Leolani: a reference machine with a theory of mind for social communication

Piek Vossen, Selene Baez, Lenka Bajčetić, and Bram Kraaijeveld

VU University Amsterdam, Computational Lexicology and Terminology Lab, De Boelelaan 1105, 1081HV Amsterdam, The Netherlands
{p.t.j.m.vossen,s.baezsantamaria,l.bajcetic,b.kraaijeveld}@vu.nl
www.cltl.nl

Abstract. Our state of mind is based on experiences and what other people tell us. This may result in conflicting information, uncertainty, and alternative facts. We present a robot that models relativity of knowledge and perception within social interaction following principles of the *theory of mind*. We implemented vision and speech capabilities on a Pepper robot to build an interaction model that stores the interpretations of perceptions and conversations in combination with provenance on its sources. The robot learns directly from what people tell it, possibly in relation to its perception. We demonstrate how the robot's communication is driven by hunger to acquire more knowledge from and on people and the objects, to resolve uncertainties and conflicts, and to share awareness of the perceived environment. Likewise, the robot can make reference to the world but also to knowledge about the world and the encounters with people that yielded this knowledge.

Keywords: robot, theory of mind, social learning, communication

1 Introduction

1.1 Motivation

People make mistakes but machines as well [9], as there is no such thing as a perfect machine. Humans and machines should therefore recognise and communicate their "imperfectness" when they collaborate. This is especially the case for robots that share our physical space. Do these robots perceive the world in the same way as we do and, if not, how does that influence our communication with them? How does a robot perceive us and can a robot trust its own perception? Can it believe and trust what humans claim to see and believe about the world? For example when a medical patient or a child gets injured, should a robot trust their judgment of the situation, should a robot trust its own perception? How serious is the injury, how much knowledge do the humans have and how urgent is the situation? How different would the communication be with a professional doctor? Human-robot communication should serve a purpose, even if it is just (social) chatting, which drives the communication. Yet, effective communication does not only depend on the purpose but also on the communication partners and the degree to which they perceive the same things, have a common understanding and agreement and trust. One of the main challenges to address in human-robot communication is therefore to handle uncertainty and conflicting information.

We address these challenges through an interaction model for a humanoid-robot based on the notion of a 'theory of mind' ([12, 7]). The 'theory of mind' concept states that children, at some stage of their development (up to 48 months), become aware that other people's knowledge, beliefs, and perceptions may be untrue and/or different from theirs. Scassellati ([14] and [13]) was the first to argue that also humanoid robots should have such an awareness. We take his work as our starting point for implementing these principles in a Pepper robot, in order to drive social interaction and communication.

Our implementation of the theory of mind heavily relies on the Grounded Representation and Source Perspective model (GRaSP, [18], [4]). GRaSP is an RDF model to represent situational information about the world in combination with the perspective of the sources of that information. The robot brain not only records the knowledge and information as symbolic interpretations but it also records from whom or through what sensory signal it was obtained. The robot acquires knowledge and information both from the sensory input as well as directly from what people tell it. The conversations can have any topic or purpose but are driven by the robot's need to resolve conflicts and ambiguities, to fill gaps, and to obtain evidence in case of uncertainty.

This paper is structured as follows. Section 2 briefly discusses related work on the theory of mind and social communication. In Section 3, we explain how we make use of the GRaSP model to represent a theory of mind for the robot. Next, Section 4 describes the implementation of the interaction model built on a Pepper robot. Finally, Section 5 outlines the next steps for improvement and explores other possible extensions to our model. We list a few examples of conversations and information gathered by the robot in the Appendix at the end.

2 Related Work

Theory of mind is a cognitive skill to correctly attribute beliefs, goals, and percepts to other people and is assumed to be essential for social interaction and for the development of children [7]. The theory of mind allows the truth properties of a statement to be based on mental states rather than observable stimuli and it is a required system for understanding that others hold beliefs that differ from our own knowledge or from the observable world, for understanding different perceptual perspectives, and for understanding pretense and pretending. Following [1], Scassellati decomposes this skill into stimuli processors that can detect static objects (possibly inanimate), moving objects (possibly animate), and objects with eyes (possibly having a mind) that can gaze or not (eye-contact) and finally a shared-attention mechanism to determine that both look at the same objects in the environment. Scassellati's work further focuses on the implementation of the visual sensory-motor skills for a robot to mimic the basic functions for object, eye-direction and gaze detection. He does not address human communication, nor the storage of the result of the signal processing and communication in a brain that captures a theory of mind. In our work, we rely on other technology to deal with the sensory data processing and add language communication and a 'brain' for storing perceptual and communicated information in the GRaSP model to reflect conflicts, uncertainty, errors, gaps, and beliefs.

