQuAD training

We experimented with 4 different language models as the basis for our training: the Base version of Multilingual BERT (cased) (Devlin et al., 2018), the Base and Large version of a Swedish BERT model trained by the Swedish Public Employment Service[2] (uncased) and XLM (Lample and Conneau, 2019) (cased). For the multilingual models we used the pretrained models provided by (Wolf et al., 2019). Their script for SQuAD training also served as the basis for our code.

The multilingual BERT model is trained on Wikipedia in 104 different languages, while the XLM model is trained on Wikipedia in 17 languages. Similarly, the Swedish models are trained on the complete Swedish version of Wikipedia. All models are thus trained on the same amount of Swedish data. It should be noted that the English Wikipedia is larger than the Swedish Wikipedia so we can expect the multilingual models to be slightly better at English than at Swedish.

4.4 Experiments

All experiments were run for 160 000 training steps with a batch size of 3, and an initial learning rate of $5 \cdot 10^{-6}$ that decreases linearly until it reaches zero at step 160 000.

For the deep learning-based projection method, we used an initial learning rate of $3 \cdot 10^{-5}$, a batch size of 8, and we trained with a linearly decreasing learning rate reaching zero after 2 epochs.

For the translation of the paragraphs, questions and answers, as well as the excerpts needed for the translation pyramid, Google Translate's free document translation service[3] was used.

The datasets used are called *en* (the original English SQuAD v1.1 dataset), *exact-sv* (the dataset produced by keeping only the answers in which there is a single exact match of the translated answer in the translated text), *exact-and-ot-sv* (the combination of *exact-sv* with the application of the optimal transport method on the answers removed by *exact-sv*), and *proj-sv* (the application of the deep learning-based projection method on the full dataset). The corresponding versions of the dev set are named similarly as *sv-dev-proj*, *sv-dev-exact* and *en-dev*. In the SQuAD training data there is only one answer per question, while in SQuAD dev each question has a list of multiple acceptable answers, averaging 3.3 per question. Because of the filtering-out of answers with no exact match, most questions in *sv-dev-exact* have fewer answers than in the original dev set, with an average of 1.1 answers per question. Therefore the performance on *sv-dev-exact* is expected to be inherently lower than on *sv-dev-proj*.

5 Results

5.1 Span Projection Methods

While there was no ground-truth dataset to evaluate the performance of the projection models, 200 random examples not used in the training were manually inspected to give statistical estimations

[2]https://github.com/af-ai-center/bert

[3]https://translate.google.com/#view=home&op=docs

of their performance. The results are summarized in Table 1 and 2, where the errors were divided into five categories. For the deep learning-based solution (Projection method II), we found no ***correct answer but wrong position*** errors and only one ***completely incorrect answer*** error, with most errors being ***partially incorrect answer*** errors. This indicates that the method is really good at finding the correct general position of the span in the text. It should also be noted that many of the ***partially incorrect answer*** errors are very small, sometimes missing only a single letter, and can often still be considered good answers to the question. Some examples were also identified as impossible to solve; ***impossible due to translation*** means that the translation has shifted words around such that a correct projection would require the span to be split into multiple parts, making a correct projection using only a single span impossible, and ***impossible due to tokenization*** means that a correct projection would require single tokens to be split into smaller pieces.

Error type	#	Percent
Correct answer but wrong position	0	0%
Partially incorrect answer	10	5%
Completely incorrect answer	1	0.5%
Impossible due to translation	3	1.5%
Impossible due to tokenization	2	1%
Total	**16**	**8%**

Table 1: Evaluation of Projection method II

Error type	#	Percent
Correct answer but wrong position	0	0%
Partially incorrect answer	37	18.5%
Completely incorrect answer	7	3.5%
Impossible due to translation	3	1.5%
Impossible due to tokenization	1	0.5%
Total	**48**	**24%**

Table 2: Evaluation of Projection method I

One issue that causes ***impossible due to tokenization*** is that definiteness is part of the word in Swedish (e.g. "hund**en**") while it is a separate article in English ("**the** dog"). Occasionally the English dataset doesn't include the definite article in the span and the Swedish dataset is not tokenized as to include the indefinite form of the word as a separate token. The model will then sometimes split on the closest boundary causing it to output a result which isn't a word.

The results for the optimal transport method (method I) on the same 200 examples are listed in Table 2, showing that the deep learning-based approach gives considerably more reliable results.

5.2 Evaluation of the Swedish QA models

The results are listed in Table 3. For each metric, we list the highest score achieved across all training checkpoints. We can see that the Multilingual BERT model significantly outperformed the XLM model in our experiments. For Multilingual BERT, the *exact-and-ot-sv* is better than using only *exact-sv*, but *proj-sv* is better than both *exact-and-ot-sv* and *exact-sv*. Additionally, we can see that adding the original English dataset to the training mix gives a small additional improvement in all metrics. Similarly, the performance on *en-dev* for these multilingual mixes is higher than *en*, i.e. the addition of Swedish data to English SQuAD improves the English performance in our experiments. Interestingly the English performance of the models trained only on Swedish data is also high, with *proj-sv* being only 1.8 F1 and 2.2 EM points lower than *en*. Also, the devoted Swedish models perform much worse than multilingual BERT, even when trained on the same Swedish-only data.

5.3 Evaluation on Spanish benchmarks

As there are no benchmarks for Swedish question answering available, we also applied the method on Spanish and evaluated the trained models on the newly introduced XQuAD (Artetxe, Ruder, and Yogatama, 2019) and MLQA (Lewis et al., 2019) benchmarks. XQuAD consists of professional translations of parts of the SQuAD dev set, while MLQA consists of thousands of new QA examples in different languages, crowdsourced from Wikipedia. For the evaluation we used the model checkpoint from the last step of the training. The results, shown in Table 4, show that our method beats the state of the art across all metrics.

6 Conclusions and Future Work

6.1 Conclusions

We have presented a method to automatically translate the question answering datasets with high quality and show that training on such datasets results in good models.

Model	Training data	F1 / EM (sv-dev-proj)	(sv-dev-exact)	(en-dev)
Multilingual BERT Base	en	75.0 / 63.9	69.2 / 57.2	88.8 / 81.5
	exact-sv	76.6 / 64.4	73.6 / 62.8	83.9 / 75.1
	exact-and-ot-sv	80.4 / 69.3	74.5 / 62.3	86.8 / 78.8
	proj-sv	81.4 / 71.5	74.9 / 62.8	87.0 / 79.3
	en + exact-and-ot-sv	80.9 / 70.0	75.1 / 63.0	89.6 / 82.6
	en + proj-sv	**81.9 / 71.6**	**75.6 / 63.3**	**89.8 / 82.8**
XLM	en + proj-sv	74.6 / 64.0	67.8 / 56.0	81.5 / 74.0
Swedish BERT Base	exact-sv	56.8 / 44.0	56.7 / 44.6	
	exact-and-ot-sv	62.6 / 49.3	60.0 / 45.8	
	proj-sv	64.3 / 51.7	60.7 / 47.1	
Swedish BERT Large	exact-sv	56.5 / 43.1	56.7 / 44.6	
	exact-and-ot-sv	62.0 / 48.2	60.0 / 45.8	
	proj-sv	63.9 / 51.4	60.3 / 46.7	

Table 3: Evaluation of the Swedish QA models

Dataset	Model	Training data	F1 / EM
XQUAD	Our models	proj-es	79.8 / 62.1
		en + proj-es	**80.4 / 62.9**
	(Carrino, Costa-jussà, and Fonollosa, 2019)	TAR-train + mBERT (SQuAD-es)	77.6 / 61.8
	XQuAD mBERT baselines	JointMulti 32k voc	59.5 / 41.3
		JointMulti 200k voc	74.3 / 55.3
		JointPair with Joint voc	68.3 / 47.8
		JointPair with Disjoint voc	72.5 / 52.5
MLQA	Our models	proj-es	70.0 / 52.2
		en + proj-es	**70.8 / 53.0**
	(Carrino, Costa-jussà, and Fonollosa, 2019)	TAR-train + mBERT (SQuAD-es)	68.1 / 48.3
	MLQA mBERT baselines	mBERT	64.3 / 46.6
		Translate-train + mBERT	53.9 / 37.4
		XLM (MLM + TLM, 15 languages)	68.0 / 49.8

Table 4: Results for the Spanish evaluation on XQuAD and MLQA

We conclude that limitations in the amount of available Swedish data for pre-training of BERT and the reduced quality of BERT that comes from this can be compensated for to some extent by having larger amounts of data in other languages such as English, which makes multi-lingual models a promising tool for applications in low-resource languages. Even when treated as a Swedish BERT model and ignoring its multi-lingual capacities, the multi-lingual BERT model is probably the best Swedish BERT model currently available, as it outperforms the devoted Swedish model to a remarkable degree (by 17.6 F1 points) in our task when fine-tuned on the same Swedish data. We would therefore recommend Swedish NLP-practitioners to use the multi-lingual BERT model rather than the existing Swedish BERT models until Swedish BERT models trained on larger Swedish datasets become available. We also conclude that when fine-tuning multilingual BERT for an end-task in a certain language (in our case Swedish or Spanish), there can be a benefit in also mixing in training data from other languages than the end-task language .

6.2 Future work

In order to alleviate the problem with poor translations, a future direction could be to try to incorporate a model for quantifying the translation certainty in order to remove the worst translations from the training set, as was proposed by (Lee et al., 2018). Uncertain projections could also be filtered out by looking at the distribution of the output logits from the projection model.

References

Artetxe, M.; Ruder, S.; and Yogatama, D. 2019. On the cross-lingual transferability of monolingual representations. *CoRR* abs/1910.11856.

Carrino, C. P.; Costa-jussà, M. R.; and Fonollosa, J. A. R. 2019. Automatic spanish translation of the squad dataset for multilingual question answering.

Conneau, A.; Lample, G.; Ranzato, M.; Denoyer, L.; and Jégou, H. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dyer, C.; Chahuneau, V.; and Smith, N. A. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 644–648.

Flamary, R., and Courty, N. 2017. Pot python optimal transport library.

Hermann, K. M.; Kociský, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. *CoRR* abs/1506.03340.

Jurczyk, T.; Zhai, M.; and Choi, J. D. 2016. Selqa: A new benchmark for selection-based question answering. *CoRR* abs/1606.08513.

Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From word embeddings to document distances. In *International conference on machine learning*, 957–966.

Lample, G., and Conneau, A. 2019. Cross-lingual language model pretraining. *CoRR* abs/1901.07291.

Lee, K.; Yoon, K.; Park, S.; and Hwang, S.-w. 2018. Semi-supervised training data generation for multilingual question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Lewis, P.; Oğuz, B.; Rinott, R.; Riedel, S.; and Schwenk, H. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Mémoli, F. 2011. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics* 11(4):417–487.

Östling, R., and Tiedemann, J. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics* 106(1):125–146.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR* abs/1606.05250.

Vayer, T.; Chapel, L.; Flamary, R.; Tavenard, R.; and Courty, N. 2018a. Fused gromov-wasserstein distance for structured objects: theoretical foundations and mathematical properties. *arXiv preprint arXiv:1811.02834*.

Vayer, T.; Chapel, L.; Flamary, R.; Tavenard, R.; and Courty, N. 2018b. Optimal transport for structured data with application on graphs. *arXiv preprint arXiv:1805.09114*.

Villani, C. 2008. *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv* abs/1910.03771.

Appendix A: 10 randomly selected examples from the generated datasets

Who was in control of the Dutch East India Company (VOC) and the Dutch West India Company (WIC)?	**The States General of the United Provinces** were in control of the Dutch East India Company (VOC) and the Dutch West India Company (WIC), but some shipping expeditions were initiated by some of the provinces, mostly Holland and/or Zeeland.
Vem hade kontroll över det nederländska East India Company (VOC) och det Dutch West India Company (WIC)?	**USA: s generalsekreterare** hade kontroll över det nederländska östindiska kompaniet (VOC) och det nederländska västindiska kompaniet (WIC), men vissa sjöfartsekspeditioner initierades av några av provinserna, främst Holland och / eller Zeeland.
¿Quién tenía el control de la Compañía Holandesa de las Indias Orientales (VOC) y la Compañía Holandesa de las Indias Occidentales (WIC)?	**Los Estados Generales de las Provincias Unidas** tenían el control de la Compañía Holandesa de las Indias Orientales (VOC) y la Compañía Holandesa de las Indias Occidentales (WIC), pero algunas de las provincias iniciaron algunas expediciones marítimas, principalmente Holanda y / o Zelanda.

How many pubs applied to be allowed to sell alcohol 24 hours a day?	The Licensing Act 2003, which came into force on 24 November 2005, consolidated the many laws into a single Act. This allowed pubs in England and Wales to apply to the local council for the opening hours of their choice. It was argued that this would end the concentration of violence around 11.30 pm, when people had to leave the pub, making policing easier. In practice, alcohol-related hospital admissions rose following the change in the law, with alcohol involved in 207,800 admissions in 2006/7. Critics claimed that these laws would lead to "24-hour drinking". By the time the law came into effect, 60,326 establishments had applied for longer hours and **1,121** had applied for a licence to sell alcohol 24 hours a day. However nine months later many pubs had not changed their hours, although some stayed open longer at the weekend, but rarely beyond 1:00 am.
Hur många pubar ansökte om att få sälja alkohol 24 timmar om dygnet?	Licenslagen 2003, som trädde i kraft den 24 november 2005, konsoliderade de många lagarna till en enda lag. Detta gjorde det möjligt för pubar i England och Wales att ansöka till kommunfullmäktige för de öppettider som de valde. Det hävdades att detta skulle avbryta koncentrationen av våld omkring klockan 11.30, när människor var tvungna att lämna puben, vilket underlättar polisarbetet. I praktiken ökade alkoholrelaterade sjukhusinläggningar efter lagändringen, med alkohol involverad i 207 800 inläggningar under 2006/7. Kritiker hävdade att dessa lagar skulle leda till "24-timmars dricka". När lagen trädde i kraft hade 60 326 anläggningar ansökt om längre timmar och **1 112** hade ansökt om tillstånd att sälja alkohol 24 timmar om dygnet. Emellertid nio månader senare hade många pubar inte ändrat sina timmar, även om vissa stannade öppna längre på helgen, men sällan efter kl.
¿Cuántos pubs solicitaron que se les permita vender alcohol las 24 horas del día?	La Ley de Licencias de 2003, que entró en vigor el 24 de noviembre de 2005, consolidó las numerosas leyes en una sola Ley. Esto permitió que los pubs en Inglaterra y Gales se postularan ante el consejo local para el horario de apertura de su elección. Se argumentó que esto terminaría con la concentración de violencia alrededor de las 11.30 p. M., Cuando la gente tenía que abandonar el pub, lo que facilitaba la vigilancia. En la práctica, los ingresos hospitalarios relacionados con el alcohol aumentaron después del cambio en la ley, con alcohol involucrado en 207,800 ingresos en 2006/7. Los críticos afirmaron que estas leyes conducirían a "beber 24 horas". Cuando entró en vigencia la ley, 60,326 establecimientos habían solicitado más horas y **1,121** habían solicitado una licencia para vender alcohol las 24 horas del día. Sin embargo, nueve meses después, muchos pubs no habían cambiado sus horarios, aunque algunos permanecían abiertos más tiempo el fin de semana, pero rara vez más allá de la 1:00 a.m.

