



Classe di Lettere

Corso di perfezionamento in

Filosofia

XXXII ciclo

**“Multimodal Event Knowledge.
Psycholinguistic and Computational
Experiments”**

Settore Scientifico Disciplinare **L-LIN/01**

Candidato

Dr. Valentina Benedettini

Relatori

Prof. Pier Marco Bertinetto

Prof. Alessandro Lenci

Anno accademico 2020/2021

T. G.

*“Non è vero che uno più uno fa sempre due,
una goccia più una goccia fa una goccia più grande.”*

Index

Introduction	6
Chapter 1: What a Verb Means	9
1.1. Actions and Participants	9
1.2. Verbs in Composition	11
1.3. Basic and Complex Actions.....	14
1.4. Perceptually Underspecified Nouns.....	16
1.5. Grounded Nouns and Verbs.....	19
1.6. Events in Cognition	22
1.7. Expectations.....	25
Chapter 2: Learn and Comprehend Events.....	30
2.1. Human and Machine Learning	30
2.2. Sentence Comprehension.....	36
2.3. Multimodal Thematic Fit	41
2.4. Computational Approaches.....	43
2.5. Visual World Paradigm	48
2.6. Anticipatory Eye Movements	54
Chapter 3: Which Object do You Expect?	58
3.1. Sentences, Pictures and Lists	59
3.2. Norming Study.....	62
3.3. Method.....	63
3.4. Results.....	69
3.4.1 Agent Time Window	70
3.4.2 Action Time Window	71
3.4.3 Patient Time Window	73
3.4.4 The Final Silent Interval	74

3.5. General Discussion	76
Chapter 4: Multimodal Event Knowledge Model	80
4.1. Model Design.....	84
4.2. Simulations and Architectures	90
4.3. Processing Dynamics and Training	94
4.4. Evaluation	101
4.5. Results.....	107
4.1.1 Event	109
4.1.2 Agent.....	111
4.1.3 Agent and Verb	113
4.1.4 Perceptually Underspecified Noun	116
4.6. General Discussion	119
Discussion and Conclusion.....	122
Theoretical Implications	124
Future Research Directions	124
References	125
Appendix	139
Eye Tracking Experiment	139
Auditory Sentences.....	139
Results	143
Analyses	164
MEK Model	183
Sequence Collections.....	183
Results	206
Analyses	256
Visual Representations	258
MEK	260
Figure Index.....	285

Table Index	286
Script Index	288

Abbreviation	Meaning	Page
AOI	Area Of Interest	68
API	Application Programming Interface	99
BLSTM	Bidirectional Long Short-Term Memory	99
CNN	Convolutional Neural Network	98
DS	Distributional Semantics	43
DSM	Distributional Semantic Model	43
ECU	Expectation Composition and Update	46
EEG	Electroencephalography	37
ERP	Event Related-Potential	37
EST	Event Segmentation Theory	26
GG	GloVe-GoogLeNet	99
LME	Linear Mixed Effects	68
LookAT	Look At That	90
LSTM	Long Short-Term Memory	90
MDM	Multimodal Distributional Model	47
MEK	Multimodal Event Knowledge	82
MUC	Memory Unification Control	37
RNN	Recurrent Neural Network	90
SVS	Structured Vector Space	45
WA	Word2Vec-AlexNet	99
WhoAct	Who did the Action	92

Introduction

A verb can denote many events. Participants define the involved actions. In this thesis, we investigated the denotational meaning of the verb from a compositional and multimodal perspective. Verb meaning is composed by the meaning of the nouns filling its thematic roles, which can be seen as lexical and conceptual constraints that reflect knowledge about specific events and entities that participate in them (McRae et al. 1998). We accounted for lexical knowledge as a web of mutual expectations linked to typical situations of the real world. People make the experience of real-world situations in the first person, by reading or listening about them or watching the television. Human semantic knowledge includes many kinds of information. Most of the times, linguistic information alone cannot lead to disambiguation of the verb meaning in terms of denoted actions. The difference between throwing a baseball ball or a football ball implies information such as the physical properties of the two types of ball, how and where playing football or baseball. Without further context, at least one of the participants has to be perceptually specified so as to individuate the actions that constitute the events. A perceptually specified noun is a lexical item that, in isolation, entails a specific type of perceptual referent, namely it cues fine-grained knowledge about the situations in which typically it appears, the events in which it is usually engaged and the entities with which commonly it interacts. Agents like *shortstop* and *quarterback* or patients like *baseball ball* and *football ball* allow disambiguating the actions denoted by the verb *throw*. In linguistic descriptions of events (sentences) often perceptually underspecified nouns such as *ball* occur because of other elements that make further specification redundant (Grice 1975): *The quarterback throws the football ball* vs *The quarterback throws the ball*. According to Altman and Mirković (2009), language comprehension includes a mapping between the unfolding sentence and the representation of the event in memory that corresponds to the real-world event. People exploit the knowledge about typical events and situations to comprehend sentences and thematic fit plays a crucial role (Zwaan and Radvansky 1998, Radvansky and Zacks 2014, Bicknell et al. 2010). Words indeed encode expectations linked to the knowledge of typical events (Ferretti et al. 2001, McRae et al. 2005, Hare et al. 2009). Thus, *quarterback* elicits the activation of information about football ball, football helmet, football field, referee as much as *baseball ball* encodes

expectations about baseball players, baseball bat, baseball uniform and so forth. According to Elman (2014), words can be referred to as cues to the knowledge of typical events. The lexical representation of the verb should include specific information for each its possible sense concerning the properties of the most probable fillers of its thematic roles (Elman 2011; Jackendoff 2002).

This thesis includes a psycholinguistic experiment performed using the eye-tracking technique and its computational simulation through an artificial neural network model. The results of the eye-tracking experiment provided empirical evidence that words are cues to the multimodal (lexical and visual) knowledge of typical events and situations. Agents like *quarterback* or *shortstop* encode expectations about the referent filling the patient role (football ball and baseball ball). Sentence comprehension involves the interplay between linguistic and visual information. The collection of physical properties that identify the referent that plays the patient role are crucial information for the disambiguation of verb meaning. The outcomes of the eye-tracking experiment report that people tend to look at the picture football ball hearing the verb of the sentence *The quarterback throws the ball* rather than baseball ball, which was instead the most looked at picture when the participants heard the sentence *The shortstop throws the ball*. Therefore, the thematic fit between the agent (*quarterback, shortstop*) and the perceptual referent of the patient role (football ball and baseball ball) is multimodal because it implies both lexical and visual information. It is guided by the knowledge of typical situations cued by the agents that is incrementally integrated with the verb selectional restrictions. Agent-verb pairs cued the information that allowed the participants to anticipate the referent denoted by the perceptually underspecified patient (*ball*).

We proposed a Long Short-Term Memory (LSTM) neural network model: Multimodal Event Knowledge (MEK). MEK is a model of the multimodal knowledge about typical events cued by words. It predicts the picture of the referent denoted by the patient of a textual event (agent, verb, patient): *quarterback throw ball-FOOTBALL BALL*. We simulated different scenarios by which people experience typical real-world situations to train MEK on the knowledge of typical events. The task of the model aims at simulating the eye-tracking experiment. To evaluate the model, we supplied the inputs that reflected the time windows analyzed in the psycholinguistic experiment. Given an agent in isolation like *quarterback*, the model reproduces the multimodal thematic fit inferring all

the objects typically linked to the input: football helmet and football ball. When the input is the agent-verb pair *quarterback-throw*, MEK infers that the most plausible referent of the patient role is a football ball. When the input is the noun *ball* in isolation, MEK predictions are affected by the kind of scenario we exploited to train it on typical events knowledge. When MEK was trained using LookAT (Look At That) collection of sequences, it infers a single picture for each word. When the model was trained exploiting WhoAct (Who Did the Action?) collection of sequences, it predicts all the possible referents of the input word: baseball ball, football ball and soccer ball. We interpret the results as evidence that the model learnt the relationship (“it is a type of”) between a perceptually underspecified noun (hypernym) and its referents.

In the first chapter, we illustrate the linguistic and psycholinguistic theoretical basis of our experiments about verb denotational meaning. Each theory was presented together with the relevant experimental evidence.

The second chapter includes a comparison between human and machine learning, followed by an overview of the sentence comprehension task and a description of the multimodal thematic fit. The review of the computational studies focused on representing the verb meaning, thematic fit and multimodal merging precedes the presentation of the visual world paradigm and the review of the eye-tracking studies related to the psycholinguistic experiment proposed in this thesis.

The third chapter describes the eye-tracking experiment. We report the results and the analyses that involved the time windows of the agent, verb (anticipatory time window), and patient. Besides, we analyzed the time window followed the listening of the perceptually underspecified patient.

The fourth chapter concerns the MEK computational model. After a description of the model design and architecture, we describe the evaluations and we discuss their results.

Chapter 1: What a Verb Means

1.1. Actions and Participants

The verb represents a linguistic tool to express properties of and to create relations between entities of the real world. Verbs describe the set of possible actions and make the real world a dynamic place. They denote events. Since a single verb can express multiple meanings, verbs are polysemous words. In isolation their meaning is incomplete. The actions that a verb can denote depends on the involved entities.

The meaning of the verb *put on* changes based on the object.

(1)

- a. *The person puts on a cap*
- b. *The girl puts on a glove*
- c. *The man puts on glasses*
- d. *He puts on a boot*

Sentence (1a) suggests the actions of grabbing the cap, raising the arms and laying the cap on the head. Sentences (1b), (1c) and (1d) imply different actions such as inserting the hand in the glove, fitting the temples of the glasses behind the ears, raising the leg, inserting the foot in the boot. The knowledge of entities is crucial in order to disambiguate the meaning of verbs in terms of the type of action.

(2)

- a. *The boy fills up a backpack*
- b. *The woman fills up a pot*

Sentence (2a) suggests that the boy grabs things such as books, rulers, pencils or waterproof jackets, swiss army knives, water bottles and tucks them in the backpack. Sentence (2b) implies that woman pours something into the pot like water, vegetables, meat or soil. Since backpack and pot are different objects, the boy and the woman execute different actions and the denotational meaning of the verb *fill up* changes.

The person who executes the action can be denoted by a word like *student*, *hiker*, *boxer*, *catcher* and so forth.

(3)

- a. *The student fills up the backpack*
- b. *The hiker fills up the backpack*

In (3a) *student* is associated with entities that typically appear in the same situations such as books, rulers or pencils, places like school, classroom, library and people like teachers. *Hiker* in (3b) evokes objects like waterproof jackets, swiss army knives, water bottles, mountains, wood, tents, sleeping bags. The information the verb, *fill up*, recalls are integrated with the information cued by the agents. The knowledge of entities that can be filled is combined with the knowledge associated with *student* and *hiker*. The resultant integration guides the comprehension of the incoming word and the definition of the corresponding referents: *backpack* is likely to refer to a school backpack in (3a) and a hiking backpack in (3b). Since the event in (3a) includes different participants with respect to the event in (3b), even if the verb is the same, the sentences describe distinct events.

(4)

- a. *The boxer puts on the glove*
- b. *The catcher puts on the glove*

The referent of *glove* in (4a) is a boxing glove, in (4b) is a baseball glove. *Put on the glove* occurs both in (4a) and (4b), however, referents change according to *boxer* and *catcher*. In (1b) the word *glove* cannot be associated with particular types of referents. *Glove* denotes the set of all possible types of gloves such as golf gloves, ski gloves, latex gloves, oven gloves etc. In (4a) and (4b) the same word denotes two kinds of glove: the gloves used to box and the gloves used to play baseball. (1b) requires that the extralinguistic context provides the missing information to disambiguate the referents of the sentence. *Boxer* and *catcher* trigger instead enough information to link the event to particular situations and to define the involved entities. The comprehension of (3) and (4) include expectations about typical situations and events. The knowledge of typical situations is an important requisite for the disambiguation of the denotational meaning of the verb.

1.2. Verbs in Composition

The types of action cued by a verb depend on the involved entities. Since the meaning of a verb in isolation is incomplete, the lexical representation of a verb includes the set of its possible patterns of arguments. The verb can be classified based on the number of syntactic arguments needed to complete its meaning. However, lexical and semantics knowledge about verbs cannot be exhaustively explained by syntactic structures. They cannot account for the semantic relations between the verb and the referents of its arguments.

(5)

- a. *The student reads the book in the library*
- b. *The boxer reads the scientific paper on the bus*
- c. *The apple reads the book on the bus*

(5a) describes not only a semantically possible event but also typical. Students are people that are able to read, they have to read books to study and usually they do that in libraries. (5b) is still a semantically plausible event because *boxer* is a person and he should be able to read. However, the described event is not typical because boxers usually train themselves in the gym to punch and participate in box matches. Reading scientific papers is an action typically associated with researchers, students, teachers and professors rather than boxers. (5c) is instead a semantically impossible event. An apple is an object and it cannot have the capacity of reading something. *Apple* cannot be part of the meaning of *read* since it does not respect the semantic constraints imposed by the verb. In (5) the same verb appears with the same syntactic pattern and the same number of arguments, but the sentences describe three distinct events. The syntactic structure alone cannot account for which event can be considered semantically acceptable. Since the number of arguments alone cannot tell a lot about the meaning of the verb, we discuss the relation between it and its arguments and we account for the interplay between their meanings during the sentence composition.

The verb imposes certain constraints on the fillers of its arguments. The restrictions concern the semantic properties that referents should have in order to participate in the event the verb describes. The semantic restrictions the verb imposes on the fillers of its arguments are called selectional restrictions (Chomsky 1965). The referents can be

described also according to the role they play in the event. Thematic roles like agent, patient and instrument help to clarify the lexical representation of the verb (Dowty 1991; Chomsky 1981). The knowledge of thematic roles derives from everyday experiences, during which people learn about the entities that tend to play certain roles in certain events. The collection of properties that identifies the filler of a thematic role results from the experience of a particular event and those similar to it. Therefore, the lexical representation of the verb includes specific information for each possible sense of the verb concerning the characteristics of the most probable fillers of its thematic roles (Elman 2011, 2014; Jackendoff 2002). The meaning of the verb is composed by the meaning of the nouns filling its thematic roles, which can be seen as lexical and conceptual constraints that reflect knowledge about specific events and entities that participate in them (McRae et al. 1998).

Polysemous words like verbs are an example of how the meaning can be affected by the context. When words are composed to build sentences, their meanings affect each other to create new meanings that say something more than their simple combination. Sentences are the result of the integration of certain aspects of the contents of words. In this thesis, we focus on the composition processes of the verb with its arguments. In particular, we investigate how the meanings of agents, verbs and patients affect each other during the comprehension of sentences. The denotational meaning of the verb can be individuated based on typical participants in the events that it describes. Agents and patients are crucial information to define the verb meanings. The notion “multimodal event knowledge” refers to the collection of linguistic and extralinguistic information the verb meaning entails. In particular, we focus on lexical and visual perceptual information of its thematic role fillers.

Patient. Among verbs of activity, we can find *fill up*, *put on*, *open* and *loosen*. They give expression to a collection of possible interactions between entities of the real world. They denote actions and, as most verbs of activity, they are polysemous words. The same verb can denote pragmatically different ways of acting. The actions the verb can denote depend on the semantic properties of the involved objects.

Open in isolation does not provide enough information to define particular movements. Specific actions can be defined once that the object is known.

(6)

- a. *Open a beer bottle* ‘uncap’
- b. *Open a toolbox* ‘handle with a lock’
- c. *Open a backpack* ‘move the slider along the rows of teeth’
- d. *Open a cooking pot* ‘lift the lid’

Loosen suggests various actions based on the patient:

(7)

- a. *Loosen the bike helmet:*

‘Manipulate the retention dial behind the head and adjust the strap so that it is snug against the chin’

- b. *Loosen the swimming goggles:*

‘Pull up on the adjustment lever or clip to release it from the hold on the straps with one hand, and use the free hand to pull out or away from the goggles the straps, release or press the clasp back in place once the straps have been altered’

- c. *Loosen the seat belt:*

‘Slide the metal end of the buckle into the latching device, and adjust the lower and shoulder straps across hips and high chest’

The collection of properties that identify the referent filling the patient role are crucial information for the disambiguation of verb denotational meaning. Verbs describes relations that involve entities. Actions are an inherent and constitutive part of the real world and contribute to defining it. Verbs and actions are ways that people use to express their intentions and goals.

Agent. Particular agents like *student, hiker, graduate, captain, boxer, catcher, quarterback* or *shortstop* contribute to defining the actions the verb denotes thanks to the link that they create between the event and the situation in which it appears. Knowledge of the situation has crucial implications on the disambiguation of verb denotational

meaning. Connection between an event and a particular situation allows to individuate the involved referents and actions. Sports represent a typical scenario.

(8)

a. *The quarterback throws the ball*

b. *The shortstop throws the ball*

(8a) and (8b) describe distinct events based on the situations in which they appear, a football match and a baseball match. The knowledge of the situation includes information about typical participants and actions. The latter constitute the denotational meaning of *throw*. *Quarterback* suggests that the referent of *ball* is a football ball. The *shortstop* plays instead with a baseball ball. Football and baseball are sports that require distinct abilities, techniques and movements of the body. Therefore, the way the quarterback throws the football ball can be different from the way the shortstop throws the baseball ball. A sport in which this aspect of the meaning of *throw* is so salient is basketball. A basketball player trains his mind and his body to achieve the best technique that allows him to throw the basketball ball into the basket. The movements involved to achieve this aim are completely different from those of a quarterback or a shortstop. However, they are expressed by the same verb, *throw*.

1.3. Basic and Complex Actions

Actions represent a class of events that involve intentions and goals. For this reason, many philosophers have focused on actions performed by people. A concept can be seen as an ability or competence to produce detailed representations of the components of the experience like agents, actions, objects and properties that support goal pursuit in the current setting (Barsalou 2008). According to Searle (2019), the structure of the concept of action consists of causal relations, “by means of”, and constitutive relations, “by way of”. They account for the hypothesis that usually people perform an action to do something else, such as raising the arms to put on the graduation cap or to throw a football ball, and they do not do the action “immediately”.

The verb *loosen* in (7a) describes an event that happens by way of or by means of the acts of manipulating the retention dial and adjusting the strap. (6d) describes the event of

opening that corresponds to grasping the cover knob and lifting the lid. The latter can be interpreted as the actions that cause the opening or as the actions that are constitutive of the event of opening a cooking pot. The constitutive relation fits the link between the predicate and the referent. In (7a), the action denoted by the verb *loosen* coincides with the operations of manipulating the retention dial behind the head and adjusting the strap. The act of opening in (6d) corresponds to grasping the cover knob and lifting a lid. Causal relations are linked to the result of the event: a loosened bike helmet is obtained by means of manipulating the retention dial behind the head and adjusting the strap. In the same way, the person should grasp the cover knob and lift the lid to open the cooking pot. Both a bike helmet and a cooking pot to be loosened and opened require actions that depend on their physical properties. Loosening a bike helmet, opening a cooking pot and throwing a ball are complex actions. Complex actions are performed by means of other actions.

They can be distinguished by basic actions: actions that do not need other actions to be performed (Danto 1963). Thus, raising the arm can be performed both for throwing a ball and for putting on a cap. Basic actions depend on the abilities of the agent or on the particular situation. What can be a basic action for a quarterback it is not for a shortstop and vice versa. While Danto based the hierarchy of complex and basic actions on the causal relation, according to Goldman (1970), another way to relate higher levels of actions to lower levels is the convention. Although throwing a ball is performed by means of raising an arm, the technique of throwing a football ball or a baseball ball or basketball ball can be referred to as a matter of convention.

The hypothesis that events like loosening a bike helmet, opening a cooking pot, throwing a football ball or a baseball ball have a meaning that includes other actions and that the latter are referred to as basic actions based on causal and conventional relations explains how the same verb can denote pragmatically different ways of acting. Conventionality depends on the co-occurrences between typical event constituents as much as typicality relies on conventional patterns of events, entities and situations.

1.4. Perceptually Underspecified Nouns

Most of the actions that people perform include interactions with objects. Linguistic descriptions are oriented around entities, their properties and relations. Two friends at a pub may remember the day of their trip by bicycle in the countryside recalling the moment in which one pierced the tire, ran over a duck near a lake and lost his helmet in the water. Two colleagues during a break may talk about the college and the day of their graduation remembering an episode like the fall of a friend who tripped over his graduation gown, or the peeled tassel of the graduation cap of a classmate.

Causal relations link entities to each other. Entities with many causal connections are likely to be the more salient in an event (Zacks and Radvansky 2014). Generic agents like *person, men, girl, boy, woman* or *he* do not cue to typical situations. When they occur, it is possible to associate the verb with different actions and particular events only if the patient role is filled by distinct words like in (1) and (2). (1) and (2) lead to a generic representation of events. In order to link *put on* and *fill up* to particular situations that can contribute to the definition of the involved actions and referents, wider linguistic and extralinguistic contexts are necessary. One component of the sentence should be perceptually specified in order to individuate the event that is consistent with the current situation.

(9)

- a. *The man puts on the baseball glove*
- b. *The boy puts on the boxing glove*
- c. *The girl puts on the graduation cap*
- d. *He puts on the uniform cap*
- e. *The person sits on the bike saddle*
- f. *The woman sits on the horse saddle*

When a hyponym fills the patient role, the sentence provides enough information to make inferences about the involved actions and referents. The hyponyms *baseball glove, boxing glove, graduation cap, uniform cap, bike saddle, horse saddle* entails a specific type of perceptual referent. They are the cues to the knowledge of situations in which they typically appear, including for instance a baseball ball, a boxer, the diploma, a captain, wheels, and a horseshoe. Semantic knowledge about participants and situations

contributes to the disambiguation of verb denotational meaning. In (9) this kind of information can be exploited only at the end of the sentence. Reading or hearing (9), a person needs to find the hyponym in order to exploit the knowledge linked to it and infer that the agents may be a quarterback, a boxer, a graduate, a captain, a cyclist and a jockey, and the situations in which the events occur may be a football and a baseball matches, a ceremony of the delivery of the diploma or a trip by bicycle or by horse in the countryside. (10)

- a. *The catcher puts on the baseball glove*
- b. *The boxer puts on the boxing glove*
- c. *The graduate puts on the graduation cap*
- d. *The captain puts on the uniform cap*
- e. *The cyclist sits on the bike saddle*
- f. *The jockey sits on the horse saddle*

The collection of properties that identifies an entity is linked to an event because they play an important role in understanding its causal connections and the functional structure of the event. The information about event participants concerns both their names and their physical properties. The sentences in (10) represent redundant constructions that tendentially do not occur in texts. The agents in (10) cue the knowledge of related situations, which includes both linguistic and perceptual information. The use of a hyponym or a hypernym depends on the context. A hyponym like *bike saddle* for *saddle* is redundant when other words, like *cyclist*, provide enough information to define the corresponding referent. According to the maxim of quantity of Grice (1975) indeed, during the communication each speaker should not make his contribution more informative than it is required for the current purpose of the exchange. In this case, the hypernym *saddle* is enough to denote a bike saddle.

Physical properties include size, colour, texture, shape, and so forth. A baseball glove (10a) has physical features that distinguish it from a boxing glove (10b); in the same way, a bike saddle (10e) is different from a horse saddle (10f). The physical properties of an entity depend on its functional relations with other entities appearing in the same events and situations. Determining the correct referent involved in an event consists of individuating the physical properties that distinguish it from other entities that the same word can denote, namely identifying the hyponym when its corresponding hypernym

stands for it. Tendentially, the set of hyponyms of a hypernym are functionally similar: a saddle has the function of providing support for a person that have to ride something like a horse or a bike. However, since a horse and a bike are completely different entities, a horse saddle and a bike saddle have different physical properties. Therefore, a bike saddle and a horse saddle have completely different physical features because of their relations with the other entities appearing in the same situations. When an agent is perceptually specified, like *cyclist* (10e) or *jockey* (10f), much information about the patient can be left implicit because the information associated with the agent allow filling missing information. Thus, the patient can be indicated through a hypernym (*saddle*). The agent is perceptually specified when it denotes a person engaged in a particular situation and, in isolation, cues fine-grained knowledge about it, the events in which he is usually involved and the entities with which commonly he interacts.

(11)

- a. *The catcher puts on the glove*
- b. *The boxer puts on the glove*
- c. *The graduate puts on the cap*
- d. *The captain puts on the cap*
- e. *The cyclist sits on the saddle*
- f. *The jockey sits on the saddle*

The sentences in (11) are the most frequent constructions in texts. The fillers of the patient role are the hypernyms *glove*, *cap* and *saddle*. They can be linked to a specific type of perceptual referent based on the agent and the verb. The knowledge about typical situations in which the agent appears and other entities he usually interacts with is integrated with the verb selectional preferences. The integration leads to identifying the referent filling the patient role and, consequently, the actions that make up the event. Knowing the specific situations associated with the agents elicits multimodal expectations about the incoming items. The expectations are multimodal because they concern both words (lexical information) and referents (perceptual information). Thus, multimodal expectations suggest inferences: for instance, if the cyclist puts up something, it may be a bike helmet, if he sits on something, it should be a bike saddle, if he rides something, it should be a bike, and if he repairs something, it plausible it is a bike tire. Identifying *cap* in (11c) as a graduation cap corresponds to individuating the correct referent of the event,

namely the most consistent event with the current situation. If a person thinks about the cap of the Raptors basketball team, then she misunderstands the meaning of the sentence and, accordingly, the event described by it.

The hypernyms in (11) can be referred to as perceptually underspecified nouns because, in isolation, they do not denote particular entities but classes of referents. The lexical item indeed does not entail a specific type of perceptual referent. Like the agent, the perceptually specified patient (*baseball glove* in (10a), *boxing glove* in (10b)), in isolation, cues fine-grained knowledge about the situations in which typically it appears, the events in which it is usually engaged and the entities with which commonly it interacts. This knowledge includes linguistic (lexical) and extralinguistic (perceptual) information. The collection of perceptual properties that identify a particular referent is crucial to establish the types of actions that constitute an event. When the agent is perceptually specified, a person can anticipate plausible nouns and referents filling the patient role. The integration with the verb selectional preferences during sentence comprehension constrains the expectations toward the most plausible filler and accounts for both lexical and visual perceptual properties that identify it.

1.5. Grounded Nouns and Verbs

We explained why the properties of event participants are crucial information to identify the actions the verb denotes. The disambiguation of verb denotational meaning relies on knowing the referents filling the agent and patient roles. Hyponyms like *baseball glove*, *boxing glove*, *graduation cap*, *uniform cap*, *bike saddle*, *horse saddle* and agents like *catcher*, *boxer*, *graduate*, *captain*, *cyclist*, *jockey* are cues to lexical and visual perceptual knowledge about specific entities, events and situations. While fine-grained information about objects (like their physical features) implies bottom-up processes in order to comprehend (9), the information about typical situations in which agents appear suggests the use of top-down processes to understand (11) correctly.

When a generic agent (*man* or *boy*) occurs with a hyponym (*baseball glove* or *boxing glove*) in the patient position, a person tends to exploit her knowledge about the referent to understand the sentence, namely to individuate the event that best fits the current situation. The knowledge about referents includes the physical properties that distinguish, for instance, a boxing glove from a baseball glove or graduation cap from a uniform cap.

Physical properties depend on functional and conventional relations between event components. Every kind of gloves have the function of cover the hands, but they have to be suitable for the numerous situations in which they can appear. What makes them suitable objects of a particular situation are their physical properties. Thus, a boxing glove is not different only from a baseball glove but also a golf glove, a ski glove, a latex glove, an oven glove, etc. It is improbable that a doctor does surgery on a patient while he is wearing oven gloves.

When a specific agent (*catcher* or *boxer*) occurs with a hypernym (*glove*) in the patient position, the former is a cue to the knowledge of typical situations, which leads to the individuation of a specific type of perceptual referent belonging to the class the hypernym denotes. If *catcher* occurs, then the situation coincides with a baseball match, and the referent of the noun *glove* is a baseball glove. In contrast, if *boxer* occurs, the situation corresponds to a boxing match, and the word *glove* refers to a boxing glove. Thinking about a perceptually specified agent means focusing on a person engaged in a particular situation. A baseball player is a person who wears the team uniform and baseball shoes, based on his role, uses equipment like baseball bats, catcher masks or baseball gloves, runs from base to base on the baseball field, receives and throws baseball balls and so forth.

Therefore, both visual perceptual and lexical knowledge are involved in sentence comprehension. The relationship between bottom-up and top-down processes is mutual. The balancing depends on the construction of the sentence, namely on the information encoded in words that compose the sentence. Words are symbols. According to Harnad (1990), there are two main processes involved in semantic comprehension, namely in the assignment of a meaning: discrimination and identification. The discrimination is the ability to judge whether two types of information are similar or different and to define the causes of their similarity or divergence. The identification corresponds to the ability to assign a unique name to a class of entities or properties or relations, namely to individuate the commonalities among them. Identification is a way to categorize. Discrimination depends on what Harnad called iconic representations that, from a linguistic perspective, coincide with hyponyms like *baseball glove* in (9a) and *boxing glove* in (9b). Identification is an effect of the discrimination because it is only thanks to the numerous occurrences of the iconic representations that people can individuate their commonalities

and differences and can decide to which particular class they belong: see for instance *glove* in (11a) and (11b). The identification implies that icons are reduced to the invariant perceptual features that distinguish which members belong to a category and which ones to another. Through identification, people create categorical representations, what in the language is denoted by a hypernym. Iconic and categorical representations are not symbolic and what assigns meaning to them is the act of interpreting. Language represents a tool used by people to assign meaning to perceptual representations. Through propositions such as *This is a glove* or *That is a baseball glove* meaning can be assigned to the non-symbolic taxonomy of iconic and categorical representations and the symbols system that corresponds to a language can be grounded in perceptual experience. Through language people systematically assign meaning to entities, properties and relations of the real world. The systematicity derives from the possibility to combine the grounded names of the taxonomy to build propositions about further category membership relations.

Thus, if a person used to make a snowman in the garden in front of her house wearing mittens during the winters of her childhood, then she has a grounded experience of the word *glove*. In particular, she should know their function, protect the hands; their causal relations in the particular situation: it was winter, during winter it is snowing, the snowman is made of snow, the snow is frozen; their physical properties like colour, texture, shape and so forth. Supposing that the same child has never seen a boxing match but her dad used to listen to boxing matches on the radio. One day the child listens to the word *boxing gloves* on the radio and asks her dad how they are made and the dad describes them as mittens only more swollen, like two pillows, in order to protect the hands of the boxer from the strength of the fists. Thanks to the grounded experience of mittens and pillows the child can identify the referent of *boxing glove* as a component of the category of gloves and discriminate them from her mittens.