More recent work on the 'theory of mind' principle for robotics appears to focus on the view point of the human participant rather than the robot. These studies reflect on the phenomenon of anthropomorphism [10] [3]: the human tendency to project human attributes to nonhuman agents such as robots. Closer to our work comes [5] who use the notion of a theory of mind to deal with human variation in response. The robot runs a simulation analysis to estimate the cause of variable behaviour of humans and likewise adapts the response. However, they do not deal with the representation and preservation of conflicting states in the robot's brain as we do. To the best of our knowledge, we are the first that complement the pioneering work of Scassellati with further components for an explicit model of the theory of mind for robots (see also [8] for a recent overview of the state-of-the-art for human-robot interactive communication).

There is a long-tradition of research on multimodal communication [11], human-computer-interfacing [2] and other component technologies such as face detection [17], facial expression and gesture detection [6]. The same can be said about multimodal dialogue systems [19], and more recently, around chatbot systems using neural networks [15]. In all these studies the assumption is made that models and systems should process signals correctly and that these signals can be trusted (although they can be ambiguous or underspecified). In this paper, we do not address these topics and technologies per se but we assume that they do their work and focus instead on the fact that they can result in conflicting information, information that cannot be trusted or that is incomplete within a framework of the theory of mind. Furthermore, there are few systems that combine natural language communication and perception to combine the result in a coherent model. An example of such work is [16] who describe a system for training a robot arm through a dialogue to perform physical actions, where the "arm" needs to map the abstract instruction to the physical space, detect the configuration of objects in that space, determine the goal of the instructions. Although their system deals with uncertainties of perceived sensor data and the interpretation of the instructions, it does not deal with modelling long-term knowledge in a "brain" but only stores the situational knowledge during training and the capacity to learn the action. As such they do not deal with conflicting information coming from different sources over time or obtained during different sessions. Furthermore, their model is limited to physical actions and the artificial world of a few objects and configurations.

3 GRaSP to model the theory of mind

The main challenges addressed in our model is the storage of the result of perception and communication in a single model and the handling of uncertainty and conflicting information. We addressed these challenges by explicitly representing all information and observations processed by the robot in an artificial brain (a triple store) using the Grounded Representation and Source Perspective model (GRaSP, Fokkens et al. 2017). For modelling the interpretation of the world, GRaSP relies on the Simple Event Model or SEM [**?**]. SEM is an RDF model for representing instances of events through Unique Resource Identifiers or URIs. RDF triples are used to relate event instances with *sem:hasParticipant*, *sem:hasPlace* and *sem:hasTime* relations to participants, places and time also represented as instances through URIs. For example, the triples [eventURI_laugh, sem:hasParticipant, John], [eventURI_laugh, sem:hasTime, 20180512] express that there was a *laugh* event involving *John* on the *12th of May 2018*. The RDF framework allows for expressing further properties of events and the participants, locations and time.

GRaSP extends this model with *grasp:denotedBy* links to express that the instances and relations in SEM have been mentioned in a specific signal, e.g. a camera signal, human speech, written news, etc. So if *Bill* told the robot "John is laughing" this expression is considered as a speech signal that mentions the entity *John* and the event instance *laughing*, while the time of the utterance is given and correlates with the tense the utterance. GRaSP also represents the signal through a URI, e.g. *chatWithBill20180512Turn1*, and each mention as an offset position and length in this signal. Likewise, we can use these offsets to identify the mentioning of entities and events in the signal:

leolaniWorld:instances {

leolaniFriends:Bill	skos:prefLabel	Bill;
leolaniFriends:Mary	skos:prefLabel	Mary;
leolaniFriends:John	skos:prefLabel	John;
	grasp:denotedBy	leolaniTalk:chatWithBill20180512Turn1#char=0,4.
leolaniWorld:eventURI_laugh	a sem:Event;	
	skos:prefLabel	laugh;
	grasp:denotedBy	leolaniTalk:chatWithBill20180512Turn1#char=8,8.