What was the European Union tasked with managing?	Italy became a major industrialized country again, due to its post-war economic miracle. The European Union (EU) involved the division of powers, with taxation, health and education handled by the nation states, while the EU had charge of **market rules, competition, legal standards and environmentalism**. The Soviet economic and political system collapsed, leading to the end of communism in the satellite countries in 1989, and the dissolution of the Soviet Union itself in 1991. As a consequence, Europe's integration deepened, the continent became depolarised, and the European Union expanded to subsequently include many of the formerly communist European countries – Romania and Bulgaria (2007) and Croatia (2013).
Vad fick EU att hantera?	Italien blev igen ett stort industrialiserat land på grund av dess ekonomiska mirakel efter kriget. Europeiska unionen (EU) involverade maktfördelningen, med beskattning, hälsa och utbildning som hanterades av nationalstaterna, medan EU hade ansvaret för **marknadsregler, konkurrens, juridiska standarder och miljöhänsyn**. Det sovjetiska ekonomiska och politiska systemet kollapsade, vilket ledde till slutet av kommunismen i satellitländerna 1989, och Sovjetunionens upplösning 1991. Som en följd av detta fördjupades Europas integration, kontinenten depolariserades och Europeiska unionen utvidgades att därefter inkludera många av de tidigare kommunistiska europeiska länderna - Rumänien och Bulgarien (2007) och Kroatien (2013).
¿Qué se encargó de gestionar la Unión Europea?	Italia se convirtió nuevamente en un importante país industrializado, debido a su milagro económico de posguerra. La Unión Europea (UE) implicó la división de poderes, con impuestos, salud y educación manejados por los estados nacionales, mientras que la UE tenía a su cargo las **reglas del mercado, la competencia, los estándares legales y el ambientalismo**. El sistema económico y político soviético se derrumbó, lo que condujo al fin del comunismo en los países satélites en 1989, y la disolución de la propia Unión Soviética en 1991. Como consecuencia, la integración de Europa se profundizó, el continente se despolarizó y la Unión Europea se expandió. para incluir posteriormente a muchos de los países europeos anteriormente comunistas: Rumania y Bulgaria (2007) y Croacia (2013).

How long was Beyonce depressed?	LeToya Luckett and Roberson became unhappy with Mathew's managing of the band and eventually were replaced by Farrah Franklin and Michelle Williams. Beyoncé experienced depression following the split with Luckett and Roberson after being publicly blamed by the media, critics, and blogs for its cause. Her long-standing boyfriend left her at this time. The depression was so severe it lasted for **a couple of years**, during which she occasionally kept herself in her bedroom for days and refused to eat anything. Beyoncé stated that she struggled to speak about her depression because Destiny's Child had just won their first Grammy Award and she feared no one would take her seriously. Beyoncé would later speak of her mother as the person who helped her fight it. Franklin was dismissed, leaving just Beyoncé, Rowland, and Williams.
Hur länge var Beyonce deprimerad?	LeToya Luckett och Roberson blev missnöjda med Mathews hantering av bandet och ersattes så småningom av Farrah Franklin och Michelle Williams. Beyoncé upplevde depression efter splittringen med Luckett och Roberson efter att ha blivit offentligt klandrad av media, kritiker och bloggar för dess sak. Hennes mångaåriga pojkvän lämnade henne just nu. Depressionen var så allvarlig att den varade i **ett par år**, under vilken hon ibland höll sig i sitt sovrum i flera dagar och vägrade att äta något. Beyoncé uttalade att hon kämpade för att tala om sin depression eftersom Destiny's Child just hade vunnit sitt första Grammy-pris och att hon fruktade att ingen skulle ta henne på allvar. Beyoncé skulle senare tala om sin mamma som personen som hjälpte henne att bekämpa den. Franklin avskedades och lämnade bara Beyoncé, Rowland och Williams.
¿Cuánto tiempo estuvo deprimido Beyonce?	LeToya Luckett y Roberson se sintieron descontentos con el manejo de la banda por parte de Mathew y finalmente fueron reemplazados por Farrah Franklin y Michelle Williams. Beyoncé experimentó depresión luego de la separación con Luckett y Roberson luego de ser acusado públicamente por los medios, críticos y blogs por su causa. Su novio de larga data la dejó en este momento. La depresión fue tan severa que duró **un par de años**, durante los cuales ocasionalmente se mantuvo en su habitación durante días y se negó a comer nada. Beyoncé declaró que le costaba hablar sobre su depresión porque Destiny's Child acababa de ganar su primer premio Grammy y temía que nadie la tomara en serio. Beyoncé luego hablaría de su madre como la persona que la ayudó a combatirlo. Franklin fue despedido, dejando solo a Beyoncé, Rowland y Williams.

What company office did Bell make the original cross country phone call from?	In January 1915, Bell made the first ceremonial transcontinental telephone call. Calling from the **AT&T** head office at 15 Dey Street in New York City, Bell was heard by Thomas Watson at 333 Grant Avenue in San Francisco. The New York Times reported:
Vilket företagskontor gjorde Bell det ursprungliga telefonsamtalet från andra länder?	I januari 1915 ringde Bell det första ceremoniella transkontinentala telefonsamtalet. Call från **AT&T** huvudkontor på 15 Dey Street i New York City hördes av Thomas Watson på 333 Grant Avenue i San Francisco. New York Times rapporterade:
¿Desde qué oficina de la compañía hizo Bell la llamada telefónica original?	En enero de 1915, Bell realizó la primera llamada telefónica ceremonial transcontinental. Llamando desde la oficina central de **AT&T** en 15 Dey Street en la ciudad de Nueva York, Thomas Watson escuchó a Bell en 333 Grant Avenue en San Francisco. El New York Times informó:

What was the name of West's fashion line for women?	On October 1, 2011, Kanye West premiered his women's fashion label, **DW Kanye West** at Paris Fashion Week. He received support from DSquared2 duo Dean and Dan Caten, Olivier Theyskens, Jeremy Scott, Azzedine Alaïa, and the Olsen twins, who were also in attendance during his show. His debut fashion show received mixed-to-negative reviews, ranging from reserved observations by Style.com to excoriating commentary by The Wall Street Journal, The New York Times, the International Herald Tribune, Elleuk.com, The Daily Telegraph, Harper's Bazaar and many others. On March 6, 2012, West premiered a second fashion line at Paris Fashion Week. The line's reception was markedly improved from the previous presentation, with a number of critics heralding West for his "much improved" sophomore effort.
Vad hette Wests modelinje för kvinnor?	Den 1 oktober 2011 hade Kanye West premiär för sin dametikett, **DW Kanye West**, vid Paris Fashion Week. Han fick stöd från DSquared2-duon Dean och Dan Caten, Olivier Theyskens, Jeremy Scott, Azzedine Alaïa och Olsen-tvillingarna, som också var närvarande under hans show. Hans debutmodeshow fick blandade till negativa recensioner, allt från reserverade observationer från Style.com till uttalande kommentarer från The Wall Street Journal, The New York Times, International Herald Tribune, Elleuk.com, The Daily Telegraph, Harper's Bazaar och många andra. Den 6 mars 2012 hade West premiär för en andra modelinje vid Paris Fashion Week. Linjens mottagning förbättrades markant från den föregående presentationen, med ett antal kritiker som vädrade West för hans "mycket förbättrade" andra ansträngning.
¿Cómo se llamaba la línea de moda de West para mujer?	El 1 de octubre de 2011, Kanye West estrenó su marca de moda femenina, **DW Kanye West** en la Semana de la Moda de París. Recibió el apoyo del dúo de DSquared2 Dean y Dan Caten, Olivier Theyskens, Jeremy Scott, Azzedine Alaïa y los gemelos Olsen, que también estuvieron presentes durante su show. Su desfile de modas debut recibió críticas mixtas y negativas, desde observaciones reservadas por Style.com hasta comentarios fascinantes de The Wall Street Journal, The New York Times, International Herald Tribune, Elleuk.com, The Daily Telegraph, Harper's Bazaar y muchos otros. El 6 de marzo de 2012, West estrenó una segunda línea de moda en la Semana de la Moda de París. La recepción de la línea mejoró notablemente de la presentación anterior, con una serie de críticos que anunciaron a West por su "mucho mejor" esfuerzo de segundo año.

Along with cabaret, striptease, bands and drama, what is a type of stage performance that can be found in pubs?	A few pubs have stage performances such as serious drama, **stand-up comedy**, musical bands, cabaret or striptease; however juke boxes, karaoke and other forms of pre-recorded music have otherwise replaced the musical tradition of a piano or guitar and singing.[citation needed]
Tillsammans med kabaret, striptease, band och drama, vad är en typ av scenuppträdande som finns på pubar?	Några pubar har scenuppträdanden som seriöst drama, **stand-up komedi**, musikband, kabaret eller striptease; men jukeboxar, karaoke och andra former av förinspelad musik har annars ersatt den musikaliska traditionen för ett piano eller gitarr och sång.
Junto con el cabaret, el striptease, las bandas y el drama, ¿cuál es un tipo de actuación en el escenario que se puede encontrar en los pubs?	Algunos pubs tienen representaciones teatrales como dramas serios, **comedias**, bandas musicales, cabaret o striptease; sin embargo, los juke boxes, el karaoke y otras formas de música pregrabada han reemplazado la tradición musical de un piano o guitarra y canto. [cita requerida]

The structure of Bern's city centre is mainly what type of buildings?	The structure of Bern's city centre is largely **medieval** and has been recognised by UNESCO as a Cultural World Heritage Site. Perhaps its most famous sight is the Zytglogge (Bernese German for "Time Bell"), an elaborate medieval clock tower with moving puppets. It also has an impressive 15th century Gothic cathedral, the Münster, and a 15th-century town hall. Thanks to 6 kilometres (4 miles) of arcades, the old town boasts one of the longest covered shopping promenades in Europe.
Strukturen i Berns centrum är främst vilken typ av byggnader?	Strukturen i Berns centrum är till stor del **medeltida** och har erkänts av UNESCO som ett kulturellt världsarv. Det kanske mest kända synet är Zytglogge (Bernese tyska för "Time Bell"), ett genomtänkt medeltida klocktorn med rörliga dockor. Den har också en imponerande gotisk domkyrka från 1500-talet, Münster och ett rådhus från 1500-talet. Tack vare 6 kilometer arkader, har gamla stan en av de längsta täckta shoppingpromenaderna i Europa.
¿La estructura del centro de la ciudad de Berna es principalmente qué tipo de edificios?	La estructura del centro de la ciudad de Berna es en gran parte **medieval** y ha sido reconocida por la UNESCO como Patrimonio Cultural de la Humanidad. Quizás su vista más famosa es el Zytglogge (alemán bernés para "Time Bell"), una elaborada torre de reloj medieval con marionetas en movimiento. También tiene una impresionante catedral gótica del siglo XV, el Münster, y un ayuntamiento del siglo XV. Gracias a 6 kilómetros (4 millas) de arcadas, el casco antiguo cuenta con uno de los paseos comerciales cubiertos más largos de Europa.

In what city are the New York Red Bulls based?	In soccer, New York City is represented by New York City FC of Major League Soccer, who play their home games at Yankee Stadium. The New York Red Bulls play their home games at Red Bull Arena in nearby **Harrison, New Jersey**. Historically, the city is known for the New York Cosmos, the highly successful former professional soccer team which was the American home of Pelé, one of the world's most famous soccer players. A new version of the New York Cosmos was formed in 2010, and began play in the second division North American Soccer League in 2013. The Cosmos play their home games at James M. Shuart Stadium on the campus of Hofstra University, just outside the New York City limits in Hempstead, New York.
I vilken stad är New York Red Bulls baserade?	I fotboll representeras New York City av New York City i Major League Soccer, som spelar sina hemmamatcher på Yankee Stadium. New York Red Bulls spelar sina hemmamatcher på Red Bull Arena i närheten av **Harrison, New Jersey**. Historiskt sett är staden känd för New York Cosmos, det mycket framgångsrika tidigare professionella fotbollslaget som var det amerikanska hemmet Pelé, en av världens mest kända fotbollsspelare. En ny version av New York Cosmos bildades 2010 och började spela i den andra divisionen North American Soccer League 2013. Cosmos spelade sina hemmamatcher på James M. Shuart Stadium på campus vid Hofstra University, precis utanför New York City gränser i Hempstead, New York.
¿En qué ciudad se encuentran los Red Bulls de Nueva York?	En el fútbol, la ciudad de Nueva York está representada por el New York City FC de Major League Soccer, que juega sus partidos en casa en el Yankee Stadium. Los Red Bulls de Nueva York juegan sus partidos en casa en el Red Bull Arena en la cercana **Harrison, Nueva Jersey**. Históricamente, la ciudad es conocida por el Cosmos de Nueva York, el exitoso ex equipo de fútbol profesional que fue el hogar estadounidense de Pelé, uno de los jugadores de fútbol más famosos del mundo. Una nueva versión del New York Cosmos se formó en 2010, y comenzó a jugar en la segunda división de la Liga de Fútbol de América del Norte en 2013. El Cosmos juega sus partidos en casa en el estadio James M. Shuart en el campus de la Universidad de Hofstra, a las afueras de New Límites de la ciudad de York en Hempstead, Nueva York.

How many people came to visit New York in 2013?	Tourism is a vital industry for New York City, which has witnessed a growing combined volume of international and domestic tourists – receiving approximately 51 million tourists in 2011, **54 million** in 2013, and a record 56.4 million in 2014. Tourism generated an all-time high US$61.3 billion in overall economic impact for New York City in 2014.
Hur många besökte New York 2013?	Turism är en viktig industri för New York City, som har sett en växande kombinerad mängd internationella och inhemska turister - med cirka 51 miljoner turister under 2011, **54 miljoner** 2013 och rekord 56,4 miljoner 2014. Turismen genererade en tid höga 61,3 miljarder US dollar i total ekonomisk påverkan för New York City 2014.
¿Cuántas personas vinieron a visitar Nueva York en 2013?	El turismo es una industria vital para la ciudad de Nueva York, que ha sido testigo de un creciente volumen combinado de turistas internacionales y nacionales: recibió aproximadamente 51 millones de turistas en 2011, **54 millones** en 2013 y un récord de 56.4 millones en 2014. El turismo generó un récord histórico alto impacto económico general de US $ 61.3 mil millones para la ciudad de Nueva York en 2014.

Appendix B: Examples of span projection errors

Partially incorrect answer (5%)

In what decade was seafloor spreading discovered?	In the **1960s**, a series of discoveries, the most important of which was seafloor spreading, showed that the Earth's lithosphere, which includes the crust and rigid uppermost portion of the upper mantle, is separated into a number of tectonic plates that move across the plastically deforming, solid, upper mantle, which is called the asthenosphere. There is an intimate coupling between the movement of the plates on the surface and the convection of the mantle: oceanic plate motions and mantle convection currents always move in the same direction, because the oceanic lithosphere is the rigid upper thermal boundary layer of the convecting mantle. This coupling between rigid plates moving on the surface of the Earth and the convecting mantle is called plate tectonics.
Under vilket decennium upptäcktes havsbotten spridning?	På **1960**-talet visade en serie upptäckter, vars viktigaste var havsbotten spridning, att jordens litosfär, som inkluderar jordskorpan och den styva översta delen av den övre manteln, är uppdelad i ett antal tektoniska plattor som rör sig över det plastiska deformerande, fast, övre mantel, som kallas asthenosfären. Det finns en intim koppling mellan plattans rörelse på ytan och konvektionen av manteln: oceaniska plattrörelser och mantelkonvektionsströmmar rör sig alltid i samma riktning, eftersom den oceaniska litosfären är det styva övre termiska gränsskiktet i konvektionsmanteln . Denna koppling mellan styva plattor som rör sig på jordens yta och konvektionsmanteln kallas plattaktonik.

Completely incorrect answer (0.5%)

Why is Warsaw's flora very rich in species?	The flora of the city may be considered very rich in species. The species richness is mainly due to the **location** of Warsaw within the border region of several big floral regions comprising substantial proportions of close-to-wilderness areas (natural forests, wetlands along the Vistula) as well as arable land, meadows and forests. Bielany Forest, located within the borders of Warsaw, is the remaining part of the Masovian Primeval Forest. Bielany Forest nature reserve is connected with Kampinos Forest. It is home to rich fauna and flora. Within the forest there are three cycling and walking trails. Other big forest area is Kabaty Forest by the southern city border. Warsaw has also two botanic gardens: by the Łazienki park (a didactic-research unit of the University of Warsaw) as well as by the Park of Culture and Rest in Powsin (a unit of the Polish Academy of Science).
Varför är Warszawas flora mycket rik på arter?	Stadens flora kan anses vara mycket rik på arter. Artrikligheten beror främst på **Warszawa**s läge inom gränsområdet för flera stora blommoregioner som omfattar betydande andelar nära vildmarksområden (naturskogar, våtmarker längs Vistula) samt åkermark, ängar och skogar. Bielany Forest, som ligger inom gränserna till Warszawa, är den återstående delen av den masoviska urskogen. Naturreservatet Bielany Forest är anslutet till Kampinos Forest. Det är hem till rik fauna och flora. Inom skogen finns tre cykel- och vandringsleder. Ett annat stort skogsområde är Kabaty Forest vid den södra stadsgränsen. Warszawa har också två botaniska trädgårdar: av Łazienki-parken (en didaktisk-forskningsenhet vid universitetet i Warszawa) samt av parken för kultur och vila i Powsin (en enhet av den polska vetenskapsakademin).