Verbs represent a class of polysemous words whose ambiguity depends on both linguistic and extralinguistic context. Without the perceptual representation of the actions that constitute an event, the lexical item cannot say a lot about the activities involved. We can imagine that a person reads (1) but has never seen someone putting up a cap or a glove or glasses or a boot. The only information that she can extract from the symbolic representations corresponding to the sentences is that caps, gloves, glasses and boots are objects that she can wear. Still, she cannot know that the verb *put on* implies different

actions based on the object that has to be worn. Reading linguistic descriptions of caps, gloves, glasses and boots the person can imagine that the movements suggested by the verb are different. Thus, if a cap is an object that is laid on the head and boot cover the feet, to wear the former an agent should not lower the arms towards the feet and to wear the latter an agent should not raise the arms towards the head. However, the symbolic description provided by the verb in isolation cannot be assigned to a particular meaning. The knowledge of how to do something is necessarily grounded in perceptual experience, which involves also grounded information about participants. Once again sports help to clarify the point. If a person has never seen a football or a baseball or a basketball match and she happens to read or listen to sentences such as (8a) and (8b) she is not able to say whether a quarterback and a catcher throw the ball in different ways, as in reality happens. These conventional relations have to be grounded in perceptual experiences of situations to assign the correct meaning to the symbolic expressions that describe them.

1.6. Events in Cognition

People transform perceptual experiences in knowledge that they can share and use in future situations. Language is the most used tool to convey information. Behind words, a world of memories of previous episodes of life opens up and updates dynamically based on the peculiarities of the current situations. In psychology, the event involves a cognitive representation just as much as objects and people. According to Barwise and Perry (1983), the way people organize the knowledge of the world mirrors elements and patterns of events because of the uniformity in the way people conceive them. Events are complex entities and across situations many of their aspects are consistent and predictable.

(12)

- a. *The cyclist puts on the helmet*
- b. *The cyclist loosens the helmet*
- c. *The cyclist sits on the saddle*
- d. *The cyclist rides the bike*
- e. *The cyclist repairs the tire*

Events in (12) may happen in different places: (12a) and (12b) in the box where the cyclist takes the bike; (12c) in the street in front of his house; (12d) in a park; (12e) at the edge

of a road. The events may happen in sequential order but with distant chronological intervals on the same day: (12a), (12b) and (12c) in the morning; (12c) and (12d) may also happen many times during the day if the cyclist made some stops; (12e) at the end of the trip, in the evening, for example, or in the middle of the trip, in the afternoon. However, some elements persist. The agent, *cyclist*. The fact that he is wearing the bike helmet while he is sitting on the bike saddle. He is wearing the bike helmet and he is sitting on the bike saddle while he is riding the bike. Tires and saddles are parts of the bike. Hence, the situation that contains the event in which the cyclist is repairing the bike tire includes also the bike saddle and the bike helmet, even if he may have removed the latter and laid it down. Thanks to the persistence of many elements through different events, it is possible to make associations and connections among them.

The distinction between episodic memory and semantic memory of Tulving (1985) reflects the hypothesis that people interiorize daily experiences and generalize the information included in them to use this knowledge in future situations. Remembering a bike trip with a friend in the countryside corresponds to recall an episode. Thinking about how to loosen the bike helmet evokes the semantic knowledge about manipulating the retention dial behind the head and adjusting the strap so that it is snug against the chin. The semantic knowledge concerning the way to loosen or tighten a bike helmet can be useful in other future situations in which the agent takes the bike to make another trip with her friend or to go to work. Remembering the graduation day is part of the episodic memory, but knowledge like the way to put up the graduation cap is part of the semantic memory: grabbing the cap, raising the arms and laying the cap on the head. The semantic knowledge linked to the way of putting on a cap is based on the experiences in which the person has worn different kinds of cap. Recognising the graduation cap as an object that belongs to the category of caps depends on the semantic memory associated with the previous situations in which a cap has appeared, such as the memory of the day in which a person went to watch a game of basketball and he has worn the cap of the Raptors basketball team.

According to Barsalou (2008), semantic memory is composed of conceptual knowledge, which supports cognitive activities like perceptual processing, the individuation of entities and events that appear in the current situation, their categorization, and the production of inferences about perceived properties of entities and events, their origin,

their interactions and what likely will happen next. People organize their knowledge around situations. A person perceives a situation from her subjective perspective, which includes a particular region of perceived space organized around a main entity in a defined chronological interval. Situations play a crucial role in sentence comprehension. Agents, patients and events are components of situations. Framing semantic knowledge in terms of situations implies that the inferences about entities, their properties and relations are specific and detailed. Entities can be associated with particular events facilitating their categorization. Moreover, knowledge about situations has a crucial role in discrimination and identification processes. The knowledge of situations cued by the agents leads to the disambiguation of the physical properties of the referents denoted by perceptually underspecified nouns filling the patient role. Without the information encoded in the agent, the latter could not be linked to any particular referent, as in (1) and (2). Words are cues to complex configurations of multimodal components organized in patterns corresponding to typical situations. When a word that denotes a component is perceived, the memories associated with it activate inferences about unperceived entities and events that could be present. The degree of plausibility and typicality of the implicit information depends on their past co-occurrence frequencies with the perceived information. The organization of semantic knowledge around typical situations leads to anticipation of incoming information in sentences comprehension. Moreover, it allows filling missing information in the sentence according to the current situation, like the physical properties of a referent denoted by a perceptually underspecified noun (hypernym) or the actions denoted by a verb.

A unique episode is the starting point for the construction of an event representation in the semantic memory. The recorded knowledge is made dynamic by integrating the information encoded in previous experiences and the peculiarities of the current situation. The semantic knowledge of the actions involved in (11c) must be generalized in order to allow the agent to perform the correct movements in case the situation is a basketball match and the cap is not a graduation cap but the cap of the Raptors basketball team. The composition of previous and new information and the dynamic update of the knowledge in semantic memory mirror the learning processes that allow people to transform their mental models (Johnson-Laird 1983). Mental models provide an explanation of how representations of specific events are used by people to accomplish tasks like sentence

comprehension. According to Zacks and Radvansky (2014), event models are representations of entities and relations useful to understand a specific state of affair. They derive from both life experiences and linguistic descriptions. Thus, a stimulus like a word may recall the knowledge about specific entities or events stored in memory, stemming from past experiences lived in the first person or via reading or hearing or watching the television.

1.7. Expectations

One of the reasons why people generalize perceptual information and transform it into semantic knowledge is the possibility to exploit it in future situations. The memory of previous experiences plays a crucial role in the comprehension of events and the definition of behaviours coherent with current situations. The ability to use the knowledge extracted from perceptual experiences in future situations is based on the expectations encoded in mental models. Expectations guide the processing of incoming information recalling the aspects of previous experiences that are coherent with it. They are connected to the knowledge of typical patterns of situations and events, namely typical participants, properties and relations. To understand an event and to anticipate correctly the incoming information is crucial knowing the components of the actions and their relations with the activity, namely whether a person is an agent or an object is the patient or the instrument. Ferretti, McRae and Hatherell (2001) investigated the conceptual content of thematic roles using single-word priming paradigm. They found that verbs activate knowledge of typical agents (*arresting-cop*), patients (*servicing-costumer*) and instruments (*stirred-spoon*) occurring in the same events. McRae et al. (2005), exploiting both short and long stimulus onset asynchrony priming paradigm, discovered that agents (*waiter*), patients (*chainsaw*), instruments (*guitar*) and locations (*cafeteria*) prime typical verbs (*servicing*, *cutting*, *strummed*, *eating*) involved in the same situations. The study of Hare et al. (2009) discovered a priming effect between nouns denoting events and things (*breakfast-egg*) and people (*sale-shopper*) typically participating in them. In the same study they found a priming effect from agents to instruments (*chef-knife*), patients (*key-door*) and from locations to people (*hospital-doctor*) and things (*barn-hay*). This study provides empirical evidence for the existence of mutual expectations between event participants.

Through expectations, people fill missing information and make predictive inferences about what will happen in the future. Talking about (12), I specified that *helmet*, *saddle* and *tire* referred to the objects bike helmet, bike saddle and bike tire. However, in (12a), (12b), (12c) and (12d) the words *helmet*, *saddle* and *tire* appear. Thanks to the knowledge of situations in which a cyclist commonly appears, and the objects with which usually he interacts, it is possible to avoid the use of a hyponym (*bike helmet*, *bike saddle*, *bike tire*) in the patient position, which would be a redundant specification of the relation between *helmet*, *saddle*, *tire* and *cyclist* (Grice 1975). The expectations linked to the meaning of the word *cyclist* allow filling missing information in (12a), (12b), (12c) and (12e) about the referents involved in the events. Moreover, when a person sees a cyclist putting on the helmet and sitting on the saddle, or she listens or reads a linguistic description of the two activities, it is plausible to suppose that she expects that the cyclist will ride the bike (12d). The person knows also that there is the possibility that the cyclist will repair the tire if he pierces it (12e).

The Event Segmentation Theory (EST) (Kurby and Zacks 2008; J. M. Zacks et al. 2007) states that ongoing perception includes predictions of the near future. Predictions are part of the representations of sensory inputs produced during the perceptual processing (Enns and Lleras 2008; Niv and Schoenbaum, 2008; Rao and Ballard 1999). EST proposes that in the presence of an event boundary the prediction error increases. Transient increases of the prediction error correspond to the updating of the event models in working memory according to the currently available sensory and perceptual information. Since the new event model is more effective than the older, the system should lead to a decrease of the prediction error and should settle into a new stable state. Individuating an event boundary corresponds to understanding whether a basic action belongs to a complex action or another, which can be sequentially ordered in the continuous perception of the sensory inputs. Neuropsychology and neurophysiology provide empirical evidence of the impairments in the online event individuation in patients with frontal lobe lesions (Zalla, Predat-Diehl and Sirigu 2003), schizophrenia (Zalla, Verlut, Franck, Puzenat and Sirigu 2004), and Alzheimer's disease (J. M. Zacks, Speer, Vettel and Jacoby 2006). The perception of event boundary in healthy people is associated with transient increases of brain activity in the posterior brain regions, posterior parietal, occipital, temporal cortex and in lateral frontal cortex (J. M. Zacks, Braver et al. 2001; J. M. Zacks, Swallow, Vettel

and McAvoy 2006). The data suggest that the beginning of an event corresponds to the beginning of a new mental representation of that event (Speer, Reynolds and Zacks 2007; J. M. Zacks et al. 2007).

The effects of the influence of the representations in the working memory on the perceptual processing stream consist in filling missing information and disambiguation of ambiguous information. In (12a), (12b), (12c) and (12d) there are elements that persist such as the agent, *cyclist*, and the bike helmet that he is wearing it while he is sitting on the bike saddle. However, the sentences describe distinct events. Once the cyclist has laid the bike helmet on the head and fitted it, he has to sit on the bike saddle. Hence, grabbing the bike, raising the leg, passing the leg above the bike, laying the foot on the pedal, bending the legs and resting the bottom on the bike saddle are actions that belong to a different causal chain defined by the event 'sit on the saddle'. The action of grabbing the bike after the cyclist had fitted the bike helmet increases the error of the prediction concerning the hypothesis that the cyclist is still adjusting the bike helmet, while the error decreases when predictions are about the hypothesis that he is performing the action in order to sit on the saddle. It is known that people are able to anticipate what will come next in sequences of events they encounter like a goalkeeper that has to adjust his position to anticipate the trajectory of a ball and successfully block a shot in a soccer match. In the same way, while people are watching other people movements shift their eyes to the predicted movements and points anticipating the whole actions. The anticipation seems to be included in the recognition of the actions that involve movements of objects in daily real-world situations (Kamide 2008). Behavioral and neurophysiological data provide evidence of the segmentation of actions into events during perception and reading. Transient increases in brain activities at event boundaries were observed in perceptive tasks as passive viewing of events (J. M. Zacks, Braver et al. 2001; J. M. Zacks, Swallow, Vettel and McAvoy 2006) and reading (Speer, Reynolds and Zacks 2007; McNerney, Goodwin and Radvansky 2011; C. Whitney et al. 2009). The study of Swallow, Zacks and Abrams (2009) used narrative films segmented previously by a group of viewers to investigate the updating effect. The clips were showed to the participants at the experiment and they were interrupted exactly five seconds after the participants had seen the objects which they are probed about. When a new event had begun during the critical five seconds, the responses mirrored the decreased availability of the objects in the

participants' working memory. The data of these studies show that performances on memory tasks reflect the interplay between event working models, event representations in long-term memory and other kinds of representations.

The collection of properties that identifies an entity play an important role in understanding the causal connections and the functional structure of an event. The functional structure depends on the mutual links between the set of properties of an object and the actions in which it can be involved. The actions that an agent can perform on an object are relations that contribute to the definition of the event. The actions that constitute an event are temporally ordered and depends on the intentions and goals of the agents (Lutz and Radvansky 1997; Magliano and Radvansky 2001). Goals and intentions are the causes of behaviours and permit mutual comprehension among people. According to Zacks, Speer and Reynolds (2009), information that belongs to the same causal chain can be interpreted as being part of the same event. When there is not a causal relation between pieces of information it is likely that they belong to distinct events. Conventional relations are another aspect that can define whether an action belongs to an event that is linked to a particular situation. Conventionality establishes also the coherence among a set of selected properties of an entity and a specific event or situation. Both caps and helmets have the function of covering the head. However, since helmets are used in situations like riding bicycles, motorbikes or horses, working in a building site, and wars, they have to have specific properties to protect the heads of cyclist, bikers, jockeys, engineers and soldiers in eventual dangerous circumstances like a fall from a horse, a fall of a building under construction or a mine explosion. Cyclists, bikers, jockeys, engineers and soldiers are involved in different events. Therefore, there are different types of helmets for each particular situation. The relation between the different kinds of caps is instead a matter of convention. Graduate students wear graduation caps, captains wear uniform caps and fans of Raptors basketball team wear another kind of cap. Their function is more related to the recognition of certain "social status" than to the safety of the agents' heads. In both cases, however, the collection of physical properties that identify particular types of helmets and caps depend on and, in turn, are important to define the events and situation in which they are involved.

Events include actions that are performed to do something else, like raising the arms to put on a cap or to throw a ball. Danto (1963), Goldman (1970) and Searle (1984, 2019)

define actions like raising the arms, which are not executed by means of some other actions but to perform other actions, basic actions. The causal and conventional relations that link basic actions and group them in sets that correspond to complex actions mirror the hypothesis of the Event Horizon Model (Zacks and Radvansky 2014), which states that the segmentation of experience into discrete events follows a hierarchical criterion which organizes sub-events and larger super-events (Newtson 1973; Hard, Tversky and Lang 2006). In this perspective, the temporally ordered and causally linked actions of grabbing the cap, raising the arms, laying the cap on the head and manipulating the retention dial, adjusting the strap correspond to sub-events of the larger super-events (11c) and (12b). Participating in an activity, reading or listening about it, maintain active the representation of the current event in the working memory and, at the same time, in long-term memory, a representation of the event is under construction. Long-term memory provides a permanent basis for the retrieval of knowledge in the future. In neuropsychology and neurophysiology, disorders of event understanding are investigated through studies of patients with action disorganization syndrome (Schwartz 2006). The data demonstrated that patients are able to execute basic actions like stirring a liquid on command, but they have difficulty in sequencing basic actions into a larger event such as preparing instant coffee. These studies suggest that the knowledge of the way particular events typically unfold is associated with specialized neural mechanisms. Patients affected by action disorganization syndrome present damage in the prefrontal cortex. Selective lesions to this region imply difficulties in producing actions in the right order and recognizing anomalous orders of events in simple stories or lists of action words (Allain, Le Gall, Etcharry-Bouyx, Aubin and Emile 1999; Fortin, Godbout and Braun 2002; Humphreys and Forde 1998; Sirigu et al. 1995, 1996). Neuroimaging studies that investigated event knowledge support the idea that the knowledge about how particular events typically unfold is linked to specialized neural mechanisms. Tasks that require people to think about the order of sub-events within a larger event selectively activate regions in the prefrontal cortex (Crozier et al. 1999; Knutson, Wood and Grafman 2004; Partiot, Grafman, Sadato, Flitman and Wild 1996).

Chapter 2: Learn and Comprehend Events

Learning processes and sentence comprehension are crucially linked to the multimodal thematic fit. This chapter is an overview of the studies focused on representing the verb meaning, thematic fit, and multimodal information merging. Moreover, we introduce the visual world paradigm and the experiments related to our study.

2.1. Human and Machine Learning

People create mental representations of situations they experience in the real world and preserve them in memory. New information is integrated into previous representations leading to the updating of the knowledge about previous ones. Mental representations are models distinguished based on the kind of information they include and their continuous updating is linked to human learning processes (Johnson-Laird 1983). Perceiving an event or playing an active role in it or thinking of it elicit a specific subset of neurons distributed in different areas of the brain of which identity defines the contents of the associated mental model. Human learning is based on the capacity of the neurons activated in a particular circumstance to modify the shape and the strength of their connections, which constitute the substratum of the memory (Dehaene 2020). The patterns of connections lead to creation of mental models and both depend on real-world experiences. People organize their knowledge of the world modelling the elements and the patterns of events and situations (Barwise and Perry 1983, Barsalou 2008, Radvansky and Zacks 2014). Event and situation models are the result of learning processes. They represent a more abstract knowledge than the superficial information that constitutes the perceptual inputs. Learning processes that lead to the creation of the mental models depend on the human capacity of generalizing perceptual experience and individuating the patterns, the rules and the causal relations that constitute it. An important function of mental representations is the interpretation of future events and situations. Mental models indeed encode the information that constitutes the expectations, which guide the processing of new incoming information recalling specific aspects of past experiences (Ferretti, McRae and Hatherell 2001, McRae et al. 2005, Hare et al. 2009). The inferences about participants, properties and relations that characterize typical situations depend on what people learnt

and how they conceptualized it (Marconi 1999). Even when the pattern of elicited neurons is turned off, the memory of the situation is present into neural circuits. A cue like a word can elicit the pattern of neurons linked to the knowledge of a previously learned event in which the entity denoted by the word was seen appear. Therefore, words can be referred to as cues to the knowledge of typical events (Elman 2014).

Deep artificial neural networks have the capacity to learn superficial statistical regularities in the data rather than higher levels of abstract concepts (Jo and Bengio 2017). While human learning mechanisms lead to creation of conceptual models of real world like event and situation models, computational learning is limited to recognition of shapes and co-occurrences in the data selected because composed of the best samples to train a network on a specific task. An algorithm able to learn from data is called machine learning algorithm:

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at task T , as measured by P , improves with the experience” (Mitchell 1997).

In computational terms, the process of learning is a device to achieve the ability to perform a particular task (Goodfellow, Bengio, Courville 2016). The abstractness that characterizes mental models mirrors the human ability to generalize based on symbolic rules rather than entirely on superficial similarity. The language is one of the most representative examples of this human capacity. The systematicity inherent to the use of language reflects the ability to exploit a finite set of words and grammatical rules to create an infinite number of sentences. According to Dehaene (2020), while the human brain can combine symbols based on the current context, artificial neural networks miss the ability to recombine what they learnt to resolve new different problems. Neural networks seem to simulate successfully the initial stages of human learning, namely the first 200ms in which the input processing is unconscious (Dehaene, Lau and Kouider 2017). However, they do not seem to be able to go beyond this stage and implement what corresponds to human reasoning, which includes the logical inferences required to catch the rules of a particular domain.

Neurophysiological adjustments are correlated to the developments in size and complexity of the events that a person can comprehend and these changes cover the duration of her entire life span. Since the first two months, infants appear to understand

temporal relations and integrate spatial information to create basic event models (Bourg, Bauer and van den Broek 1997, Baillargeon 1986, 1987; Baillargeon, Spelke and Wasserman 1985). According to Spelke and Kinzler (2007), among the core domains in which the child's knowledge emerges early and almost independently of the details of experience there are inanimate objects motion, spatial relations and social interactions. Thus, during childhood, learning processes lead to generating models of the perceptual experience of the real world like the representations of objects and the possible actions which involve them. The collection of properties that identify a sippy cup constitutes a representation of a specific type of cup that is typically used by toddlers. This knowledge is connected to the information concerning the possible actions can be performed with it like fill, empty, grab, open, close and so on. The creation of the model that includes information about the sippy cup and the actions of filling, emptying, grabbing or opening derives from the situations the toddler experienced in the real world. The first time the toddler might have seen a sippy cup might have been in the kitchen, while he was sitting on the high chair and looking at his mother was preparing a cake using a measuring cup. The child might have made the experience of the sippy cup also during an afternoon at the park when he asked for something to drink to his mother and she gave to him his sippy cup filled up by his favourite juice. In this situation, the toddler might have seen a friend asking the same thing to his mother and receive another sippy cup but with different physical properties: the sippy cup of the first child might have been green with dinosaurs depicted on it, while the sippy cup of his friend might have been blue with Nemo depicted on it. The mental models of the two situations (kitchen and park) in which the same sippy cup appeared encode the mapping between the parameters that make it up that are represented by other items appearing in same situations (juice, mother, high chair, bib or children playground) and the corresponding real entities. In both situations, the language played a crucial role. In the kitchen, the child may have heard his mother utter the word *cup* to refer to both the measuring cup and the sippy cup. What the toddler may have learnt from the first situation is that also his mother was using a cup but, because she did not drink something from it and, instead, she used it to pour the milk in a bowl, the measuring cup is another type of *cup* with different functional properties, that is, it is used to do something different from drinking. From the second situation the child might instead have learnt that *sippy cup* refers to all cups with that particular physical shape, but

different colour, used by toddlers to drink. Therefore, the child learnt the concept of cup (hypernym) and the distinction between different types of cups (hyponyms) based on their functions, which are strongly associated with the agents that use them: a toddler tends to use a sippy cup to drink, a mother usually uses a measuring cup to pour an ingredient in the bowl while she is cooking. The learning of conceptual hierarchical relations like “it is a type of” allows the use of the related expectations in future situations that require similar information processing to what was previously seen.

The conceptual and semantic knowledge derived from perceptual data is linked to the ability to pay attention to certain components of the experience and their particular properties, like what makes a sippy cup different from a measuring cup. This capacity is needed to comprehend the current event and to integrate new information in an already existing mental model. Event and situation models include a selection of information learnt through the use of attentional mechanisms, which improve their performances wider and more complex is the experiences that a person lives (Radvansky and Zacks 2014). A toddler may have seen his mother cooking using a measuring cup and he may have interacted only with his sippy cup. However, suppose that one day the mother may have decided to teach her child how to prepare biscuits. She may have explained to her son that, before mixing all the ingredients in a bowl, they should have measured their quantity. To obtain the right quantity, she may have made the child pour the ingredients in a measuring cup, may have made him notice the indications depicted on the cup and may have told him to stop when each ingredient reached a certain number. Thanks to the indications of his mother, the toddler should have focused the attention toward the numbers depicted on a measuring cup and on the fact that the object can be called through the same name that he uses to refer to his sippy cup. Thanks to intentions of his mother to teach him how to prepare biscuits, the attentions of the child may have focused on the details that make a measuring cup different from a sippy cup and she may have helped him to understand through an experience in which the child participated in the first person that the same word (*cup*) can refer to certain objects that can be distinguished based on their physical properties and that, in turn, their physical properties depend on the function of the object in a particular situation: a measuring cup has numbers depicted on it because it is used to measure the quantities of the ingredients, while a sippy cup is used by toddlers

to drink and it has a particular kind of cap that prevents the liquid from pouring on their clothes.

Learning new words is an example of the salient role played by the interplay between the intentions of the agent and the attention of the learner and of what distinguishes the learning processes of a human brain and an artificial neural network. The higher level of knowledge that the human brain can reach assuming the existence of abstract rules allows the child to go beyond the set of possible hypotheses limited to observable data and makes the human learning faster than the computational learning (Dehaene 2020). Human interactions play a crucial role during the learning of new words and meanings. When a mother, looking at a cup, says to her child “this is a cup” there is the same number of probabilities that the word *cup* refers to the single cup or to a class of cups. Only after a few instances of the object in different contexts a child learns that the word *cup* refers to a category of objects. The convergence of the few instances of cups to a single word makes the child learn that they are associated with the meaning of the word *cup*. An artificial neural network requires instead a huge number of instances before it learns to associate an entity to the right word. A crucial variable in learning new meanings is represented by the intention of the speaker. A child learns the meaning of a word if he has understood the intentions of the speaker, who usually uses cues like the direction of gaze or the pointing a finger toward something to contribute to the comprehension. The experiment of Ma and Xu (2013) demonstrated that a two- or three-years-old baby remember the name of an object only if an adult looking at a new toy says something like “Oh, a wog!” rather than by listening to the same sentence spoken by an artificial speaker. To learn and comprehend words correctly the listener has to identify himself in the speaker and try to comprehend the intentions encoded in his words (Carpenter et al. 1998, Lohmann and Tomasello 2003). Behaviours like looking at an object or indicating it with a finger emphasise the role of attention mechanisms in learning. The child selects a subset of information that involve his perceptual senses focusing his attention on that portion of reality which the speaker refers to. Paying attention to the same entity is a crucial component of the communication because allows people to make assumptions about what other people may think about it and, consequently, to comprehend the meaning of the words that express the intentions of the speaker. During language comprehension, individuating what a person refers to when she is speaking is important as much as

understanding what she thinks about it. If two friends are talking about the last baseball match of Toronto Blue Jays and one talks about the day of his childhood when his father bought him the team's cap, then the listener cannot think that cap refers to a graduation cap. Otherwise, it means that he did not comprehend correctly the situation they are talking about. Therefore, during the communication, the speaker should make sure that the listener pays proper attention to the entities he refers to. In the case of the Toronto Blue Jays cap, the discourse is about the last baseball match of the team. Thereby, the listener should focus the attention on the collection of properties that identify a typical cap of a baseball team rather than another type of cap that the word cap could denote, such as a graduation cap.

Learning the meaning of a type of word like a verb entails recognizing the event it denotes. Since verbs tend to be polysemous, the same lexical item may denote different events, and it often refers to a causal chain of actions. Eleven-month infants are able to identify sequences of two or more actions joining them coherently based on temporal relations, causal associations and the individuation of the intentions of the agents (Nelson and Gruendel 1986, Hudson 1988). According to Baldwin and Baird (1999), infants segment sequences of human actions depending on physical cues and statistical regularities, but, mostly, through the ability to create associations between an action and the corresponding intentions of the agents. From the first early stages of the learning, children can extract the intentions of the agent from the actions that he performs and many studies demonstrated that infants can individuate motion events similarly to adults (Wynn 1996, Baldwin et al. 2001, Saylor et al. 2007, Sommerville and Woodward 2005, Saffran 2003, Fiser and Aslin 2002). Learning and comprehending verbs require extra-linguistic information encoded by the entities involved in the events and linked to the intentions of the agent or the speaker. Event and situation models can be seen as useful strategies to conceptualize the information about verbs and actions. Focusing the attention toward the participants leads the learner to interpret the current event correctly, including verb denotational meaning. Moreover, it supports the expectations required to comprehend future events or fill missing information in the sentence, like in the case of a perceptually underspecified noun (hypernym) instead of its corresponding hyponym.

2.2. Sentence Comprehension

People tend to anticipate linguistic information during language comprehension because they have expectations cued by previous words that compose texts and sentences. What distinguishes language from the other experiences is that linguistic descriptions unfold sequentially and information are not simultaneously present. Thus, during language comprehension, people integrate information cued by words in an incremental way. One of the earliest studies about mental model creation of Ehrlich and Johnson-Laird (1982) investigated the ability to create coherent models during the linguistic descriptions of spatial layout. They presented a set of sentences that were labelled as “continuous descriptions” like *The knife is in front of the pot*, *The pot is on the left of the glass* and *The glass is behind the dish*. Another set of sentences were named “discontinuous descriptions”: *The knife is in front of the pot*, *The glass is behind the dish* and *The pot is on the left of the glass*. Because of the positions of the referents in the “discontinuous description”, participants had more difficulties in mapping incoming information with previous knowledge to create the correct model than in the case of the “continuous descriptions”. The experiment illustrated that the creation of event models through language needs an incremental understanding of the described circumstances. When the composition of words in a sentence is structurally ambiguous or semantically underspecified, a person has to exploit elaborated inferences to hang on several ideas and integrate current information with recorded knowledge in memory to obtain a representation of the event consistent with the current situation. Language processing plays an important role in the construction of mental models (Johnson-Laird 1983, 1989; van Dijk and Kintsch 1983). Because of the strong connection between models inferred from sensorimotor experiences and models extracted from linguistic usage many researchers believe that linguistic and experience models share most of their properties. The representation of events of the world constrains the composition of words in sentences, which can be seen as descriptions of the entities, their properties and relations. Lexical items give access to semantic knowledge associated with people and objects and the context defines which aspects of their meanings are coherent with the current situation. In cognitive semantics, there is a fleeting distinction between encyclopedic or world knowledge and lexical knowledge (Marconi 1999). The lexical meaning of words like *saddle*, *helmet*, *cap*, *backpack*, *bottle* or *tire* should contain information such as their

form, their constitution, their aim, possible actions that can be performed on them, and temporary features. As we saw above, the meaning of a verb is incomplete in isolation. Thus, the lexical representation of a verb includes sense-specific information regarding the properties of the nominals that best fit its thematic roles (Elman 2011). The preferences for the role fillers of a verb reflect specific knowledge of the events that can be associated with it.

In neuropsychology and neurophysiology, N400 represents a component of time-locked Electroencephalography (EEG) signals defined Event-Related Potentials (ERP). N400 is a negative deflection which peaks around 400 milliseconds after the onset of the stimulus. Its amplitude is linked to generalizations across input modalities. The regularities among N400 properties and sensory, conceptual and linguistic factors suggest indeed that the effects are modality sensitive but not modality specific (Kutas and Federmier 2011). According to Baggio, Van Lambalgen and Hagoort (2012), N400 represents the consequence of the construction process of compositional semantic representations. Compositionality in language depends on the balancing between the recorded knowledge in memory and the processing of new information encoded in the input (Baggio and Hagoort 2011). Lexical items are the cues to multimodal knowledge of events. However, when linguistic descriptions do not correspond to typical situations experienced in real world, new relations between words have to be activated in order to disambiguate the meaning of the sentence and more cognitive effort is needed to combine the information. Memory, Unification and Control (MUC) framework is a model of sentence comprehension that takes into account the balancing between the knowledge recorded in memory and the processing of new inputs (Hagoort 2005, 2013, 2016). In MUC, the memory component is language-specific and it corresponds to the recorded knowledge of typical constructions. Typical constructions can be seen as a set of constraints for each level of linguistic representation (Goldberg 2006). The unification component has the aim to combine the elements of the memory in wider structures. It is not a language-specific component and it includes also extralinguistic knowledge related to the context. The unification is a parallel process that combines information at all levels of representation and affects the composition of words in sentences. The control component has to establish a connection between the action and the social context.

According to Zacks and Radvansky (2014), language comprehension involves three levels of representation: surface form, propositional representation and situation model.