}

GRaSP further allows to express properties of the signal such as the source and the perspective of the source towards the content or claim. In the case of robot interactions, the source of a spoken utterance is the person identified by the robot through face recognition, so in this example the utterance is attributed to *Bill*.[1] Other properties are the instance relation of the utterance to the type *prov:Turn* and the moment in time it was perceived. Finally, we use a specific URI, *chatWithBill20180512Turn1#attr1*, to store properties for the perspective of the source to the claimed content of the utterance: what emotion is expressed, how certain is the source, does the source confirm or deny it. In

[1] For this we follow the PROV-O model: `https://www.w3.org/TR/prov-o/`

this example, we express that *Bill confirms* John's laughing and that he is *uncertain* and *scared*. The perspective subgraph resulting from the conversation would look as follows, where the claim *John's laughing* is represented through the URI *chatWithBill20180512Turn1#claim1*:

grasp:perspectives {

leolaniTalk:chatWithBill20180512Turn1		
	a	prov:Turn;
	sem:hasTime	leolaniTime:20180512T17:56:23;
	prov:wasAttributedTo	leolaniFriends:Bill .
leolaniTalk:chatWithBill20180512Turn1#attr1		
	rdf:value	grasp:uncertain , grasp:scared , grasp:confirm;
	grasp:wasDerivedFrom	leolaniTalk:chatWithBill20180512Turn1;
	prov:isAttributionFor	leolan-iTalk:chatWithBill20180512Turn1#claim1.

}

We then extended GRaSP with a layer of claims that contain the actual SEM relations that represent the claim as the relations for the event:

grasp:claims {

leolaniTalk:chatWithBill20180512Turn1#claim1		
	grasp:subject	leolaniWorld:eventURI_laugh;
	grasp:predicate	sem:hasActor;
	grasp:object	leolaniFriends:John.

}

Now assume that *Mary* is also in the room and she denies that John is laughing by saying: "No, John is not laughing". This utterance then gets a unique identifier e.g. *chatWithMary20180512Turn1*. Our Natural Language Processing will derive exactly the same claim as before from this utterance involving the same relation between the instance *John* and the same event. The new information is the mentioning by *Mary* and her perspective to this claim. We can then express this perspective on *chatWithBill20180512Turn1#claim1* as follows:

leolaniTalk:chatWithMary20180512Turn1		
	a	prov:Turn;
	sem:hasTime	leolaniTime:20180512T17:58:11;
	prov:wasAttributedTo	leolaniFriends:Mary.
leolaniTalk:chatWithMary20180512Turn1#attr1		
	rdf:value	grasp:certain , grasp:denial;
	grasp:wasDerivedFrom	leolan-iTalk:chatWithMary20180512Turn1;
	prov:isAttributionFor	leolan-iTalk:chatWithBill20180512Turn1#claim1.

In addition to representing the new utterance and perspective, we further need to add *grasp:denotedBy* links to the offsets in Mary's utterance indicating the new mentions of *John* and the *laughing*. For the rest, the knowledge graph will not change. Along the same lines, if *Bill* now agrees with *Mary* by saying "Yes, you are right", we model this by adding only another utterance of *Bill* and his revised perspective to the same claim, as shown below.[2]

[2] There are now two perspectives of *Bill* on the same claim (he changed his mind), expressed in two different utterances

leolaniTalk:chatWithBill20180512Turn2		
	a	prov:Turn;
	sem:hasTime	leolaniTime:20180512T17:59:01;
	prov:wasAttributedTo	leolaniFriends:Bill.
leolaniTalk:chatWithBill20180512Turn2#attr2		
	rdf:value	grasp:certain , grasp:denial ;
	grasp:wasDerivedFrom	leolaniTalk:chatWithBill20180512Turn2;
	prov:isAttributionFor	leolan- iTalk:chatWithBill20180512Turn1#claim1.

In the above examples, we only showed information given to the robot through conversation. GRaSP can however deal with any signal and we can therefore also represent sensor perceptions as making reference to the world or people that the robot knows. Assuming that the robot also sees and recognises *John* about whom *Bill* and *Mary* are talking, this can be represented as follows, where we now include all the other mentions from the previous conversations:

leolaniFriends:John	skos:prefLabel	John;
	grasp:denotedBy	leolan- iTalk:chatWithBill20180512Turn1#char=0,4;
	grasp:denotedBy	leolan- iTalk:chatWithMary20180512Turn1#char=4,4;
	grasp:denotedBy	leolaniSen- sor:observation20180512FaceRecognition1.

If a facial expression detection system is added to the model, it is possible to store the result of this sensor data as a perspective on *John* by the robot, e.g. [*leolaniSensor:observation20180512FaceRecognition1, rdf:value, grasp:sad*]. This then reflects agreement in perspective of the robot with *Mary* and *Bil* about John's state of mind.