Impossible due to translation (1.5%)

Note how the first word of the English answer, "attacked" (in Swedish "attackerade"), has been moved many words back in the Swedish translation, making it impossible to include it without also including many words that are not in the English answer.

During withdrawal from Fort William Henry, what did some Indian allies of French do?	French irregular forces (Canadian scouts and Indians) harassed Fort William Henry throughout the first half of 1757. In January they ambushed British rangers near Ticonderoga. In February they launched a daring raid against the position across the frozen Lake George, destroying storehouses and buildings outside the main fortification. In early August, Montcalm and 7,000 troops besieged the fort, which capitulated with an agreement to withdraw under parole. When the withdrawal began, some of Montcalm's Indian allies, angered at the lost opportunity for loot, **attacked the British column**, killing and capturing several hundred men, women, children, and slaves. The aftermath of the siege may have contributed to the transmission of smallpox into remote Indian populations; as some Indians were reported to have traveled from beyond the Mississippi to participate in the campaign and returned afterward having been exposed to European carriers.
Vad gjorde några indiska allierade franska under utträdet från Fort William Henry?	Franska oregelbundna styrkor (kanadensiska speider och indier) trakasserade Fort William Henry under första halvan av 1757. I januari övergick de bakhåll mot brittiska räknare nära Ticonderoga. I februari inledde de en vågig raid mot positionen över den frusna sjön George och förstörde lagerhus och byggnader utanför huvudbefästningen. I början av augusti beleirade Montcalm och 7000 trupper fortet, som kapitulerade med ett avtal om att dra sig tillbaka under parol. När tillbakadragandet började, attackerade några av Montcalms indiska allierade, arga över den förlorade möjligheten till **plundring, den brittiska kolumnen** och dödade och fångade flera hundra män, kvinnor, barn och slavar. Efterdyningarna av belägringen kan ha bidragit till överföringen av smittkoppor till avlägsna indiska populationer; som vissa indier rapporterades ha rest från utanför Mississippi för att delta i kampanjen och återvända efter att ha blivit utsatta för europeiska transportörer.

Impossible due to tokenization (1%)

Note that the original answer is incorrectly missing the final digit of the act, while the Swedish version has added it back. Even though it improves the final dataset, it is nevertheless an error in the context of cross-lingual projection. Since 1855 is treated as a single token we consider this "impossible due to tokenization".

What document formed the Parliament of Victoria?	Victoria has a written constitution enacted in 1975, but based on the 1855 colonial constitution, passed by the United Kingdom Parliament as the **Victoria Constitution Act 185**5, which establishes the Parliament as the state's law-making body for matters coming under state responsibility. The Victorian Constitution can be amended by the Parliament of Victoria, except for certain "entrenched" provisions that require either an absolute majority in both houses, a three-fifths majority in both houses, or the approval of the Victorian people in a referendum, depending on the provision.
Vilket dokument bildade Victoria parlamentet?	Victoria har en skriftlig konstitution som antogs 1975, men baserad på den koloniala konstitutionen 1855, som antogs av Storbritanniens parlament som **Victoria Constitution Act 1855**, som fastställer parlamentet som statens lagstiftande organ för frågor som kommer under statligt ansvar. Den viktorianska konstitutionen kan ändras av Victoria parlament, med undantag för vissa "förankrade" bestämmelser som kräver antingen en absolut majoritet i båda husen, en tre femtedelars majoritet i båda husen eller godkännande av det viktorianska folket i en folkomröstning, beroende på bestämmelsen.

Word Sense Distance in Human Similarity Judgements and Contextualised Word Embeddings

Janosch Haber and **Massimo Poesio**
Queen Mary University of London
{j.haber|m.poesio}@qmul.ac.uk

Abstract

Homonymy is often used to showcase one of the advantages of context-sensitive word embedding techniques such as ELMo and BERT. In this paper we want to shift the focus to the related but less exhaustively explored phenomenon of polysemy, where a word expresses various distinct but related senses in different contexts. Specifically, we aim to i) investigate a recent model of polyseme sense clustering proposed by Ortega-Andrés and Vicente (2019) through analysing empirical evidence of word sense grouping in human similarity judgements, ii) extend the evaluation of context-sensitive word embedding systems by examining whether they encode differences in word sense similarity and iii) compare the word sense similarities of both methods to assess their correlation and gain some intuition as to how well contextualised word embeddings could be used as surrogate word sense similarity judgements in linguistic experiments.

1 Introduction

Homonymy, the linguistic phenomenon of a word taking on a different meaning based on its context, such as *match* in (1), is often used to showcase one of the advantages of context-sensitive word embedding techniques such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2018) and GPT-2 (Radford et al., 2019) over their traditional word-vector counterparts such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), which are unable to encode context-dependent meaning.

(1) a. The match burned my fingers.
b. The match ended without a winner.

In this paper we want to shift the focus to the related but less exhaustively explored phenomenon of polysemy. We speak of polysemy when a word takes on different distinct but related senses given its context (Lyons, 1977), such as *school* in the various contexts of (2).[1]

(2) a. The school [building] is on fire.
b. The school [rules] has prohibited wearing hats in the classroom.
c. I have talked to the school [director, staff] about it already.
d. The school [participants] went for a visit to the cathedral.

Specifically, we aim to investigate a recent model of polyseme sense clustering proposed by Ortega-Andrés and Vicente (2019), suggesting that similarity differences in polysemic senses could lead to a grouping in their representation in the Generative Lexicon (Pustejovsky, 1991), addressing and attempting an explanation for processing differences observed within the seemingly homogeneous group of polysemes.

Through a range of surveys we collect word sense similarity judgements for a set of polysemes to provide empirical data for an investigation of word sense clustering as proposed by Ortega-Andrés and Vicente. We then aim to extend the linguistic evaluation of context-sensitive word embeddings by examining whether their contextualised encodings of polysemes show signs of word sense grouping, and whether these groupings correlate with the patterns observed in the human judgements. If this is the case, contextualised word embeddings could be used as surrogate word sense indicators in linguistic experiments.

1.1 Processing of Polysemes

While on a first glance homonymy and polysemy seem to be two closely related phenomena, poly-

[1] Examples taken from Ortega-Andrés and Vicente (2019)

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 128–145

semy should not be viewed as a simple extension of homonymic ambiguity: While the interpretation of a homonym requires the selection of one and only one specific meaning, polysemes have been found to activate multiple sense interpretations simultaneously and in many cases accommodate for sense shifting without additional processing cost. Frazier and Rayner (1990) for example showed that late disambiguating contexts can cause processing difficulties for homonyms but not so for polysemes. This observation led them to postulate a fully specified mental representation for homonymic meaning (i.e. one entry per meaning), but an un- or under-specified representation of polysemic sense. Studies like Klepousniotou (2002); Pylkkänen et al. (2006) and Klepousniotou et al. (2012) later revisited this experiment with the support of MEG and EEG readings, observing significant priming effects in homonyms but not so for polysemes. This led them, too, to postulate a principled processing difference in the interpretation of homonyms and polysemes.

A second case for a systematic difference between homonymy and polysemy has been made using so-called co-predication tests. In co-predication, two different meanings or senses of a word are simultaneously invoked by the context. In the case of homonymy, co-predication will always result in an infelicitous sentence, like for example in (3). For polysemous words on the other hand, co-predication with different senses seems to be felicitous in principle (e.g. example (4)).

(3) # The match burned my fingers but ended without a winner.

(4) Lunch was delicious but took forever. [food/meal]

1.2 Representation of Polysemes

A variety of linguistic models, including the Generative Lexicon (Pustejovsky, 1991) and Type Theory with Records (TTR, e.g. Cooper and Ginzburg, 2015), have been proposed to accommodate the observed processing differences between homonyms and polysemes. Specifically, Gotham (2014) proposed methods for addressing co-predication, quantification and individuation of polysemic senses in TTR, and Asher and Pustejovsky (2006) and Asher (2011, 2015) augmented the Generative Lexicon model by proposing that the various senses of a polyseme are represented by so-called dot-objects, complex objects that distinguish the different aspects, facets and types of polysemic sense interpretations, arguing that a word's context selects for the appropriate sense from within that representation.

Opposing a unified, under-specified representation of polysemic sense, a growing body of work however also collected a range of observations indicating that there might be significant and potentially systematic differences between various polysemic interpretations as well. Dating back to at least Apresjan (1974), for example, stems the idea that polysemes should be sub-divided into two types, *regular* (or *systematic*), and *irregular* polysemy, based on whether a polyseme's set of interpretations is idiosyncratic or shared among a group of similar words (also see Falkum (2015)). Supporting this principled split, Klepousniotou et al. (2012) report that their experiments indicate that regular polysemes might be represented differently than their irregular counterparts, arguing that in their processing, irregular polysemes more resemble homonymic meaning alterations than the sense alterations in regular polysemes. Furthering this discussion, Dölling (Forthcoming) recently collected a fine-grained distinction of 19 different patterns of polysemic sense alteration within the set of systematic polysemes, begging the question whether even regular polysemes form a homogeneous group and share a common representation, or whether these, too, require a more structured distinction than previously assumed.

Other evidence comes from an ongoing series of co-predication studies (Antunes and Chaves (2003); Traxler et al. (2005); Zobel (2017), and Filip and Sutton (2017); Sutton and Filip (2018); Schumacher (2013) for observations and models specifically concerning content/container alterations), showing that not all polysemic senses can be co-predicated either, and that the co-predication of some polysemic interpretations can lead to infelicitous and zeugmatic expressions, too (see example (5)).[2]

(5) a. # The newspaper fired its editor in chief and got wet from the rain. [publisher/publication]

b. # They took the door off its hinges and walked through it. [object/aperture]

[2]Examples from Cruse (1995)

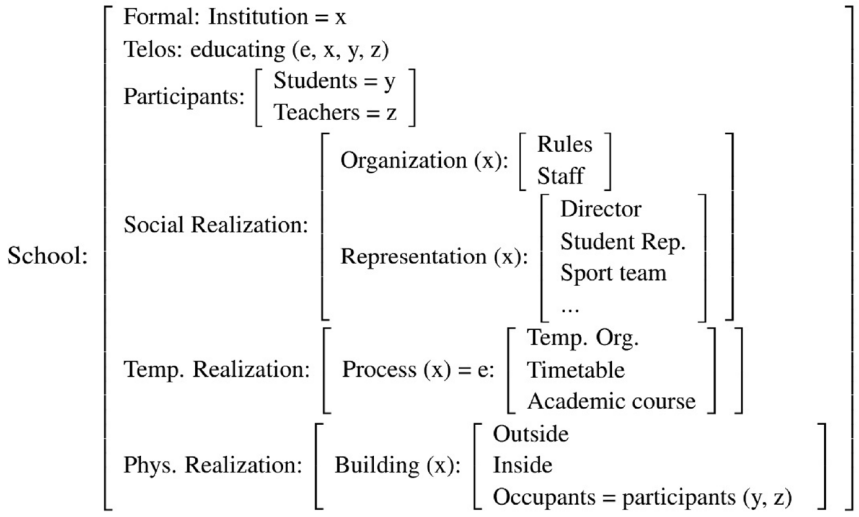


Figure 1: Knowledge structure for polyseme *school* proposed by Ortega-Andrés and Vicente (2019). Figure replicated from ibid., page 5.

To account for processing differences among polysemic senses, Ortega-Andrés and Vicente (2019) recently proposed an extension to Asher and Pustejovsky's model by postulating a hierarchical representation of polysemic sense that groups target word senses based on their similarity, creating so-called activation packages. Sense shifting is assumed to be automatic and free of processing costs within the under-specified representation of an activation package, but will lead to processing difficulties and infelicitous co-predication when moving outside of it. Figure 1 shows a proposed ordering of the hierarchical structure and resulting co-activation packages for polyseme *school* according to this model.

In summary, a number of recent observations, hypotheses and models concerning polysemy point towards a continuum of sense (or meaning) similarity between truly polysemous expressions (or, in Pinkal's terms, p-type ambiguity) and homonymic ambiguity (or h-type ambiguity, see Pinkal (1985) and Poesio (Forthcoming)) where some senses might be more closely related to one another than others. In this paper we aim to provide additional empirical data for investigating this claim by i) collecting graded word sense similarity judgements to assess the notion of word sense grouping as a driving factor in determining the representation of polysemic sense and accounting for differences in their processing costs and co-predication acceptability. In addition, we ii) investigate if clustering the representations of polysemes as generated by contextualised word-embedding techniques such as ELMo and BERT develops a word sense grouping and whether this grouping correlates with that derived from the collected sense similarity judgements. If this is the case, contextualised word embeddings could be used as surrogate word sense indicators in linguistic experiments.

2 Method

In order to generate the clearest results possible for investigating potential distances between different polysemic sense interpretations, we decided to use custom samples instead of resorting to corpus samples in this study. By creating the samples ourselves, we can construct contexts that invoke a certain polysemic sense as clearly as possible, and we can create sentence pairs that combine any of the different interpretations in order to have annotators judge their similarity directly. In addition, a preliminary investigation of context-sensitive polyseme representations obtained from ELMo revealed that factors such as i) the position of the target expression in the sentence, ii) its syntactic function and iii) the overall sentence length all significantly influence the resulting embedding and might overshadow the differences in encoding stemming from interpretation differences.[3] Designing custom samples helps us to control for these factors.

[3] See Appendix A

2.1 Samples

As target expressions for our samples we decided to focus on regular polysemes, as they are more likely to produce the clearest results possible due to their canonical division of sense interpretations. We selected ten of the systematic polysemy types compiled in Dölling (Forthcoming), with target expressions having between two and four clearly distinct but related senses, and picked one of the most frequently used expressions representing each class. We then created a sample set for each of the ten polysemes, containing two sample sentences for each of the target expression's interpretations.[4] The samples were created such that i) the target expression is the subject of the sentence, ii) the context is kept as short as possible, and iii) the context invokes a certain sense as clearly as possible without mentioning that sense explicitly.[5] As an example, consider the six sample sentences for polyseme *newspaper*, generated for its three senses (1) *organisation/institution*, (2) *physical object* and (3) *information/data*:

1a The newspaper fired its editor in chief.,
1b The newspaper was sued for defamation.
2a The newspaper lies on the kitchen table.,
2b The newspaper got wet from the rain.
3a The newspaper wasn't very interesting.,
3b The newspaper is rather satirical today.

All sample sentences were rated to be acceptable by annotators recruited from Amazon Mechanical Turk (AMT)[6] in a validation experiment. Individual sample sentences were then combined into pairs invoking all possible combinations of sense interpretations (i.e. creating nine sentence pairs for *newspaper*) and distributed over books so that no target expression appears twice in any book. In total, we generated 67 target pairs and distributed them over 15 books. We then followed Lau et al. (2014) by adding one of 15 sentence pairs containing homonyms and one of 15 sentence pairs containing synonyms to each book to create test items for spotting spammers, and further filled the books with random combinations of filler sentence pairs in order to disguise the focus on polysemes and present objectively low similarity items to calibrate the annotator's ratings.

[4] See Appendix B for details.
[5] As in "The school is an old building." for sense *building*
[6] https://www.mturk.com/

2.2 Human Judgements

We used AMT to collect word sense similarity judgements by highlighting (polysemic) target expressions in the sentence pairs and asking workers to rate the highlighted expressions using a slider labelled with "The highlighted words have a completely different meaning" on the left hand side and "The highlighted words have completely the same meaning" on the right.[7] The submitted slider positions are translated to a similarity score between 0 and 100 and stored in combination with a workers unique ID. To improve judgement quality, we required workers to have obtained a US high school degree and reached the "AMT Master" qualification.[8] Workers were paid 0.35 USD for every completed book.