(13)

- a. *The doctor uncaps the bottle*
- b. *The bartender uncaps the bottle*

The information included in the surface form level concern morphological, syntactical and phonological or phonetical knowledge about words occurring in sentences. The number of arguments in (13a) and (13b) is the same, but the same verb-patient pair, *uncap bottle*, occurs with different agents, *doctor* and *bartender*. *Doctor* and *bartender* play the same role in the event but they are semantically different. Semantic information exploited during the comprehension belong to the propositional level. In cognitive sciences, a proposition consists of a unit of thought composed by the verb and its arguments. A propositional representation derives indeed from the surface form but it is a more abstract compositional representation that captures the meaning of the linguistic units than the surface representation. Situation model is the highest level of representation in language processing (Zwaan and Radvansky 1998). The construction of a situation model involves linguistic information, inferences associated with general world knowledge and memories of previously related experiences. A situation model corresponds to the event described in the sentence but it is more abstract than information included in surface and propositional levels. Different interpretations are distinguished in the cognitive structures associated with the meaning. *Doctor* and *bartender* are cues to many kinds of information associated with them. They are about typical situations and events in which they appear in the real world. Knowledge about situations and events includes typical participants and part of their properties. Thus, *doctor* makes people think that he may use a syringe or he may describe the contents of a pills bottle or he may wear the white coat or he may work in the hospital or he may use the stethoscope to visit patients or he may use scalpels to operate a patient. *Bartender* evokes information like that he may prepare a cocktail or he may serve beer bottles or he may work in pubs or he may fill up glasses or he should know different types of alcohol beverages or he may set up chairs and tables at the end of his daily turn and so forth. Situation model implies the selection among many kinds of information and their composition is based on other elements of the sentence. The composition of the information cued by the agent with the selectional preferences of the

verb has as result a representation that is richer than the information which can be extracted from each component of the sentence because *doctor* and *bartender* are linked to information that belong to different situation models. According to Altman and Mirkovic (2009), the main function of situation models in language comprehension consists in enabling predictions about the incoming information. A prediction can be seen as a change in the state of the language processing system based on the context prior to the availability of new input (Kuperberg and Jaeger 2016). The expectations encoded in lexical items cue the knowledge that guides the predictions, which allow people to anticipate incoming linguistic elements. The anticipation of information makes the comprehension more efficient and compensate for missing or ambiguous information during processing. The expectations of *doctor* and *bartender* in (13a) and (13b) concern typical events and situations in which the two agents appear. The following word is the verb *uncap*. The composition of the expectations of *doctor* and *uncap* in (13a) and *bartender* and *uncap* in (13b) corresponds to an incremental updating of the situation model, which now has enough information about the current context to anticipate a plausible incoming patient, *bottle*. Pickering and Garrod (2007) state that word-level predictions include not only particular individual words but also their features like gender, semantic field and grammatical category. However, according to Zwan and Radvansky (2014), event models play a salient role in evoking semantic features of words rather than grammatical properties. The ERP experiment carried out by van Berkum, Brown, Hagoort and Zwitserlood (2003) showed that reading a vignette like *As agreed upon, Jane was to wake her sister and her brother at five o'clock in the morning. But the sister had already washed herself, and the brother had even got dressed. Jane told the brother that he was exceptionally ...* When followed by the word *slow* produces a larger N400 than when the following word was *quick*. Even if both *slow* and *quick* are congruent with the meaning of the sentence, the context implies that *slow* was a surprising ending with respect to *quick*, that was instead the correct word to be predicted based on the context. The data provided empirical evidence of the influence of situation models in the processing of semantic information of words.

According to the Event Indexing Model, the clues that a working model is under update during the reading of narrative texts are momentarily slower readings (Zwaan, Langston and Graesser 1995; Zwaan, Magliano and Graesser 1995; Zwaan and Radvansky 1998).

The reader tends to update the representations in the working model when salient dimensions of the situation change, such as entities and causal breaks. According to J. M. Zacks, Speer and Reynolds (2009), people update working models when situational dimensions change because the changes in the situational frame make the activities in the narrative less predictable. Zwaan, Langston and Graesser (1995) investigated the hypothesis that event boundaries define the separation of the information in memory. They proposed some texts in which appeared various event boundaries. After the reading, people were presented with a set of verbs from the texts and were asked to sort them. The results were confronted with the event boundaries in texts. Verbs that belonged to different events were sorted in different lists. Verbs included in the same event were often placed in the same sorting list. The experiment of Glenberg, Meyer and Lindem (1987) investigated the persistence of entities from an event to another, namely shared information between old and new event models. Participants had to read short narratives in which critical objects were either associated with or dissociated from the protagonist of the story. A sentence like *John was arranging a bouquet for the table* was followed by *He put the last flower in his buttonhole, then left the house to go shopping for groceries* or *He put the last flower in the vase, then left the house to go shopping for groceries*. Reading that John left the house suggests the creation of a new event model. When the flower was in John's buttonhole, it should have had more chance to be part of the new model. When the flower remained in the vase, it should have not been in the new model. Results confirmed that readers had more difficulties in recognising the word *flower* or in reading an expression referred back to the flower when it had been left behind in the previous event.

2.3. Multimodal Thematic Fit

We explained how the knowledge of typical situations and events linked to words affect their composition in sentences and the construction of models in semantic memory. To be adequate to the current situation a model has to capture some of the perceptual properties of the experience described by the language (Zwaan 1999). Thus, if we are talking about doctors, we should infer that the bottle typically associated with them, pills bottle, has physical properties which distinguish it from the bottle that a bartender may serve, beer bottle, or that the tables of a pub are different from the table on which a doctor may use the scalpel on a patient. In (13) *uncap* denotes the action of opening an object, like a bottle, that is closed by a cap. The action may involve several movements such as grab the cap; twist the cap in a counter-clockwise direction; remove the cap; push the cap down until it rotates; use the grooves around the cap to get a good grip on it, squeeze and turn the cap; use the palm to push down the tab and turn the cap until it opens; hold the bottle by the neck; wedge the sharp edge of the bottle opener under the cap; lift the handle of the bottle opener up and so forth. The action of uncapping the bottle is performed by means of temporally ordered movements that are grouped based on the goals and intentions of the agent and the collection of physical properties that identifies the type of bottle. Intentions and goals help to clarify why certain objects are engaged in particular events. In turn, the collection of the properties that identify an object are crucial for the comprehension of the actions in which they are involved.

According to Long et al. (1990), the creation of an event model during language comprehension includes the ability to incorporate information about the features that entities may have but, most of the times, the properties of the entities must be inferred. The expectations about typical situations associated with *doctor* and *bartender* include linguistic and perceptual information. Among the properties that identify the objects that appear in the same situations of doctors and bartenders there are physical features. Thus, the composition of the expectations of *doctor* and *uncap* in (13a) and *bartender* and *uncap* in (13b) permit to anticipate not only that the patient of *uncap* is the word *bottle* but also that the corresponding referent is a pills bottle in (13a) and a beer bottle in (13b). Since pills bottle and beer bottle have different physical properties, also their caps are different. The links between the word *doctor* and the referent pills bottle, *bartender* and beer bottle allow the identification of the actions involved in the event *uncap the bottle* leading to

the disambiguation of meaning of the verb *uncap*. Thus, a pills bottle may have a cap that requires actions like twisting it in a counter-clockwise direction or pushing it down or using the grooves around the cap and squeezing or using the palm to push down the tab and turning the cap. A beer bottle instead can be opened by holding the bottle by the neck, wedging the sharp edge of the bottle opener under the cap and lifting the handle of the bottle opener up.

The relations between the nouns filling the agent and patient roles, the agent and the referent of the patient role, the noun filling the patient role and its referent are exploited by people during sentence comprehension. Lexical knowledge can be seen as a web of mutual expectations linked to typical real-world situations. According to Altman and Mirković (2009), language comprehension includes a mapping between the unfolding sentence and the representation of the event in memory that corresponds to the real-world event. Sentence comprehension is an incremental process of updating the expectations encoded in lexical items based on the described events. The fillers of the verb thematic roles encode lexical and perceptual constraints that mirror knowledge about specific situations and entities, which participate in events that a verb denotes. Each sense of a verb is composed of knowledge about a particular pattern of fillers (Elman 2014, McRae et al. 1998). Thematic fit plays an important role in sentence comprehension because it represents the amount of semantic coherence among the components of an event. Thematic fit affects combinations of nouns and verbs and relations between nouns that typically fill the agent and the patient roles of the same verbs. Since sentence comprehension is an incremental process, the thematic fit among the agent and the patient roles represents the amount of influence that the already filled agent role has on the unfolding patient role not already filled. Bicknell et al. (2010) conducted an Event Related Potential (ERP) experiment to investigate the effect of the thematic fit during online sentence comprehension. They found that typical agent-patient pairs such as *journalist-spelling* and *mechanic-brakes* in the sentences *The journalist checks the spelling* and *The mechanic checks brakes* elicited reduced N400s as compared to possible combinations but unexpected like *The journalist checked the brakes* and *The mechanic checked the spelling*.

A noun can be referred to as perceptually specified when, in isolation, it entails a specific type of perceptual referent. It cues fine-grained knowledge about the situations in which

it typically appears, the events in which it is usually engaged and the entities with which it is commonly interacting. This knowledge includes lexical and perceptual information. Multimodal thematic fit corresponds to the degree of coherence between an agent and a specific type of perceptual referent filling the patient role. Multimodal thematic fit mirrors the expectations that constitute the multimodal event knowledge, composed of lexical and perceptual information about its components.

Most of the times, the words that occur in sentences are perceptually underspecified, and the multimodal thematic fit based on the multimodal event knowledge have to be exploited. Thus, in (12) expectations encoded in *cyclist* permit to anticipate plausible fillers of the patient role in terms of lexical items and corresponding referents: if the cyclist puts up something, it may be a bike helmet; if he sits on something, it may be a bike saddle; if he rides something, it may be a bike, and if he repairs something, it plausible it is a bike tire. Even when in the sentence hypernyms like *helmet*, *saddle* or *tire* occur, multimodal thematic fit links *cyclist* to objects like bike helmet, bike saddle and bike tire. The multimodal event knowledge and multimodal thematic fit guide sentence comprehension, individuating the most consistent event with the current situation, identifying the involved actions and, accordingly, defining the correct meaning of the sentence.

2.4. Computational Approaches

Distributional Semantics (DS) represents words as vectors. DS is a usage-based method to investigate the meaning of words. In DS the representation of word meaning is based on the distributional hypothesis, which states that lexemes that occur in the same contexts have similar meanings (Harris 1954, p. 156). Distributional Semantic Models (DSMs) provide a quantitative representation of the meaning in terms of co-occurrence statistics between words. They are based on the assumption that word meaning can be learned from the linguistic environment. DSMs showed good performances in performing different semantic tasks (Clark 2015; Mikolov et al. 2013, Turney and Pantel 2010). However, it is still an open question whether statistical co-occurrences alone are enough to address deep semantic questions or they are only a surrogate representation of the lexical meaning (Lenci 2018).

According to Kintsch (2001), the combination of a verb with an argument (predication) leads to the creation of new meaning-in-context. The predication implies the selection of peculiar properties of the arguments that are appropriate for the selected meaning of the verb. In turn, the senses of a verb emerge through its co-occurrences with particular arguments. Each sense reflects the expectations associated with a particular context, namely a subset of properties that are contextually appropriate for the specific argument. According to Erk and Padò (2008), the interpretation of a word in context is guided by expectations about typical events that mirror plausible co-occurrences between words. A model of word meaning should provide representations that encode typical co-occurrences between arguments and verbs. The meaning of a verb should be handled accounting for the compositional processes that link it to its arguments. The compositional perspective mirrors the expectations between the event and its typical participants. In distributional semantics, the representation of a composed expression (sentence) is a vector. Different approaches among which linear algebraic operations like addition and multiplication, are used to project word vectors to phrase vectors. However, both the sum and the product of vectors are symmetric operations and do not take into account word order. The additive composition combines the content of the two constituents without that the contribution of one is affected by the contribution of the other. The multiplicative function selects only the contents of a vector that is relevant for the combination with another. Thus, a representation can be considered affecting the other (Lenci 2011). According to Mitchell and Lapata (2010), a model of the semantic composition should generate novel meanings through the selection and the modification of specific aspects of the involved elements. In distributional terms, a phrase vector produced by two constituent vectors should be the representation of a multiword vector that encodes a new meaning. The authors proposed a composition function based on the assumption that a model of semantic similarity that accounts for composition should handle with the combination of the semantic content of words in relation to their positions in the sentence. The idea is that the meaning of a proposition can be obtained by a function that computes the combination of two words, the relation that exists between them, and any knowledge involved in the compositional process. Erk and Padò (2008) integrated the information about multi-word contexts in a single distributional representation handling the expectations encoded in lexical items and involved in compositional

processes. The authors proposed the Structured Vector Space (SVS) model. In SVS the meaning of an individual word (a) in a context (v) is obtained through the combination of the vector of a with the vectors of the lexical expectations of v based on the specific semantic relation between a and v . Chersoni et al. (2016) proposed a DSM that account for verb meaning representations handling the contexts as joint syntactic dependencies. The authors used the definition of Melamud et al. (2014) of joint context, namely a word window of order n around a target word, to introduce the syntactic joint contexts, which take advantage from syntactic dependencies instead of linear word window. The verb vector representation corresponds to a typical verb-argument combination that mirrors the knowledge of typical event participants. The extraction of a collection of verb-argument dependencies from a parsed corpus is followed by the identification of all direct dependencies for each verb from the sentence of occurrence. A joint context feature is generated for each sentence by joining all the dependencies for a grammatical relation of interest.

The composition of words in sentences concerns the relation between a verb and its arguments. In many cases, the presence of more than one argument makes the relation between them crucial for the comprehension of the sentence. People indeed can determine the plausibility of a noun as filler of a thematic role based on how the other roles have been already filled. The thematic fit between the agent and the patient roles contributes to defining an event. Some studies accounted for both the verb selectional restrictions and the thematic fit between its arguments. Baroni and Lenci (2010) measured the thematic fit of an argument comparing its vector with a prototype vector obtained by the average over the vectors of the most typical arguments of the verb. The basic computational assumption was that the thematic fit of a noun as argument of a verb can be measured by the similarity in a vector space between the noun and the set of nouns that occur in the same role of it. Sayeed and Demberg (2014) and Sayeed et al. (2015) exploited the same method but they assigned the roles through the use of the semantic role labeler SENNA (Collobert et al. 2011). Greenberg et al. (2015) proposed a hierarchical agglomerative clustering algorithm to create the prototype representation. Clustering together typical fillers, the algorithm split them into multiple prototypes based on the sense of the verb. Tilk et al. (2016) generated probability distributions over selectional preferences for each thematic role using two neural network architectures that exploited role-labeled corpora

to optimize the distributional representations used for the thematic fit modelling. Santus et al. (2017) proposed a distributional method for modelling thematic fit that involves the use of a syntax-based DSM to build the prototype representation of the verb roles, the extraction of the second order contexts for each role and the computation of the thematic fit as a weighted overlap between the top features of the candidates as fillers and the prototype. Lenci (2011) proposed a computational model of dynamic composition and update of verb-argument expectations: Expectation Composition and Update (ECU). The main assumption of ECU is that part of the semantic content of a word consists of expectations about likely co-occurring words. In ECU online sentence comprehension consists of a dynamic updating of the argument expectations and thematic fit relations that integrates various type of knowledge about events and their participants. The relation between arguments guides expectations during sentence comprehension and makes it a dynamic process. Expectations reflect the variability of the meaning in context. ECU model addresses both thematic fit and compositional phenomena assuming that nouns and verbs are linked in a web of mutual expectations. When words are composed their expectations are integrated and updated. The semantic combination of an agent with a verb updates the expectations cued by the verb about a plausible filler of the patient role. The author proposed a function, $EX_{PA}(\langle n_{AG}, v \rangle)$, for the composition and update of expectations that involves: the expectations of a verb v about plausible patients, $EX_{PA}(v)$, and the expectations about typical events and patients associated with the agent, $EX(n_{AG})$.

$$EX_{PA}(\langle n_{AG}, v \rangle) = f(EX(n_{AG}), EX_{PA}(v)) \quad (1)$$

The thematic fit of a patient n_{PA} as patient of $\langle n_{AG}, v \rangle$ is measured by the cosine between the vector of n_{PA} and the prototype vectors of the top-k expected objects belonging to $EX(n_{AG}, v)$.

DSMs are completely based on linguistic information. However, most of the knowledge included in mental models and involved in sentence comprehension is implicit in language. According to Glenberg and Robertson (2000), traditional DSMs suffer from a lack of grounding in extralinguistic modalities. The necessity of grounding linguistic information in the perceptual environment led to the development of Multimodal

Distributional Models (MDMs), which integrate textual information extracted from corpora of texts with visual perceptual features automatically induced from collections of pictures (Feng and Lapata 2010; Bruni et al. 2012, 2014; Silberer and Lapata 2014; Kiela and Bottou 2014; Lazaridou et al. 2015; Chrupała et al. 2015).

Bruni et al. (2014) proposed a multimodal distributional semantic model in which the creation of textual and image-based vectors for the same word correspond to independent processes. After the creation of linguistic and visual representations, they exploited the Singular Value Decomposition to concatenate the two representations. Silberer and Lapata (2014) used visual representations with high-level visual attributes annotations and stacked autoencoders for the multimodal fusion. Kiela and Bottou (2014) adopted a concatenation strategy that exploited Convolutional Neural Networks (CNNs) to extract the visual features and the Skip-gram model (Mikolov et al. 2013) for the creation of the textual vectors. According to Lazaridou et al. (2015), the construction of linguistic and visual representations of the same concepts through two different steps represents a drawback for MDMs. This method does not include the generalization across modalities during the training stage. It leads to the assumption that linguistic and visual information is available for all words. The authors proposed the Multimodal Skip-gram model. It is based on the approach of Mikolov et al. (2013) but, differently from it, Multimodal Skip-gram model includes visual information for some instances of a subset of words. The model learns jointly linguistic and visual representations. This learning method simulates a typical scenario in which a person hears words together with concurrent visual stimuli. Kádár et al. (2017) proposed a method for analyzing the activation patterns of Recurrent Neural Networks (RNNs) to explore the learnt linguistic structures using a multi-task gated RNN architecture with two parallel pathways that shared word embeddings: IMAGINET (Chrupała et al. 2015). A visual pathway was trained on predicting the representations of the visual scene that corresponded to an input sentence. A textual pathway had to predict the next word in the same sentence. IMAGINET projects both the linguistic and the visual information in a joint semantic space.

2.5. Visual World Paradigm

The experiment exposed in the third chapter exploits the eye-tracking technique and the visual world paradigm. The visual world paradigm allowed us to explore the interplay between language and visual perception. The manipulation of the multimodal stimuli based on the kind of information under investigation permits to observe its involvement during language comprehension. The visual world paradigm concerns the use of linguistic and visual stimuli to explore the nature of the cognitive processes involved in language comprehension. The technique provides data on time-locked eye movements (fixation and saccades) toward particular positions of the visual scene during the perception of auditory linguistic stimuli. In the visual scene appear target, differently related and unrelated entities to the auditory sentences. There is empirical evidence that the strategies exploited in the visual world paradigm reflect normal language processing rather than strategies of explicit name retrieval based on the pictures in the visual scene. According to Huettig and McQueen (2007), fixations would present a random behaviour if they were based on preactivated names that fail to match the auditory words. In addition, the relations between the words and competitors suggest that participants' behaviour is not determined solely on the limited contents of the visual environment (Dahan et al. 2001a). Some studies explored the effects of the presence and absence of target object in the visual scene. When the target appears, people prefer the matches that involve it; in target-absent condition instead, the effects of the semantic competitor are stronger than conditions in which both the target and semantic competitor are present (Huettig and McQueen 2007). Other studies manipulated fine-grained details of the auditory linguistic stimuli and left unchanged the visual scenes to explore how linguistic stimuli modulate eye movements (Dahan et al. 2001, McMurray et al. 2002, Salverda et al. 2003, Dahan and Tanenhaus 2005, Shatzman and McQueen 2006a, 2006b). The resulted data showed that fixations into the visual world paradigm can be regarded as empirical evidence of the normal operations of the spoken-word and picture-recognition systems.

Temporary deflection of the visual attention towards the competitors in the visual scene was interpreted as the index of the activation of the kinds of knowledge that link the competitor to the target like the phonological knowledge when the names overlap, or the lexical and semantic knowledge when the target and the competitor belong to the same conceptual category (Huettig and McQueen 2007). One of the first experiments which

exploited the eye-tracking technique and the visual world paradigm demonstrated that people during the listening of words tend to look at the referents that auditory words denote (Cooper 1974). Thus, hearing the word *dog* people are more likely to fixate the picture of a dog than the picture of an apple. Cooper observed that the tendency concerns not only the pictures of the referents of the word but also semantically related entities as in the case of the word *lake* and the picture of a sailboat. According to Meyer and Schvaneveldt (1971), those data suggest a sort of visual semantic priming. Huettig and Altmann (2005) investigated if the visual semantic priming was based on semantic relatedness rather than semantic association. They created three versions for each sentence-visual scene combination. The sentence *Eventually, the man agreed hesitantly, but then he looked at the piano and appreciated that it was beautiful* was combined with three different scenes composed of four pictures. The first version was called “target” and included the target object, a piano, and three distractors. In the second version, the semantic competitor was an object that belonged to the same conceptual category of the target (a trumpet) and it appeared together with three distractors. The competitor was semantically related but not semantically associated with the target object. The third version was named “target & competitor” because both the piano and the trumpet were present together with two distractors. When the second version was proposed more looks were directed towards the trumpet upon hearing the word *piano*, between 200-300 and 800ms, than towards the distractors. The authors interpreted the results as an overlapping of the semantic information encoded in the word *piano* and the semantic information included in the mental representation of the trumpet. During the “target & competitor” version the piano was more looked at than the other objects without any preferential fixation toward the semantic competitor. The data provided empirical evidence that lexical information guides visual attention toward the objects that are semantically related. Thereby, eye movements mirror the conceptual similarity between the object on the visual scene and the target object denoted by the lexical item. Eye movements mediated by language can be referred to as a measure of the overlap between conceptual information conveyed by words and conceptual knowledge of the visual objects.

In his study Cooper (1974) noted that people tend to look at objects that present the physical properties evoked by the referent of the lexical item that occurs in the auditory sentence. Thus, participants tend to look at a snake listening to a sentence where the word

wormed appears (*just as I had wormed my way on my stomach*). In Dahan and Tanenhaus (2005), the data about eye fixations reported that participants looked at the competitors that presented the typical shape of the objects denoted in the linguistic stimuli. When they were instructed to move a snake from a location to another, indeed, the visual-shape competitor, a rope, was fixated less than the snake but more than the other unrelated objects in the visual scene. The differences in eye movements happened approximately between 200-300 and 1100ms after the onset of the critical word. Huettig and Altmann (2007) investigated how visual shape of objects affect eye movements during language comprehension. The outcomes of their study reported that people focused their attention toward the picture of a cable during the listening to the word *snake*. Since a cable and a snake can be associated with each other thanks to their shape, the authors concluded that the visual shape of objects plays a salient role in guiding eye movement during language comprehension. Huettig and Altmann (2004) studied whether the influence of the physical properties of referents on eye movements was linked to the current perceptual information on the screen or the stored semantic knowledge about them. They focused on colour relations because an entity can be associated with a prototypical colour but it can appear differently coloured in the current visual environment. Each sentence was combined with four pictures. In one condition the referent of a word in the sentence was present (frog). In another condition, the referent did not appear but an entity with the same prototypical colour was present in the scene (lettuce). Both frog and lettuce are typically associated with the green colour. In one condition the competitor was coloured with the prototypical colour of the target, in another condition it did not appear coloured. They found that only when the colour of the competitor appeared, namely when the lettuce was green, participants, listening to the critical word, looked more at it than other distractors. When the referent of the word was not associated with its prototypical colour instead participants looked at the picture that was associated with the prototypical colour of the referent. In this case, the perception of the current visual context rather than the stored knowledge seemed to guide the visual attention of the participants. The authors concluded that the probability of fixating an object into the visual environment suggests the interplay between the stored knowledge about its physical properties and the visual features extracted from the current perception.

Tanenhaus et al. (1995) proposed a study that exploited the eye-tracking technique to investigate the interplay between linguistic and visual information during the comprehension of syntactically ambiguous sentences. They instructed the participants to act through two types of sentences. A set of sentences presented syntactical ambiguity. Another set included unambiguous syntactical patterns. The sentence *Put the apple on the towel in the box* presents the ambiguity after the word *apple*. *On the towel* specified the location of the object to be picked up but without the specification *that's*. Before to listen to *in the box*, participants interpreted the instruction as if the towel was the destination. Hearing *in the box* they had to resolve the ambiguity linked to the fact that the box was the destination, namely the location where the apple had to be put. The sentence *Put the apple that's on the towel in the box* represented an unambiguous control condition. There were four different sentence-pictures combinations. The two sentences were combined with two different visual scenes. The first condition named "one-referent visual context" included pictures of an apple on a towel, an empty towel, a box and a pencil. It supported the destination interpretation because only one apple appeared. The second condition was called "two-referent visual context" because the picture of a pen was replaced by the picture of an apple on a napkin. Since two apples appeared, *on the towel* should have been interpreted as the specification of the apple that had to be moved. Fixations patterns revealed that in the one-referent condition, listening to *on the towel* participants initially interpreted it as the destination. In the two-referent condition instead, *on the towel* was correctly interpreted as the modifier of *apple*. When the one-referent condition was combined with the ambiguous instruction the pictures of the apple was looked at for 500ms after the hearing of the word *apple*. Then participants looked at the empty towel, the incorrect destination, 55% of the time. When the same visual condition was combined with the unambiguous instruction, participants never looked at the incorrect destination. In the two-referent condition, after the listening of *apple*, participants often looked at both apples in the scene. When it was combined with the ambiguous instruction participants looked at the incorrect destination 62% of times. The time exploited to establish the reference correctly in the two-referent condition did not differ for both ambiguous and unambiguous instructions. The authors concluded that during the earliest moments of language comprehension people have the tendency to establish the reference based on their goals and intentions. In addition, relevant referential

visual information immediately affects the comprehension of the structure of linguistic information.

Language comprehension is an incremental process. The time-course of the processing plays a crucial role when people have to handle incremental inputs. Huettig and McQueen (2007) explored the time-course of the retrieval of phonological, visual-shape and semantic information during online language comprehension. They performed four experiments. In two of them, they proposed sentences in which the critical word occurred in a neutral sentence, such as *Eventually she looked at the beaker that was in front of her*. The critical word was *beaker*. The corresponding visual scene consisted of pictures of a beaver, the phonological competitor; a bobbin, the visual-shape competitor; a fork, the semantic competitor, and an umbrella, the distractor. Other sentences were built so that the critical word was not predictable like, for example, *He thought of a word that rhymed with [...]*, *He dreamt that night about a [...]*, *She turned round and saw the [...]*. All sentences were in Dutch. In the first experiment, the pictures appeared on the screen at the start of the auditory presentation of the sentence. In the second experiment, the visual scene was presented only 200ms before the onset of the critical word. In the first experiment, the results revealed that participants focused the attention on the phonological competitors before to shift the eyes toward the visual-shape and the semantic competitors. Under the conditions of the second experiment, the visual-shape competitor was the most fixated together with the semantic competitor, meanwhile, fixations toward the phonological competitor did not show significant differences from the distractor. According to the authors, the candidates consistent with the acoustic-phonetic information of the auditory stimuli are involved in a parallel process which includes also the phonological level of representation. However, the stored knowledge about words concerning the physical properties of their referents and their semantic attributes interfere before that the phonetic-phonological level is completed. Moreover, the data provided evidence that there is a fleeting distinction between the perceptual and the conceptual components of the semantic knowledge. Thus, the knowledge about a concept like bean includes both visual properties (shape) and functional attributes (it is edible). Eye movements during the visual inspection of the scene mediated by the language depend on the interplay between the visual and the linguistic current contexts and the corresponding mutual interferences can involve phonological, visual-features and

semantic levels. The employment of each one is linked to the time point of the presentation of the auditory stimuli in which the visual scene is displayed.

Huetting and McQueen (2007) demonstrated that the information extracted from the current visual context (visual-shape and semantic knowledge) rapidly affect eye movements during language comprehension. The stored semantic knowledge is evoked through visual and linguistic stimuli and it is rapidly updated to provide the most coherent behaviour to the current situation. However, through the exploiting the eye-tracking technique may be difficult to individuate and distinguish the factors that contribute to the comprehension and the amount of their influence. Thus, some studies manipulated the visual stimuli to build an alternative visual world with the aim to confront the rapidity of the integration of the current linguistic and visual information with the time-course of the intervention of the stored real-world knowledge (Knoeferle and Crocker 2006, 2007).

According to Knoeferle and Guerra (2016), the outcomes of those studies provided evidence of a referential preference or priority during language comprehension. The notion of referential stands for the relationship between a noun and the denoted object as much as that between a verb and the denoted actions (Jackendoff 2002). In the eye-tracking perspective, more looks at an object in the visual context that is named rather than an object that is semantically related to the linguistic input are the evidence of the nominal referential preference. The referential preference that relates a verb to a depicted action is crucially linked to the agent, which can be stereotypical or unusual. Thus, eye gaze toward the agent in the visual scene who is executing the action denoted by the verb was interpreted as a referential priority for that action even when it is executed by a non-prototypical agent (Knoeferle and Crocker 2006). Other studies manipulated the time perspective of the actions described by the sentences (Knoeferle and Crocker 2007; Knoeferle et al. 2011; Abashidze et al. 2014), the presence/absence in the visual context of the described event (Altmann and Kamide 2009) or exploited the coercion phenomena (Scheepers, Keller, and Lapata 2008) to explore the influence of the current visual context on sentence comprehension, the referential preference. The referential preference highlights the relation between the fixations toward a referent and both its lexical representation (Tanenhaus, Magnuson, Dahan, and Chambers 2000) and its conceptual representation extracted from both linguistic and visual contexts (Altmann and Kamide 2007). The notion of visually situated language comprehension defines a field of studies

that concerns the interplay between language comprehension, attention, and non-linguistic visual context and exploits methods like the eye-tracking technique to explore it (Knoeferle and Guerra 2016). In this perspective, the chronological interval between the beginning of the processing of a word and the shifting of eye gaze to its referent mirrors the establishing of a reference. The data about eye movements seem to confirm that referential relations take priority over other relations between language and real-world knowledge during language comprehension (Huettig and Altmann 2005). The tendency of people is first to check the visual scene looking for referential relations based on the components of the auditory sentences.