As all data are represented as RDF triples, we can query all the claims made by people and all the properties stored by the robot on instances of the world. We can also query for all the signals (expressions and sensor data) in which these instances are mentioned and all the perspectives that are expressed. The model further allows to store certainty values for the observations and claims as well as the result of any emotion detection in addition to the content of utterances (e.g. through modules for facial expression detection or voice-emotion detection). Finally, all observations and claims can be combined with other background knowledge on objects, places and people that is available as RDF and linked open data.

The things observed by the robot in the environment and the things mentioned in conversation can thus be stored as unified data in a single storage or "brain". This brain contains the identified people with whom the robot communicates as friends, the perceived objects about which they communicated[3] as well as any property that has been identified or stated of these objects or people. Given this model, we can now design a robot communication model in combination with sensor processing on top of a theory of mind. In the next section, we

[3] The robot continuously detects objects but we only store them in memory when they are referenced by humans in the communication

explain how we this model has been implemented and what kind of conversations can be held.

4 Model Overview

Our conversation model consists of four layers: the Signal Processing layer, the Conversation Flow layer, the Natural Language Processing layer, and the Knowledge Representation layer, which are summarised in Figure 1. Signal Processing (I) establishes the mode of input through which the robot acquires experiences (vision and sound) but also knowledge (communication). The communication layer (II) acts as the controller, as it determines the communicative goals, how to interpret human input, and whether the robot should be proactive. Layer III is the natural language processing layer that processes utterances and generates expressions. Both can be questions or statements. Incoming statements are stored in the brain, while questions are mapped to SPARQL queries to the brain. SPARQL queries to the brain can also be initiated by the controller (layer II) on the basis of sensor data (e.g. recognising a face or not) or the state of the brain (e.g. uncertainty, conflicts, gaps) without a human asking for it. The next subsections will briefly describe the four layers individually. We illustrate the functions through example dialogues that are listed in the Appendix. Our robot has a name *Leolani*, which is Hawaiain for *voice of an angel*, and a female gender to make the conversations more natural. No further character traits are implemented for the robot.

4.1 Signal Processing

Signal processing is used to give the robot awareness of its (social) surroundings and to enable conversation. Face Recognition is used to identify the identity of the person

Includes the following modules:

1. Context
   1.1. Face detection
   1.2. Eye contact detection
   1.3. Speech detection
   1.4. Object recognition
2. Addressee
   2.1. Face recognition
   2.2. Gender recognition
   2.3. Speech recognition
   2.4. Name recognition
   2.5. Friend recognition

Items related to the contexts run continuously as Leolani tries to learn and recognize her surroundings. Items related to an addressee are triggered when she is in a conversation.

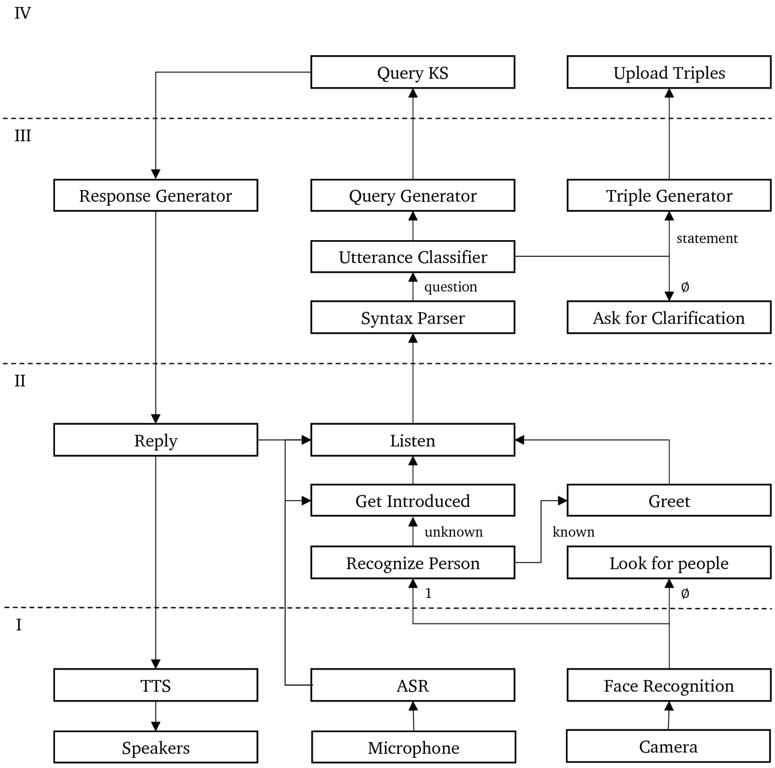

Fig. 1: The four layer conversation model, comprised of I. Sensory layer, II. Conversation Flow layer, III. Natural Language layer and IV. Knowledge Representation layer.