We collected 20 judgements for each book. A total of 65 individual workers contributed to the study, with HITs taking an average of 133.4 seconds (median of 90.0). Through filtering out any books where the homonym sentence pair or a filler pair was given a similarity score higher than 60, or where the synonym sentence pair was rated lower than 50, we removed a total of 51 books and obtained an average of 16.6 judgements per sentence pair (min = 13).

2.3 Word Embeddings

Models of polysemy have previously been proposed in distributional semantics (see for example Boleda et al. (2012)), but for the most part, such models found limited application in computational linguistics. With the recent development of context-sensitive models of word embeddings such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), the field however obtained a new tool to capture polysemic sense alterations, leading to a demonstrated improvement in various NLP systems. ELMo was developed explicitly to capture a target word's context, processing sentences with a two-layered, bi-directional LSTM and computing the weighted sum of their hidden states depending on the task at hand to create contextualised word embeddings. BERT on the other hand is a language model that borrows and stacks the encoder architecture of the Transformer (Vaswani et al., 2017), an attention mecha-

[7] See Figure 7 for a screenshot of the AMT HIT interface.
[8] According to AMT's website, "[T]hese Workers have consistently demonstrated a high degree of success in performing a wide range of HITs across a large number of Requesters," https://www.mturk.com/worker/help

nism for learning the contextual relations between words, and adds a masking technique that allows for processing sentences in a non-directional fashion with minimised interference among the layers. While BERT's output, either an array of embeddings or a single pooled one, is normally fed to a further model to process a language-based task, our aim is to see whether it is able to capture any differences in polysemic sense and use its outputs directly.

For the ELMo analysis we used a pretrained model available on TensorFlow Hub[9] and extracted target word vectors from the LSTM's second layer hidden state, which has previously been shown to encode more semantic information than the character-level first layer or the LSTM's first layer (and consequently the ELMo output layer that combines them. See Appendix A, and Ethayarajh (2019)). For the investigation of BERT's embeddings we used the output of a pretrained cased model from the same repository[10] with 12 layers, a hidden state size of 768 and 12 attention heads. We extract i) sub-word vectors before pooling, ii) use the pooled sentence vector or iii) the embedding of the special [CLS] token.

3 Results

3.1 Similarity Differences in Judgements

As a first step, we calculated the overall means of word similarity judgements for all polyseme, homonym, synonym and filler sentence pairs in the dataset to determine any principled differences among these groups. The polyseme sentence pairs obtained a mean similarity rating of 87.12 (std=20.92), synonym sentence pairs a mean of 92.38 (std=10.35), homonym pairs a mean of 3.76 (std=8.37) and filler sentence pairs a mean of 2.71 (std=7.19). We then used Student's T-Test to compare the distributions of judgements; The polyseme and synonym distributions each are significantly different from all other distributions ($p<0.05$). This means that annotators rated synonyms to be overall more similar to each other than different uses of polysemes - a first indicator that word sense interpretations might not be perceived as carrying identical meaning.

Because the ten different polysemous target expressions used in this study each represent a different type of regular polysemy, we next split the collected judgements based on their target expression and calculated the mean sense similarity judgements for same-sense and cross-sense sentence pairs. Table 1 displays these numbers, showing that same-sense means are consistently higher than the cross-sense ones, and except for *glass* and *construction* range above 95 (i.e. higher than the synonym mean). This means that barring these two outliers, the generated same-sense pairs were rated as invoking an almost identical interpretation of the polysemic target expression. The average similarity of cross-sense pairs often ranges between 80 and 90, showing a high similarity still, but indicating that not all cross-sense pairs seem to be perceived as invoking the same sense.

Polyseme	**Same-sense**		**Cross-sense**	
	mean	std	mean	std
Newspaper (3)	99.17	2.36	77.71	30.08
Hemingway (2)	96.26	16.64	85.64	24.71
War and Peace (3)	99.55	2.65	91.78	22.73
Lunch (2)	96.15	11.98	80.35	24.51
Door (2)	99.33	2.27	95.88	9.73
DVD (3)	95.56	12.34	88.12	20.58
School (4)	96.30	8.57	88.08	22.97
Wine (2)	99.85	0.50	92.30	17.25
Glass (2)	70.39	35.02	65.03	38.02
Construction (2)	86.49	21.65	59.93	33.44

Table 1: Polysemic target expression (number of senses), and means and standard deviations of the same-sense and cross-sense samples' pairwise similarity ratings.

Turning to a more qualitative analysis of the results obtained for each individual polyseme, we investigated the similarity ratings obtained for sentence pairs containing a specific target expression to assess whether the collected data provides any evidence for sense clustering as proposed by Ortega-Andrés and Vicente (2019). Since it is difficult to collapse results over the different types of polysemes tested, we here exemplify our analyses through a summary of the observations concerning polyseme *newspaper* and draw parallels to other test items where possible. As mentioned above, the polyseme *newspaper* was taken to invoke one of three distinct but related senses; (1) *organisation/institution*, (2) *physical object* and (3) *information/data*, and creating all combinations of senses generates the

[9] https://tfhub.dev/google/ELMo/3

[10] https://tfhub.dev/google/bert_cased_L-12_H-768_A-12/1

following nine sentence pairs:[11]

11 organisation/organisation
22 physical/physical
33 information/information
12 organisation/physical
21 physical/organisation
13 organisation/information
31 information/organisation
23 physical/information
32 information/physical

Figure 2 shows the mean word similarity judgements for these nine sentence pairs. The three same-sense pairs 11, 22, and 33 (red) receive mean similarity ratings close to 100, showing that in these cases annotators indeed perceive the target word contexts to invoke exactly the same sense in both sample sentences. This effect can be observed for all tested polysemes except for *glass*, where one of the same-sense pairs does not actually seem to elicit the same sense (rated at a similarity of 48) and a same-sense pair for *construction* which only received a similarity score of 82 (being higher still than the cross-sense pairs). Returning to *newspaper*, all six cross-sense pairs receive lower ratings than the same-sense pais: Both, the *organisation/physical* sentence pairs 12 and 21 (yellow), and the *organisation/information* sentence pairs 13 and 31 (green) receive significantly lower similarity ratings than the same-sense pairs. The similarity ratings for the *physical/information* pairs 23 and 32, (blue) are ranging between 90 and 100, being significantly higher than the ratings for pairs 12, 21, 13, but significantly lower than same sense-sense pair 22. This indicates that at least between the *organisation* and *physical* sense interpretation there seems to be a notable difference in meaning, while the *information* readings are judged to be relatively similar to either - however not to a level that same-sense sample pairs are similar to each other. We see a similar but less pronounced effect for the tested polysemes with two senses, where cross-sense pairs are rated as being less similar than the same-sense pairs, as well as in some of the senses of target expressions with three or four interpretations, with significant differences between the *building* and *administration* and *institution* senses of polyseme *school*.

[11] see Appendix G for the full list of sample sentences.

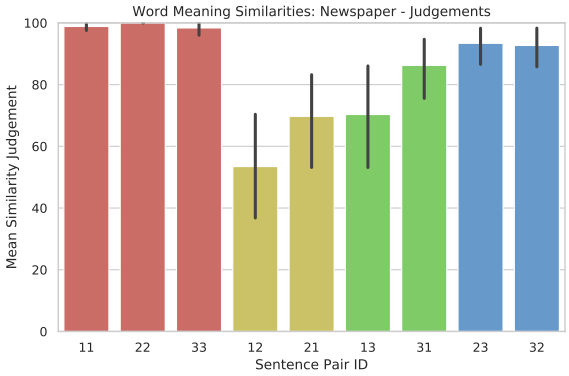


Figure 2: Similarity judgements for sentence pairs containing the polyseme *newspaper*. The two numbers in the sentence pair IDs indicate the combination of senses. The first three bars thus indicate same-sense pairs, the other three groups the different variations of cross-sense samples. The full set of similarity judgements graphs can be found in Appendix D.

Returning to the *newspaper* samples, a second point of interest are the notable though non-significant differences in similarity ratings for sentence pairs 12 and 21, and 13 and 31, respectively. Since these sentence pairs were created to invoke the same pair of (cross-sense) interpretations, it is noteworthy that their ratings differ so much. This difference can be the result of two factors: i) the sentence pairs contain different sample sentences, which within the same sense interpretation could evoke interpretation differences, and ii) the order of presentation for the two sentence pairs is different, and presentation order is known to induce biases and affect acceptability in co-predication studies. To control for the latter, we repeated our experiments with the same set of samples, but inverting the presentation order within the sentence pairs. Based on an average of ten judgements, only one of the 67 sentence pairs' similarity ratings changed significantly, indicating that the observed difference in similarity ratings is not an effect of presentation order, but indeed due to subtle interpretation differences in the contexts used to elicit a certain sense.

3.2 Correlation with Embedding Techniques

Observing noticeable differences in the word sense similarity ratings between some of the sample sentences invoking different interpretations of a polyseme - and in some cases even within sentence pairs that were designed to invoke the same

	Newsp.	Hemingw.	W&P	Lunch	Door	DVD	School	Wine	Glass	Constr.
BERT WE	0.383	0.692	**0.235**	**0.899**	0.079	0.409	0.259	0.459	-0.739	**0.623**
BERT SE	0.591	**0.999***	-0.159	0.316	**0.449**	0.355	0.092	0.458	-0.973*	-0.115
BERT CLS	0.317	0.960*	0.017	0.152	-0.202	**0.517**	0.084	0.216	-0.933	-0.492
ELMo WE	**0.919***	0.916	-0.310	-0.278	0.018	-0.167	**0.332**	0.442	-0.666	0.648
Word2Vec SE	0.576	0.126	0.089	-0.923	0.177	0.361	-0.310	**0.795**	-0.614	0.117

Table 2: Correlations between human sense similarity judgements and the similarities in the representations derived from different contextualised word embedding techniques as measured with Pearson's r. Highest correlating model output in bold font, significant correlations ($p<0.05$) starred.

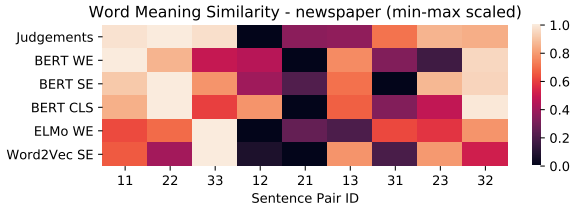


Figure 3: Comparison of word sense similarity ratings based on annotator judgements and ELMo and BERT context-sensitive word embeddings, min-max normalised to amplify the visibility of effects. Brighter indicates higher similarity.

sense - we proceeded to investigate whether the contextualised embeddings of polysemes generated by ELMo and BERT also exhibit word sense similarity differences. To this end we used the raw embeddings as returned by the models and calculate their similarity based on cosine. If a target expression contains multiple words (or sub-word tokens in the case of BERT), we average the embeddings of all parts. In addition to ELMo and BERT word embeddings for the target expressions alone, we also consider BERT's pooled sentence embedding, the embedding of the special [CLS] token, and a sentence embedding by Word2Vec created by averaging all word embeddings in the sentence. Table 2 displays the correlations between the human sense similarity ratings and the cosine similarities of the target expressions (or sentences) given these different embedding techniques. With only a fraction of the correlations being significant,[12] none of the embedding techniques appears to capture the similarity patters observed in the human judgements consistently, with each of the methods achieving the strongest correlation for one or two of the target expressions, but also showing instances of negative or no correlation for some samples.

[12]Note that the compared similarity vectors are of length 4-16 only

Moving to a more qualitative analysis of the contextualised embeddings, we created heat maps to display the similarity patterns for the different polysemic expressions tested. The resulting heat map for *newspaper* is shown in Figure 3, displaying on a more accessible level the difference in correlation between the human judgements and contextualised embeddings.[13] While in some cases the cosine similarities between the contextualised embeddings seem to reflect the human judgements - especially so for sense interpretations rated to be highly similar (e.g. 11, 22 and 32) or dissimilar (12, 21) - overall the differences in embeddings do not consistently resemble the human judgements. This observation is replicated throughout the ten polysemes tested in this study, with some of the 2-sense samples also exhibiting more consistent patterns.

While the similarities between contextualised embeddings do not consistently match the collected sense similarity ratings, the patterns in their embeddings indicate that they do differentiate between the different contexts. We further investigated this intuition by applying a non-linear function to reduce the dimensionality of 15 different word embeddings for polyseme *newspaper* produced by ELMo using t-SNE (van der Maaten and Hinton, 2008) and visualising the result in the two-dimensional scatter plot displayed in Figure 4. The samples for this experiment were created to invoke the polyseme's three senses *organisation* - red (1-5), *physical* - yellow (6-10), and *information* - green (11-15).[14] And while no clear grouping into different sense clusters seems to emerge, we do observe a similar pattern to that found in Figure 2, namely that the *physical* interpretations seem to be more similar to the *information* senses,

[13]The heat maps for the full set of tested polysemes can be found in Appendix E.

[14]See Appendix F for the list of samples.

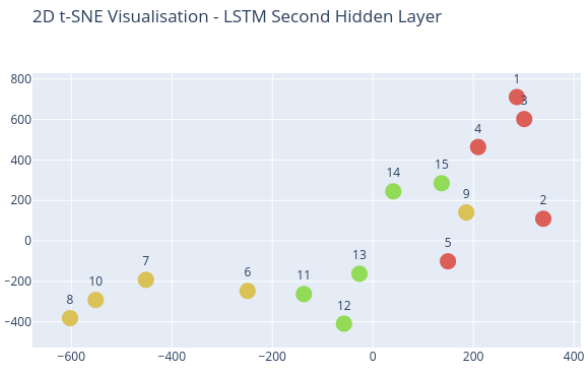


Figure 4: t-SNE scatter plot of the reduced ELMo embeddings for 15 instances of polyseme *newspaper* presented with disambiguating contexts for its three senses: *organisation* - red, *physical* - yellow, and *information* - green. Sample sentences in Appendix F.

which are occupying the space between *physical* and *organisation* readings. Note however that due to the working of t-SNE's dimensionality reduction algorithm, results can change between iterations, and the observed pattern is not always visible. This means that from this test alone it is unclear whether polysemic sense is indeed encoded in the ELMo embeddings.

4 Conclusion

While our results are difficult to collapse as they survey different types polysemes, there are some overarching conclusions to be drawn from the data collected in this study. First of all, we provide empirical evidence that readers are indeed sensitive to differences between polysemic word senses. Polysemic target expressions in contexts designed to invoke the same sense interpretation are consistently rated as highly similar, while similarity ratings of cross-sense pairs receive ratings on a spectrum ranging from highly similar to significantly dissimilar. It thus seems that some sense interpretations are perceived to be more similar to each other than others, providing support for a similarity-based grouping of word senses like the one as proposed by Ortega-Andrés and Vicente (2019). In some cases, distances in sense interpretations correspond to intuitive groupings of senses, but the collected judgements also reveal a notion of gradedness that usually is not assumed to be present in canonical samples (see e.g. Lau et al. (2014)). Given these observations, we see merit in exploring a more structured representation of polysemic sense, since a fully underspecified, single-entry approach would be insufficient to fully account for them. Having investigated only one target expression for a small set of systematic polysemes, we acknowledge that more empirical research is needed to investigate potential patterns within polysemy types or in the much larger set of irregular polysemes in order to determine whether there are any systematic effects - or whether each and every polysemic expression requires its own idiosyncratic representation structure. The graded word sense judgements obtained through our data collection however also indicate to us that a categorical approach to word sense disambiguation (WSD) such as implied by a number of recent models (e.g. Levine et al. (2019); Wiedemann et al. (2019); Blevins and Zettlemoyer (2020)) might be geared more towards the distinction of homonymic ambiguity and fall short of capturing the full spectrum of polysemic sense alterations. Current approaches focusing on graded word sense similarity and word sense shifting (see for example Armendariz et al. (2019)) on the other hand might produce new insights in mapping out the intricacies of polysemic word sense interpretation and representation.