2.6. Anticipatory Eye Movements

During the perception of sequences of events people can anticipate what will come next. This ability concerns actions and movements in real-world situations as much as linguistic entities during sentence comprehension. Since language implies incremental and continuous processes there are pro and cons about the hypothesis that people exploit anticipation during the comprehension. On the one hand, anticipation is a mechanism that allows people to get ready about the future. On the other hand, the result of the anticipation may be wrong and, in that case, the greater cognitive cost of the elaboration of alternative answers consistent with the current situation may not be advantageous. Thus, anticipation seems to be exploited only when the benefits outweigh the costs. Anticipation concerns certain aspects of the incoming information included in the linguistic input and it can be referred to as prediction. The observation of anticipation mechanisms provides precious empirical evidence of the incrementality of the language, which is in agreement with the psycholinguistic and computational theories about the continuous mapping between incoming items and mental representations under constructions. Thanks to anticipation people can build a representation of the incoming items without delay and integrate it into the previous representations. This perspective suggests that anticipation mechanisms presuppose the incremental processing of linguistic information. There are many studies in the eye-tracking literature concerning the notion of predictability. They addressed the link between eye movements toward an item of the visual environment and the corresponding forthcoming item in the linguistic input. However, some studies provided a different definition of the notion of

predictability, which was interpreted as contextual cohesion or lexical co-occurrences probabilities. In addition, the anticipation mechanisms were often interpreted as the integration of the congruent incoming items into the preceding context (van Berkum et al. 2005, DeLong et al. 2005, Federmeier 2007). In the psycholinguistic literature, the priming effect was associated with anticipation. Even though most of the studies about the priming effect focused on the relationship between the prime and the target items in order to demonstrate that the prime facilitates the processing of the target, some studies linked the priming effect to prediction mechanisms. Ferretti, McRae and Hatherell (2001), McRae et al. (2005), Hare et al. (2009) explored the expectations encoded in lexical items. McRae et al. (2005) suggested that the results of their experiments supported the hypothesis that the semantic processing of nouns leads to anticipatory computation of the verbs that may follow in the sentence during the comprehension. Other studies used sentence reading methods to investigate anticipatory mechanisms in sentence comprehension (Rayner et al. 1983, Taraban and McClelland 1988, Altmann 1999). Since both the priming effect paradigm and sentence reading methods did not provide a clear distinction between integration and prediction processes, other researchers exploited the eye-tracking technique to study the anticipation mechanisms (Altmann and Kamide 1999; Kamide et al. 2003). In these studies, anticipatory eye movements correspond to the relatively frequent eye movements toward the predicted objects before the onset of the referring expression. The main assumption is that the data about eye movements recorded through the visual world paradigm provide empirical evidence of the interplay between linguistic and real-world knowledge. Both auditory sentences and visual contexts are cues to the knowledge about real-world situations. Thus, the studies about the anticipation mechanisms focused the attention on which type and amount of contextual information are necessary to be considered as predictors of a certain incoming item (Creel, Aslin, and Tanenhaus 2008).

Knoeferle and Crocker (2006) explored the interplay between the current visual context and the recorded knowledge about typical events during online sentence comprehension. They compared the relation of a verb with both its current referential action executed by an unusual agent and its stereotypical agent engaged in an unusual action. The sentences *The detective will soon spy on the pilot* and *The wizard will soon spy on the pilot* were combined with pictures of a wizard looking at a pilot through the telescope, a detective

serving the pilot some food, a pilot and a tree. The first condition corresponded to a stereotypical agent-verb relation because the detective typically spies, but he did not do it in the visual scene. The second condition stood for the current action referent relation because in the visual scene the action of spying was executed by the wizard. The authors found that during the verb time window (*spy*) in the second condition participants looked more often at the wizard, even if spying is an action typically executed by the detective. Since the visual scenes provided information that conflicted with typical event knowledge stored in memory, the authors interpreted the outcomes as a confirm of the hypothesis that listeners exploit information extracted from the current visual context during online comprehension to establish a referential relation between a verb and its corresponded action. Since participants inspected the action denoted by the verb and performed by an unusual agent (wizard-spy) more often than the prototypical agents performing an unusual action (detective-serve), the data provided evidence of the priority of verb-action reference over the expectations about actions that the stereotypical agents might perform. Altmann and Kamide (1999) investigated the hypothesis that people tend to predict which object will fit the patient role after hearing the verb. The sentence *The boy will eat the cake* was combined with pictures of a boy, a birthday cake, a toy car, a toy train and a ball. Results reported that the participants fixated the single edible object in the scene, birthday cake, more often than the other depicted objects before hearing the critical word *cake*. By contrast, when participants heard *The boy will move the cake* together with the same visual scene, they looked at all of the movable objects without statistically significant differences among them. The authors concluded that the outcomes provided empirical evidence that the selectional preferences of verbs constrain the set of possible objects that may follow them.

Kamide, Altmann and Haywood (2003) investigated the hypothesis that agent-verb pairs elicit anticipatory eye movements toward entities that may fit the patient role. The sentences *The man will ride the motorbike*, *The girl will ride the carousel*, *The man will taste the beer*, *The girl will taste the sweet* were combined with pictures of a motorbike, a carousel, a beer and a sweet. Anticipatory eye movements on the predicted objects were triggered at the time in which participants listened to the verbs. Listening to the combination *man-ride* participants looked more at the motorbike than the other objects in the visual scene while hearing the combination *girl-ride* participants focused the attention

on the picture of the carousel. *Man-taste* and *girl-taste* combinations guide the eye movements of the participants respectively toward the pictures of the beer and the sweet. The results were consistent with the assumption that expectations associated with agent-verb pairs guide people's eyes toward the most plausible entity that may fill the incoming patient role.

Chapter 3: Which Object do You Expect?

In this chapter we will describe an experiment exploiting the eye-tracking technique and the visual world paradigm. We expect that the incremental composition of the information cued by the agent and verb elicits expectations concerning the referent filling the patient role. The expectations depend on the information cued by words and depicted in the visual scene. The current visual environment included targets and competitors that typically appear in the same situations as the agent and the verb. Agent-verb pairs and object pictures allowed participants to fill in missing information in the auditory sentences about the perceptually underspecified noun (hypernym) in the patient position. The study proposed here differs from the study of Kamide et al. (2003) in two crucial details. Firstly, while they used generic agents like *man* and *girl*, we proposed sentences where specific agents occur such as *student*, *hiker*, *boxer*, *catcher*, *cyclist*, *jockey* or *biker*. Secondly, in Kamide et al. (2003) the patient role presented different nominal fillers based on the agent (*man-beer/motorbike*, *girl-candies/carousel*). We instead proposed two conditions in which the agent-patient pairs differed in the two specific agents but were composed of the same hypernym as nominal filler of the patient role: *backpack*, *glove*, *saddle*, *helmet* and *bottle* appeared in combinations like *hiker-backpack*, *student-backpack*, *boxer-glove*, *catcher-glove*, *cyclist-saddle*, *jockey-saddle*. As Kamide et al. (2003), we expected anticipatory eye movements towards targets, however, our aim was different. We wanted to demonstrate that the representation of the verb meaning includes multimodal information about typical participants that play a role in events it denotes. Among multimodal information about entities, their physical properties are crucial in individuating the actions denoted by verbs. According to Elman (2014), the time-course of information processing plays a crucial role in defining what should be included in the lexical representation. We assumed that the speed with which people integrate information about the multimodal thematic fit between the agent and the patient should be considered a clue of the relatedness of information about event participants to the verb semantic representation. In our experiment, anticipatory eye movements were interpreted as empirical evidence that information associated with typical agents and patients should be included in the verb representation. As described in the first chapter, agents and patients constrain the situation and, consequently, lead to the individuation of the actions

constitute the involved events. Multimodality is mirrored in the interplay between linguistic and visual information. In our experiment, it did not only depend on the visual world method but, most of all, on co-occurrences in the visual scene of targets with action-related objects denoted by the same hypernym that fills the patient role in the auditory sentences.

3.1. Sentences, Pictures and Lists

Each trial consists of a combination of linguistic and visual information. The participants listened to auditory sentences while looking at four pictures (the visual scene). We created ninety trials consisting of sixty experimental and thirty filler trials. In the experimental trials, each sentence describes a typical event performed by a particular agent.

(14)

- a. *The seamstress turns on the machine*
- b. *The bartender turns on the machine*

Seamstress and *bartender* can be referred to as perceptually specified agents because they elicit information about typical situations and events in which they usually appear (see the first chapter). Differently from the studies of Altmann and Kamide (1999) and Kamide, Altmann and Haywood (2003), the agents of this study provide the information to individuate particular situations, which constrain the set of the incoming plausible objects. In (14a) and (14b) the verb is the same, *turn on*, because we wanted to focus on the fact that the two agents, in completely different situations, can perform different actions denoted by the same verb. Even the nominal filler of the patient role is the same, a perceptually underspecified noun or hypernym (*machine*). Since we wanted to account for the influence of visual information during sentence comprehension, we made sure that the information exploited during the disambiguation of the patient role depended on knowledge of typical situations cued by the agent integrated with the verb selectional restrictions and the extra-linguistic information extracted from the pictures.

The visual scene was composed of the pictures of four objects shown during listening to the sentences. One image pictured an object that usually appears in the same situations and events of the agent. We called it Agent-Related. The object related to *seamstress* was a thread; *bartender* was related to a mug. Another image depicted what we defined an

Action-Related object because it fitted the verb selectional restrictions but it was typically non-involved in the same situations of the agent. The Target object was the correct referential filler of the patient role based on the semantic constraints of both the agent and verb. We switched the Action-Related and the Target objects according to the agent. The Action-Related object associated with *seamstress* was an espresso machine, while, the Action-Related object of *bartender* was a sewing machine. Because they correspond to the Target objects in the inverted sense (*seamstress*-sewing machine and *bartender*-espresso machine) we were sure that both could have been a plausible patient of the verb *turn on*. Hence, in the experimental trials the agent performs an action that could be associated with two pictures in the visual scene, the Target and the Action-Related. The patient role was filled by a perceptually underspecified noun that could refer to both objects (*machine*). A fourth image was the same for both events and was unrelated to the events described by the sentences: it was called distractor or Unrelated picture. In the examples (14) it was a lock. See Figure 1, which shows the combination of sentences and pictures in the *seamstress* and the *bartender* conditions.



Figure 1 First and second lists trials.

The trials were split into two lists to present only one type of verb-patient pair, Target and Action-Related pictures to each participant (*turn on-machine*, sewing machine and espresso machine in Figure 1). In the experimental trials, there are two agents for each

type of verb-patient pair. We assigned the agents that appear with the same verb-patient pair to different lists. Hence, if the *turn on-machine* pair co-occurs with *seamstress*, the sentence appears in the first list; if *bartender* is the agent of the *turn on-machine* pair, the sentence is in the second list. Figure 1 shows that participants assigned to the first list listened to sentence (14a): *The seamstress turns on the machine*. Participants assigned to the second list listened to sentence (14b): *The bartender turns on the machine*. Crucially, however, they saw the two set of pictures in Figure 1, where:

- The Target and Action-Related pictures exchange their role.
- The Agent-Related pictures are different.
- The Unrelated picture (distractor) is the same.

Each list was composed of thirty experimental and thirty filler trials. The filler trials were the same in the first and the second lists. The trials were shown in random order. For a complete list of the auditory sentences refer to Auditory Sentences in the Appendix.

The filler trials aim to avoid the participants discovering the relationship between the event described by the sentence and the objects depicted in the visual scene, namely the link between the agent, the Target and the Agent-Related objects, and the connection between the verb-patient pair, the Target and the Action-Related objects. In fifteen filler trials, the sentence was associated with a visual scene (four pictures) in which two objects could be denoted by the same word, but were Unrelated to the sentence content. The filler of the patient role referred instead to a third object.

(15)

The man does not like candies

was combined with pictures of a candy, a fishing hook, a coat hook and a candelabra. In the given example, the word *hook* applies to two different images.

Additional fifteen filler sentences had various syntactic structures with one word referring to one of the pictures in the visual scene.

For instance:

(16)

Karen made the tea with her new pot

was combined with pictures of a teapot, a marble, a picture frame, a mitten.

We used four practice trials to familiarize participants with the experiment.

3.2. Norming Study

Before to experiment, we carried out a norming study to test our experimental stimuli through human judgements. In particular, the norming study involved the agents that appeared in sentences and the Target objects showed in the visual scene. We measured the strength of the relatedness between the agents and the predicted object images. We used the Figure-Eight crowdsourcing platform¹ to create a task in which participants evaluated how likely it was that the agent and the object appeared in the same situation, using a scale that ranged from 1, which indicated “not very likely”, to 7 that stood for “very likely”. We asked participants to read the name of the agent and click on a link that opened the image that depicted the corresponding Target object. Participants read the name of an agent appearing on our linguistic stimuli like *doctor* and opened the link for the Target object picture, pills bottle. They had to rate “How likely is it that the person and the object appear in the same situation?”. The mean ratings of the answers were 6.3 and the 95% confidence interval was 0.1. We interpreted the outcomes as the evidence that the agents and the objects typically co-occur in the same real-world situations.

¹ www.figure-eight.com

How likely is that the person and the object appear in the same situation?

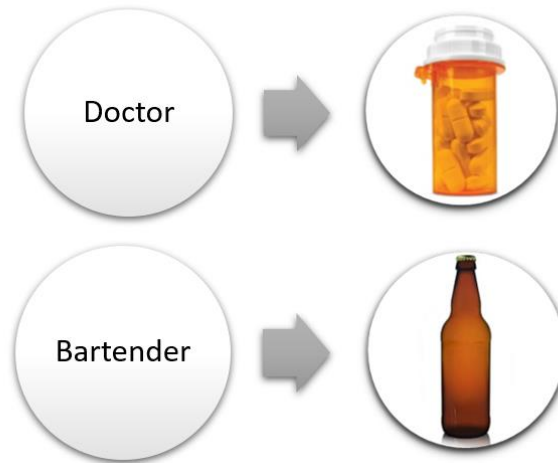


Figure 2 Norming study task.

3.3. Method

Participants. Participants were recruited through the SONA System of the Psychology Department Undergraduate Research Participant Pool. It is available to students enrolled in undergraduate psychology courses. Twenty-four University of Western Ontario undergraduate students participated in the experiment. They ranged in age from 19 to 28 years. All participants had normal or corrected to normal visual acuity and self-reported English as their native language. Self-reportedly, participants had never endured a traumatic brain injury or illness and were not currently diagnosed with any major psychiatric illness. They were compensated \$10 for their participation.

Auditory Stimuli. The same female native English speaker recorded all the sentences. The files were recorded using a Sennheiser e845S mic with Audacity Cross-Platform Sound Editor 2.2.2 (released February 20 2018), in a 4” thick sound-proof booth (Model CL-13 LPMR), with a Sound Devices USB Pre2 preamp on a MacBook Air OSX. We annotated the files by marking relevant points of the sentence using a customized script in Praat, Version 6.0.37² (retrieved February 3 2018). For each sentence we set a pointer at the start of the sentence; the agent onset; the agent offset, which corresponded to the

² <https://www.fon.hum.uva.nl/praat/>

verb onset; the verb offset, which corresponds to the second article onset; the second article offset, which corresponds to the patient onset; the patient offset; the end of the sentence. The agent offset/verb onset was normalized in all auditory files at 1200ms. The sound files were played through Logitech X-120 speakers (120V ~ 60Hz) using a PC computer with Windows XP and an Automedia 2 soundcard.

Visual Stimuli. All images were presented at 300x300 pixels in colour. Each picture was placed in a different quadrant of the screen at a 45-degree angle from the centre. The location of the four images was randomized across trials and participants. The pictures were selected from BOSS³ and KONKLAB⁴ Image Corpora.

Eye Tracker. We used a desktop mounted Eyelink 1000 eye tracker to record eye movements, and Experiment Builder, Version 1.10.1241 software⁵ (SR Research Ltd.) to coordinate and present the stimuli. The camera lens was positioned approximately 60 cm from the participant's head at an approximately 35-degree angle to the participant's eyes. Participants were positioned 70 cm away from a 16-inch monitor displaying the visual stimuli (resolution set to 1024 x 768 dpi). Calibration was performed before the start of the experiment, as well as at any time the equipment registered significant head movement. At the beginning of each trial, a fixation point was presented as a calibration check to ensure that in case the camera ever lost the pupil the program automatically would have gone to camera set up to allow for calibration to be completed.

Procedure. At the beginning of each trial, a fixation cross was presented for a maximum of ten seconds. When the time limit was reached the participant was redirected to calibration. After three seconds during which the participant fixated the cross, this was replaced by the four trial images, one for each quadrant. Participants had one second to become familiar with the images before the auditory stimulus was presented. After the preview period, a series of red circles were flashed in the centre of the screen to bring the participant's attention back to the fixation cross. Once the focus on the fixation cross was

³ <https://sites.google.com/site/bosstimuli/>

⁴ <https://konklab.fas.harvard.edu/>

⁵ <https://www.sr-research.com/experiment-builder/>

registered, this signalled the program to begin playing the sentence. The four pictures remained on the screen while the sentence was presented and participants' eye movements were recorded. An additional 300ms of silence followed the end of the sentence before the images disappeared and the next trial began showing the fixation cross. Before starting the session, participants were assigned to a list. Each list contained three trial blocks. At the start of the experiment, participants received the following instructions:

“You will see a display with four pictures while hearing a sentence. There is no task involved; just look at the pictures and listen to the sentences. We'll start with some practice trials to see how it works.”

The first block contained four practice trials to get participants used to the task. Thereafter, participants saw:

“This is the end of the practice sessions for part one. Do you have any questions before the experiment begins?” The other two trial blocks contained the experimental and filler trials randomly presented for each participant. An equal number of experimental and filler items were presented in each list. Instructions were repeated at the start of each block. Participants were given a short break between blocks to rest their eyes.

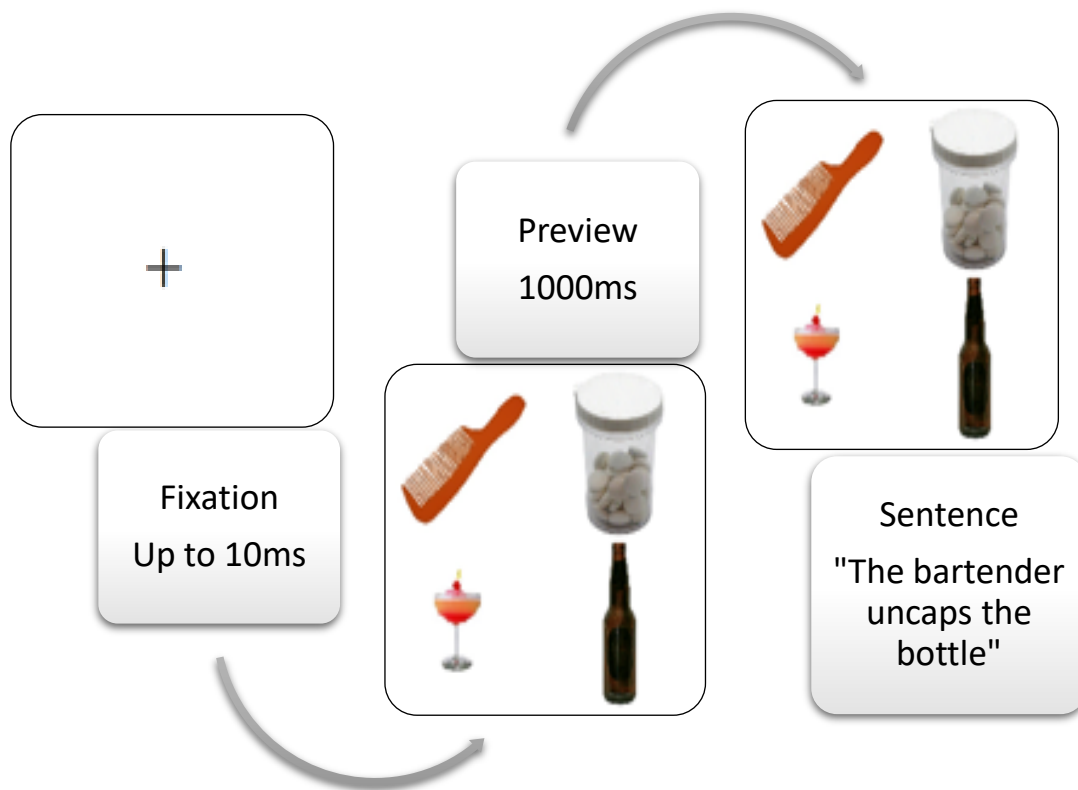


Figure 3 Procedure.

Predictions. In the first chapter, we described how typical fillers of thematic roles affect the verb meaning in terms of denoted actions. We discussed the influence of the models of typical events and situations on sentence comprehension and how they influence the thematic fit between the agent and the patient roles. Perceptually specified agents are cues to the knowledge of situations in which they typically appear, the events in which they are usually engaged and the information about the entities with which they commonly interact. Hence, we expect that listening to the agent, participants focus on the pictures of the objects typically present in the same situations: Target and Agent-Related.

The verb plays a crucial role in the individuation of the specific event in which the agent is engaged. Verb selectional restrictions allow to answer the questions “what is the agent doing?” and “which object could the agent use to do that?”. Hence, the verb constrains the set of entities previously cued by the agent to only those objects that can be used to perform the actions it refers to. Therefore, hearing the verb, participants’ attention should

focus on the object that typically appears in the situations in which also the agent is present and fits the verb selectional restrictions: Target.

The four pictures were selected with the following goals in mind. We assumed that the Target picture would attract the participants' attention for self-explaining reasons. Importantly, if the Target is the most fixated object before listening to the word that denotes it (perceptually underspecified patient), we can interpret the results as anticipatory eye movements. They are evidence of the incremental integration of information cued by the agent and the verb. However, the comparison with the other pictures is crucial in order to avoid plausible confounds. The Agent-Related picture plays a crucial role because it rules out the possibility that participants look at the Target only because of the knowledge of typical situations cued by the agents. Similarly, the Action-Related picture is motivated by the intention of investigating the interplay between linguistic information and the current visual context. The Action-Related picture represents an entity that, like the Target, fits the verb selectional restrictions and belongs to the set of the hyponyms of the same hypernym that fills the patient role. The presence of the Action-Related picture constrains the participants to focus on the object in the visual scene with the physical features congruent with the situation and the event denoted by the agent-verb pair. Therefore, if the integration of the knowledge about typical situations cued by the agent and the verb selectional restrictions supply enough multimodal information to identify both the nominal and the referential fillers of the patient role, the Target should remain the most fixated object even in the presence of another entity denoted by the same perceptually underspecified noun. We expect that hearing the perceptually underspecified patient, the Target remains the most fixated picture. Still, the Action-Related object should receive more attention than the Agent-Related object.

Analyses. The analyses concerned the proportions of eye fixations towards the Target, the Agent-Related, the Action-Related and the Unrelated pictures during the auditory time course. Before analyzing the data, all blinks and fixations to anywhere other than the images on the screen were removed. Fixations were then averaged over 10ms time bins that specified what proportion of the fixations within each bin was spent looking at each image. Then we split this information into time windows based on the sentence critical

times to conduct our analyses. We focused on the differences in proportions of eye fixations towards the four pictures during the agent, verb and patient time windows. The critical times were calculated based on the averages of their onsets and offsets. See Table 1, which shows the averaged onset and offset values. A silence of 456.8ms preceded the onset of the first article. A silence of 300ms followed the offset of the patient: the value 2523.7ms in Table 1. The first time window coincided with the agent (*seamstress, bartender, doctor, hiker, student*). The first article (*the*) was not included. See Table 1, which reports the agent onset averaged value 610.4ms. The second time window included the verb and the second article: *turn on the, uncap the, fill up the* and so forth. The second article was part of the second time window because it represented the anticipatory time window, namely the interval that followed the agent and preceded the critical word. The latter coincided with the third time window filled by the perceptually underspecified noun (hyponym) that occupied the patient position (*machine, bottle, backpack*). In what follows, these time windows will be designated as: agent time window, action time window, patient time window.

We recorded the proportions of eye fixations toward the pictures (Target, Agent-Related, Action-Related and Unrelated) and compared them. In the analyses, we referred to the quadrant of the screen containing each picture as Area Of Interest (AOI).

Time Windows	Onset*	Offset*	Duration*
Sentence	456.8	2823.7	2366.98
Agent	610.4	1200	589.68
Verb	1200	1898.8	698.81
Patient	1898.8	2523.7	624.91

*(ms)

Table 1 Critical time points of the auditory sentence time course.

The analyses were conducted with RStudio Version 1.1.463 (2009-2018). We fitted one Linear Effects Mixed Model (LMER) for each time window using the “lmer” function from the linear mixed-effects package “lme4” (Bates et al., 2015; Baayen et al., 2008; Barr et al., 2013). Linear Mixed Effects (LME) models are called also “hierarchical regression” or “multilevel regression”. LME method allows modelling the effects of items and participants simultaneously in a single analysis (Baayen et al. 2008, Barr et al. 2013). LME method is similar to multiple linear regression but its focus is on repeated measures

analyses and it includes terms that account for the variability above and beyond the experimental manipulations (Baayen 2008).

The four AOIs and the two lists were the fixed effects. We calculated two random slopes accounting for random effects: subjects and trials. Fixed and random effects remained stable for each model and during all the analyses conducted on the dataset. For each time window, we calculated the estimated means of proportions, the Standard Errors, the t-values, and the p-values of the AOIs comparisons.

3.4. Results

In this section, we report the analyses of the two lists. We describe the comparisons among the proportions of eye fixations toward the four AOIs (Target, Agent-Related, Action-Related and Unrelated) in the agent, the action and the patient time windows. Subsequently, we compare the proportions of eye fixations toward the four AOIs during the final silent interval. As we expected, there were no significant differences between the two lists. See Figure 29, Figure 32 and Figure 35 in the Appendix to compare the first and the second lists in the agent, the action and the patient time windows.

The successive subsections will present the following data:

- A table shows the comparisons of the proportions of eye fixations toward the four AOIs cumulatively considered. We collapsed the data of the two lists.
- A table shows comparisons of the proportions of eye fixations toward the four AOIs disentangling the data of the first and the second lists.

See Results in the Appendix for the analyses between the two lists and the data representations.

3.4.1 Agent Time Window

The agent time window begins at the agent onset and ends at the verb onset. See Table 1, which shows the onset and the offset averaged values: 610ms and 1200ms. The total duration was 589ms. Participants had already seen the visual scene that had appeared 1000ms before the onset of spoken sentences. We expected that participants would focus on the Agent-Related and the Target objects because they typically appear in the same situations as the agent.

List	Comparison	Estimate	SE	t-value	p-value
In 1 & 2	Target-Action Related	0.01	0.02	0.51	0.62
	Target-Agent Related	0	0.02	0.21	0.83
	Target-Unrelated	0.04	0.02	1.68	0.11
	Action Related-Agent Related	-0.01	0.02	-0.31	0.76
	Action Related-Unrelated	0.03	0.02	1.26	0.22
	Agent Related-Unrelated	0.03	0.02	1.56	0.13
	List1-List2	0.03	0.03	1.07	0.3

Table 2 Analyses in the two lists.

List	Comparison	Estimate	SE	t-value	p-value
1	Target-Action Related	0.03	0.03	0.87	0.39
	Target-Agent Related	-0.01	0.03	-0.49	0.63
	Target-Unrelated	0.07	0.03	2.39	0.03*
	Action Related-Agent Related	-0.04	0.03	-1.36	0.19
	Action Related-Unrelated	0.05	0.03	1.63	0.12
	Agent Related-Unrelated	0.09	0.03	2.97	0.01*
2	Target-Action Related	0	0.03	-0.16	0.88
	Target-Agent Related	0.02	0.03	0.79	0.44
	Target-Unrelated	0	0.03	-0.01	0.99
	Action Related-Agent Related	0.03	0.03	0.92	0.37
	Action Related-Unrelated	0	0.03	0.14	0.89
	Agent Related-Unrelated	-0.02	0.03	-0.76	0.45

* p-value < 0.05

Table 3 Analyses in the first and second lists.

The comparisons among the four AOIs have no statistical relevance in the agent time window. However, the estimated proportions of eye fixations show already some differences. By inspecting the data, we can notice that the Target and the Agent-Related

AOIs were more looked at than the Action-Related and the Unrelated AOIs. See the column “Estimate” in Table 2. Table 3 shows that the dissimilarities in the proportions of eye fixations toward the Target and the Agent-Related AOIs compared with the proportions of eye fixations toward the Unrelated AOI are statistically relevant in the first list: see the p-values 0.03 and 0.01.

The above data confirm the proportions of fixations we expected in the agent time window. They indicate more fixations toward the Target and the Agent-Related pictures than the Action-Related and the Unrelated pictures. The AOIs comparisons do not show statistical relevance at this stage of sentence processing. However, the fixations toward the Agent-Related AOI suggest the influence of the typical situation knowledge, like the fixations toward the Target, which imply an anticipatory sentence comprehension process.

3.4.2 Action Time Window

The action time window begins with the verb onset and ends at the offset of the second article, which is also the patient onset. See Table 1, which shows the onset and the offset averaged values: 1200ms and 1899ms. The total average duration was 699ms. We expected anticipatory eye-movements toward the Target AOI. Even though the Action-Related object fits the verb selectional restrictions, the knowledge cued by the agent should guide the participants' attention toward the object with the physical properties congruent with the current situation: Target.

List	Comparison	Estimate	SE	t-value	p-value
In 1 & 2	Target - Action Related	0.17	0.03	5.92	3.84e-06*
	Target - Agent Related	0.08	0.02	4.74	8.00e-05*
	Target - Unrelated	0.22	0.03	8.05	2.18e-08*
	Action Related - Agent Related	-0.09	0.02	-3.72	0.0010*
	Action Related - Unrelated	0.05	0.02	3.18	0.0016*
	Agent Related - Unrelated	0.14	0.02	6.19	1.29e-06*
	List1 - List2	0.05	0.02	1.85	0.0764

* p-value < 0.05

Table 4 Analyses in the two lists.

List	Comparison	Estimate	SE	t-value	p-value
1	Target - Action Related	0.23	0.04	5.69	6.78e-06*
	Target - Agent Related	0.13	0.02	5.15	2.81e-05*
	Target - Unrelated	0.31	0.04	8.01	2.40e-08*
	Action Related - Agent Related	-0.11	0.03	-3.14	0.0043*
	Action Related - Unrelated	0.08	0.02	3.52	0.0005*
	Agent Related - Unrelated	0.18	0.03	5.81	3.50e-06*
2	Target - Action Related	0.11	0.04	2.68	0.0130*
	Target - Agent Related	0.04	0.02	1.55	0.1339
	Target - Unrelated	0.13	0.04	3.38	0.0024*
	Action Related - Agent Related	-0.07	0.03	-2.13	0.0433*
	Action Related - Unrelated	0.02	0.02	0.97	0.3318
	Agent Related - Unrelated	0.09	0.03	2.94	0.0066*

* p-value < 0.05

Table 5 Analyses in the first and second lists.

The differences in the proportions of eye fixations toward the four AOIs have statistical relevance in the action time window. The Target turned out to be the most looked at AOI, followed by the Agent-Related AOI. The Action-Related AOI was less looked at as compared with the Agent-Related AOI, but more than the Unrelated AOI.