4.2 Conversation Flow

In order to guide and respond during a one-to-one conversation, the robot needs to reason over its knowledge (about itself, the addressee, and the world) while taking into account its goals for the interaction. As previously mentioned, the robot is currently designed to be hungry for social knowledge. This includes intentions like asking for social personal information (name, profession, interests, etc.), or asking for knowledge to resolve uncertainties and conflicts.

4.3 Natural Language Processing

Includes the following modules:

1. Initiate conversation (Greetings and Introduction)
2. Syntax parser
3. Name entity recognition [4]
4. Utterance classifier

[4] NOTE: TODO, should include?

5. Sentiment analysis [5]
6. Answer generation
7. End conversation (Goodbye)

4.4 Knowledge Representation

Knowledge representation refers to the field of Artificial Intelligence that is concerned with how knowledge can be represented symbolically and thus, enabling machines *reason* over it and *manipulate* it in an automated fashion.

In order to represent the knowledge acquired by the robot during conversations, we represent the factual statements with the use of the Resource Description Framework (RDF), a data model used to describe resources on the web in the form of triples (subject-predicate-object).
Since we work with a machine, using a framework that adds semantics to otherwise plain strings or text, adds a lot of value and enables us to implement the theory of mind model. We use Unique Resource Identifiers (URI's) to represent the statements that are produced by the Natural Language Processing model, described in section 4.3. However, this is not the only thing we are able to represent.

With the use of ontologies, many different domains and concepts can be represented, understood and reasoned over. Ontologies define domain concepts and relations between them, thus enabling new knowledge to be generated via a mechanism known as inferencing.

Within the robot's brain we can see the entanglement of different domain knowledge, such as social (*Persons*), spatial (*Places*), and conversational (*Statements* and *Opinions*). Thus, the use of ontologies becomes imperative in modeling the world from the robots perspective.

Semantic Web technologies What technologies are used ?

Includes the following modules:

1. Statement processing
   1.1. Create triples and update brain
2. Question processing
   2.1. Factual knowledge (Wolfram)
   2.2. Build SPARQL query and query brain

5 Conclusions and Future Work

We described a model and implementation for a robot to perceive the world and communication about the world and to store the result in an RDF repository that models a theory of mind. This allows us to deal with uncertainties,

[5] NOTE: Need to discuss

conflicting information, gaps, and different perspectives on the source of the information. The state of the information and knowledge that the robot obtains can drive the communication, e.g. to resolve conflicts or uncertainties. Currently, the communication scripts and the type of knowledge are limited. For the future, we plan to make our implementation more modular so that we can include different conversational models, NLP modules and types of knowledge resources more easily. Furthermore, we want to extend the platform so that we can empirically evaluate the performance of the system. Finally, we want to make the system functional through task-based dialogues and support the processing of the sensors and the utterance knowledge and situation-aware.

6 Acknowledgement

This research was funded by the VU University of Amsterdam and by the Netherlands Organization for Scientific Research (NWO) via the Spinoza grant awarded to Piek Vossen in the project *Understanding Language by Machines*. We also thank Bob van der Graft for supporting this research.

References

1. Baron-Cohen, S.: Mindblindness: An essay on autism and theory of mind. MIT press (1997)
2. Card, S.K.: The psychology of human-computer interaction. CRC Press (2017)
3. Epley, N., Waytz, A., Cacioppo, J.T.: On seeing human: a three-factor theory of anthropomorphism. Psychological review **114**(4), 864 (2007)
4. Fokkens, A., Vossen, P., Rospocher, M., Hoekstra, R., van Hage, W.: Grasp: Grounded representation and source perspective. In: Proceedings of KnowRSH, RANLP-2017 workshop, Varna, Bulgaria (2017)
5. Hiatt, L.M., Harrison, A.M., Trafton, J.G.: Accommodating human variability in human-robot teams through theory of mind. In: IJCAI Proceedings-International Joint Conference on Artificial Intelligence. vol. 22, p. 2066 (2011)
6. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on. pp. 46–53. IEEE (2000)
7. Leslie, A.: Pretense and representation: The origins of "theory of mind.". Psychological review (4) (1987)
8. Mavridis, N.: A review of verbal and non-verbal human–robot interactive communication. Robotics and Autonomous Systems **63**, 22–35 (2015)
9. Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., Tscheligi, M.: To err is robot: How humans assess and act toward an erroneous social robot. Frontiers in Robotics and AI **4**, 21 (2017)
10. Ono, T., Imai, M., Nakatsu, R.: Reading a robot's mind: A model of utterance understanding based on the theory of mind mechanism. Advanced Robotics **14**(4), 311–326 (2000)
11. Partan, S.R., Marler, P.: Issues in the classification of multimodal communication signals. The American Naturalist **166**(2), 231–245 (2005)

12. Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences **4**, 515–526 (1978)
13. Scassellati, B.: Theory of mind for a humanoid robot. Autonomous Robots **12**(1), 13–24 (2002)
14. Scassellati, B.M.: Foundations for a Theory of Mind for a Humanoid Robot. Ph.D. thesis, Massachusetts Institute of Technology (2001)
15. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A.C., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: AAAI. vol. 16, pp. 3776–3784 (2016)
16. She, L., Chai, J.: Interactive learning of grounded verb semantics towards human-robot communication. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1634–1644 (2017)
17. Viola, P., Jones, M.J.: Robust real-time face detection. International journal of computer vision **57**(2), 137–154 (2004)
18. Vossen, P., Agerri, R., Aldabe, I., Cybulska, A., van Erp, M., Fokkens, A., Laparra, E., Minard, A.L., Aprosio, A.P., Riga, G.: Newsreader: using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. Knowledge-Based Systems (2016), `http://www.sciencedirect.com/science/article/pii/S0950705116302271`
19. Wahlster, W.: SmartKom: foundations of multimodal dialogue systems, vol. 12. Springer (2006)

Appendix: Dialogues

In the dialogues, *L* preceding an utterance stands for Leolani, other letters preceding utterances stand for various people. Perceptions of the robot of people and objects are marked using square brackets, e.g. [Sees a new face].

Meeting a new person	
L:	[Sees the face of an unknown woman]
L:	Hi there, I would like to know you.
L:	My name is Leolani.
L:	What is your name?
H:	My name is Selene.
L:	I hope I am correct and your name is: Selene.
L:	Nice to meet you Selene. Now I have a new friend.
L:	Where are you from?
H:	I am from Mexico.
L:	Now I know 1 person from Mexico.

Handling conflicting information	
L:	Hi Lenka, nice to see you, where are you from?
H:	I am from Serbia.
L:	Nice, I did not know anybody from Serbia yet.
H:	Where is Bram from?
L:	Bram is from the Netherlands.
H:	Do you know that Bram likes romantic movies.
L:	Oh, I did not know that.
L:	[Sees Lenka leave and Bram arrive.]
L:	Hi Bram.
B:	I like science fiction movies.
L:	I am surprised.
L:	Bram likes romantic movies, says Lenka.
L:	Bram likes science fiction movies, says Bram.
L:	humans are confusing.

Handling conflicting information	
L:	Greetings Bram. Nice to see you again
B:	Do you know where I am from?
L:	You are from the Netherlands, you said.
B:	Do you also know Lenka?
L:	Yes I know her, she is a very good friend of mine.
B:	Where is she from?
L:	Lenka is from Serbia, Lenka said
B:	And do you believe Lenka?
L:	I believe her.

Observing the environment	
L:	[Sees a rabbit but detects it as a cat]
L:	Greetings Bram. Nice to see you again?
L:	Guess what I just saw a cat!
B:	That is not a cat but a rabbit.
L:	[Now labels the image as a rabbit and stores it in her brain]
B:	I like rabbits.
L:	[Sees Bram leave and Selene arrive]
L:	Hi Selene. Greetings.
L:	Guess what I just met a rabbit.
S:	Rabbits bite.
S:	I like cats.
S:	Have you ever seen a cat.?
L:	No I have never seen a cat.
S:	What animals did you see?
L:	I saw a rabbit and a panda.
S:	What do rabbit do?
L:	Rabbits bite, Selene said.
S:	Who likes rabbits?
L:	Bram likes rabbits, Bram said.

Asking for a dump of her brain	
S:	How many things do you know??
S:	Who have you talked to?
S:	How many friends are from Amsterdam?
S:	Tell me everything I told you.