Concerning the encoding of polysemic sense in the contextualised word embeddings of ELMo and BERT, we do observe differences in the representation of polysemic expressions invoking different sense interpretations which could indicate the encoding of context-specific information, but similarities between word embeddings do not consistently correlate with the collected human similarity judgements. While the raw embeddings thus cannot directly be used to distinguish polysemic senses to the same degree as human judgements do, they still could contain word sense information that requires non-linear functions in order to be accessed. Failing to provide conclusive answers in this respect, we hope that future work will help to determine to what extend - and how exactly - polysemic sense is represented in contextualised embeddings to shed more light into the black box processes that improve so many NLP systems.

Acknowledgements

The work presented in this paper was supported by the DALI project, ERC Grant 695662. The authors would like to thank Derya Çokal and Andrea Bruera for their input, and the anonymous reviewers for their feedback.

References

Sandra Antunes and Rui Pedro Chaves. 2003. On the Licensing Conditions of Co-Predication. In *Proceedings of the 2nd International Workshop on Generative Approaches to the Lexicon*.

Juri D. Apresjan. 1974. Regular polysemy. *Linguistics*, 12:5–32.

Carlos Santos Armendariz, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, Marko Robnik-Šikonja, Mark Granroth-Wilding, and Kristiina Vaik. 2019. Cosimlex: A resource for evaluating graded word similarity in context.

Nicholas Asher. 2011. *Lexical Meaning in Context: A Web of Words*. Cambridge University Press.

Nicholas Asher. 2015. Types, meanings and coercions in lexical semantics. *Lingua*, 157:66–82.

Nicholas Asher and James Pustejovsky. 2006. A type composition logic for generative lexicon. *Journal of Cognitive Science*, 6(1).

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss-informed biencoders.

Gemma Boleda, Sabine Schulte im Walde, and Toni Badia. 2012. Modeling regular polysemy: A study on the semantic classification of catalan adjectives. *Computational Linguistics*, 38(3):575–616.

Robin Cooper and Jonathan Ginzburg. 2015. *Type Theory with Records for Natural Language Semantics*, chapter 12. John Wiley & Sons, Ltd.

Alan D. Cruse. 1995. Polysemy and related phenomena from a cognitive linguistic viewpoint. In Patrick Saint-Dizier and EvelynEditors Viegas, editors, *Computational Lexical Semantics*, Studies in Natural Language Processing, page 33–49. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Johannes Dölling. Forthcoming. Systematic Polysemy. In Daniel Gutzmann, Lisa Matthewson, Cécile Meier, Hotze Rullmann, and Thomas Ede Zimmermann, editors, *The Blackwell Companion to Semantics*. Wiley.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Ingrid Lossius Falkum. 2015. Polysemy: Current perspectives and approaches. *Lingua*, 157:1–16.

Hana Filip and Peter Sutton. 2017. Singular count NPs in measure constructions. In *Semantics and Linguistic Theory*, volume 27, pages 340–357.

Lyn Frazier and Keith Rayner. 1990. Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*.

Matthew Graham Haigh Gotham. 2014. *Copredication, quantification and individuation*. Ph.D. thesis, UCL (University College London).

Ekaterini Klepousniotou. 2002. The Processing of Lexical Ambiguity: Homonymy and Polysemy in the Mental Lexicon. *Brain and Language*, 81(1-3):205–223.

Ekaterini Klepousniotou, G. Bruce Pike, Karsten Steinhauer, and Vincent Gracco. 2012. Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and Language*.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2014. Measuring Gradience in Speakers' Grammaticality Judgements. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society (CogSci 2014)*.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2019. Sensebert: Driving some sense into bert.

John Lyons. 1977. *Semantics*, volume 2. Cambridge University Press.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

Marina Ortega-Andrés and Agustín Vicente. 2019. Polysemy and co-predication. *Glossa: a journal of general linguistics*, 4(1).

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Manfred Pinkal. 1985. *Logik Und Lexikon: Die Semantik des Unbestimmten*. De Gruyter.

Massimo Poesio. Forthcoming. Ambiguity. In Daniel Gutzmann, Lisa Matthewson, Cécile Meier, Hotze Rullmann, and Thomas Ede Zimmermann, editors, *The Blackwell Companion to Semantics*. Wiley.

James Pustejovsky. 1991. The Generative Lexicon. *Comput. Linguist.*, 17(4):409–441.

Liina Pylkkänen, Rodolfo Llinás, and Gregory L Murphy. 2006. The representation of polysemy: Meg evidence. *Journal of cognitive neuroscience*, 18(1):97–109.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Petra Schumacher. 2013. When combinatorial processing results in reconceptualization: toward a new approach of compositionality. *Frontiers in Psychology*, 4:677.

Peter R Sutton and Hana Filip. 2018. Counting Construcions and Coercion: Container, Portion and Measure Interpretations. *Oslo Studies in Language*, 10(2).

Matthew J. Traxler, Brian McElree, Rihana S. Williams, and Martin J. Pickering. 2005. Context effects in coercion: Evidence from eye movements. *Journal of Memory and Language*, 53(1):1–25.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings.

Sarah Zobel. 2017. The sensitivity of natural language to the distinction between class nouns and role nouns. In *Semantics and Linguistic Theory*, volume 27, pages 438–458.

Appendices

A Analysis of ELMo Embeddings

In order to determine what factors would have to be taken into account when generating test samples for analysing word sense differences through their embedding, we ran a series of preliminary experiments comparing the embeddings of target words in different context settings. We conducted these experiments using ELMo embeddings obtained by accessing specific target words from sentence embeddings created based on these different context conditions. Using canonical co-predication examples at first, we quickly realised that the position and function of the target word within the sentence has a significant effect on the resulting embedding and thus potentially overshadows any effects caused by sense shifting. To control for this, we next created a set of sample sentences that fixed the position and function of the target word and generated four levels of context: 1) the absolute minimal context to invoke a certain sense, 2) compact context, 3) extensive but descriptive context, 4) extensive, natural context. Using polyseme *newspaper*, with senses a) *physical*, b) *information*, and c) *organisation* we generated the following samples according to these guidelines:

1a The newspaper is folded.
1b The newspaper is boring.
1c The newspaper is famous.
2a The newspaper is lying on the table.
2b The newspaper is listing job openings.
2c The newspaper is struggling financially.
3a The newspaper is made up of 40 sheets of thin, recycled paper, has three columns of text and only a few colour images.
3b The newspaper contains reports on national and international incidents, the daily weather report and sports results.
3c The newspaper fired its editor in chief after her new business strategy caused the company to lose important partners.
4a The newspaper got wet from the sprinklers because the paper boy hadn't thrown it far enough to reach the front porch.
4b The newspaper wasn't very interesting but got the local obituaries and job offers which were read by almost everyone.
4c The newspaper was attacked over its populist coverage of the recent events surrounding the general election in May.

We then calculated the cosine similarities between the embeddings of the target word *newspaper* for all sentence pairs using the LSTM's first layer's hidden state, the LSTM's second layer's hidden state and the ELMo output. See Figure 5 for results. The embeddings of sample sentences 1-6 seem to form a cluster of high similarity compared to the rest of the pairwise comparisons in all of the embedding layers. It thus seems that the extensive context of samples 7-12 causes the target word embeddings to be noticeably different from those of the short context samples. As we aim to analyse the differences between the different senses of a target word and solely need context to invoke these different senses, we propose to keep the context in the test samples for our experiments as short and descriptive as possible.

To determine which output layer provides the most sensitivity to word sense, we calculated the similarity of each sense cluster's mean to the other cluster mean vectors to establish the overall distances between embedding vectors of different senses, i.e. the amount of variance in the outputs. We propose that if this variance is higher, the embeddings are easier to differentiate and different senses therefore might be identified more easily. The result of this experiment is shown in Figure 6, revealing the the second layer's hidden state exhibits the largest differences among the three sense cluster means. We therefore decided to use this embedding layer output for our experiments.

B Sample Creation

Instead of creating sentence pairs by combining every sample sentence for a given polyseme with every other one, we decided to create two pairs for every cross-sense sentence condition only and just one pair for every same-sense condition. This was done by combining the selected first sample sentence for every sense (a) with the selected second sample sentence for every sense (b). For a polyseme with two senses, this results in the four sentence pairs

1a - 1b (ID=11)	2a - 2b (ID=21)
1a - 2b (ID=12)	2a - 1b (ID=22)

By analogy, a polyseme with three distinct senses generates nine samples:

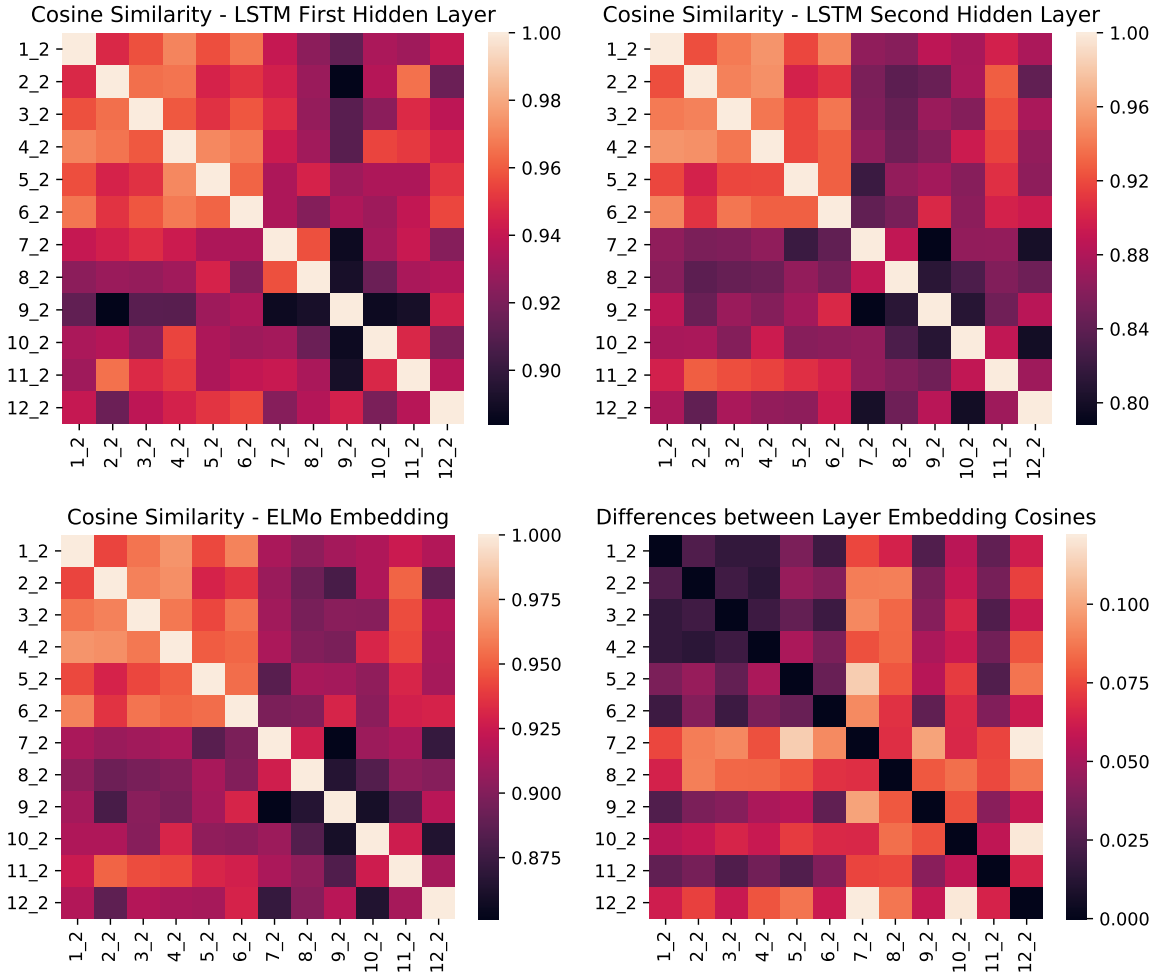


Figure 5: Heat maps of the pairwise cosine similarity of target word embeddings using a given ELMo layer, and a heat map of the differences in cosine similarity between the first and second LSTM layers' hidden state representation

1a - 1b (ID=11)
1a - 2b (ID=12)
1a - 3b (ID=13)
2a - 2b (ID=22)
2a - 1b (ID=21)
2a - 3b (ID=23)
3a - 3b (ID=33)
3a - 1b (ID=31)
3a - 2b (ID=48)

We leave it to the reader to apply this system to generate the 16 pairs for polysemes with four senses. Note that this procedure creates cross-sense pairs with each of the two senses being the first one in the pair once.

C AMT Interface

A screenshot of the AMT user interface can be found in Figure 7.

D Word Sense Similarity Graphs

Graphs of the word sense similarity judgements for the ten regular polysemes tested can be found in Figure 8.

E Comparison of Similarity Ratings

Graphs of the correlation between human word sense similarity judgements and ELMo and BERT embeddings for the ten polysemes tested in this study can be found in Figure 9.

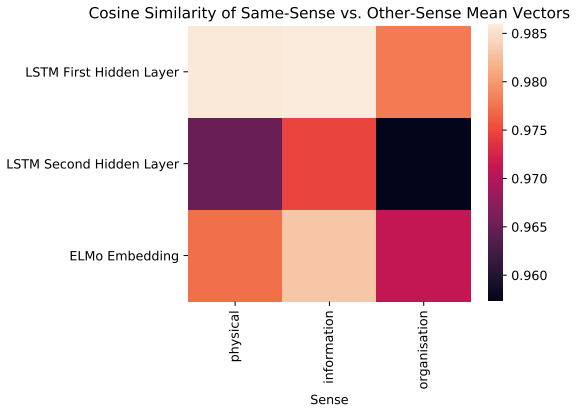


Figure 6: Cosine similarities of sense cluster means to the other senses' means - a measure of overall sense embedding differences in the ELMo layers.

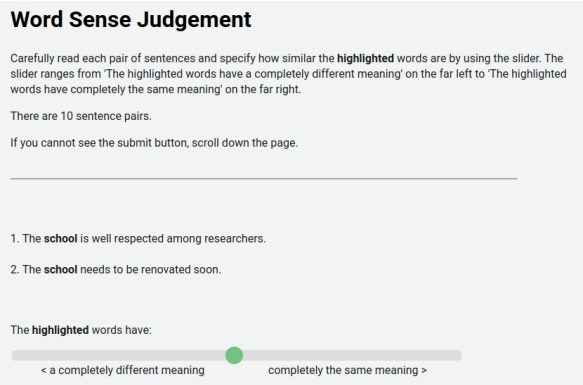


Figure 7: Screenshot of the Amazon Mechanical Turk (AMT) user interface designed to collect the word sense similarity judgements.

F Sense Clustering Samples

The newspaper fired its editor in chief.
The newspaper struggles financially.
The newspaper hired a new designer.
The newspaper was sued for defamation.
The newspaper has around 150 employees.
The newspaper has large coffee stains.
The newspaper lies on the kitchen table.
The newspaper got wet from the rain.
The newspaper weighs less than yesterday.
The newspaper fell behind the counter.
The newspaper contains advertisements.
The newspaper listed all affected stores.
The newspaper has a large sports section.
The newspaper wasn't very interesting.
The newspaper is rather satirical today.

G Full Sample List

Using the procedure described in 2 and Appendix B, the following sentence pairs were created:

Newspaper
11: newspaper: organisation/organisation,
The newspaper fired its editor in chief.,
The newspaper was sued for defamation.
12: newspaper: organisation/physical,
The newspaper fired its editor in chief.,
The newspaper got wet from the rain.
13: newspaper: organisation/information,
The newspaper fired its editor in chief.,
The newspaper is rather satirical today.
21: newspaper: physical/organisation,
The newspaper lies on the kitchen table.,
The newspaper was sued for defamation.
22: newspaper: physical/physical,
The newspaper lies on the kitchen table.,
The newspaper got wet from the rain.
23: newspaper: physical/information,
The newspaper lies on the kitchen table.,
The newspaper is rather satirical today.
31: newspaper: information/organisation,
The newspaper wasn't very interesting.,
The newspaper was sued for defamation.
32: newspaper: information/physical,
The newspaper wasn't very interesting.,
The newspaper got wet from the rain.
33: newspaper: information/information,
The newspaper wasn't very interesting.,
The newspaper is rather satirical today.