In detail, the high statistical relevance of the Target AOI stands in sharp contrast with the low statistical relevance of the other AOIs. See Table 4, which shows p-values of the comparisons with the Agent-Related (8.00e-05), the Action-Related (3.84e-06) and the Unrelated (2.18e-08) AOIs. The Agent-Related AOI was the second most fixated AOI after the Target: see Table 4 for the p-values of comparing with the Action-Related and the Unrelated AOIs: 0.0010 and 1.29e-06. Finally, the Action-Related AOI received more attention than the Unrelated AOI: see Table 4, which shows the p-value of 0.0016.

As expected, in the action time window, the Target AOI was the most looked at. The results show anticipatory eye movements toward the Target that mirror the incremental integration of the information cued by the agent-verb pair, namely the knowledge of typical situations and the verb selectional restrictions. The Target was indeed the only object in the visual scene that fits both the agent and the verb semantic constraints.

3.4.3 Patient Time Window

The patient time window begins with the second article offset and ends with the offset of the perceptually underspecified noun. See Table 1, which shows the onset and the offset averaged values 1899ms and 2523ms. The patient was filled by a perceptually underspecified noun referred to both the Target and the Action-Related AOIs. Still, we expected that the former would remain the most fixated AOI because of the knowledge of typical situations cued by the agent. The Action-Related AOI should receive more attention than the Agent-Related AOI because of the hypernym and the influence of the previously heard verb.

List	Comparison	Estimate	SE	t-value	p-value
In 1 & 2	Target - Action Related	0.35	0.04	8.29	1.65e-08*
	Target - Agent Related	0.27	0.04	6.96	3.39e-07*
	Target - Unrelated	0.40	0.04	10.67	1.18e-10*
	Action Related - Agent Related	-0.08	0.02	-3.36	0.0026*
	Action Related - Unrelated	0.05	0.02	3.10	0.0028*
	Agent Related - Unrelated	0.13	0.02	6.55	2.96e-07*
	List1 - List2	0.04	0.02	1.72	0.0978

* p-value < 0.05

Table 6 Analyses in the two lists.

List	Comparison	Estimate	SE	t-value	p-value
1	Target - Action Related	0.42	0.06	7.10	2.38e-07*
	Target - Agent Related	0.35	0.05	6.48	1.05e-06*
	Target - Unrelated	0.50	0.05	9.41	1.41e-09*
	Action Related - Agent Related	-0.07	0.03	-2.04	0.0523
	Action Related - Unrelated	0.08	0.02	3.24	0.0018*
	Agent Related - Unrelated	0.15	0.03	5.11	1.70e-05*
2	Target - Action Related	0.27	0.06	4.62	0.0001*
	Target - Agent Related	0.18	0.05	3.36	0.0026*
	Target - Unrelated	0.30	0.05	5.68	7.12e-06*
	Action Related - Agent Related	-0.09	0.03	-2.71	0.0121*
	Action Related - Unrelated	0.03	0.02	1.14	0.2582
	Agent Related - Unrelated	0.12	0.03	4.15	0.0002*

* p-value < 0.05

Table 7 Analyses in the first and second lists.

In the patient time window, the comparisons among the four AOIs have statistical relevance. In line with our expectations, the Target was the most looked at AOI. The Agent-Related received more fixations than the Action-Related AOI, which was more looked at as compared with the Unrelated AOI.

Table 6 shows the p-values of the Target AOI when compared with the Agent-Related ($3.39e-07$), the Action-Related ($1.65e-08$), and the Unrelated ($1.18e-10$) AOIs. The Agent-Related AOI received more attention than the Action-Related AOI. However, the statistical relevance of the comparison was lower than the action time window. See Table 6, which shows a p-value of 0.0026. The p-value of the same comparison in the anticipatory time window was 0.0010 (see Table 4). More exactly, in the first list, the dissimilarities in the proportions of eye fixations toward the Agent-Related and the Action-Related AOIs has no statistical relevance: see the p-value 0.0523 in Table 7.

As expected, the Target AOI was the most looked at as compared with the other AOIs. We interpret the statistical relevance of the comparison between the Agent-Related and the Action-Related AOIs as evidence of the interplay between the linguistic information encoded by the verb-patient pair and the visual information provided by the Action-Related AOI in the scene.

3.4.4 The Final Silent Interval

The pictures remained on the screen for 300ms after the auditory sentence was over. See Table 1, which shows the onset and the offset averaged values 2523ms and 2823.7ms. We compared the proportions of eye fixations toward the four AOIs during the final silence to investigate the interplay between information cued by the hypernym in the auditory sentence and information elicited by the current visual context. We expected the Action-Related AOI to be more fixated than the Agent-Related AOI because it is a plausible referent of the patient and fits the verb selectional restrictions like the Target AOI. However, the Target AOI should remain the most fixated AOI because of integrating the typical situation knowledge cued by the agent with the verb selectional restrictions.

List	Comparison	Estimate	SE	t-value	p-value
In 1 & 2	Target - Action Related	0.35	0.05	6.90	3.89e-07*
	Target - Agent Related	0.39	0.05	7.58	8.04e-08*
	Target - Unrelated	0.50	0.04	12.12	7.88e-12*
	Action Related - Agent Related	0.04	0.03	1.16	0.26
	Action Related - Unrelated	0.14	0.02	5.98	1.48e-06*
	Agent Related - Unrelated	0.11	0.02	4.63	5.82e-05*
	List1 - List2	0.02	0.01	1.69	0.10

* p-value < 0.05

Table 8 Analyses in the two lists.

List	Comparison	Estimate	SE	t-value	p-value
1	Target - Action Related	0.42	0.07	5.86	4.83e-06*
	Target - Agent Related	0.45	0.07	6.24	1.88e-06*
	Target - Unrelated	0.58	0.06	10.06	3.59e-10*
	Action Related - Agent Related	-0.07	0.03	-2.04	0.523
	Action Related - Unrelated	0.16	0.03	4.69	5.67e-05*
	Agent Related - Unrelated	0.13	0.03	3.96	3.93e-04*
2	Target - Action Related	0.28	0.07	3.90	6.76e-04*
	Target - Agent Related	0.33	0.07	4.48	1.56e-04*
	Target - Unrelated	0.41	0.06	7.08	2.31e-07*
	Action Related - Agent Related	0.04	0.04	0.98	0.34
	Action Related - Unrelated	0.13	0.03	3.77	7.10e-04*
	Agent Related - Unrelated	0.08	0.03	2.59	1.44e-02*

* p-value < 0.05

Table 9 Analyses in the first and second lists.

The dissimilarities in the proportions of eye fixations toward four AOIs have statistical relevance during the final silent interval. The Target AOI received more attention than the other AOIs. The Action-Related AOI was more looked at than the Agent-Related AOI. The Unrelated AOI was the least looked at.

In line with our predictions, the Target AOI remained the most fixated AOI. See Table 8 for the p-values of the comparisons with the Agent-Related (8.04e-08), the Action-Related (3.89e-07), and the Unrelated (7.88e-12) AOIs. Table 8 shows that the Action-Related AOI received more attention than the Agent-Related AOI: see the value 0.04 in the column “Estimate”. The comparison has no statistical relevance (0.26). Table 9 shows

that the comparisons between the Agent-Related and the Action-Related AOIs have no statistical relevance in either list: see the p-values 0.523 and 0.34.

As expected, the focus remained on the only object with the properties that best fit both the agent and verb semantic constraints: Target. However, as predicted, after listening to the perceptually underspecified noun filling the patient role, the participants turned their attention to the other object in the visual scene that the hypernym could have denoted: the Action-Related. The lack of statistical relevance in the comparison between the Action-Related and the Agent-Related AOIs witnesses the interplay between the linguistic information encoded by the sentence and the visual information depicted in the visual scene.

3.5. General Discussion

The eye-tracking experiment demonstrated that the typical situations knowledge associated with specific agents elicits multimodal expectations about the entities typically involved in them. In our study, this kind of expectations was represented by the Agent-Related and the Target AOIs. The participants rapidly integrated the expectations encoded by the agent and the verb. The semantic constraints encoded by the verb restricted the set of previously activated entities to the objects that fit its selectional restrictions. The result of the incremental knowledge integration led to anticipatory eye movements toward the Target AOI. The anticipatory eye movements are in line with the hypothesis that specific agent-verb pairs encode expectations about the referent filling of the incoming patient role. The perceptually underspecified noun in the patient position referred to the Target and Action-Related AOIs. However, there was no relationship between the Action-Related object and the situation knowledge elicited by the agent. Since the Target physical properties were congruent with the specific situation, it received more attention than the Action-Related object. The proportions of eye fixations toward the Target in the patient time window and during the final silent interval mirror the interplay between the linguistic information encoded by the auditory sentence and the extra-linguistic information derived from the current visual context during sentence comprehension.

Time Window	Comparison (Target)	Estimate	SE	t-value	p-value
Agent	Action-Related	0.01	0.02	0.51	0.62
	Agent-Related	0.00	0.02	0.21	0.83
	Unrelated	0.04	0.02	1.68	0.11
Action	Action-Related	0.17	0.03	5.92	3.84e-06*
	Agent-Related	0.08	0.02	4.74	8.00e-05*
	Unrelated	0.22	0.03	8.05	2.18e-08*
Patient	Action-Related	0.35	0.04	8.29	1.65e-08*
	Agent-Related	0.27	0.04	6.96	3.39e-07*
	Unrelated	0.40	0.04	10.67	1.18e-10*

* p-value < 0.05

Table 10 Comparisons among eye fixations proportions toward the Target AOI with respect to eye fixations proportions toward the other AOIs in the agent, the action and the patient time windows.

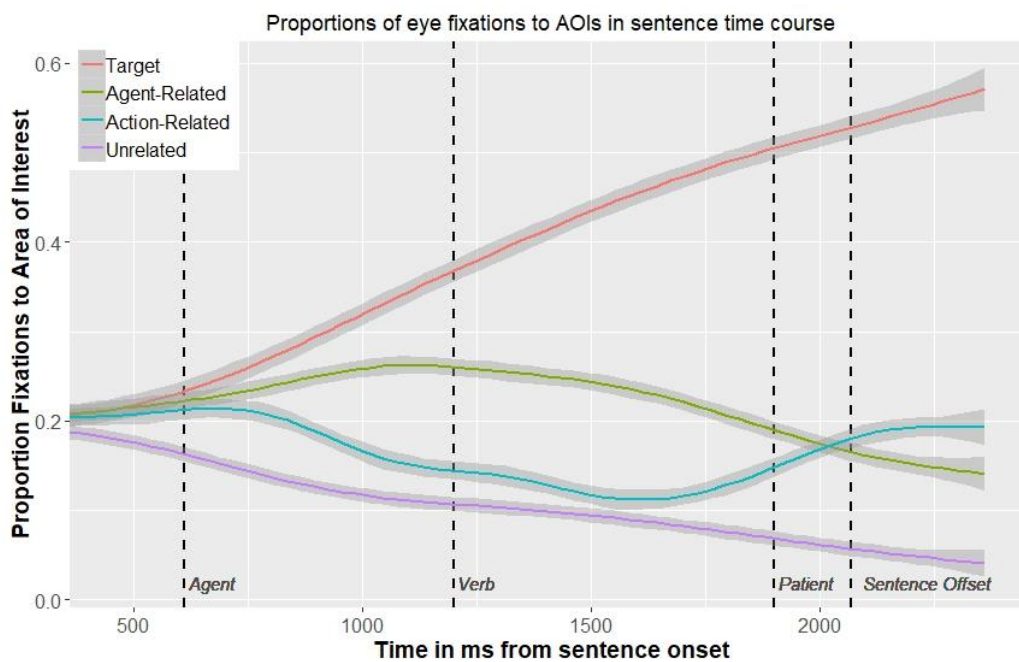


Figure 4 Time course of the AOIs eye fixations proportions. The plot shows the agent, the verb and the patient onsets, and the sentence offset.

Table 10 shows the comparisons among the proportions of eye fixations toward the Target AOI and the competitors (Action-Related and Agent-Related AOIs) as well as the distractor (Unrelated AOI) in the agent, the action and the patient time windows. It sums up the analyses shown in Table 2, Table 4 and Table 6 concerning the agent, action, and patient time windows in the two lists. Figure 4 shows the time course of the eye fixations

proportions toward the four AOIs while listening to the sentence. The onsets of the agent, the action and the patient time windows are indicated.

Listening to the agent, the proportions of eye fixations toward the Target AOI did not show statically significant differences with respect to the Action-Related (0.62), the Agent-Related (0.83) and the Unrelated (0.11) AOIs. However, Figure 4 shows that eye fixations proportions toward the four AOIs present already some dissimilarities. Approximately around the last 200ms of the agent time window, the two objects associated with the agent (Target and Agent-Related) were more looked at than the Unrelated and the Action-Related objects. The preview period allowed the participants to identify the depicted objects before the onset of the auditory sentence. Hence, the directions of the eye gaze were not random. The data confirm the hypothesis that specific agents cue the multimodal knowledge of the situations in which they typically appear, guiding the participants' attention toward the entities typically involved in them.

Table 10 shows that the Target was the most looked at AOI as compared with the Action-Related ($3.84e-06$), the Agent-Related ($8.00e-05$) and the Unrelated ($2.18e-08$) AOIs hearing the verb and the following article. The differences in the proportions of eye fixations toward the Target and the other AOIs have statistical relevance. The Target represented the correct referential filler of the patient role. The anticipatory eye movements toward the Target before hearing the critical word that denoted it, the perceptually underspecified noun, mirror the incremental integration of the typical situation knowledge cued by the agent with the verb selectional restrictions. Figure 4 shows that, during approximately the last 200ms of the anticipatory time window, the Action-Related AOI increasingly received more attention. The number of fixations toward the Agent-Related picture decreased. The results confirm the hypothesis that the verb selectional restrictions play a crucial role during sentence comprehension. The Target and the Action-Related objects were the only entities that fit the verb semantic restrictions. Thus, even if the Target continued to be the most looked at AOI because of the information cued by both the agent and the verb, the Action-Related AOI received a substantial share of fixations because of the verb influence.

The perceptually underspecified noun constitutes the last time window under investigation. The hypernym in the patient position allowed investigating the interplay between linguistic and visual information during sentence comprehension. The Action-

Related object was a competitor of the Target object because both fit the verb selectional restrictions and are hyponyms of the same hypernym. Table 10 reports that the Target AOI was the most looked at AOI as compared with the Action-Related ($1.65e-08$), the Agent-Related ($3.39e-07$) and the Unrelate ($1.18e-10$) AOIs. The comparisons have statistical relevance. However, Figure 4 shows that during the last milliseconds of the patient time window the Action-Related AOI received an increasing number of fixations. The Agent-Related AOI was instead less and less observed.

The visual world paradigm allowed us to explore the relationship between the information cued by words composing the auditory sentences and the referents depicted in the visual scene. We obtained empirical evidence that the event knowledge cued by lexical items is incrementally processed during comprehension, and it implies both linguistic and extra-linguistic information.

In detail, the multimodal knowledge of typical events concerns:

- a. The degree of coherence between the nominal filler of the agent role and the referent of the patient role, namely the multimodal thematic fit.
- b. The verb selectional restrictions, which determine the set of properties that identify a congruent referent.
- c. The link between the hypernym, a perceptually underspecified noun, and a referent representing the most suitable entity for the situation cued by the agent.

The results demonstrated that the multimodal knowledge about typical situations cued by the agent is incrementally integrated with the verb selectional restrictions. The resulting integration encodes information about the unmentioned patient, leading to the anticipation of its referential fillers.

Chapter 4: Multimodal Event Knowledge Model

In line with the purpose of the thesis, which concerns the study of verb meaning from a compositional and multimodal perspective, we modelled verb selectional restrictions based on typical event participants exploiting their linguistic and visual information. The disambiguation of the verb denotational meaning depends on linguistic and extra-linguistic information about the event constituents (see the first chapter). Physical properties of the entities that typically fill verb thematic roles represent crucial information about events. The shape, colour, texture or dimension of constituents are the information needed to disambiguate and define the actions that are part of an event. According to Elman and McRae (2019), the main dimensions by which event knowledge can be described concerns “the component pieces of activities that are part of an event”. Although the disambiguation of verb meaning depends on the involved entities' physical features, they have to be inferred most of the times. If the verb *fasten*, for example, co-occurs with the agent *handyman*, a person expects that the patient has the properties of an object that usually appears in the same situations in which also a handyman is present and can be fastened, such as a tool belt. When the agent of *fasten* is *driver*, the expectations concern a different kind of belt because the agent is a person who will go somewhere by driving a car. In this case, the patient role is likely filled by a referent that corresponds to a seat belt. The expectations encoded by words are constrained by the context and include extra-linguistic information (see Expectations in the first chapter). In sentences, hypernyms or perceptually underspecified nouns like *belt* tend to occur more frequently than *tool belt* or *seat belt* when anticipated by agents like *handyman* and *driver*. They indeed encode expectations about the situations in which they typically appear, including multimodal information about other involved entities that allow distinguishing between, for example, a tool belt and a seat belt even if denoted by the perceptually underspecified noun *belt*. See Perceptually Underspecified Nouns in the first chapter, which describes typical usages of hypernyms and hyponyms in sentences.

The eye-tracking experiment demonstrated the impact of the multimodal knowledge of typical situations and events on sentence comprehension (see General Discussion in the third chapter).

It can be described in terms of:

- a. The multimodal thematic fit between the noun filling the agent role and the referent of the patient role.
- b. The verb selectional restrictions.
- c. The link between a perceptually underspecified patient and an agent-congruent referent.

As for point (a), the degree of coherence between the fillers of the agent and patient roles is determined by linguistic (lexical) and extra-linguistic (visual perceptual) information. Thus, for example, *handyman* cues information about both the word *belt* (hypernym or perceptually underspecified noun) and the corresponding object tool belt. The agent *driver* elicits expectations that concern the noun *belt* and the referent seat belt. The thematic fit is based on expectations that mirror the knowledge of typical situations and events (see Multimodal Thematic Fit in the second chapter).

Sentence comprehension is an incremental process. As for point (b), the expectations encoded by the verb constrain the set of patient role fillers to the entities that can fit its selectional restrictions, leading to the specific event individuation. The verb *fasten*, for instance, tends to appear only with objects that can be fastened. However, the same verb can denote different activities based on the involved entities. A tool belt requires different actions to be fastened with respect to a seat belt. Fastening a tool belt implies actions like taking the end of the belt and pushing it through the frame of the buckle and, when the belt feels tight enough, push the prong through the closest hole at the end of the belt. Fastening a seat belt denotes actions like pulling the belt and inserting the buckle into the latching device until the click. See Verbs in Composition in the first chapter, which concerns the influence of event components on the disambiguation of verb denotational meaning. Multimodal thematic fit between the agent and patient roles leads to the individuation of the event constituents and, consequently, to the disambiguation of the performed actions.

In agreement with Grice's principle of quantity (1975), if the agent provides enough information for the individuation of the current situation, the patient can be expressed by a perceptually underspecified noun to avoid redundancy. However, the individuation of the referent suitable for the current situation mirrors the comprehension of the correct meaning of the sentence. For instance, if a person listening to the sentence *The handyman fastens the belt* thinks of a seat belt, she did not understand its meaning correctly. As for point (c), the relation between a perceptually underspecified noun and the referent consistent with the situation cued by the agent is crucial in sentence comprehension. The multimodal thematic fit implies the link between the perceptually underspecified noun and the most suitable referent for the current situation.

In line with the eye-tracking experiment, we focused on the hypothesis that sentences composed of a specific agent, a verb and a perceptually underspecified patient encode expectations about the referent of the patient role. The integration of typical situation knowledge cued by the agent with verb selectional restrictions makes up for unmentioned information about the patient, namely the extra-linguistic properties that distinguish a particular referent from the other entities denoted by the same hypernym. We propose a computational model of typical events multimodal knowledge which reproduces the expectations cued by words during sentence comprehension: Multimodal Event Knowledge (MEK). The model aims at mirroring the eye-tracking experiment. The participants listening to sentences composed of a specific agent, a verb and a perceptually underspecified patient individuated the patient role referent among four pictures differently related to the event described by the auditory stimulus (see Sentences, Pictures and Lists in the previous chapter). People's capacity to identify the referent of the patient even when it is expressed by a perceptually underspecified noun derives from their experience of real-world situations, which most of the time implies first-person experiences and interactions with other involved entities.

MEK predicts the image of the referent filling the patient role of a textual event. The model learns the internal structure of typical events, including the multimodal relations between their constituents (a, b and c), from sequences of activities. MEK infers unmentioned multimodal information about the patient and incrementally deals with information coming from the environment reproducing multimodal expectations exploited by people during sentences comprehension.

In this chapter, we will present the model MEK. In particular, the description of the model design focuses on what makes MEK a reproduction of the eye-tracking experiment. In what follows, we introduce the model architecture illustrating its input, hidden and output layers. In detail, we will introduce the simulations of real-world scenarios through which people experience situations and events.

The section Processing Dynamics and Training will include:

- The processing dynamics exploited by the model to learn the internal structure of the typical events encoded by the sequences.
- The explanation of how MEK distinguishes between textual and visual information.
- The numbers of agents, verbs and patients appearing in the sequence collection.
- The salient information about the set up of the model during training.

A description of the model evaluation precedes the results. The evaluation refers to the eye-tracking experiment analyses. We proposed different types of input to evaluate MEK. In particular, we evaluated if the model predicts the correct referent of the patient role given a textual event (agent, verb and perceptually underspecified patient) in the input. Moreover, we tested if MEK learnt multimodal thematic fit between the agent and patient roles, verb selectional restrictions, and the relationship between perceptually underspecified nouns and their referents. The event, agent and agent-verb pair inputs mirror the time windows analyzed in the eye-tracking experiment. Hence, we report and discuss the results.

4.1. Model Design

MEK is a computational model of multimodal expectations about typical situations and events cued by words. During comprehension, the expectations cued by words depend on the multimodal knowledge of typical situations, which, in turn, affects the inferences about incoming words in sentences (see General Discussion in the third chapter). MEK learns the internal structure of typical events and, relying on multimodal thematic fit between the agent and patient roles and verb selectional restrictions, predicts the picture of the referent denoted by the perceptually underspecified noun (hypernym) filling the patient role. As described in Perceptually Underspecified Nouns in the first chapter, hypernyms in isolation do not entail a specific type of perceptual referent. They do not cue fine-grained knowledge about situations in which typically they appear, events in which they are usually engaged and entities with which they commonly interact. In line with the hypothesis that specific agent-verb pairs cue multimodal expectations that lead to the anticipation of the referent filling the patient role even when a hypernym denotes it, MEK predictions depend on the agent-verb pairs.

According to Elman and McRae (2019), an event knowledge model should be able to perform "pattern completion" along two dimensions: across time and in the moment. MEK can predict the referent of the incoming patient role given an agent-verb pair. Moreover, when the input is a textual event (agent, verb and perceptually underspecified patient), the model infers the referent of the patient modelled as an image.

The learning process does not include a priori definitions or templates of typical events. MEK extracts regularities in the data about agents, verbs, patients and the referents of the patient role. It incrementally learns the relationships that make an event typical, namely its internal structure: the degree of coherence between the agent and patient roles, the verb selectional restrictions, and the relationship between the perceptually underspecified patient and its referent.

The data used to train the model are simulations of real-world scenarios in which people experience situations and events. We created linguistic sequences composed of a specific agent, a verb and a perceptually underspecified patient. They stand for descriptions of typical events and derive from a subset of the visual world experiment sentences. See Sequence Collections in the Appendix for a complete list of the sequential data. We aimed to represent all the links between the event denoted by the auditory sentence and the four

pictures in the visual scene (See Sentences, Pictures and Lists in the third chapter): Target, Agent-Related, Action-Related and Unrelated.

To reproduce the Agent-Related relationship, we had to create sequences in which the Agent-Related object fills the patient role.

Therefore, the data consist of:

- Events in which a Target object fills the patient role, like *handyman fasten belt*.
- Events in which an Agent-Related object fills the patient role, such as *handyman grab screwdriver*.

We extracted the stimuli of the eye-tracking experiment from both the first and the second lists. Hence, the Action-Related relationship depends on the sequences composed of the same verb-patient pair but different agents, such as *handyman fasten belt* and *driver fasten belt*. The former is associated with the tools belt picture; the latter is linked to the seat belt picture. The tools belt is the Action-Related object of *driver fasten belt*. The seat belt is the Action-Related object of *handyman fasten belt*.

In what follows, we describe how we modelled the training data to provide the model with a representation of the four types of sentence-picture connections (Target, Agent-Related, Action-Related, Unrelated) exploited in the eye-tracking experiment.

Target. When a Target object fills the patient role, each agent-patient pair of the auditory sentence subset of the eye-tracking experiment appears with three different verbs in the training data.

(17)

- a. *The gardener fills up the pot*
- b. *Gardener grab pot*
- c. *Gardener fill pot*
- d. *Gardener empty pot*

(17a) is the auditory sentence proposed in the visual world experiment while on the screen appeared the pictures of a plant pot (Target), a cooking pot (Action-Related), a rake (Agent-Related) and a mirror (Unrelated). (17b-d) are the sequences used to train MEK.

The agent-patient pair *gardener-pot* (17a), co-occurs with three different verbs: *grab*, *fill* and *empty* (17b-d).

The picture linked to the training sequences represents a plant pot:



Figure 5 Plant pot.

Agent-Related. In the eye-tracking experiment, the sentences did not consist of lexical items that denoted Agent-Related entities. The agent was the only link to the Agent-Related pictures. A subset of sequences had to include words referring to the Agent-Related entities to represent this type of relation in MEK training data. The model, indeed, can associate the agent with its Agent-Related object exploiting their occurrence in the same sequence. The Agent-Related objects in the patient position support the hypothesis that an agent can be associated with many entities typically present in the same situations: what we defined as the multimodal thematic fit.

In the eye-tracking experiment, the Agent-Related objects did not fit the verb selectional restrictions. Hence, in the MEK training data, the verbs that appear with agent-Target pairs (*gardener-pot* in (17)) are different with respect to the verbs co-occurring with the pairs composed of the same agent and the Agent-Related object (*gardener-rake* in (18)). Using different verbs based on the Target and the Agent-Related conditions will allow us to reproduce the verb selectional restrictions effect and, consequently, the incrementality inherent to the sentence comprehension process.

The information cued by the Agent-Related picture in the visual world experiment were provided to the model through sequences in which the Agent-Related object filled the patient role.

Here follows a list of sentences which illustrate the above indications.

This subset of events includes two sequences for each type of object.
(18)

- a. *Gardener clean rake*
- b. *Gardener hold rake*

(18) are linked to a rake picture:



Figure 6 Rake.

Action-Related. The collection includes sequences like
(19)

- a. *Cook grab pot*
- b. *Cook fill pot*
- c. *Cook empty pot*

(19) is linked to the picture of a cooking pot:









Figure 7 Cooking pot.

Sequences like (19) and (17b-d) present the same verb-patient pair, *fill/grab/empty-pot*, but different agents: *gardener* in (17b-d) and *cook* in (19).

They represent the connection among the perceptually underspecified noun in the auditory sentences, the Target and the Action-Related pictures of the eye-tracking experiment data. In particular, the connection corresponds to the relationship between a hypernym and the set of its possible instances (see Sentences, Pictures and Lists in the third chapter). Both referents, plant pot in Figure 5 and cooking pot in Figure 7, are denoted by the same perceptually underspecified noun *pot* and fit the verb selectional restrictions (*fill, grab, empty*).

The presence of two referents denoted by the same word should help MEK recognise the agent's influence on the hypernym-referent relation in typical events. *Pot* denotes a plant pot when the agent is *gardener* (17) and a cooking pot if *cook* fills the agent role (19). See Multimodal Thematic Fit in the second chapter, which describes how a thematic role already filled (the agent) elicits multimodal expectations about how an incoming role will be filled (the patient).

Unrelated. The subset of pictures that never appeared with the components of a sequence represents the Unrelated objects to the denoted event. In the eye-tracking experiment, the participants saw four pictures for each auditory sentence, of which only one Unrelated to the described event. MEK has to predict the object denoted by the perceptually underspecified noun filling the patient role, selecting it from a set that includes all the pictures provided during the training stage. Hence, Unrelated objects correspond to the subset of pictures that did not occur with any input sequence components.

Text-Picture Relationship	Agent	Verb	Patient	Referent
Target		Grab	Pot*	
		Fill		
		Empty		
Agent-Related	GARDENER	Hold	Rake	
		Clean		
Action-Related				
Target		Grab	Pot*	
		Fill		
		Empty		
Agent-Related	COOK	Wear	Apron	
		Adjust		
Action-Related				

* Perceptually Underspecified Noun or Hypernym

Table 11 Text-Picture relationships in MEK training data.

4.2. Simulations and Architectures

In this section, we present two simulations of real-world scenarios where people experience situations and events. The simulations are collections of sequences describing the same events. They differ in the information included in each sequence. The simulations correspond to two methods through which we trained MEK on the multimodal knowledge of typical events. Depending on the method, the input in the training stage changes.

MEK is a Long Short-Term Memory (LSTM) neural network model (Hochreiter and Schmidhuber 1997) that predicts the referent of the patient role given an event as input. An LSTM network is a special kind of Recurrent Neural Network (RNN) because it can use recorded knowledge about previous events to inform later ones exploiting, differently from traditional RNNs, long-term dependencies between the data (Bengio et al. 1994). We decided to use an LSTM neural network model because MEK has to learn the internal structure of typical events extracting event components relationships from different sequences of activities.

Look At That. We called the first simulation LookAT (Look At That). In LookAT, the perceptually underspecified noun is followed by the picture of its referent. Hence, the noun and the referent filling the patient role appear in the same sequence.

(20)

- a. *Doctor open bottle* PILLS BOTTLE
- b. *Bartender open bottle* BEER BOTTLE

When the agent is *doctor*, the word *bottle* co-occurs with the picture PILLS BOTTLE (20a). When the agent is *bartender*, the perceptually underspecified noun *bottle* is followed by the picture BEER BOTTLE (20b).

LookAT simulates a real-world scenario in which linguistic information coexists with the corresponding entities, which are part of the extra-linguistic context. The scenario consists of a speaker who, looking at something or pointing it with a finger, makes the listener focus on a particular object involved in the event the speaker is talking about. See Human and Machine Learning in the second chapter, which explains the strategies exploited by people to learn new meanings. This condition is similar to the learning

method exploited by Lazaridou et al. (2015) for constructing the Multimodal Skip-gram model. The real entities denoted by some lexical items are present in the current environment.

The collection includes 144 sequences in which Targets fill the patient role and 96 sequences in which Agent-Related referents occur in the patient position. LookAT is constituted of 240 events.