Hemingway
11: Hemingway: person/person,
Hemingway was born in Illinois.,
Hemingway won a Nobel prize.
12: Hemingway: person/work,
Hemingway was born in Illinois.,
Hemingway is not suitable for children.
21: Hemingway: work/person,
Hemingway is still widely read today.,
Hemingway won a Nobel prize.
22: Hemingway: work/work,
Hemingway is still widely read today.,
Hemingway is not suitable for children.

War and Peace
11: War and Peace: work/work,
War and Peace was finally published in 1869.,
War and Peace won a range of international awards.
12: War and Peace: work/content,
War and Peace was finally published in 1869.,
War and Peace describes a number of historic battles.
13: War and Peace: work/physical,
War and Peace was finally published in 1869.,
War and Peace is bound in black embossed leather.
21: War and Peace: content/work,
War and Peace chronicles the period of 1805 to 1820.,
War and Peace won a range of international awards.
22: War and Peace: content/content,
War and Peace chronicles the period of 1805 to 1820.,
War and Peace describes a number of historic battles.
23: War and Peace: content/physical,
War and Peace chronicles the period of 1805 to 1820.,
War and Peace is bound in black embossed leather.
31: War and Peace: physical/work,
War and Peace gathers dust on the top shelf.,
War and Peace won a range of international awards.

32: War and Peace: physical/content,
War and Peace gathers dust on the top shelf.,
War and Peace describes a number of historic battles.
33: War and Peace: physical/physical,
War and Peace gathers dust on the top shelf.,
War and Peace is bound in black embossed leather.

Lunch
11: lunch: food/food,
Lunch was exceptionally delicious today.,
Lunch got cold while we waited for someone.
12: lunch: food/event,
Lunch was exceptionally delicious today.,
Lunch is great for socialising and networking.
21: lunch: event/food,
Lunch took more than an hour yesterday.,
Lunch got cold while we waited for someone.
22: lunch: event/event,
Lunch took more than an hour yesterday.,
Lunch is great for socialising and networking.

Door
11: door: physical/physical,
The door was turned into a table top.,
The door splintered when they hit it.
12: door: physical/aperture,
The door was turned into a table top.,
The door connects the two rooms.
21: door: aperture/physical,
The door leads to a long hallway.,
The door splintered when they hit it.
22: door: aperture/aperture,
The door leads to a long hallway.,
The door connects the two rooms.

DVD
11: DVD: physical/physical,
The DVD has some scratches but looks fine.,
The DVD got stuck in the player yesterday.
12: DVD: physical/content,
The DVD has some scratches but looks fine.,
The DVD wasn't very entertaining somehow.
13: DVD: physical/medium,
The DVD has some scratches but looks fine.,
The DVD has won the battle against VHR.
21: DVD: content/physical,
The DVD is a low resolution home movie.,
The DVD got stuck in the player yesterday.
22: DVD: content/content,
The DVD is a low resolution home movie.,
The DVD wasn't very entertaining somehow.
23: DVD: content/medium,
The DVD is a low resolution home movie.,
The DVD has won the battle against VHR.
31: DVD: medium/physical,
The DVD will be replaced by BluRay soon.,
The DVD got stuck in the player yesterday.
32: DVD: medium/content,
The DVD will be replaced by BluRay soon.,
The DVD wasn't very entertaining somehow.
33: DVD: medium/medium,
The DVD will be replaced by BluRay soon.,
The DVD has won the battle against VHR.

School
11: school: building/building,
The school was painted during the holidays.,
The school needs to be renovated soon.
12: school: building/administration,
The school was painted during the holidays.,
The school informed parents about this year's events.
13: school: building/institution,
The school was painted during the holidays.,
The school recently got a more modern website.
14: school: building/students,
The school was painted during the holidays.,
The school went on a field trip last summer.
21: school: administration/building,
The school requires students to wear a uniform.,
The school needs to be renovated soon.
22: school: administration/administration,
The school requires students to wear a uniform.,
The school informed parents about this year's events.
23: school: administration/institution,
The school requires students to wear a uniform.,
The school recently got a more modern website.
24: school: administration/students,
The school requires students to wear a uniform.,
The school went on a field trip last summer.
31: school: institution/building,
The school is well respected among researchers.,
The school needs to be renovated soon.
32: school: institution/administration,
The school is well respected among researchers.,
The school informed parents about this year's events.
33: school: institution/institution,
The school is well respected among researchers.,
The school recently got a more modern website.
34: school: institution/students,
The school is well respected among researchers.,
The school went on a field trip last summer.
41: school: students/building,
The school developed an important algebraic proof.,
The school needs to be renovated soon.
42: school: students/administration,
The school developed an important algebraic proof.,
The school informed parents about this year's events.
43: school: students/institution,
The school developed an important algebraic proof.,
The school recently got a more modern website.
44: school: students/students,
The school developed an important algebraic proof.,
The school went on a field trip last summer.

Wine
11: wine: container/container,
The wine lay in a padded wooden box.,
The wine is a little dusty from storage.
12: wine: container/content,
The wine lay in a padded wooden box.,
The wine tastes great with fish.
21: wine: content/container,
The wine had a beautiful red tint.,
The wine is a little dusty from storage.
22: wine: content/content,
The wine had a beautiful red tint.,
The wine tastes great with fish.

Glass
11: glass: container/container,
The glass broke when she dropped it.,
The glass fits about 200 ml of liquid.
12: glass: container/content,
The glass broke when she dropped it.,
The glass was absolutely refreshing.
21: glass: content/container,
The glass had a thick layer of foam.,
The glass fits about 200 ml of liquid.
22: glass: content/content,
The glass had a thick layer of foam.,

The glass was absolutely refreshing.

Construction

11: construction: process/process,
The construction took far longer than expected.,
The construction will begin in early September.
12: construction: process/product,
The construction took far longer than expected.,
The construction is larger than most in the city.
21: construction: product/process,
The construction has a solid steel frame.,
The construction will begin in early September.
22: construction: product/product,
The construction has a solid steel frame.,
The construction is larger than most in the city.

Homonyms

0: Homonym: bat,
The bat came in through the open window.,
The bat broke when he hit the fence with it.
1: Homonym: match,
The match burned my fingers.,
The match ended without a winner.
2: Homonym: club,
The club only admits women older than 50.,
The club felt very heavy and unwieldy.
3: Homonym: bank,
The bank was washed out by the current.,
The bank increased the interest rate.
4: Homonym: mole,
The mole dug tunnels all throughout the garden.,
The mole needs to be removed as it is cancerous.
5: Homonym: pitcher,
The pitcher threw a number of perfect curveballs.,
The pitcher broke when the waiter dropped it.
6: Homonym: rocket,
The rocket left the atmosphere at 2AM tonight.,
The rocket was bitter taste and ruined the pizza.
7: Homonym: tank,
The tank could easily fit 500 litres of water.,
The tank could easily shoot further than 3 miles.
8: Homonym: watch,
The watch slipped off his hand while he was swimming.,
The watch reported troop movements on the south border.
9: Homonym: yard,
The yard equals exactly three feet.,
The yard is just over 10 feet wide.
10: Homonym: stall,
The stall barely fit the large bull.,
The stall didn't have any toilet paper.
11: Homonym: spring,
The spring in the garden feeds the little pond with fresh water.,
The spring in the ballpen lets you open it with a simple click.
12: Homonym: mine,
The mine had to close after the accident.,
The mine could be defused by an expert.
13: Homonym: order,
The order welcomed the new members.,
The order was shipped two weeks late.
14: Homonym: jumper,
The jumper broke a long-standing record.,
The jumper didn't really fit her that well.

Synonyms

0: Synonym: answer/reply,
The answer came after more than a month.,
The reply arrived within a couple of minutes.
1: Synonym: street/road,
The street leads to a small town in the mountains.,
The road ends at a beautiful hut made from wood.
2: Synonym: world/planet,
The world is heating up because of CO2 emissions.,
The planet is heading towards a serious climate crisis.
3: Synonym: computer/PC,
The computer suddenly turned off.,
The PC needs to be replaced soon.
4: Synonym: problem/issue,
The problem was solved by replacing a cable.,
The issue couldn't be resolved without tools.
5: Synonym: capability/ability,
The capability of modern computers is astonishing.,
The ability to read and write is crucially important.
6: Synonym: area/space,
The area was roped off by the police.,
The space was littered with rubbish.
7: Synonym: audience/crowd,
The audience was very quiet during the concert.,
The crowd was cheering on the football team.
8: Synonym: note/memo,
The note on the fridge read "clean me!".,
The memo simply said "Meeting at 1PM".
9: Synonym: advice/tip,
The advice wasn't very good.,
The tip helped to fix the TV. 10: Synonym: photo/image,
The photo was of a picturesque lake.,
The image shows a red muscle car.
11: Synonym: building/structure,
The building burned down last week.,
The structure collapsed years ago.
12: Synonym: company/organisation,
The company had to find a new office building.,
The organisation expanded to Eastern Europe.
13: Synonym: plank/board,
The plank was torn out of the floor.,
The board covered up a crack in the wall.
14: Synonym: sea/ocean,
The sea was much colder than the beach.,
The ocean looked beautiful in the sunset.

Fillers

11: Filler: bottle/The Guardian,
The bottle fell off the kitchen counter.,
The Guardian contains advertisements.
12: Filler: War of the Worlds/The Guardian,
War of the Worlds is made up of five consecutive parts.,
The Guardian is rather satirical today.
13: Filler: Dickens/university,
Dickens was born in Portsmouth.,
The university was closed during the holidays.
14: Filler: CD/Dinner,
The CD broke when I accidentally sat on it.,
Dinner got cold while we waited for someone.
15: Filler: Dickens/CD,
Dickens didn't really grip me.,
The CD sparked discussions about copyright laws.
16: Filler: War of the Worlds/The Guardian,
War of the Worlds is made up of five consecutive parts.,
The Guardian wasn't very interesting.
17: Filler: Dinner/War of the Worlds,
Dinner was moved to 7:00 PM earlier today.,
War of the Worlds only took a few months to be completed.
18: Filler: The Guardian/beer,
The Guardian has around 150 employees.,
The beer is a little dusty from storage.
19: Filler: The Guardian/War of the Worlds,
The Guardian hired a new designer.,
War of the Worlds is an expensive,

signed first edition.
20: Filler: The Guardian/War of the Worlds,
The Guardian contains advertisements.,
War of the Worlds won a range of international awards.
21: Filler: War of the Worlds/beer,
War of the Worlds just wouldn't fit on the new shelves.,
The beer had a hand-drawn label.
22: Filler: The Guardian/War of the Worlds,
The Guardian fired its editor in chief.,
War of the Worlds only took a few months to be completed.
23: Filler: War of the Worlds/record,
War of the Worlds is an expensive,
signed first edition.,
The record contained times and dates.
24: Filler: The Guardian/bottle,
The Guardian fired its editor in chief.,
The bottle had a hand-drawn label.
25: Filler: Dickens/The Guardian,
Dickens is full of satire and caricature.,
The Guardian listed all affected stores.
26: Filler: The Guardian/Dinner,
The Guardian wasn't very interesting.,
Dinner was moved to 7:00 PM earlier today.
27: Filler: The Guardian/War of the Worlds,
The Guardian has around 150 employees.,
War of the Worlds was first published in 1898.
28: Filler: The Guardian/university,
The Guardian was sued for defamation.,
The university was closed during the holidays.
29: Filler: Dickens/milk,
Dickens advocates Children's rights.,
The milk had a red cow on the label.
30: Filler: picture/War of the Worlds,
The picture was propped up on the mantelpiece.,
War of the Worlds was used to weigh down the mail.
31: Filler: Dickens/The Guardian,
Dickens grew up very poor.,
The Guardian hired a new designer.
32: Filler: Dinner/War of the Worlds,
Dinner was exceptionally delicious today.,
War of the Worlds describes an alien attack on Earth.
33: Filler: War of the Worlds/Dinner,
War of the Worlds only took a few months to be completed.,
Dinner was held in a restaurant in London.
34: Filler: Dinner/War of the Worlds,
Dinner was so spicy that it made me cry.,
War of the Worlds is an expensive,
signed first edition.
35: Filler: The Guardian/War of the Worlds,
The Guardian has around 150 employees.,
War of the Worlds is made up of five consecutive parts.
36: Filler: The Guardian/War of the Worlds,
The Guardian listed all affected stores.,
War of the Worlds is bound in black embossed leather.
37: Filler: hatch/Dinner,
The hatch leads to a long tunnel.,
Dinner is great for socialising and networking.
38: Filler: War of the Worlds/bottle,
War of the Worlds was used to weigh down the mail.,
The bottle had a hand-drawn label.
39: Filler: The Guardian/university,
The Guardian struggles financially.,
The university went on a field trip last summer.
40: Filler: War of the Worlds/Dinner,
War of the Worlds was first published in 1898.,
Dinner was exceptionally delicious today.
41: Filler: Dinner/The Guardian,
Dinner was hastily devoured before the meeting.,
The Guardian is rather satirical today.
42: Filler: The Guardian/Dickens,
The Guardian hired a new designer.,
Dickens is about social equality.
43: Filler: university/War of the Worlds,
The university recently got a more modern website.,
War of the Worlds won a range of international awards.
44: Filler: The Guardian/Dickens,
The Guardian contains advertisements.,
Dickens didn't really grip me.
48: Filler: The Guardian/Dickens,
The Guardian struggles financially.,
Dickens didn't really grip me.
46: Filler: beer/The Guardian,
The beer has a rich golden tint.,
The Guardian wasn't very interesting.
47: Filler: The Guardian/War of the Worlds,
The Guardian wasn't very interesting.,
War of the Worlds was adapted as a movie multiple times.
48: Filler: The Guardian/Dickens,
The Guardian fired its editor in chief.,
Dickens advocates Children's rights.
49: Filler: milk/Dinner,
The milk tastes a little bitter today.,
Dinner got cold while we waited for someone.
50: Filler: bottle/War of the Worlds,
The bottle fell off the kitchen counter.,
War of the Worlds is an expensive,
signed first edition.
51: Filler: The Guardian/War of the Worlds,
The Guardian contains advertisements.,
War of the Worlds describes an alien attack on Earth.
52: Filler: The Guardian/beer,
The Guardian contains advertisements.,
The beer lay in a padded wooden box.
53: Filler: War of the Worlds/bottle,
War of the Worlds gathers dust on the top shelf.,
The bottle has a modern screw-on cap.
54: Filler: War of the Worlds/The Guardian,
War of the Worlds describes an alien attack on Earth.,
The Guardian listed all affected stores.
55: Filler: War of the Worlds/The Guardian,
War of the Worlds is bound in black embossed leather.,
The Guardian struggles financially.
56: Filler: picture/The Guardian,
The picture stands on the living room table.,
The Guardian was sued for defamation.
57: Filler: War of the Worlds/Dinner,
War of the Worlds is an expensive,
signed first edition.,
Dinner was so spicy that it made me cry.
58: Filler: War of the Worlds/picture,
War of the Worlds is made up of five consecutive parts.,
The picture was glued into a photo album.
59: Filler: Dinner/bottle,
Dinner was moved to 7:00 PM earlier today.,
The bottle lay in a padded wooden box.

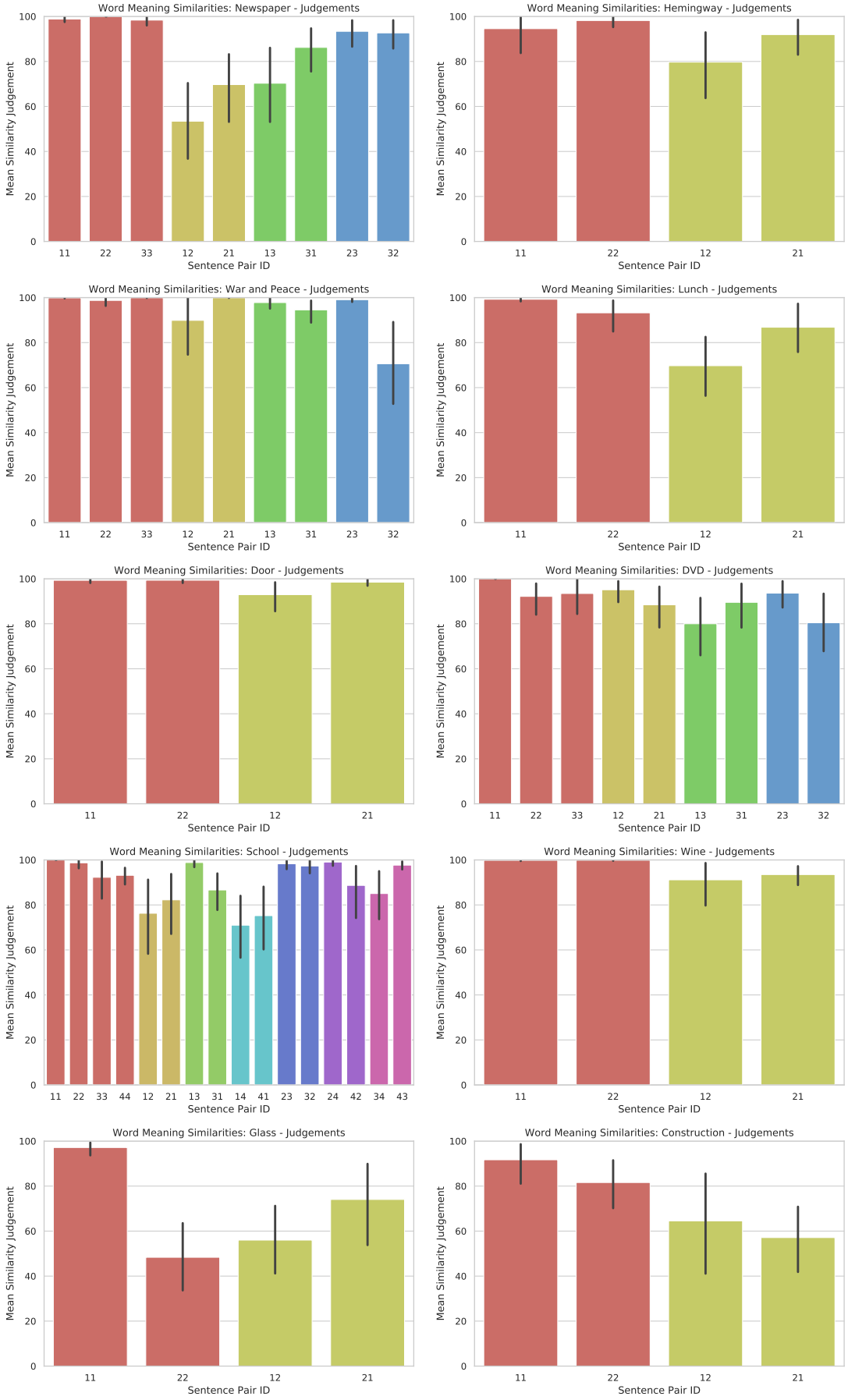


Figure 8: Word sense similarity judgements for the ten tested types of regular polysemy.

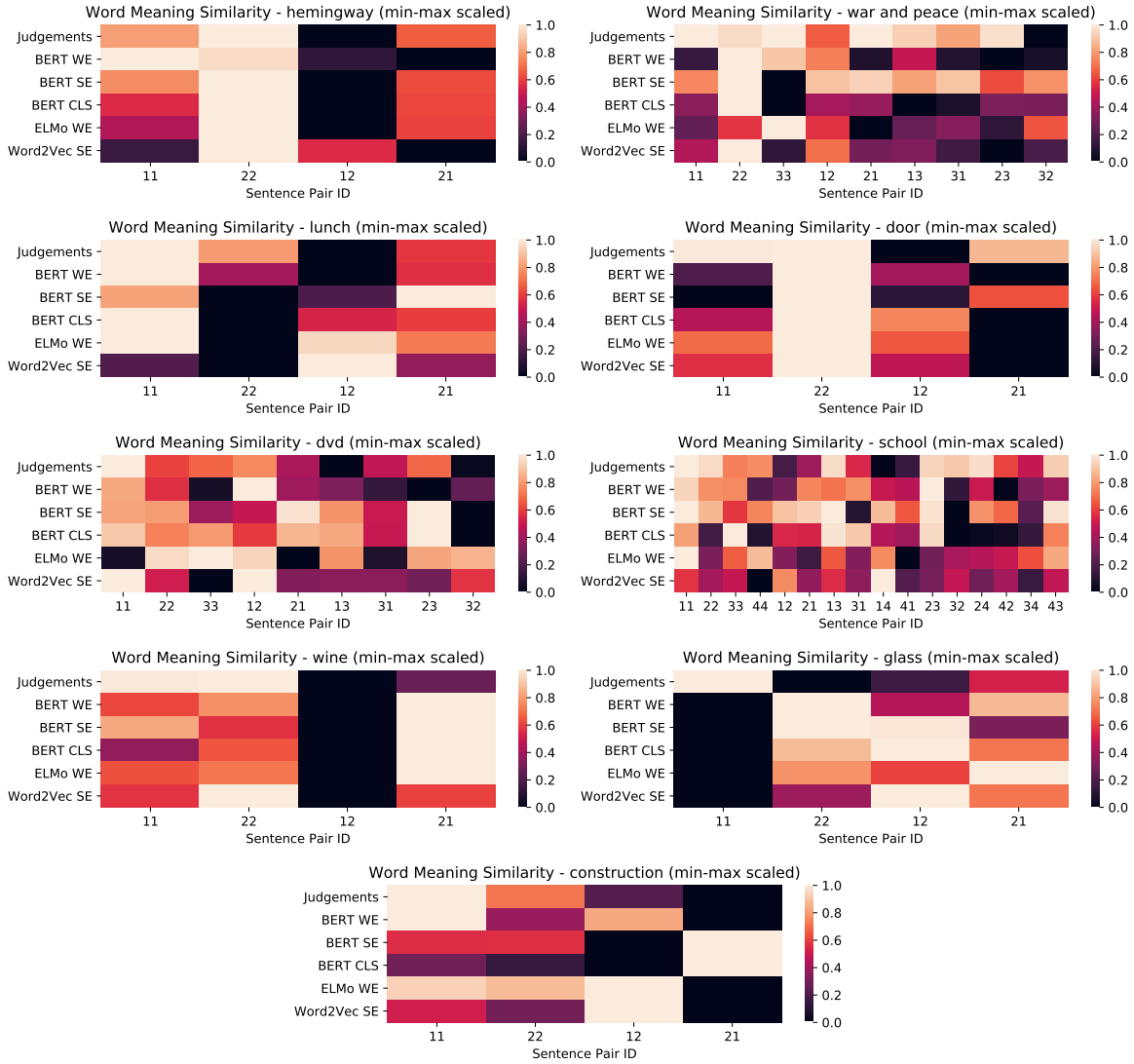


Figure 9: Correlations between human word sense similarity judgements and ELMo and BERT embeddings for the regular polysemes tested in this study.

Short-term Semantic Shifts and their Relation to Frequency Change

Anna Marakasova
Faculty of Informatics
TU Wien
Vienna, Austria
anna.marakasova@tuwien.ac.at

Julia Neidhardt
Faculty of Informatics
TU Wien
Vienna, Austria
neidhardt@ec.tuwien.ac.at

Abstract

We present ongoing research on the relationship between short-term semantic shifts and frequency change patterns by examining the case of the refugee crisis in Austria from 2015 to 2016. Our experiments are carried out on a diachronic corpus of Austrian German, namely a corpus of newspaper articles. We trace the evolution of the usage of words that represent concepts in the context of the refugee crisis by analyzing cosine similarities of word vectors over time as well as similarities based on the words' nearest neighbourhood sets. In order to investigate how exactly the contextual meanings have changed, we measure cosine similarity between the following pairs of words: words describing the refugee crisis, on the one hand, and words indicating the process of mediatization and politicization of the refugee crisis in Austria proposed by a domain expert, on the other hand. We evaluate our approach against expert knowledge. The paper presents the current findings and outlines the directions of the future work.

1 Introduction

Words are prone to change their meaning over time. Automatic detection and analysis of such change is beneficial for a number of theoretical and computational linguistics tasks, as well as for other research areas, i.e., digital humanities, social and political science. Semantic change occurs due to various linguistic and extra-linguistic factors. Although it might be rather challenging to distinguish between these two causes of change, the latter offer an opportunity to gain insights into temporal dynamics of social, cultural, political phenomena reflected in texts.

The discussion about the relationship between semantic evolution of words and their frequencies has been already introduced in the studies of computational semantic change detection (Kulkarni et al., 2015; Hamilton et al., 2016b; Kahmann et al., 2017; Tahmasebi, 2018; Yao et al., 2018; Del Tredici et al., 2019; Vylomova et al., 2019; Haider and Eger, 2019). Although in linguistics, semantic change was shown to be associated with frequency rise (Feltgen et al., 2017), Bowern (2019) emphasized that a change in frequency can be a precursor to semantic change, but does not capture the change itself. Therefore, when studying semantic shift it is important to consider the effect of frequency change on the observed shift.

We contribute to the field by analysing short-term semantic shifts and their relation to frequency change. Furthermore, we propose an evaluation approach as in the case of short-term change previously proposed strategies cannot be adopted. Unlike large-scale semantic shifts that span across decades or centuries, short-term semantic shifts within monthly or yearly time slices reflect changes in usage rather than in meaning as it is understood in lexicography. Therefore, our work aims to answer the following question: with respect to frequency, what patterns that represent contextual meaning change emerge when dealing with short time slice data? More specifically, does a significant increase in frequency indicate a semantic shift towards broadening or narrowing of usage contexts? And the other way round, if relative frequency does not fluctuate much, can we expect stability of contextual meaning? We address these questions by investigating the case of the refugee crisis in Austria from 2015 to 2016 as covered in the Austrian media. We evaluate our approach relying on the evidence from a domain expert.

The rest of the paper is organized as follows: In Section 2 the related work is discussed. In Section 3 the data, on which our analysis is based, is presented. In Section 4 we introduce our method-

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 146–153

ological approach. Details about our experiments are provided in Section 5 and the results are discussed in Section 6. Finally, Section 7 contains our conclusions.

2 Related Work

Recently, there has been seen an increased interest in computational semantic change detection methods. Hamilton et al. (2016a) proposed to distinguish between two types of semantic change: linguistic drift and cultural shift. According to them, comparison of word's vectors in different time periods (global measure) reflects linguistic drift, while comparison of word's nearest neighbors (local measure) is suitable for identifying cultural shifts. Furthermore, Kutuzov et al. (2017) pointed out that when detecting short-term semantic shifts, the neighbourhood based model outperforms successful alignment methods of long-term semantic change. However, Hamilton et al. (2016a) mentioned that a global measure is sensitive to slight change in word usage, hence we investigate both measures.

With a focus on tracing contextual variability over time, we assume that rising similarity of contextual meaning indicates homogenization of overall word use (narrowing), while a decrease in similarity signifies diversification (broadening) (Haider and Eger, 2019). In order to bring light on the direction of narrowing of a word's usage, we look at the semantic change of pairs of words which was shown to be reliable in different experiment settings (Rosin et al., 2017; Kutuzov et al., 2017; Orlikowski et al., 2018; Dridi et al., 2019; Sommerauer and Fokkens, 2019). More specifically, we study temporal dynamics between the words of interest and their most close semantic associates.

One of the studies of contextual variability that explore correlation with frequency is by Cafagna et al. (2019). They explored the effect of frequency on quantifying synchronic semantics shifts between words in two Italian newspaper corpora and found that contextual variability is larger if a word is relatively more frequent in the second corpus. On the other hand, Kahmann et al. (2017); Del Tredici et al. (2019); Vylomova et al. (2019) showed that significant change in frequency does not necessarily imply semantic context change.

Hamilton et al. (2016b) tackled the task of semantic change detection by applying several algorithms on English, German, French and Chinese data, and conducting evaluation against known diachronic change. According to the law of conformity proposed by them, high frequency words are less likely to undergo semantic change. Later research supported (Tahmasebi, 2018; Cafagna et al., 2019; Rodina et al., 2019) or denied the law by showing that frequency does not correlate with semantic change (Dubossarsky et al., 2016).

Dubossarsky et al. (2017) proved that due to the training procedure semantic change models capture both the change in meaning and noise. They demonstrated that in controlled conditions, the reported meaning change effects largely disappear or become considerably smaller. Therefore, when tracing semantic change there is a risk to over interpret differences in word's meaning representations that actually stem from noise. Frequency difference among words possibly accounts for such noise.

3 Data

The Austrian Media Corpus (AMC) is a diachronic text corpus that contains Austrian newspapers, magazines, press releases, transcribed television interviews, news stories from television, etc. from the last thirty years (Ransmayr et al., 2013). With over 44 million articles, it is one of the largest text corpora for German and the largest for Austrian German. The language data is tokenized, part-of-speech tagged and lemmatized. In total, it contains 10.500 billion tokens.

Although the corpus spans over thirty years, we use the data covering ten years, from 2008 to 2017, and only newspaper articles were taken for our analysis. We split the data into yearly spanned subcorpora, thus obtaining ten corpora for the study with their sizes ranging from 141 to 152 million tokens. Next, we extracted lemmas representing common and proper nouns, verbs, adjectives and applied stop words filtering (numerals, names of months and days of the week). A frequency threshold of 100 minimum counts was applied to each subcorpus.

4 Methods

Distributional semantic methods adopt the hypothesis that meaning of a word is conveyed in its co-occurrence relationships (Harris, 1954; Firth, 1957). The semantic similarity of two words is then approximated by the cosine similarity (the

value ranges from 0 to 1) between their vectors that capture information about its co-occurrence statistics. Recent studies mainly make use of dense word representations, usually prediction-based word embeddings models. However, reliability of word2vec based approaches was considered questionable by some studies (Antoniak and Mimno, 2018; Wendlandt et al., 2018; Sommerauer and Fokkens, 2019; Hellrich, 2019; Dubossarsky et al., 2019). Furthermore, Schlechtweg et al. (2019) evaluated various word's meaning representations and showed that although the best run of the word2vec based model strongly outperforms other methods, its mean performance measured over several runs is comparable with the two count-based representations, namely word vectors from a weighted matrix with positive pointwise mutual information (PPMI) scores and truncated singular value decomposition (SVD) of a PPMI matrix[1]. In the light of the above and since PPMI vectors were previously demonstrated to be less affected by frequency effects (Dubossarsky et al., 2017), we opt for using an average score of 50 runs of the PPMI-based model.

As follows, we construct sparse co-occurrence matrices with the window of seven words for each subcorpus and apply PPMI weighting with the parameters suggested by Levy et al. (2015). A PPMI score with a smoothing parameter α of a word w and its context word c is calculated by the following formula:

$$PPMI_\alpha(w,c) = \max(\log(\frac{P(w,c)}{P(w)P_\alpha(c)}), 0)$$

We employ vector and local neighbourhood similarity measures to get an idea on the overall semantic stability/instability and a time point where the possible change occurs (i.e., allows to discover what and when has changed), while pairwise similarity measure provides information on the direction of context shift (i.e., allows to discover how it has changed). In order to compare to what extent an obtained similarity time series correlates with the frequency time series of a particular word we apply autocorrelation based dissimilarity measure which is the Euclidean distance between simple autocorrelation coefficients of the two given time series (Montero et al., 2014). Autocorrelation distance is preferred to a simple correlation metric due to the fact that autocorrelation implicitly normalizes scores which is helpful when dealing with the time series data of different scale.

4.1 Vector similarity measure

To obtain a vector similarity (VS) score for a target word and two given time periods, we first align the corresponding co-occurrence matrices by intersecting their columns. This allows us to then compute cosine similarities of a word between its representation in two subcorpora. For each word, we compute a time series to trace temporal dynamics of contextual meaning. Similarities between each two subsequent time periods are measured. We interpret similarity values as follows: the lower the similarity, the broader the usage of a word.