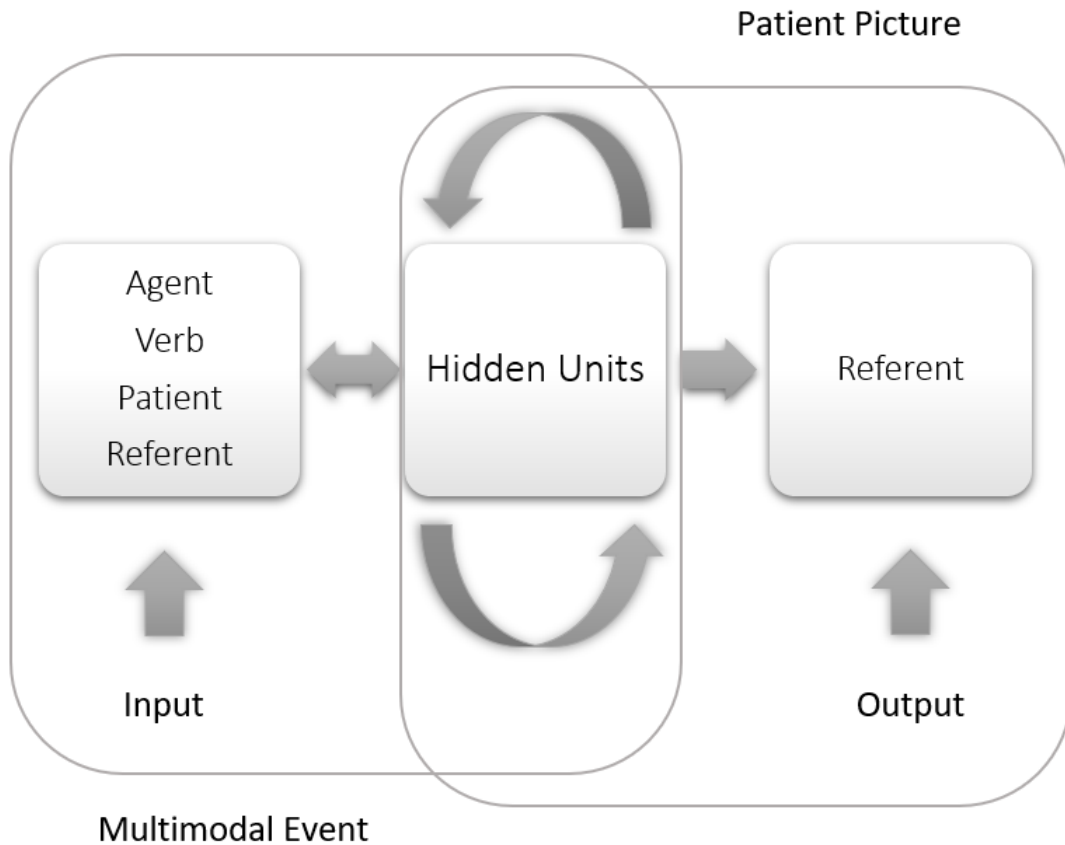


Figure 8 LookAT Architecture.

Figure 8 shows the input-output combination during the training of MEK on multimodal knowledge of typical events using the LookAT sequences. The model receives in the input an event constituted of an agent, a verb, a perceptually underspecified patient and its referent. MEK processes the information encoded in the sequences in its hidden units and provides in the output the picture of the referent filling the patient role. Each perceptually underspecified noun in the patient position co-occurs in the training data

with two or more referents. Therefore, the correct prediction mirrors multimodal thematic fit between the agent and the referent of the patient role, and verb selectional restrictions.

Who Did the Action? The second simulation was defined as WhoAct (Who Did the Action). In WhoAct, the same agent-verb pair is followed by the perceptually underspecified noun or the referent filling the patient role. The agent-verb pair appears with both the noun and referent but in different sequences.

In LookAT, MEK links each perceptually underspecified noun to a referent based on their co-occurrence. In WhoAct, the model would learn the relationship between the name of a class of objects (hypernym) and its instances relying on the agent-verb pair that precedes them. Hence, MEK has to individuate the relationship between the perceptually underspecified noun and its referent between sequences.

(21)

- a. *Doctor open bottle*
- b. *Doctor open PILLS BOTTLE*
- c. *Bartender open bottle*
- d. *Bartender open BEER BOTTLE*

In (21) *bottle* occurs with both *doctor-open* (21a) and *bartender-open* (21c) pairs. However, *doctor-open* appears also with the picture PILLS BOTTLE (21b), *bartender-open* occurs with the picture BEER BOTTLE (21d). MEK has to learn the hypernym-referent relation (*bottle-PILLS BOTTLE/BEER BOTTLE*), exploiting the co-occurrence with the same agent-verb pair (*doctor-open* and *bartender-open*).

WhoAct mixes two real-world scenarios where people experience situations and events. In (21a), the referent pills bottle is denoted by the word *bottle*, but the object is absent in the extra-linguistic environment. In (21b), the referent pills bottle is present, but its name is not pronounced.

(21a) and (21c) correspond to a person who reads a book, a newspaper, or listens to the radio. Words are the cues to the multimodal information linked to mental models of typical situations. The agent-verb pair provides the information needed to link the perceptually underspecified noun filling the patient position to its referent.

(21b) and (21d) correspond to a scenario where the speaker, after pronouncing the agent and the verb, points the finger toward the object filling the patient role without pronouncing its name. The listener sees the object, but he does not hear its name. The agent-verb pair justifies the presence of the object in the current extra-linguistic environment. This condition recalls the proportions of eye fixations toward the Target AOI in the action time window of the eye-tracking experiment. The participants looked at the Target AOI before hearing the perceptually underspecified noun that should have denoted it. The anticipatory eye movements reflect the multimodal expectations encoded by the agent-verb pair (see Action Time Window in the third chapter).

The collection includes 288 sequences with Target referents and 192 events in which the Agent-Related entities fill the patient role. There are 240 different events, but the total of sequences is 480 because two descriptions are provided for each event: textual sequences, like (21a) and (21c); multimodal sequences, like (21b) and (21d).

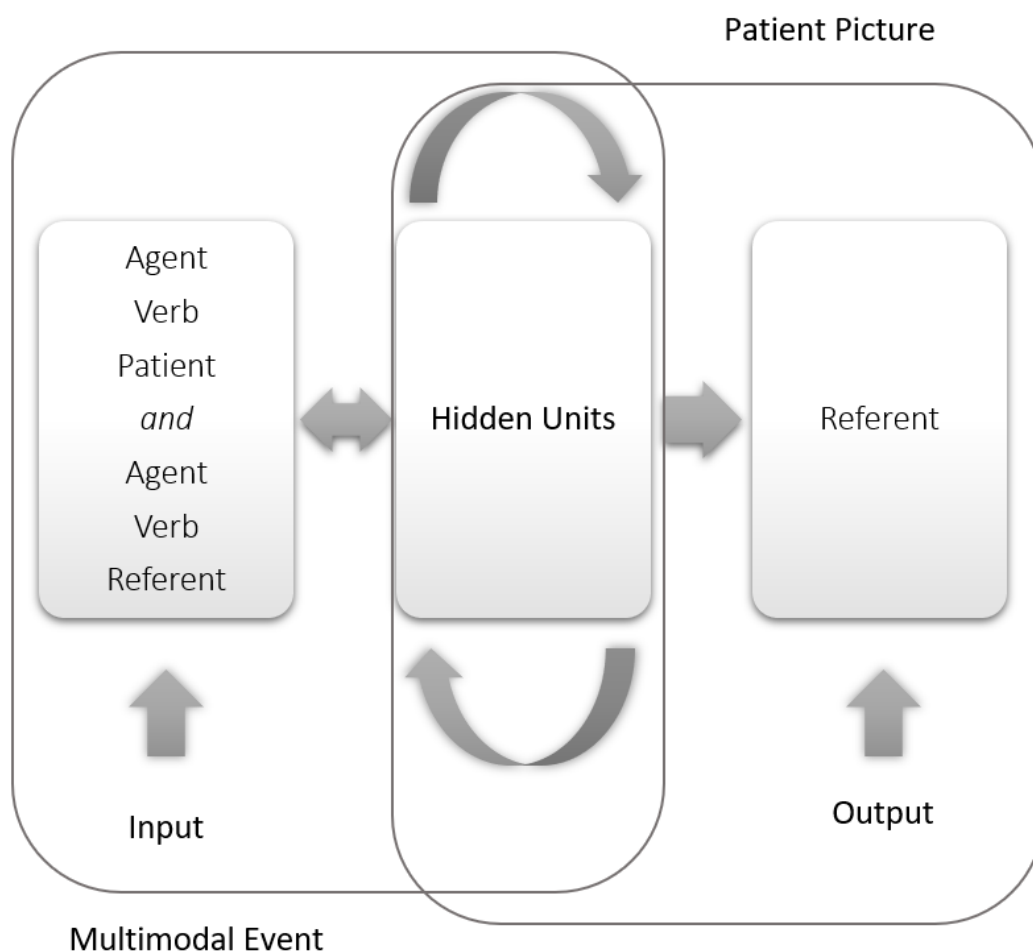


Figure 9 WhoAct Architecture.

Figure 9 shows the input-output combination during the training of MEK on multimodal knowledge of typical events exploiting the WhoAct sequences. The input consists of two descriptions of the same event (21a and (21b) distinguished based on the information included in the sequence. The textual description is constituted of an agent, a verb and a perceptually underspecified noun in the patient position. The multimodal description comprises a textual agent and a verb followed by the referent of the patient role. MEK processes the information encoded in the sequences in its hidden units and predicts the picture of the referent filling the patient role of both the textual and multimodal events. MEK does not see the referent of the patient role in textual descriptions (*doctor open bottle* (21a) vs *doctor open PILL BOTTLE* (21b)). Its predictions depend on the agent-verb pair (*doctor-open*) that co-occurs with both the perceptually underspecified noun (*bottle*) and the picture (PILL BOTTLE). Moreover, a perceptually underspecified noun co-occurs with two or more agent-verb pairs (*doctor-open* in (21a) and *bartender-open* in (21c)). Thus, the correct predictions mirror multimodal thematic fit between the agent and the referent of the patient role and verb selectional restrictions.

4.3. Processing Dynamics and Training

In this section, we describe what we intend for the internal structure of a typical event. A general description of the training sequence collection follows. We then explain how MEK distinguishes between textual and visual information. Finally, we illustrate the model settings during the training.

Typical Event Structure. MEK learns multimodal knowledge of typical events from sequences of activities composed of an agent, a verb and a patient. The latter is expressed by a perceptually underspecified noun or a picture or both.

(22)

- a. *Quarterback throw ball* FOOTBALL BALL
- b. *Shortstop throw ball* BASEBALL BALL

In (22) FOOTBALL BALL and BASEBALL BALL (Figure 10) are the pictures of the referents of the perceptually underspecified noun (hypernym) *ball* filling the patient role in the sequences.



Figure 10 Football ball and baseball ball.

The relationships between the hypernym (*ball*) and its two possible referents corresponding to different subtypes of balls (football ball and baseball ball) depend on *quarterback* and *shortstop*. Recording the co-occurrences between the event components in (22), we expect MEK to find out the links among:

- *Quarterback/shortstop* and *ball*, namely the thematic fit between the nouns filling the agent and patient roles.
- *Quarterback-FOOTBALL BALL* and *shortstop-BASEBALL BALL*, which corresponds to the multimodal thematic fit between the noun filling the agent role and the referent of the patient.
- *Ball* and FOOTBALL BALL/BASEBALL BALL, which is the relationship between the perceptually underspecified noun and its referents.

MEK learns that the agent of *throw* can be *quarterback* (22a) or *shortstop* (22b). The patient can be expressed by the same perceptually underspecified noun (*ball*), but the visual information associated with the event change depending on the agent.

Sequence Collection in Numbers. Training data includes 43 agents. *Hiker, chef, swimmer* and *cyclist* appear with more than one Target and Agent-Related referents.

(23)

- a. *swimmer wear suit* BATHING SUIT
- b. *swimmer try suit* BATHING SUIT
- c. *swimmer adjust suit* BATHING SUIT
- d. *swimmer wear goggles* SWIMMING GOGGLES
- e. *swimmer loosen goggles* SWIMMING GOGGLES
- f. *swimmer tighten goggles* SWIMMING GOGGLES

- g. *swimmer grab slippers* BATHING SLIPPERS
- h. *swimmer remove slippers* BATHING SLIPPERS
- i. *swimmer fold towel* TOWEL
- j. *swimmer spread towel* TOWEL

BATHING SUIT and SWIMMING GOGGLES in (23 a-f) are the Target objects associated with *swimmer*. BATHING SLIPPERS and TOWEL in (23 g-j) correspond to the Agent-Related objects.

There are 48 Target and 48 Agent-Related objects for a total of 96 referents filling the patient role. The perceptually underspecified patients in the Target condition are 23 because each hypernym refers to two referents: *bottle-beer bottle/pills bottle, backpack-hiking backpack/school backpack*⁶. See Action-Related in Model Design and Sentences, Pictures and Lists in the third chapter that describe how we reproduced the hypernym-referent relation in MEK training data and the eye-tracking experiment, respectively.

⁶ *Box* refers to toolbox, fuse box, ring box and mail box.

There are 35 verbs in the collection. The verbs that co-occur with the Targets (24) differ from the verbs that appear with the Agent-Related (25) referents associated with the same agent (see Model Design).

(24)

- a. *Quarterback grab ball* FOOTBALL BALL
- b. *Quarterback hold ball* FOOTBALL BALL
- c. *Quarterback throw ball* FOOTBALL BALL
- d. *Shortstop grab ball* BASEBALL BALL
- e. *Shortstop hold ball* BASEBALL BALL
- f. *Shortstop throw ball* BASEBALL BALL

(25)

- a. *Quarterback wear helmet* FOOTBALL HELMET
- b. *Quarterback adjust helmet* FOOTBALL HELMET
- c. *Shortstop shake bat* BASEBALL BAT
- d. *Shortstop clean bat* BASEBALL BAT

The sequences in (24) include agent-Target pairs: *quarterback*-FOOTBALL BALL and *shortstop*-BASEBALL BALL. The same verb-patient pairs (*grab-ball*, *hold-ball*, *throw-ball*) co-occur with different agents, *quarterback* and *shortstop*.

FOOTBALL HELMET and BASEBALL BAT are the Agent-Related objects. The verbs in (24) are different from those in (25): *grab*, *hold*, *throw* versus *wear*, *adjust*, *shake*, *clean*.

In the Target condition 18 verbs appear. Each verb co-occurs with eight different agent-patient pairs. There are 144 different events in the Target condition.

The Agent-Related condition includes 31 verbs. Each type of pair composed of the agent and the Agent-Related referent occurs with two different verbs (25a-d) for a total of 96

events.⁷ See Sequence Collections in the Appendix, which reports a complete list of the sequences exploited to train MEK.

Textual and Visual Representations. We did not provide MEK with a priori knowledge about the internal structure of a typical event. The model had to extract it case-by-case from sequences of activities that include multimodal information. The latter consists of textual and visual information. Textual information regards agents (*quarterback*, *shortstop*), verbs (*grab*, *hold*, *throw*) and perceptually underspecified nouns (*ball*) in the patient position. Visual information concerns referents of the patient role (FOOTBALL BALL and BASEBALL BALL). Depending on the structure of the sequences (LookAT or WhoAct) and the representations (textual or visual) of their components, MEK learnt the multimodal relations that constitute a typical event: the thematic fit between the agent and patient roles, verb selectional restrictions and the relationship between the perceptually underspecified noun and its referents. In detail, the first step of MEK is associating each sequence constituent with a vector representation, which indicates if the component is a word or a picture.

We used two types of textual representations: Word2vec (Mikolov et al. 2013) vectors⁸ and GloVe⁹ (Pennington et al. 2014) pre-trained vectors.

We created the visual representations exploiting GoogLeNet (Szegedy et al. 2015) and AlexNet (Krizhevsky et al. 2012) Convolutional Neural Networks (CNNs)¹⁰. We choose

⁷ All the verbs in the Target condition also appear with Agent-Related referents associated with other agents, except for *try*, *fasten*, *loosen*, *tighten*. The verbs that appear only in Agent-Related condition are *turn*, *shake*, *roll*, *fold*, *strand*, *moor*, *hit*, *kick*, *ride*, *exhibit*, *paste*, *lick*, *run*, *garnish*, *spread*, *collect*, *play*.

⁸ For the creation of Word2vec vectors, we used the EnWik9 text corpus and the Skip-gram algorithm.

⁹ <https://nlp.stanford.edu/projects/glove/>

¹⁰ The CNNs AlexNet (2010) and GoogLeNet (2014) were evaluated in the IMAGENET Large Scale Visual Recognition Challenge (IMAGENET LSVRC) context. IMAGENET LSVRC (Russakovsky et al. 2015) is a competition to estimate the content of photographs for retrieval and automatic annotation. A sub-set of 1000 categories of the hand-labelled ImageNet dataset (Deng et al. 2009) is used for the evaluation. The categories on which the CNNs AlexNet and GoogLeNet were evaluated correspond to hypernyms such as *backpack*, *glove* and *bottle*. However, the visual vectors encode the physical properties of hyponyms like *hiking backpack* and *school backpack*, *baseball glove* and *boxing glove*, *beer bottle*, and *pills bottle* in MEK. Therefore, we implemented the transfer learning method (using a MATLAB algorithm) to obtain the hyponyms visual vectors. The collection of pictures used to perform the transfer learning derives from

GoogLeNet and AlexNet CNNs¹¹ because of the results obtained by Rotaru and Vigliocco (2019) in comparing them with other models in the task of predicting subjective similarity/relatedness ratings.

We created two matrices distinguished on the basis of the included representations. Word2vec textual vectors and AlexNet visual vectors constitute the matrix defined “WA”. We called “GG” the matrix composed of GloVe and GoogLeNet vectors.

Training. MEK was created using the Application Programming Interface (API) of Keras¹². The model has a hierarchical structure composed of three layers: embedding, Bidirectional-LSTM (BLSTM) and prediction. BLSTM is firstly trained on the input sequence as it is and, secondly, on its reversed copy. This method should provide additional context to the model leading to a better learning on the task (Goodfellow, Bengio and Courville 2016).

During the training and the testing stages, the data were randomly split into subsets through the cross-validation method¹³ to avoid overfitting¹⁴. The dropout¹⁵ (Srivastava et

Google Image Search Engine. According to Fergus et al. (2005), Google images are competitive hand prepared datasets for training object recognition.

¹¹ See Visual Representations Accuracy in the Appendix, which concerns the accuracy of the CNNs in predicting the correct label for each picture.

¹² <https://keras.io/>

¹³ https://scikit-learn.org/stable/modules/cross_validation.html

¹⁴ The overfitting phenomenon corresponds to a situation in which both learning and testing stages exploit the same prediction function parameters. The model repeats the classes of the already seen samples leading to a perfect score that does not mirror its real learning because if it was provided with new yet-unseen data, it could fail to predict the correct classes.

¹⁵ The dropout randomly selects a set of neurons that have to be ignored (dropped-out) during training. Their contribution to the activation of downstream neurons is temporally removed on the forward pass, and any update of the weights are not applied to the neuron on the backward pass. The neural network learning process implies that each neuron weight settles into its context and is tuned for specific features that make it specialized. Neighbouring neurons rely on this specialization. This neighbourhood relationship can lead to a lack of generalization and too much specialization on training data (overfitting). Since dropout ignores some neurons randomly during training, the remaining neurons have to make predictions for the missing neurons, leading to multiple independent internal representations learning. The resulting network is less sensitive to the specific weights of the neurons, and its performance is a photograph of how the network generalized on the data.

al. 2014) is a regularization technique that avoids overfitting like the cross-validation method. We set the dropout¹⁶ parameter to 0.2: the nodes are randomly selected to be dropped-out with a probability of 20% each weight update cycle. We exploited the Adam optimization algorithm¹⁷, an extension to stochastic gradient descent¹⁸ that updates network weights iterative based on training data (Diederik and Ba 2015; Ruder 2016).

MEK was created to handle a multi-class classification predictive modelling problem. The model predicts an object image for each event in the input, selecting it from a set that includes all the pictures provided during the training stage, namely 96. Hence, we exploited the categorical cross-entropy loss function.¹⁹ It calculates a score that summarizes the average difference between the actual and the predicted probability distributions for all classes in the problem.²⁰

During the training stage, MEK sees each event ten times (Elman and McRae 2019). It reflects the assumption that usually, a person experiences the same event or similar events many times during her life span. The repetition of the number of times MEK processes the same event should improve its performance in memorizing the semantic relations among the event components and inferring the correct referent of the patient role.

After training, the weights encoding information that MEK derived from the multimodal descriptions of typical events were frozen and used to evaluate the model. The main task of MEK is predicting the referent filling the patient role of a textual event. It aims at simulating the eye-tracking experiment task. MEK was evaluated using different events as compared with events used to train it. The evaluation involved also different input types in order to check if the model learnt multimodal thematic fit between the agent and patient roles, verb selectional restrictions and the relationship between a perceptually

¹⁶ https://keras.io/api/layers/regularization_layers/dropout/

¹⁷ <https://keras.io/api/optimizers/adam/>

¹⁸ Since neural networks are trained using a stochastic gradient descent optimization algorithm, the error of the current state of the model is estimated repeatedly. The error function is commonly called the loss function, and it has the aim to estimate the loss of the model to update the weights reducing the loss on the next evaluation. The loss function must be appropriated to the mapping between the network inputs and outputs, namely the specific predictive modelling problem.

¹⁹ https://keras.io/api/losses/probabilistic_losses/#categorical_crossentropy-function

²⁰ The score is minimized, and the perfect cross-entropy value is 0. The softmax activation predicts the probability for each class.

underspecified noun and its referents. They indeed mirror the multimodal expectations cued by words during sentence comprehension and entail the multimodal knowledge of an event.

4.4. Evaluation

In this section, we illustrate the types of input we provided to the model to evaluate it and explain what they represent.

Event. MEK predicts the referent of the patient role, having received in input a textual event composed of a specific agent, a verb and a perceptually underspecified patient. MEK selects the referent (Target) from a collection of pictures that includes Agent-Related, Action-Related and Unrelated objects to the event. There are 96 pictures in the set. See Model Design, which describes the relations included in the training data.

The input corresponds to the patient time window of the eye-tracking experiment. Listening to the perceptually underspecified patient, since the participants had already heard the agent and the verb, they focused the attention on the object that fits both the agent and verb semantic constraints: Target. The latter represents an object usually present in the situations in which the agent typically appears and fits the verb selectional restrictions. See Patient Time Window in the third chapter.

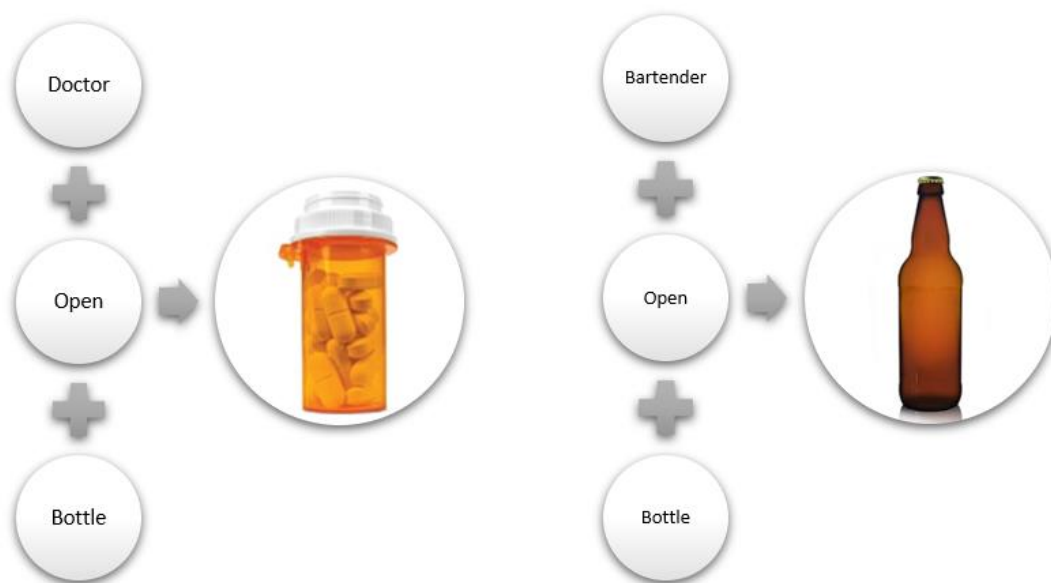


Figure 11 Textual Event and Target Picture.

Figure 11 illustrates two examples of MEK input-output combination. When the input corresponds to the textual sequence *doctor open bottle*, MEK would predict the pills bottle picture. If *bartender open bottle* is the input, then the model would predict the beer bottle picture. The two examples are linked to each other because they correspond to two sentences respectively of the first and the second list of the eye-tracking experiment. See Sentences, Pictures and Lists in the third chapter.

Based on the knowledge learnt during the training stage, MEK should identify the multimodal thematic fit between the agent (*doctor* and *bartender*) and the referent filling the patient role (pills bottle and beer bottle), even if the latter is denoted by the same perceptually underspecified noun in the patient position (*bottle*). The verb is the same (*open*), but the denoted actions differ based on the patient. See the first chapter, which describes how the verb denotational meaning disambiguation depends on the multimodal information about the entities involved in the denoted event.

Agent. The agent in isolation in the input corresponds to the agent time window of the eye-tracking experiment. See Agent Time Window in the third chapter. We used agents in isolation to evaluate whether MEK learnt the multimodal thematic fit between the agent and patient roles. The model has to predict the pictures that appear in the same sequences of the agent. In the training data, the agents occur with at least two patient pictures, Target

and Agent-Related. Except for *hiker*, *chef*, *swimmer* and *cyclist* because they appear with more than two patients. When the agent is *hiker* or *chef* or *swimmer*, MEK has four out of 96 chances of predicting the correct picture with respect to the two out of 96 of the other agents and the six out of 96 of *cyclist*. The distribution of the agents in the collections of training sequences reflects the hypothesis that an agent cues the knowledge of situations in which typically it appears, which includes multimodal information about the other involved entities. Thus, an agent can be indeed associated with more than one object.

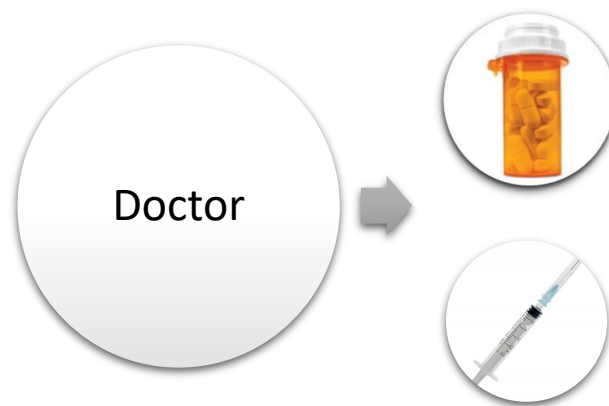


Figure 12 Multimodal thematic fit.

Figure 12 shows an example of multimodal thematic fit between the agent and patient roles. If the input corresponds to the agent *doctor*, MEK would predict one of the pictures that appeared in the same training sequences. Since *doctor* occurs in the sequences *doctor close bottle*, *doctor open bottle*, *doctor empty bottle*, which are linked to the pills bottle picture (Target), and *doctor grab syringe*, *doctor fill syringe*, which are associated with the syringe picture (Agent-Related), both the pills bottle and the syringe represent two correct predictions. See Sequence Collections in the Appendix, which reports the complete list of the sequences exploited to train MEK on typical events knowledge.

Agent and Verb. In sentence comprehension, once the situation has been individuated, the verb leads to identifying the event. The expectations cued by the verb constrain the set of entities previously associated with the agent to the referent that fits its selectional restrictions.

We evaluated whether MEK learns the verb multimodal selectional restrictions providing it with agent-verb pairs in the input. The latter allows us to figure out whether the model processes the information incrementally. The agent-verb pair in the input corresponds to the action time window of the eye-tracking experiment. See Action Time Window in the third chapter.

Thanks to the idea to use in the training data different verbs based on the Target and the Agent-Related conditions, we can reproduce the action time window. There is only one correct picture for each input pair as well as in the eye-tracking experiment visual scene only the Target fitted both the agent and the verb semantic constraints. See Model Design that explains why the verbs that occur with the agent-Target pair differ with respect to the verbs that appear with the pair composed of the same agent and the Agent-Related patient.



Figure 13 Multimodal verb selectional restrictions.

Figure 13 shows two input-output combinations of MEK evaluation. The input is constituted of an agent and a verb in both cases. The agent is the same (*doctor*). The verb changes: *open* and *fill*. As we explained above, the two situations represent the Target and the Agent-Related conditions in the training data. Since *doctor* co-occurs with *open* only when associated with the pills bottle pictures, the latter corresponds to the correct prediction. The verb *fill* appears in the same sequences of *doctor* if the associated picture

represents a syringe. Hence, the syringe picture is the correct prediction if the input is the agent-verb pair *doctor-fill*.

Perceptually Underspecified Noun. The sequences of activities correspond to multimodal descriptions of typical events because they include textual and visual information. The pictures that appear in the sequences do not belong to objects randomly chosen as in Lazaridou et al. (2015). We used the visual properties of referents that fill the patient role in particular events. The sequences encode the relationships between the agent, the verb, and the patient. They constitute the internal structure of a typical event: thematic fit between the agent and patient roles, verb selectional restrictions, and the relationship between a hypernym (perceptually underspecified noun) and its referents. The relationships among the event components imply multimodal information. See the first chapter, which tells about the hypothesis that at least one of the event constituents has to be perceptually specified to individuate a specific situation, identify the involved entities, disambiguate the actions denoted by the verb and, consequently, understand the event and comprehend the sentence that describes it. As we explained in *Processing Dynamics and Training*, the link between the perceptually underspecified noun filling the patient role and its referent depends on the agent. Specific agents cue typical situations knowledge that includes multimodal information about the other involved entities.

We provide perceptually underspecified nouns in isolation in the input to evaluate whether MEK learnt the multimodal relation between the name of a class of entities (hypernym or perceptually underspecified noun) and its possible instances. The input differs from the patient time window of the eye-tracking experiment, which is instead reproduced by the event in the input (agent, verb and patient). When the participants listened to the patient, they had already heard the agent and the verb. MEK instead does not receive information about them. The isolated perceptually underspecified noun corresponds to the real-world scenario in which a person listens to a hypernym in isolation, such as *helmet*, *ball* or *glove*. Without further linguistic and extra-linguistic information about the current context, she tends to not focus on a specific object. She would think of all possible referents that the word can refer to, like bike helmet, motorbike helmet, baseball ball, football ball, soccer ball, boxing glove, baseball glove and so forth.

The training data include two referents for each perceptually underspecified noun in the Target condition and a correspondence one-to-one in the Agent-Related condition, except for the nouns *bottle*, *ball*, *helmet*, *camera* that appear in both the Target and Agent-Related conditions and *box* that is associated with four different referents. When the input corresponds to *bottle* or *ball* or *helmet* or *camera*, MEK has three out of 96 chances of predicting the correct picture with respect to the two out of 96 of the other hypernyms and the four out of 96 of the word *box*.



Figure 14 Relationships between a perceptually underspecified noun (hypernym) and its referents.