4.2 Neighbourhood similarity measures

Nearest neighbourhood measures allow to track the change in a set of words semantically related to a target word based on the idea that two words are similar if they are related to similar words (Jeh and Widom, 2002). These methods give a fine-grained idea of meaning dynamics and were shown to be particularly efficient for change point detection (Shoemark et al., 2019). We employ two kinds of neighbourhood measures:

- Second-order cosine similarity (SOCS). Following Hamilton et al. (2016a); Shoemark et al. (2019); Schlechtweg et al. (2019), we create two vectors whose length is the size of the union of a word's most similar terms in the given time slices, cosine similarities between a word and each term provide scores of the vectors. To build the vectors we take 50 most relevant words in each year. We assume that high cosine similarity between these vectors represents stability of contexts in corresponding time slices.

- Rank discounted cumulative gain similarity (RDCG)[2]. RDCG is a improved version of the normalized discounted cumulative gain measure described in Katerenchuk and Rosenberg (2016). We apply RDCG to the set of 50 most similar words for each target in each year ranking words based on their cosine similarity values. The score of 1 means

[1] Please, note that there is no available test set to define the best model run for our data.

[2] Jaccard similarity index and Kendall's tau coefficient were tested as well.

that two sets are identical, while close to 0 values signify high difference in ranking.

4.3 Pairwise similarity measure

We estimate the degree of relatedness of two words by computing cosine similarities between their vectors in each time period, thus producing a time series. We believe that in the case of short-term semantic change, increasing similarity of contexts in different time periods means a preference of a specific usage.

5 Experiments

We represent the discourse of the refugee crisis by the following list of words identified by a linguistics specialist: *Asyl* 'asylum', *Asylwerber* 'asylum seeker', *Ausländer* 'foreigner', *Flüchtling* 'refugee', *Flüchtlingskrise* 'refugee crisis', *Flüchtlingsstrom* 'refugee flow', *Migrant* 'migrant', *Migration* 'migration', *Schlepper* 'illegal migrants smuggler'[3], *Zuwanderung* 'immigration'. Among other parts of speech, we use only nouns as they are particularly sensitive to cultural shifts in meaning which represent the scope of the current work[4]. We explore semantic development of these concepts in the context of the refugee crisis as covered in the media. To enhance the semantic analysis based on these terms and to evaluate the methods we compile two supplementary lists of words which are described below. We expect our approach to show the rise in similarity between between the "seed" concepts and supplementary terms in the years 2015, 2016, 2017 (the refugee crisis period, plus one subsequent year when the flow of refugees declined but the topic of migrants was still widely discussed in the media).

Our evaluation is based on the statement of an expert in critical discourse analysis that addresses the way societal power relations are established and reinforced through language use (Wodak, 2001). The statement focuses on the shift to the far-right in the Austrian party landscape, mainly the Austrian People's Party (*ÖVP*) and the Freedom Party of Austria (*FPÖ*). During the 2016-2017 election campaign, one of the governing parties (*ÖVP*) adopted extreme-right *FPÖ's* policy towards migration and refugees. According to the expert analysis of the political and media discourse, this shift to the right is reflected in language by referring to the migrants as a threat to the welfare of the nation and instigating fear among citizens. Moreover, it was proposed to build a border fence and set the maximum limit for asylum applicants. This is pointed to by the increased use of the silent concepts such as *Angst* 'fear', *Bedrohung*, 'threat', *Grenze* 'border', *Grenzzaun* 'border fence', *Obergrenze* 'upper limit', *Richtwert* 'guiding number', *Terrorismus* 'terrorism' in the refugee crisis discourse (Rheindorf and Wodak, 2018). These seven terms constitute our first supplementary list.

Surprisingly, there was no outrage following the changed attitude towards the refugees even though some of the *FPÖ* claims are clearly against universal human rights. The expert contends that with the purpose of legitimation of unprecedented policy, certain mobilizing and politicizing concepts, i.e., *Humanität* 'humanitarianism', *Protektionismus* 'protectionism', *Sicherheit* 'security', *Vielvalt* 'diversity' were deployed by politicians and the media (Rheindorf and Wodak, 2018). These words form our second supplementary list.

6 Results

6.1 Frequency correlation

First of all, we explore a degree of correlation between frequency time series of the "seed" words and their related time series of the semantic shift detection methods described in Section 4. Table 1 presents the results. One can clearly notice that the vector similarity measure exhibits high correlation with words' frequency change, while neighbourhood measures show remarkably lower correlation. Moreover, unlike vector similarity measure that gives high frequency correlation for almost all words, results of the neighbourhood measures present a range of patterns (see Table 1).

6.2 Semantic shift analysis

We are particularly interested in the change of the words' self-similarities values in the years 2015, 2016, 2017. Context instability implies that the meaning (or usage) of a word is either broad, or undergoing a process of development. When the similarity between a word's semantic representation of consecutive time slices stays stable and low, that means word's usage is rather broad. The increase of similarity values indicates the process of narrowing of the contexts diversity. This might

[3]The meaning is relevant to the context of the refugee crisis.

[4]See (Hamilton et al., 2016a) for details.

Lemma	VS	SOCS	RDCG
Asyl	0.27	0.62	0.55
Asylwerber	0.24	1.0	0.23
Flüchtlingskrise	0.16	0.83	0.19
Flüchtlingsstorm	0.51	0.49	0.35
Flüchtling	0.36	0.49	0.67
Schlepper	0.41	0.9	1.0
Ausländer	0.48	0.75	0.82
Zuwanderung	0.41	0.87	0.8
Migrant	0.26	0.74	0.93
Migration	0.48	0.75	0.5
mean	**0.36**	**0.69**	**0.72**

Table 1: Autocorrelation distance between frequency time series and semantic similarity time series (VS, SOCS and RDCG) for the "seed" words.

happen, particularly, due to the deliberate usage of only specific contexts of the word's meaning.

Frequency change patterns are different for the selected "seed" terms. Thus, most of the words, namely *Asyl*, *Asylwerber*, *Flüchtling*, *Flüchtlingskrise*, *Flüchtlingsstrom*, *Schlepper* have a noticeable peak in 2015 or 2015 and 2016. Other words (*Zuwanderung*, *Ausländer*) have relatively stable frequency counts over time, or show a steady increase in frequency in the last three years of consideration (*Migrant*, *Migration*). We trace semantic dynamics with relation to these patterns; henceforth we refer to them as "peak", "steady", and "stable" patterns respectively.

6.2.1 Vector similarity measure

Most of the "seed" words exhibit an increased similarity of contexts within the years 2015, 2016, 2017. The most characteristic example is the word *Flüchtlingskrise* itself. It appeared in our semantic representation in 2014, showed rather broad usage until 2016 since when the context narrowed down and stayed stable. One exception is *Ausländer* which probably was not used exclusively in the context of the refugee crisis as one would expect (its frequency also stays constant).

The "peak" frequency pattern has its highest similarity scores in 2016 that stay stable or slightly drop in 2017. In contrast, their frequency significantly falls after 2015 or after 2016 (in the case of *Asylwerber*, *Flüchtlingskrise*). *Flüchtling* *Asylwerber* undergo the same degree of semantic change (their mean variance is 0.017, only *Flüchtlingskrise* has higher, 0.019) and overall are rather stable (have high mean similarity scores, 0.62 and 0.52 respectively). In comparison, *Flüchtling* is 3.5 times more frequent than *Asylwerber* which frequency is comparable with other words.

For the words *Flüchtlingsstrom*, *Schlepper* self-similarity starts rising from 2014, while frequency stays the same as in 2013 (increasing only in 2015), the word *Zuwanderung* holds very similar semantic tendency, but has "stable" frequency pattern. There is no considerable difference in semantic change degree between such terms as *Asyl*, *Zuwanderung*, and *Migration* that represent three different frequency patterns.

Figure 1: Frequency time series and scaled semantic similarity time series for the word *Asyl*.

6.2.2 Neighbourhood similarity measures

Time series produced by different neighbourhood measures positively correlate with each other and show less correlation with vector similarity measure. In general, neighbourhood statistics also detect an increased context similarity of the "seed" concepts, but have certain distinctive features when comparing to the vector similarity measure. First, the mean similarity scores among words of "peak" frequency pattern is rather smaller, especially, in the last years which are relevant to the refugee crisis, and is around 0.6. Second, while *Flüchtlingskrise*, *Asyl*, *Flüchtlingsstrom* show a comparable rise of self-similarity values, *Asylwerber*, *Flüchtling*, *Schlepper* are found to be relatively stable with the degree of change similar to the one of *Ausländer*, *Zuwanderung*, *Migration*. Third, all "stable" and "steady" frequency pattern terms clearly indicate a change of usage in the year 2015 which is followed by narrowing of the con-

text. Furthermore, the neighbourhood measures captures the change in usage before a drastic frequency rise happens which is illustrated in the example of the word *Asyl* (see Figure 1).

6.3 Pairwise similarity measure

Overall, the pairwise similarity measure shows an increased relatedness of the "seed" concepts and supplementary terms during the period of interest (2015-2017). More precisely, similarity with the first supplementary list words mainly rises in the years 2015-2016 (the actual years of the refugee crisis), whereas similarity with the second supplementary list words becomes higher in 2016-2017 (see Figure 2). We find this results rather reasonable since the first list terms serve as an indication of the direction of the semantic shift of the refugee crisis related concepts, while the second list represents the mediatization process of the restricted policy towards refugees which was particularly apparent during the 2017 election campaign.

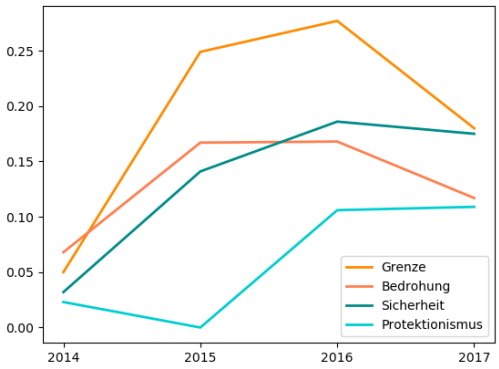


Figure 2: Pairwise similarity for the word *Flüchtlingskrise* and the words from the supplementary lists.

6.4 Evaluation

We match each of the "seed" words with the words from the supplementary lists. In total, there are 110 pairs for evaluation. Next, for every subcorpus we compute cosine similarities between terms that constitute a pair. Recall that according to the expert knowledge, supplementary lists words were widely used in the refugee crisis related discourse in the years 2015 to 2017, i.e., they became more related to the concepts that represent the refugee crisis.

We compare the mean similarity values of the aforementioned period against the mean values of the period of the seven years preceding it (2008-2014). The difference is expected to be positive if our approach goes along with the expert assessment. Indeed, the negative statistically significant difference is only observed in 9% of pairs (10 out of 110). It is rather small (average value for these ten pairs is -0.0087) and found among the pairs with the concepts that do not exhibit specific semantic shift during the refuge crisis (mostly, pairs with the concept *Ausländer*, but also *Zuwanderung-Vielfalt*, *Migrant-Vielfalt*, *Migration-Vielfalt*) which are also related to the discussion of migration in the years 2010-2012.

7 Conclusion

Short-term semantic shift is a complex phenomenon, and understanding the nature of it implies a lot of challenges. Our findings suggest that significant frequency increase is not necessarily followed by significant change in usage, and relatively constant frequencies over time do not imply stability of contextual meaning. We showed that dynamics of word usage is a possible indicator of wider socio-cultural or political shifts.

Since the paper presents an ongoing study, there is a space for enhancement in many aspects. First, more methods should be compared and comprehensive fine-tuning of the models performed with the careful control for randomness and instability that these models feature. Second, one could experiment with the choice of time slices, whether they could be defined empirically or represented in continuous way intersecting one another. Third, candidates for semantic shift could be selected in a robust way and fine-grained annotated data would be beneficial for thorough evaluation and scaling up the experiments for providing more evidence of the phenomenon under discussion. Fourth, control conditions should be introduced in order to ensure verification of the obtained results.

8 Acknowledgments

The research was supported by the project Diachronic Dynamics of Lexical Networks (DYLEN) funded by the ÖAW go!digital Next Generation grant (GDNG 2018-020).

References

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Claire Bowern. 2019. Semantic change and semantic stability: Variation is key. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 48–55, Florence, Italy. Association for Computational Linguistics.

Michele Cafagna, Lorenzo De Mattei, and Malvina Nissim. 2019. Embeddings shifts as proxies for different word use in italian newspapers.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

Amna Dridi, Mohamed Medhat Gaber, R Muhammad Atif Azad, and Jagdev Bhogal. 2019. Leap2trend: A temporal word embedding approach for instant detection of emerging scientific trends. *IEEE Access*, 7:176414–176428.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. Association for Computational Linguistics.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2016. Verbs change more than nouns: a bottom-up computational approach to semantic change. *Lingue e linguaggio*, 15(1):7–28.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1136–1145.

Quentin Feltgen, Benjamin Fagard, and J-P Nadal. 2017. Frequency patterns of semantic change: corpus-based evidence of a near-critical dynamics in language change. *Royal Society open science*, 4(11):170830.

John R Firth. 1957. A synopsis of linguistic theory 1930–1955. instudies in linguistic analysis (pp. 1–32).

Thomas Haider and Steffen Eger. 2019. Semantic change and emerging tropes in a large corpus of new high German poetry. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 216–222, Florence, Italy. Association for Computational Linguistics.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 2116. NIH Public Access.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Johannes Hellrich. 2019. *Word embeddings: reliability & semantic change*, volume 347. IOS Press.

Glen Jeh and Jennifer Widom. 2002. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543.

Christian Kahmann, Andreas Niekler, and Gerhard Heyer. 2017. Detecting and assessing contextual change in diachronic text documents using context volatility. In *KDIR*.

Denys Katerenchuk and Andrew Rosenberg. 2016. Rankdcg: Rank-ordering evaluation measure. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 978-2-9517408-9-1. European Language Resources Association (ELRA).

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2017. Tracing armed conflicts with diachronic word embedding models. In *Proceedings of the Events and Stories in the News Workshop*, pages 31–36, Vancouver, Canada. Association for Computational Linguistics.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Pablo Montero, José A Vilar, et al. 2014. Tsclust: An r package for time series clustering. *Journal of Statistical Software*, 62(1):1–43.

Matthias Orlikowski, Matthias Hartung, and Philipp Cimiano. 2018. Learning diachronic analogies to analyze concept change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–11.

Jutta Ransmayr, Karlheinz Mörth, and Matej Ďurčo. 2013. Linguistic variation in the Austrian Media Corpus: Dealing with the challenges of large amounts of data. In *Proceedings of the 5th International Conference on Corpus Linguistics (CILC 2013)*. Procedia - Social and Behavioral Sciences 95.

Markus Rheindorf and Ruth Wodak. 2018. Borders, fences, and limits—protecting austria from refugees: Metadiscursive negotiation of meaning in the current refugee crisis. *Journal of Immigrant & Refugee Studies*, 16(1-2):15–38.

Julia Rodina, Daria Bakshandaeva, Vadim Fomin, Andrey Kutuzov, Samia Touileb, and Erik Velldal. 2019. Measuring diachronic evolution of evaluative adjectives with word embeddings: the case for english, norwegian, and russian. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 202–209.

Guy D. Rosin, Eytan Adar, and Kira Radinsky. 2017. Learning word relatedness over time. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1168–1178, Copenhagen, Denmark. Association for Computational Linguistics.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 732–746. Association for Computational Linguistics.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76.

Pia Sommerauer and Antske Fokkens. 2019. Conceptual change and distributional semantic models: an exploratory study on pitfalls and possibilities. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 223–233.

Nina Tahmasebi. 2018. A study on word2vec on a historical swedish newspaper corpus. In *DHN*, pages 25–37.

Ekaterina Vylomova, Sean Murphy, and Nicholas Haslam. 2019. Evaluation of semantic change of harm-related concepts in psychology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 29–34.

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102, New Orleans, Louisiana. Association for Computational Linguistics.

Ruth Wodak. 2001. The discourse-historical approach. *Methods of critical discourse analysis*, 1:63–94.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the eleventh acm international conference on web search and data mining*, pages 673–681.

Author Index