Figure 14 shows an example of the input-output combination proposed to evaluate whether MEK learnt the relationship between the name of a category (hypernym or perceptually underspecified noun) and its possible instances. Since the perceptually underspecified noun *bottle* can denote many entities, many outputs can be referred to as correct. In particular, *bottle* denotes a pills bottle in the sequences *doctor close bottle*, *doctor open bottle*, *doctor empty bottle*; a beer bottle in *bartender close bottle*, *bartender open bottle*, *bartender empty bottle*. They correspond to the Target objects associated with *doctor* and *bartender*. Moreover, *bottle* denotes a water bottle in the training sequences *hiker empty bottle* and *hiker hold bottle*, which correspond to the Agent-Related object of *hiker*. Therefore, the hypernym *bottle* is linked to three objects in the

training sequences: pills bottle, beer bottle and water bottle. All of them correspond to correct predictions.

4.5. Results

In what follows, we report the results of the evaluations. The inputs will be called event, agent, agent-verb, perceptually underspecified noun. In reporting the results, we will specify the training method used: LookAT and WhoAct (see Simulations and Architectures).

We used two types of textual (GloVe and Word2Vec) and visual (AlexNet and GoogLeNet) representations, which we collected in the WA and GG matrices (see Processing Dynamics and Training). As we expected, there were no significant differences between them. Therefore, accuracy, precision, recall, and f1-score measures related to the WA and GG matrices will be reported in the Appendix²¹. The ratings of the

²¹ Accuracy, precision and recall measures imply four types of data representing a fine-grained description of the model predictions: True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). True positives are predictions considered correct by the model that are actually right. True negatives are referents classified as wrong that are actually wrong. False positives are referents predicted as correct but actually wrong. False negatives are referents considered wrong by the model but actually correct. The accuracy indicates the set of model correct predictions. It is calculated as the relation between the total of the correct predictions and the total of the predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

The precision (or positive predicted value) is the fraction of the correct predictions reported by the model (Goodfellow, Bengio, Courville 2016). The model correct predictions are based on the truly correct answers (TP) and those resulted as correct but actually wrong (FP).

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

model depend on the quantity of data used to train MEK on the multimodal knowledge of typical events. LookAT and WhoAct include indeed 240 types of events (see Simulations and Architectures). Although, we focused on the structure of the data. The ratings demonstrate that LookAT and WhoAct simulations are adequate representations of the multimodal relationships between the event components that constitute multimodal event knowledge and mirror multimodal expectations exploited by people in sentence comprehension.

We will illustrate MEK activation patterns through confusion matrices. A confusion matrix shows the activation degrees of each referent, providing a visual representation of MEK predictions. The different shades of blue indicate the strength of activation. The darker blue coloured cells represent stronger activations corresponding to a particular referent. On the abscissa axis are indicated the referents predicted by MEK. On the ordinate axis are reported the referents actually correct. On the diagonal, there are the referents that MEK predicted correctly. The numbers that appear on the two axes indicate the referents. Since the total of referents is 96, including all of them in the plot is impossible. The numbers correspond to the referents in alphabetical order. See Referents Report in the Appendix for the tables corresponding to each confusion matrix. The numbers are linked to the corresponding labels in the tables, where each activation degree is indicated.

More than one picture can be a correct prediction when the input corresponds to the agent and the perceptually underspecified noun in isolation. The confusion matrix allows us to check whether MEK activations correspond to the set of predictions consistent with

The recall, known also as sensitivity, is the ability of the model to individuate the correct referents taking into account the truly correct answers and those resulted wrong but actually correct (FN). It is the fraction of true events that were predicted (Goodfellow, Bengio, Courville 2016).

$$Recall = TP / (TP + FN) \quad (4)$$

F-score is the harmonic mean of precision and recall.

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

multimodal thematic fit between the agent and the referent of the patient role and the relationship between the perceptually underspecified noun (hypernym) and its referents.

4.1.1 Event

MEK should predict the referent that fits both the agent and verb semantic constraints when the input is a textual event (agent, verb, and perceptually underspecified patient). In particular, the referent should appear in the same situations where the agent is also present, and it should be consistent with verb selectional restrictions. The output should correspond to the object the perceptually underspecified noun in the patient position refers to (see Evaluation).

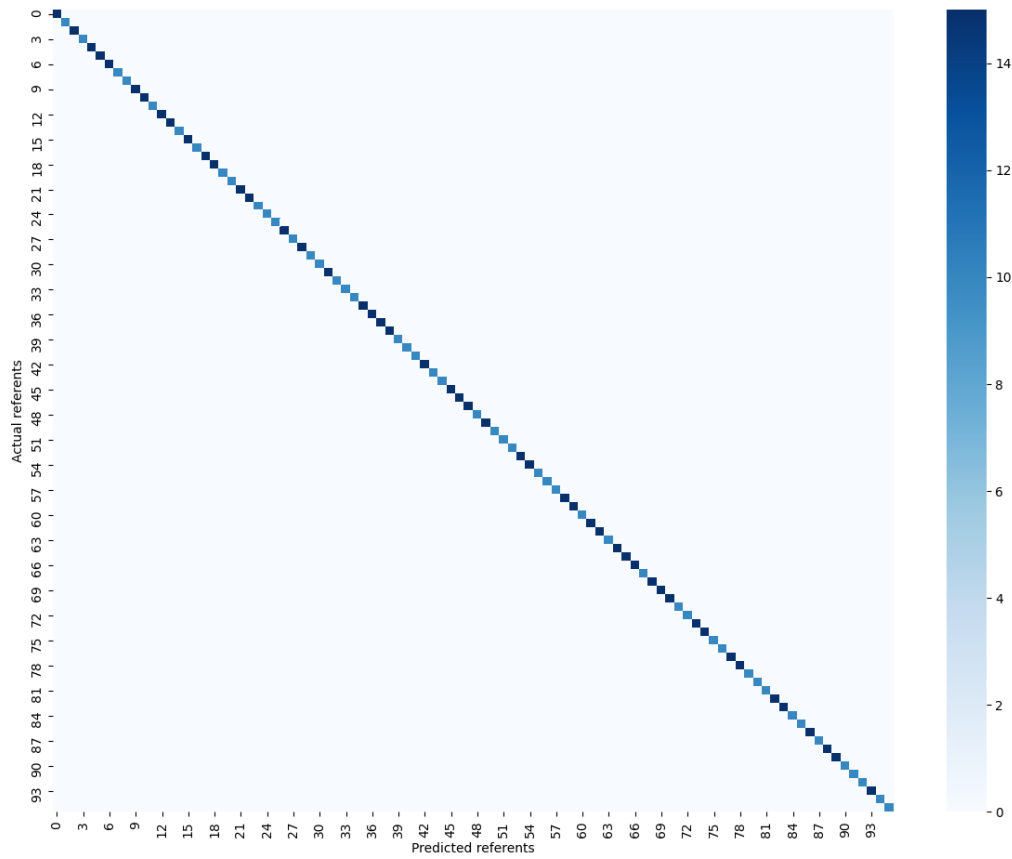


Figure 15 LookAT Confusion Matrix.²²

²² Only some of the numbers representing the referents appear on the axes due to the high number of the latter (96). The numbers correspond to the referents in alphabetical order. A legend of the number-referent combination can be found in Referents Report in Appendix: Table 22 and Table 23.

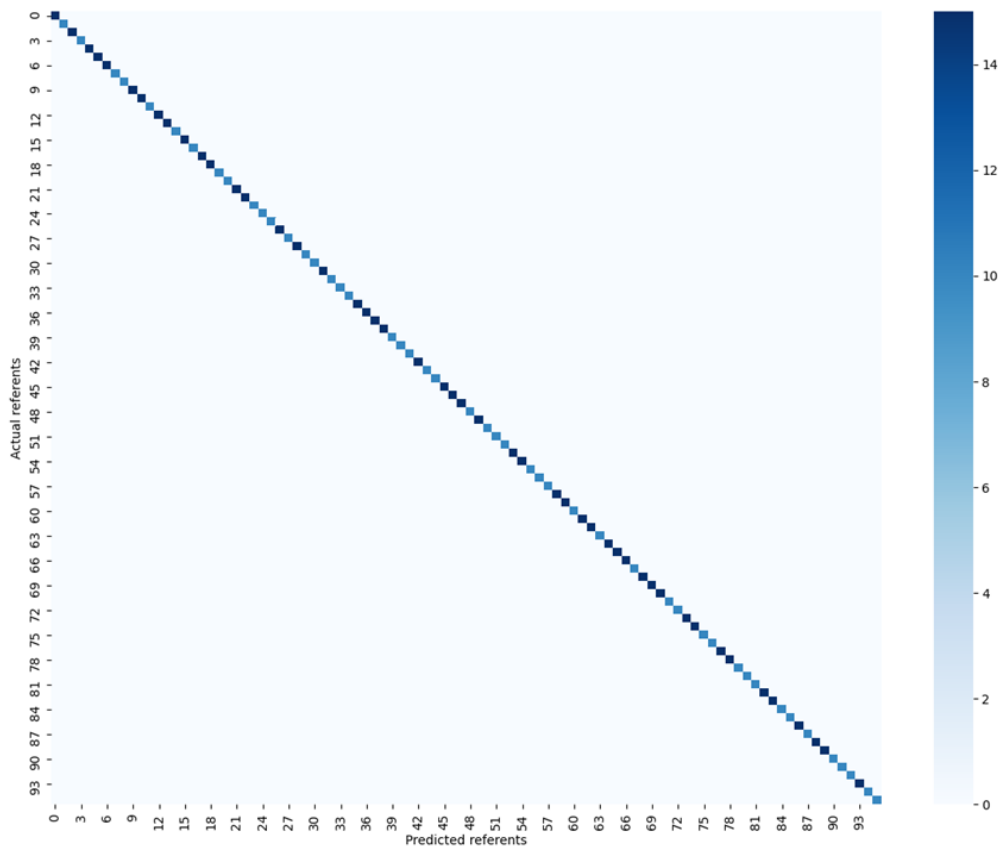


Figure 16 WhoAct Confusion Matrix.²³

Figure 15 and Figure 16 show that MEK predicted the pictures of the correct referents. All the activations are located indeed along the diagonals of the confusion matrices. Hence, for instance, given *gardener fill pot* in the input, MEK predicted the picture of a plant pot. When the input was *cook fill pot*, the output corresponded to the cooking pot picture. The results reflect the proportions of eye fixations toward the Target AOI in the patient time window of the eye-tracking experiment (see Patient Time Window in the third chapter).

²³ Only some of the numbers representing the referents appear on the axes due to the high number of the latter (96). The numbers correspond to the referents in alphabetical order. A legend of the number-referent combination can be found in Referents Report in Appendix: Table 24 and Table 25.

4.1.2 Agent

A specific agent cues the knowledge of situations in which it typically appears. Typical situation knowledge includes multimodal information about other involved entities. We expect that, given an agent in isolation in the input, MEK would predict at least one of the referents that occur in the same sequences (see Evaluation).

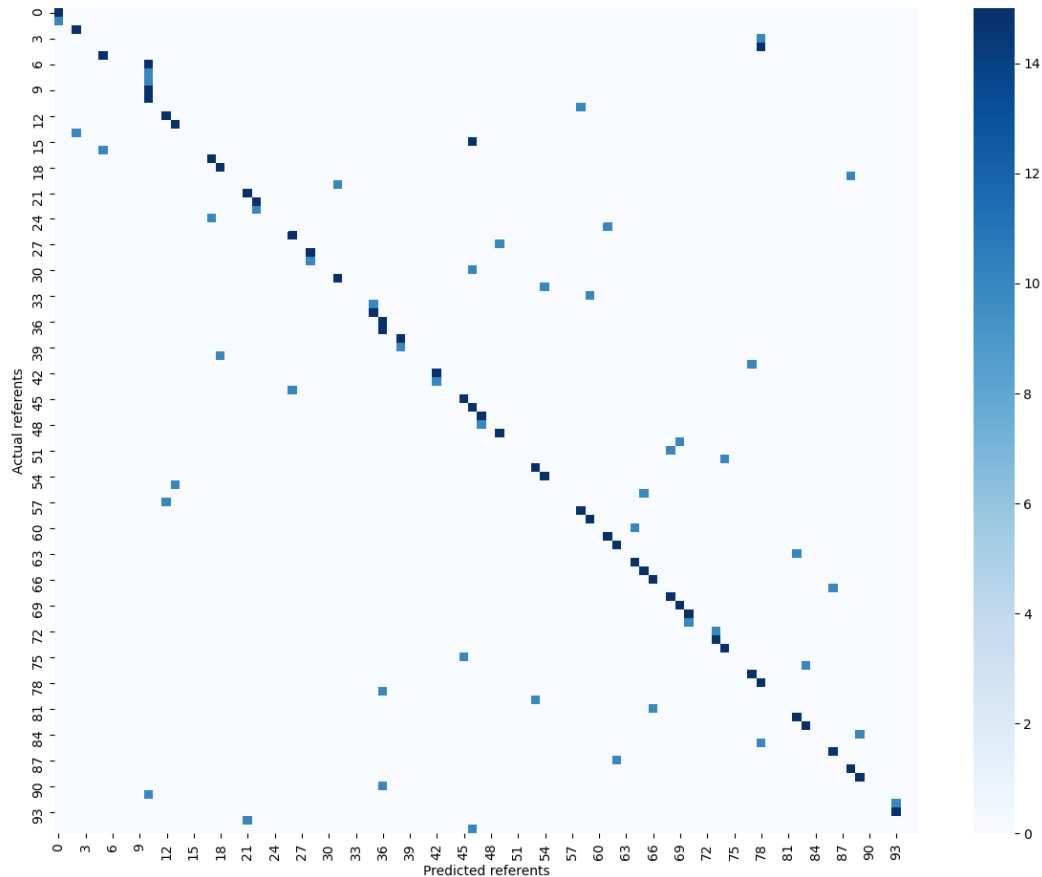


Figure 17 LookAT Confusion Matrix.²⁴

²⁴ Only some of the numbers representing the referents appear on the axes due to the high number of the latter (96). The numbers correspond to the referents in alphabetical order. A legend of the number-referent combination can be found in Referents Report in Appendix: Table 26 and Table 27.

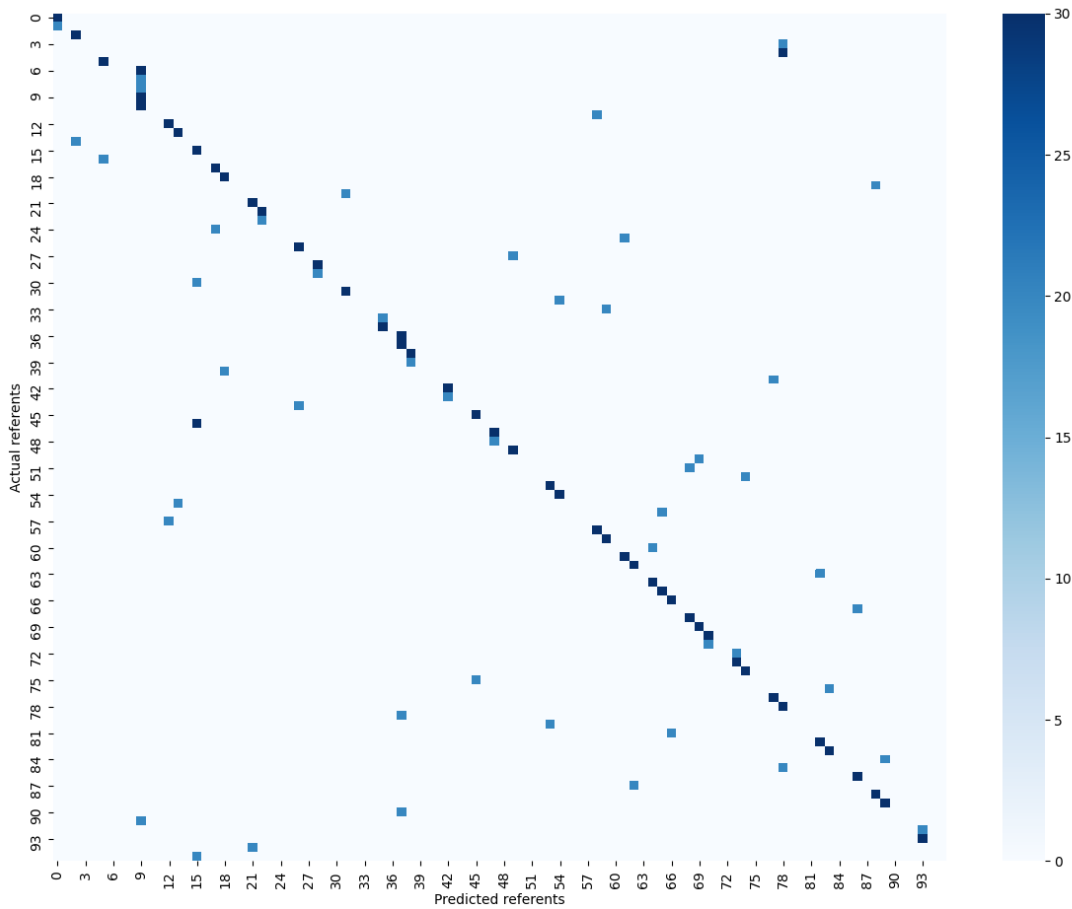


Figure 18 WhoAct Confusion Matrix.²⁵

Figure 17 shows that darker blue cells correspond to the actual referents. Some lighter blue cells are around the diagonal. They indicate that MEK individuated more than one object associated with each agent.

Giving some examples: *driver* activated the Agent-Related road sign (60) and the Target seat belt (64); *tourist* activated the Target video camera (89) and the Agent-Related tourist map (84); *cyclist* activated the activations of bike helmet (6), bike lock (7), bike pump (8), bike saddle (9) and bike tire (10); *player* activated the Agent-Related soccer ball (72) and the Target soccer shoe (73).

²⁵ Only some of the numbers representing the referents appear on the axes due to the high number of the latter (96). The numbers correspond to the referents in alphabetical order. A legend of the number-referent combination can be found in Referents Report in Appendix: Table 28 and Table 29.

As expected, Figure 18 shows activations both along and around the diagonal: *biker* activated the Target motorbike helmet (47) and the Agent-Related motorbike (48); *graduate* activated the Agent-Related diploma (34) and the Target graduation cap (35); *quarterback* activated the Target football ball (28) and the Agent-Related football helmet (29).

Therefore, the blue cells that are not located along the diagonal represent consistent predictions. They show that MEK identified the objects associated with a particular agent. The results are in line with the hypothesis that agents are cues to typical situations knowledge that includes multimodal information about other involved entities. The data mirror the proportions of eye fixations toward the Target and the Agent-Related AOIs in the agent time window of the eye-tracking experiment (see Agent Time Window in the third chapter).

4.1.3 Agent and Verb

The results of the proportions of eye fixations toward the four AOIs in the Action Time Window (see the third chapter) showed anticipatory eye movements toward the Target before listening to the critical word, namely the perceptually underspecified noun in the patient position. The anticipatory eye movements mirrored the incremental integration of the agent multimodal expectations and the verb selectional restrictions.

Agent-verb pairs in the input aim to reproduce the anticipatory time window. The input allows evaluating if MEK incrementally deals with information coming from the environment and reproduces the multimodal expectations exploited by people during sentence comprehension. We expect that MEK would predict the referent that occurs in the same sequence of the agent and the verb (see Evaluation).

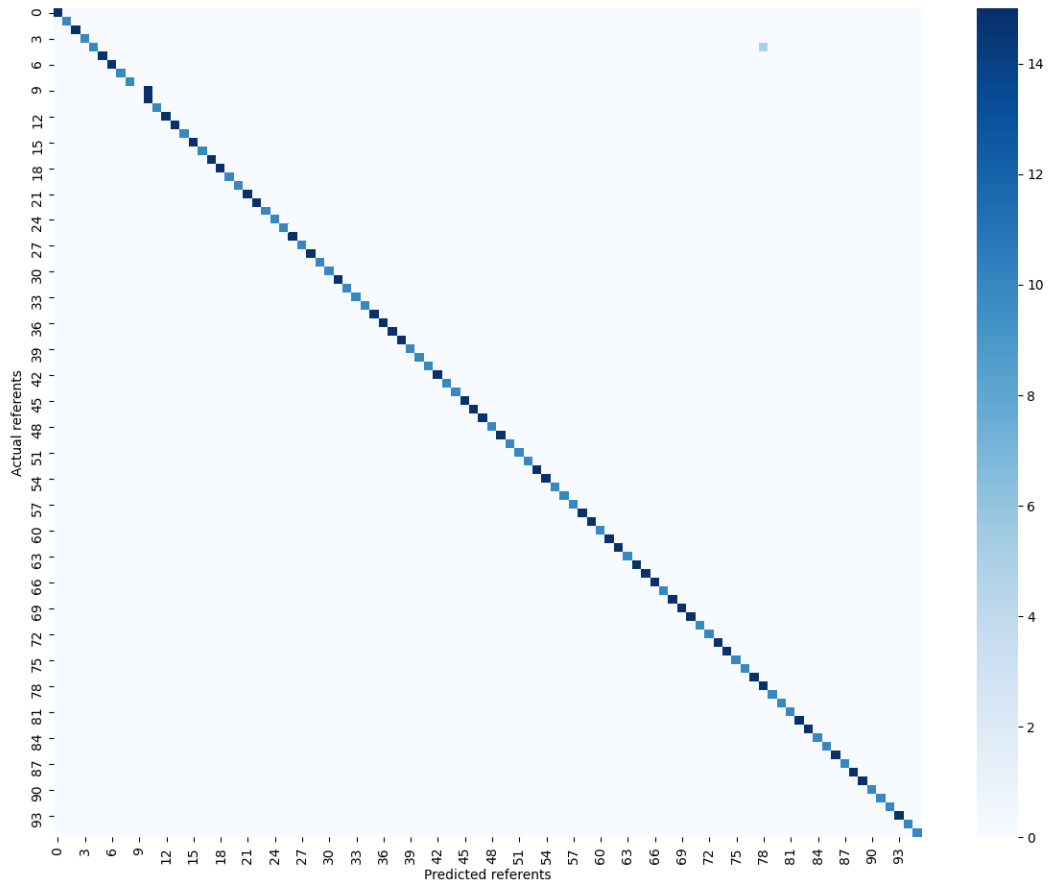


Figure 19 LookAT Confusion Matrix.²⁶

²⁶ Only some of the numbers representing the referents appear on the axes due to the high number of the latter (96). The numbers correspond to the referents in alphabetical order. A legend of the number-referent combination can be found in Referents Report in Appendix: Table 30 and Table 31.

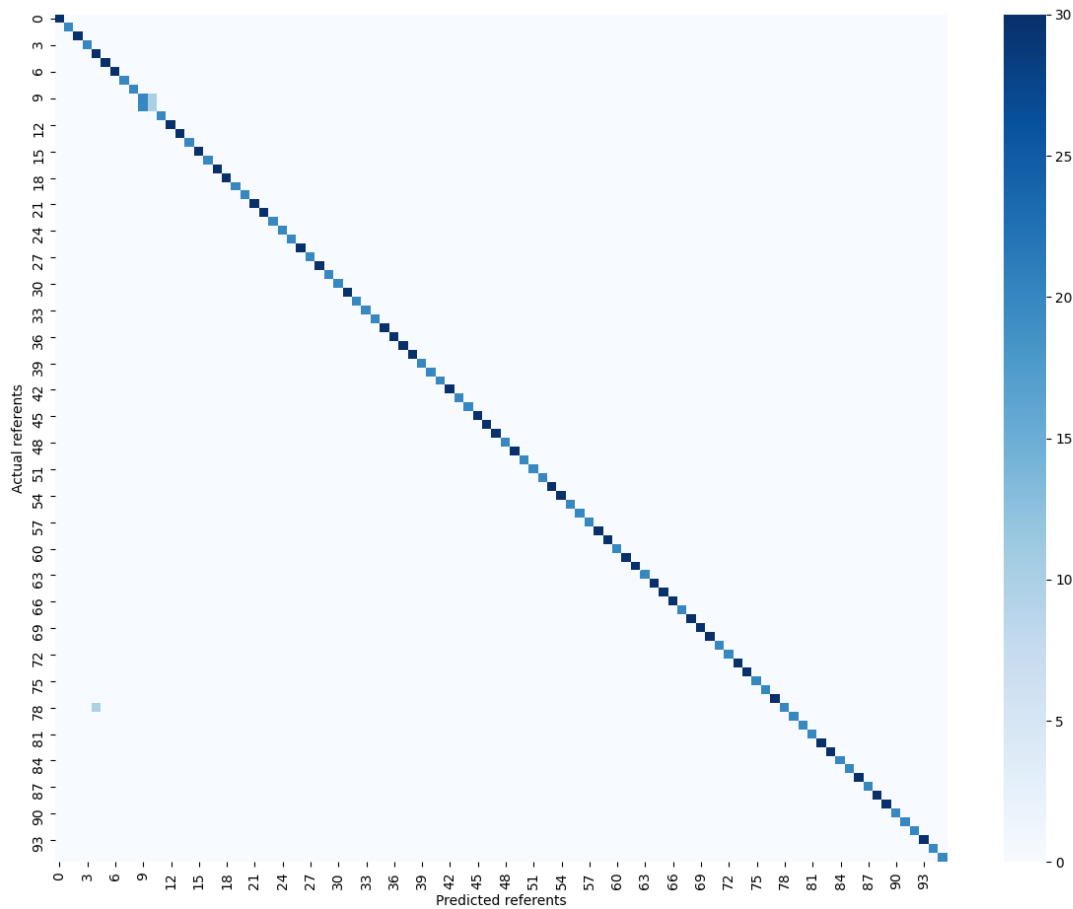


Figure 20 WhoAct Confusion Matrix.²⁷

As expected, Figure 19 and Figure 20 show the predicted referents corresponded to the actual referents. Only in correspondence of bathing suit (4) and bike saddle (10), the confusion matrices show the slight activation of swimming goggles (78) and bike tire (9), respectively. As reported in Sequence Collections in the Appendix, bathing suit and swimming goggles appear in the same sequences of *swimmer*; bike saddle and bike tire are associated with *cyclist*.

Therefore, swimming goggles and bike tire are consistent predictions. They represent a little mismatch due to the agent, and they are a clue of the multimodal thematic fit effect. The results show that MEK processes incrementally the information derived from the

²⁷ Only some of the numbers representing the referents appear on the axes due to the high number of the latter (96). The numbers correspond to the referents in alphabetical order. A legend of the number-referent combination can be found in Referents Report in Appendix: Table 32 and Table 33.

environment. The model predictions mirror the anticipatory eye movements toward the Target AOI in the Action Time Window of the eye-tracking experiment.

4.1.4 Perceptually Underspecified Noun

As described in Evaluation, the input corresponding to the perceptually underspecified noun in isolation is not equivalent to the patient time window of the eye-tracking experiment. When the participants listened to the patient, they had already heard the agent and the verb. Hence, eye movements in the patient time window were affected by the expectations encoded in the agent and the verb.

MEK receives the hypernyms in isolation. The input reproduces the scenario where, listening to perceptually underspecified nouns in isolation, such as *ball*, *belt* or *bottle*, a person tends to not focus on a specific referent because she does not have further linguistic and/or extra-linguistic information about the current situation. See Perceptually Underspecified Nouns in the first chapter for a broad discussion about why hypernyms in isolation do not entail a specific type of perceptual referents but classes of entities. Providing MEK with perceptually underspecified nouns in isolation, we evaluate if MEK learnt the relationship between a hypernym and the set of all its possible instances.

What distinguishes LookAt and WhoAct simulations is how we modelled the relationship between the hypernym and its referents. In LookAT, the referent appears in the same sequence of the agent, verb and perceptually underspecified patient (see the example (20)). In WhoAct simulation, the perceptually underspecified noun and its referents appear in different sequences. MEK would derive the referents of the perceptually underspecified nouns from their co-occurrences with the same agent-verb pair. See the example (21). See Processing Dynamics and Training that reports how many referents appear for each perceptually underspecified noun in the training data.

We expect that MEK activations (see Evaluation) correspond to:

- The referents that co-occurred with the hypernym in the LookAT sequences.
- The referents that appeared with the agent-verb pair that co-occurred with the perceptually underspecified noun in the WhoAct sequences.

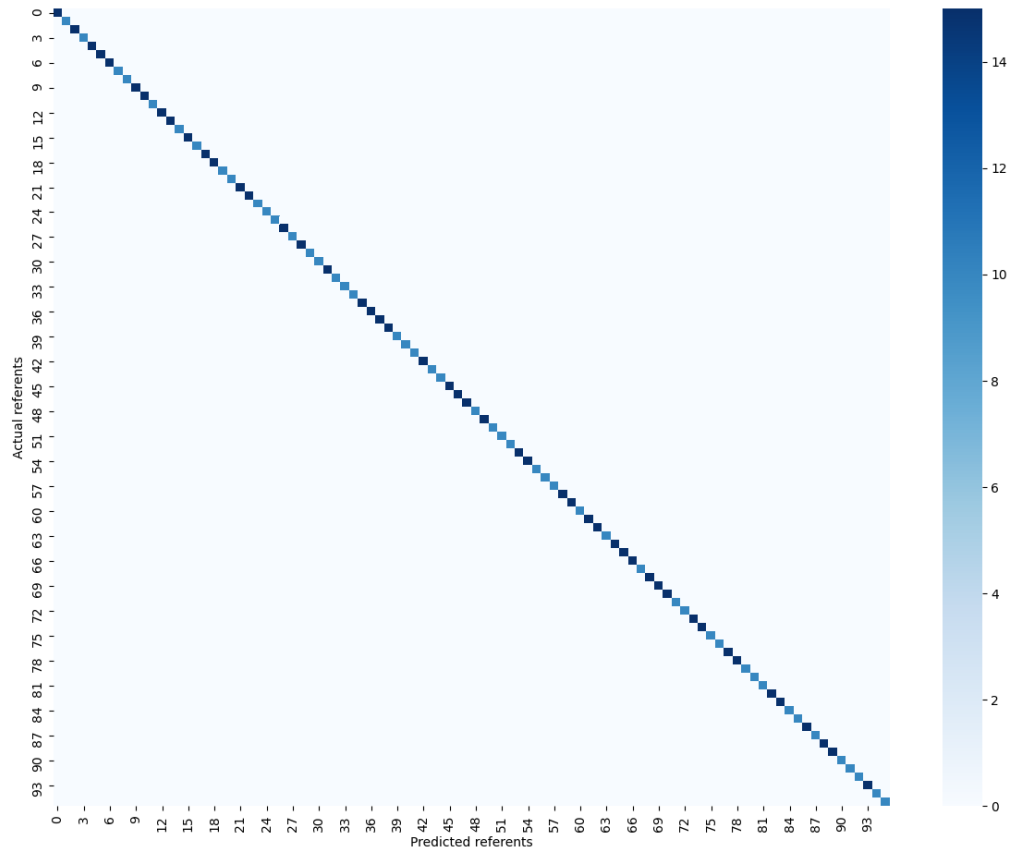


Figure 21 LookAT Confusion Matrix.²⁸

²⁸ Only some of the numbers representing the referents appear on the axes due to the high number of the latter (96). The numbers correspond to the referents in alphabetical order. A legend of the number-referent combination can be found in Referents Report in Appendix: Table 34 and Table 35.

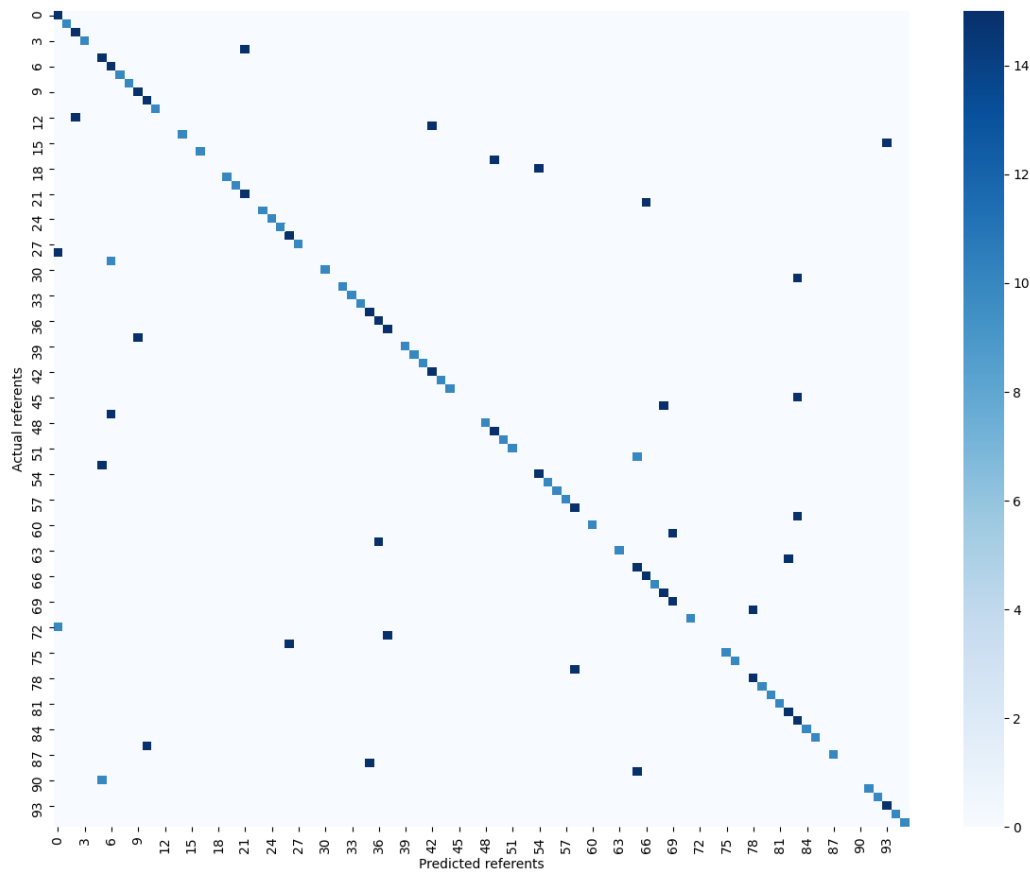


Figure 22 WhoAct Confusion Matrix.²⁹

Figure 21 shows that the activations are all located along the diagonal. Unlikely what expected, the data present a single activation for each type of input. However, Figure 22 shows many activations around the diagonal. The perceptually underspecified nouns denote at least two Target objects in the training data. Therefore, if the activation is not located along the diagonal but corresponds to a plausible referent of the hypernym in the input, it is a consistent prediction.

Giving some examples, Figure 22 shows that: *camera* activated video camera (89) and security camera (65), the Target objects of *police* and *tourist*; *boot* activated sky boot (69) and rubber boot (61), the Targets of *olympian* and *fisherman*; *backpack* activated hiking backpack (36) and school backpack (62), the Targets of *hiker* and *student*; *glove* activated

²⁹ Only some of the numbers representing the referents appear on the axes due to the high number of the latter (96). The numbers correspond to the referents in alphabetical order. A legend of the number-referent combination can be found in Referents Report in Appendix: Table 36 and Table 37.

baseball glove (2) and boxing glove (12), Targets of *catcher* and *boxer*; *cap* activated uniform cap (88) and graduation cap (35), Targets of *captain* and *graduate*; *helmet* activated football helmet (29) and motorbike helmet (47), the Agent-Related object of *quarterback* and the Target of *biker*.

The results show that the model derived the relationship ("it is a type of ") between a perceptually underspecified noun and the set of its possible instances from the WhoAct sequences. When trained on the multimodal knowledge of typical events using the WhoAct sequences, MEK recognised that a hypernym is a class denotator. In isolation, a perceptually underspecified noun, indeed, does not entail a specific type of perceptual referent but refers to a set of possible instances (see Perceptually Underspecified Nouns in the first chapter).

4.6. General Discussion

MEK is a model of multimodal knowledge about typical events cued by words. The model reproduces the multimodal expectations elicited by the event components denoted by words in sentences and integrates them incrementally. MEK infers the unmentioned information about the patient role of an event based on multimodal knowledge of typical events it previously learnt.

In line with the eye-tracking experiment, MEK reproduces:

- The multimodal thematic fit between the agent and the referent of the patient role.
- The multimodal verb selectional restrictions.
- The relation between a perceptually underspecified noun and its referents.

We evaluated the model providing it with different types of input that mirror the time windows analyzed in the eye-tracking experiment: event, agent, agent-verb. See General Discussion in the third chapter.

Moreover, we supplied the model with hypernyms in isolation in the input to test if MEK learnt the relationship between a perceptually underspecified noun and its referents.

The training stage aims at simulating different real-world scenarios where people experience situations and events (LookAT and WhoAct). They include sequences of

activities that describe the same typical events but differently presented (see Simulations and Architectures).

The textual event input (agent, verb, perceptually underspecified nouns) aims at reproducing the patient time window of the eye-tracking experiment. Listening to the perceptually underspecified noun, the participants focused on the object that fitted both the agent and verb semantic preferences: Target. See Patient Time Window in the third chapter.

MEK predicts:

- The referent that co-occurs in the same sequence, if the model was trained using the LookAT event descriptions (see Simulations and Architectures).
- The referent that appears with the agent-verb pair that co-occurs with the perceptually underspecified noun filling the patient role, if the model was trained using the WhoAct sequences (see Simulations and Architectures).

MEK predictions are in line with the hypothesis that particular agent and verb combinations cue expectations about the referent of the patient role.

We provided agents in isolation in the input to reproduce the agent time window of the eye-tracking experiment (see Agent Time Window in the third chapter) and check if MEK learnt the multimodal thematic fit between the agent and the referent of the patient role. MEK predicts the referents that appear in the same sequences of the agents. In particular, the model predicts both Target and Agent-Related objects appearing in the training data. The results are in line with the hypothesis that agents are cues to typical situation knowledge that includes multimodal information about other involved entities. MEK predictions reflect the proportions of eye fixations toward the Target and the Agent-Related AOIs in the agent time window.

Agent-verb pairs in the input aim at reproducing the anticipatory time window of the eye-tracking experiment (see Action Time Window in the third chapter). We tested if MEK incrementally deals with information coming from the environment. The model integrates the information encoded in the agent and verb, reproducing the multimodal expectations exploited by people during sentences comprehension. MEK predicts the referent that appears with the agent-verb pair in training data. The results show that MEK predicts the referent that appears in the same situations in which the agent is typically present and fits verb selectional restrictions.

The perceptually underspecified nouns in isolation in the input do not correspond to the patient time window of the eye-tracking experiment (see Patient Time Window in the third chapter). When the participants listened to the patient, they had already heard the agent and the verb. MEK received the hypernym in isolation. We evaluated if MEK learnt the relationship (“it is a type of”) between a noun denoting a class of entities (perceptually underspecified noun or hypernym) and the set of its possible instances (see Perceptually Underspecified Nouns in the first chapter).

MEK predicts:

- The referent that co-occurs with the perceptually underspecified noun in the LookAT sequences (see Simulations and Architectures).
- The referent that appears with the agent-verb pair co-occurring with the perceptually underspecified noun in the WhoAct sequences (see Simulations and Architectures).

Unlike the previous evaluations, there were significant dissimilarities between the LookAT and the WhoAct simulations. The results in the LookAT condition showed a single activation for each lexical item in the input. In the WhoAct sequences, where the hypernyms in the patient position and the corresponding referents do not co-occur (see the example (21)), the individuation of the relation between the perceptually underspecified noun and its referents relies on the co-occurrence with the same agent-verb pair. When MEK was trained on multimodal knowledge of typical events through the WhoAct sequences, it predicted the set of plausible instances of the hypernym in the input. The model indeed individuated the set of its possible referents. The results mirror the scenario where, in the absence of further linguistic and/or extra-linguistic information about the current situation, listening to a word like *helmet*, *cap*, *pot* or *bottle*, a person tends to not focus on a specific instance because she knows that the hypernym might recall another plausible instance of the word.

Discussion and Conclusion

We investigated verb denotational meaning accounting for typical fillers of its thematic roles and their physical properties. Words encode expectations linked to the knowledge of typical real-world events and situations. The main hypothesis of this thesis states that lexical items are cues to multimodal knowledge of typical events and situations, including the collection of physical properties that identify event components. Visual perceptual features of the fillers of verb thematic roles are crucial information to disambiguate the actions involved in events.

We exploited the eye-tracking technique and visual world paradigm to study if information encoded by agents and verbs include multimodal expectations about the referents filling the patient role during sentence comprehension. The results showed that specific agents are cues to knowledge about the situations in which they typically appear, the events in which they usually are engaged and information about other entities with which they commonly interact.

The participants looked at Target and Agent-Related pictures listening to the agent. The proportions of eye fixations reflect the multimodal thematic fit, namely the degree of coherence between the agent and a specific type of perceptual referent filling the patient role. Multimodal thematic fit relies on the knowledge of typical real-world situations.

Sentence comprehension is an incremental process. The information cued by the agent are integrated with the expectations encoded by the verb. The latter restricts the set of entities previously cued by the agent to only those items that fit its selectional restrictions. The results show anticipatory eye movements toward Target picture before listening to the critical word (perceptually underspecified noun or hypernym.) referred to it.

The Target remained the most looked at picture until the end of the auditory sentence. It was indeed the only object in the visual scene consistent with expectations cued by the agent and the verb. The proportions of eye fixations toward Action-Related picture compared to Agent-Related picture in the patient time window and the final silence were clues of the influence of the current visual context during sentence comprehension.

Moreover, we created a computational model that aims at simulating multimodal thematic fit between the agent and the referent of the patient role, verb selectional restrictions and the relationship between perceptually underspecified nouns and the set of its plausible instances. We called it MEK (Multimodal Event Knowledge). MEK is a model of

multimodal knowledge about typical events cued by words. We did not provide a priori definitions of typical events. The model learnt case-by-case the internal structure of a typical event. The training stage involved two simulations of real-world scenarios where people experience situations and events: LookAt and WhoAct. MEK derived the internal structure of typical events from a subset of stimuli of the eye-tracking experiment that we manipulated in order to represent the relations between the sentences and the pictures in the visual scene: Target, Agent-Related, Action-Related and Unrelated.

We evaluated MEK providing it with inputs corresponding to the eye-tracking experiment time windows: event, agent and agent-verb pairs. MEK predicts the referent of the patient role of a textual event. The results show that MEK learnt the multimodal thematic fit between the agent and a specific type of perceptual referent filling the patient role, infers unmentioned multimodal information about the patient and incrementally deals with information coming from the environment reproducing multimodal expectations exploited by people during sentences comprehension.

Moreover, we provided the model with perceptually underspecified noun in isolation in the input. When the model was trained on WhoAct sequences, the results showed that MEK learnt the relationship between a perceptually underspecified noun and its referents. The activation patterns reproduced the real-world scenario where a person, hearing hypernyms like *glasses*, *glove* or *helmet* without further linguistic and/or extra-linguistic information about the current situation, tend to not focus on a specific type of perceptual referent.

The eye-tracking experiment provided evidence that multimodal expectations cued by agents and verbs lead to the individuation of the situation and event that guide people' attention toward specific referents of the incoming patient role. MEK predictions mirror the eye-tracking experiment results. However, only when the model was trained on the multimodal knowledge of typical events using the WhoAct sequences, it derived the relationships between a perceptually underspecified noun and the set of its possible instances.

Theoretical Implications

Verb denotational meaning is a good example of the binding between language and information people perceive through senses, like vision. Language comprehension needs the disambiguation of the relations between verbs and actions, nouns and real-world entities. The disambiguation of the actions denoted by a verb requires that language is grounded in perceptual experience. Human semantic knowledge is composed of many types of information derived from real-world experience. They interact during language comprehension. The assumption that typical events and situations guides learning processes and sentences comprehension implies that verb meaning depends on the typical fillers of its thematic roles. Meanings of verbs, agents, and patients include mutual multimodal expectations that mirror real-world events patterns. This thesis demonstrated that the meaning of a word could not be exhaustively explained relying only on linguistic information. A world of daily life experiences of an entire person's life span hides behind the use of words. Combining lexical items in sentences leads to new meanings whose correct interpretation requires knowledge about what is typical and belongs to the extra-linguistic environment, namely what we call the real world.

Future Research Directions

We exploited information about event components to investigate verb denotational meaning and demonstrate the importance of extra-linguistic properties of the entities in disambiguating the involved actions. We based our experiments on the assumption that at least one of words occurring in sentences has to be perceptually specified because a specific situation can be cued. The knowledge about the situation leads to the individuation of the events that constrains the set of participants. The collection of properties that identify event components are crucial information for the disambiguation of the actions denoted by the verb. Since a single verb usually denotes a continuous stream of movements connected through causal and conventional relations, it could be interesting to focus future investigations on concepts like basic and complex actions, manipulating the stimuli to study how this theoretical aspect can affect the representation of verb meaning. New studies could focus on the relationship between how people percept real-world basic and complex actions and how this knowledge is expressed through language.

References

- Abashidze, D., Knoeferle, P., & Carminati, M. N., (2014). How robust is the recent event preference? In Bello, P., Guarini, M., McShane, M., & Scassellati, B., *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Allain, P., Le Gall, D., Etcharry-Bouyx, F., Aubin, G., & Emile, J., (1999). Mental representation of knowledge following frontal-lobe lesion: Dissociations on tasks using scripts. *Journal of Clinical and Experimental Neuropsychology*, *21*, 643-665.
- Altmann, G. T. M., (1999). Thematic role assignment in context. *Journal of Memory and Language*, *41*, 124-45.
- Altmann, G. T. M., & Kamide, Y., (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*, 247-64.
- Altmann, G. T. M., & Kamide, Y., (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*, 502-518.
- Altmann, G. T. M., & Kamide, Y., (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition*, *111(1)*, 55-71.
- Altmann, G. T. M., & Mirković, J., (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*, 583-609.
- Baayen, R. H., Davidson, D. J., & Bates, D. M., (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59(4)*, 390-412.
- Baillargeon, R., (1986). Representing the existence and the location of hidden objects: Object permanence in 6- and 8- month old infants. *Cognition*, *23*, 21-41.
- Baillargeon, R., (1987). Young infants' reasoning about the physical and spatial properties of a hidden object. *Cognitive Development*, *2*, 179-200.
- Baillargeon, R., Spelke, E. S., & Wasserman, S., (1985). Object permanence in five-month-old infants. *Cognition*, *20*, 191-208.
- Baldwin, D. A., & Baird, J. A., (1999). Action analysis: A gateway to intentional inference. In p. Rochat (Ed.), *Early Social Cognition*, Hillsdale, NJ: Erlbaum, 215-240.

- Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A., (2001). Infants parse dynamic action. *Child Development*, 72, 708-717.
- Barr, D. J., (2013). Random effects structure for testing interactions in linear mixed effects models. *Frontiers in psychology*, 4.
- Bates, D., Maechler, M., Bolker, B., & Walker, S., (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.
- Baggio, G., & Hagoort, P., (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes*, 26(9), 1338-1367.
- Baggio, G., Van Lambalgen, M., Hagoort, P., (2012). The processing consequences of compositionality. In M., Werning, W., Hinzen, & E., Machery, *The Oxford Handbook of Compositionality*, 32, Oxford University Press.
- Baroni, M., & Lenci, A., (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4), 673-721.
- Barsalou, L. W., (2008). Situating concepts. *Cambridge Handbook of Situated Cognition*, 14, 236-263.
- Barwise, J., & Perry, J., (1983). *Situations and Attitudes*. Cambridge, MA: MIT-Bradford.
- Bengio, Y., Simard, P., Frasconi, P., (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2), 157-166.
- Bicknell, K., Elman, J. L., Hare, M., McRae, K. & Kutas, M., (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, 63(4), 489-505.
- Bourg, T., Bauer, P. J., & van den Broek, P., (1997). Building the bridges: The development of event comprehension and representation. In P. W. van den Broek, P. J. Bauer, & T. Bourg (Eds.), *Developmental spans in event comprehension and representation: Bridging fictional and actual events*. Philadelphia, PA: Psychology Press, 385-407.
- Bruni, E., Boleda, G., Baroni, M., & Tran, N. K., (2012). Distributional semantics in Technicolor. In *Proceedings of ACL*, Jeju Island, Korea, 136-145,
- Bruni, E., Tran, N. K., & Baroni, M., (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49, 1-47.

- Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social Cognition, Joint Attention, and Communicative Competence from 9 to 15 Months of Age. In *Monographs of the Society for Research in Child Development*, 63(4), I-174.
- Chersoni, E., Santus, E., Lenci, A., Blache, P., & Huang, C-R., (2016). Representing verbs with rich contexts: an evaluation on verb similarity. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Austin, Texas, 1967-1972.
- Chomsky, N., (1965). *Aspects of the Theory of Syntax*. Cambridge, MA, The Mit Press.
- Chomsky, N., (1981). *Lectures on Government and Binding*. Dordrecht, Foris.
- Chrupała, G., Kádár, A., & Alishahi, A., (2015). Learning language through pictures. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China, 112-118.
- Clark, S., (2015). Vector space models of lexical meaning. In S. Lappin & C. Fox (Eds.) *Handbook of Contemporary Semantics*, Oxford, UK: Wiley-Blackwell, 493-522.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P., (2011). Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research (JMLR)*.
- Cooper, R. M., (1974). The control of eye fixation by the meaning of spoken language: a new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6, 84-107.
- Creel, S. C., Aslin, R. N., & Tanenhaus, M. K., (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, 106, 633-664.
- Crozier, S., Sirigu, A., Lehericy, S., Moortele, P.-F. v. d., Pillson, B., Grafman, J., et al., (1999). Distinct prefrontal activations in processing sequence at the sentence and script level: An fMRI study. *Neuropsychologia*, 37, 1469-1476.
- Dahan, D., Magnuson, J. S., & Tanenhaus, M. K., (2001). Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cognitive Psychology*, 42, 317-367.
- Dahan, D., & Tanenhaus, M., (2005). Looking at the rope when looking for the snake: conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin & Review*, 12, 453-459.

- Danto, A., (1963). What we can do. *Journal of Philosophy*, 60, 435-445.
- Dehaene, S., Lau, H., Kouider, S., (2017). What is consciousness, and could machines have it? In *Science*, 358, 6362, 486-492.
- Dehaene, S., (2020). *How We Learn. Why Brain Learn Better Than Any Machine ... For now*. Viking, USA.
- DeLong, K. A., Urbach, T. P., & Kutas, M., (2005). Probabilistic word pre-activation during comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8(8), 1117-21.
- Deng, J., Dong, W., Socher, R., Li, L-J., Li K., & Fei-Fei, L., (2009). ImageNet: A large-scale hierarchical image database. *IEEE Computer Vision and Pattern Recognition (CVPR)*.
- Diederik, P. K., & Ba, J., (2015). Adam: A Method for Stochastic Optimization. In *ArXiv:1412.6980*.
- Dowty, D., (1991). Thematic Proto-Roles and Argument Selection. *Language*, 67(3), 547-619.
- Ehrlich, K., & Johnson-Laird, P. N., (1982). Spatial descriptions and referential continuity. *Journal of Verbal Learning and Verbal Behaviour*, 21, 296-306.
- Elman, J. L., (2011). Lexical knowledge without a lexicon? *The Mental Lexicon*, 60, 1-33.
- Elman, J. L., (2014). Systematicity in the lexicon: On having your cake and eating it too. In P. Calvo & J. Symons (Eds.). *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*. Cambridge, MA, MIT Press, 115-145.
- Elman, J. L., & McRae, K. (2019). A model of event knowledge. *Psychological Review*, 126(2), 252-291.
- Enns, J. T., & Lleras, A., (2008). What's next? New evidence for prediction in human vision. *Trends in Cognitive Sciences*, 12, 388-396.
- Erk, K., & Padó, S., (2008). A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, 897-906.
- Federmeier, K. D., (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, 44, 491-505.

- Feng, Y., & Lapata, M., (2010). Visual information in semantic representation. In *Proceedings of HLT-NAACL*, Los Angeles, CA, 91-99.
- Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A., (2005). Learning Object Categories from Google's Image Search. In *Proceedings of the 10th International Conference on Computer Vision (ICCV)*.
- Ferretti, T. R., McRae, K., & Hatherell, A., (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44, 516-547.
- Fiser, J., & Aslin, R. N., (2002). Statistical learning of higher-order temporal structure from visual shape-sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 458-467.
- Fortin, S., Godbout, L., & Braun, C. M., (2002). Strategic sequence planning and prospective memory impairments in frontally lesioned head trauma patients performing activities of daily living. *Brain and Cognition*, 48, 361-365.
- Glenberg, A. M., Meyer, M., & Lindem, K., (1987). Mental models contribute to foregrounding during text comprehension. *Journal of Memory and Language*, 26, 69-83.
- Goldberg, A. E., (2006). *Constructions at Work. The Nature of Generalization in Language*. Oxford University Press, New York.
- Glenberg, A., & Robertson, D., (2000). Symbol grounding and meaning: A comparison of high dimensional and embodied theories of meaning. *Journal of Memory and Language*, 3(43), 379-401.
- Goldman. A. I., (1970). *A Theory of Human Action*. Englewood Cliffs, NJ: Prentice-Hall.
- Goodfellow, I., Bengio, Y., Courville, A., (2016). *Deep Learning*. Cambridge, MA: MIT Press.
- Greenberg, C., Demberg, V., & Sayeed, A., (2015a). Verb Polysemy and Frequency Effects in Thematic Fit Modeling. In *Proceedings of NAACL Workshop on Cognitive Modeling and Computational Linguistics*.
- Greenberg, C., Sayeed, A., & Demberg, V., (2015b). Improving Unsupervised Vector-space Thematic Fit Evaluation via Role-filler Prototype Clustering. In *Proceedings of HLT-NAACL*.
- Grice, P., (1975). Logic and conversation. In Cole, P., Morgan, J. *Syntax and semantics 3: Speech acts*. Academic Press, New York, 41-58.

- Hagoort, P., (2005). On Broca, brain, and binding: a new framework. *Cognitive Sciences*, 9(9), 416-423.
- Hagoort, P., (2013). MUC (Memory, Unification, Control) and beyond. *Frontiers in Psychology*, 4, 416.
- Hagoort, P., (2016). MUC (Memory, Unification, Control): A model on the neurobiology of language beyond single word processing. *Neurobiology of Language*, 28, 339-347.
- Hard, B. M., Tversky, B., & Lang, D., (2006). Making sense of abstract events: Building event schemas. *Memory & Cognition*, 34, 1221-1235.
- Hare, M., Jones, M. N., Thomson, C., Kelly, S., & McRae, K., (2009). Activating event knowledge. *Cognition*, 111, 151-167.
- Harnad, S., (1990). The symbol grounding problem. *Physica D*, 42, 335-346.
- Harris, Z. S., (1954). Distributional Structure, *WORD*, 10:2-3, 146-162.
- Hochreiter, S., & Schmidhuber, J., (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.
- Hudson, J. A., (1988). Children's memory for atypical actions in script-based stories: Evidence for a disruption effect. *Journal of Experimental Child Psychology*, 46, 159-173.
- Huettig, F., & Altmann, G. T. M., (2004). The online processing of ambiguous and unambiguous words in context: evidence from head-mounted eye-tracking. In Carreiras, M., & Clifton, C., *The on-line study of sentence comprehension: Eyetracking, ERP and beyond*. New York, NY: Psychology Press, 187-207.
- Huettig, F., & Altmann, G. T. M., (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm, *Cognition*, (96), B23-B32.
- Huettig, F., & McQueen, J. M., (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language*, 57, 460-482.
- Huettig, F., & Altmann, G. T. M., (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Visual Cognition*, 15(8), 985-10.
- Humphreys, G. W., & Forde, E. M. E., (1998). Disordered action schema and action disorganization syndrome. *Cognitive Neuropsychology*, 15, 771-811.

- Jackendoff, R., (2002). *Foundations of Language. Brain, Meaning, Grammar, Evolution*. Oxford University Press, New York.
- Jo, J., Bengio, Y., (2017). Measuring the tendency of CNNs to learn surface statistical regularities. In *ArXiv: 1711.11561*.
- Johnson-Laird, P. N., (1983). *Mental Models*. Cambridge, MA: Harvard University Press.
- Johnson-Laird, P. N., (1989). Mental models in M. I. Posner (Ed.), *Foundations of Cognitive Science*. Cambridge, MA: MIT Press.
- Kádár, A., Chrupała, G., & Alishahi, A., (2017). Representation of linguistic form and function in Recurrent Neural Networks. *Computational Linguistics*, 43 (4).
- Kamide, Y., Altmann, G. T. M., & Haywood, S. L., (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, 49(1), 133-156.
- Kiela, D., & Bottou, L., (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of EMNLP*, Doha, Qatar, 36-45.
- Kiela, D., Hill, F., Korhonen, A., & Clark, S., (2014). Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of ACL*, Baltimore, MD, 835-841.
- Kintsch, W., (2001). Predication. *Cognitive Science*, 25, 173-202.
- Knoeferle, P., & Crocker, M. W., (2006). The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive Science*, 30, 481-529.
- Knoeferle, P., & Crocker, M. W., (2007). The influence of recent scene events on spoken comprehension: Evidence from eye movements. *Journal of Memory and Language*, 57, 519-543.
- Knoeferle, P., Carminati, M. N., Abashidze, D., & Essig K., (2011). Preferential Inspection of Recent Real-World Events Over Future Events: Evidence from Eye Tracking during Spoken Sentence Comprehension. *Frontiers in Psychology*, 2(376).
- Knoeferle, P., & Guerra, E., (2016). Visually situated language comprehension. *Language and Linguistic Compass*, 10(2), 66-82.
- Knutson, K., M., Wood, J. N., & Grafman, J., (2004). Brain activation in processing temporal sequence: An fMRI study. *NeuroImage*, 29, 1299.

- Krizhevsky, A., Sutskever, I., & Hinton, G. E., (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097-1105.
- Kuperberg, G. R., & Jaeger, T. F., (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31:1, 32-69.
- Kurby, C. A., & Zacks, J. M., (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, 12, 72-79.
- Kutas, M., & Federmeier, K. D., (2011). Thirty years and counting: Finding meaning in the N400 component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, 62, 621-47.
- Lazaridou, A., Pham, N. T., & Baroni, M., (2015). Combining Language and Vision with a Multimodal Skip-gram Model. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, Denver, Colorado, 153-163.
- Lenci, A., (2008). Distributional semantics in linguistic and cognitive research. In *Rivista di Linguistica*, 20 (1), 1-31.
- Lenci, A., (2011). Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 58-66.
- Lenci, A., (2018). Distributional models of word meaning. *Annual Review of Linguistics*, 4, 151-71.
- Lohmann, H., & Tomasello, M., (2003). The role of language in the development of false belief understanding: A training study. In *Child Development*, 74(4), 1130-1144.
- Long, D. L., Golding, J. M., Graesser, A. C., & Clark, L. F., (1990). Goal, event, and state inferences: An investigation of inference generation during story comprehension. *Psychology of Learning and Motivation*, 25, 89-102.
- Lutz, M. F., & Radvansky, G. A., (1997) The fate of completed goal information in narrative comprehension. *Journal of Memory and Language*, 36, 293-310.
- Ma, L., & Xu, F., (2013). Preverbal infants infer intentional agents from the perception of regularity. *Developmental Psychology*, 49, 7, 1330-1337.
- Magliano, J. P., & Radvansky, G. A., (2001). Goal coordination in narrative comprehension. *Psychonomic Bulletin & Review*, 8, 372-376.

- Marconi, D., (1999). *La Competenza Lessicale*. Roma-Bari, Laterza.
- McMurray, B., Tanenhaus, M. K., & Aslin, R. N., (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, 86, B33–B42.
- McNerney, M. W., Goodwin, K. A., & Radvansky, G. A., (2011). A novel study: A situation model analysis of reading times. *Discourse Processes*, 48, 453-474.
- McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38, 283-312.
- McRae, K., Hare, M., Elman, J. L., & Ferretti, T. R., (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, 33, 1174-1184.
- Melamud, O., Dagan, I., Goldberger, J., Szpektor, I., & Yuret, D., (2014). Probabilistic modeling of joint-context in distributional similarity. In *CoNLL*, 181-190.
- Meyer, D. E., & Schvaneveldt, R. W., (1971). Facilitation in recognizing pair of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2), 227-234.
- Mikolov, T., Chen, K., Corrado, G. S., Dean, J., (2013a). Efficient estimation of word representations in vector space. In *ArXiv:1301.3781*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J., (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th Conference on Advances in Neural Information Processing Systems*, 3111-19.
- Mikolov, T., Yih, W-T., Zweig, G., (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg, 746-51.
- Mitchell, T., (1997). *Machine Learning*. McGraw Hill.
- Mitchell, J., & Lapata, M., (2010). Composition in distributional models of semantics. *Cognitive Science*, 34, 1388-1429.
- Nelson, K., & Gruendel, J., (1986). Children's scripts. In K. Nelson (Ed.). *Event knowledge: Structure and function in development*, Hillsdale, NJ: Erlbaum, 21-46.
- Newtson, D., (1973). Attribution and the unit of perception of ongoing behaviour. *Journal of Personality and Social Psychology*, 28, 28-38.

- Niv, Y., & Schoenbaum, G., (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences*, 12, 265-272.
- Partiot, A., Grafman, J., Sadato, N., Flitman, S., & Wild, K., (1996). Brain activation during scripts event processing. *Neuroreport*, 7, 761-766.
- Pennington, J., Socher, R., & Manning, C. D., (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543.
- Radvansky G. A., & Zacks, J. M., (2014). *Event Cognition*. Oxford University Press, NY.
- Rao, R. P., & Ballard, D., (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2, 79-87.
- Rayner, K., Carlson, M., & Frazier, L., (1983). The interaction of syntax and semantics during sentence processing. *Journal of Verbal Learning and Verbal Behavior*, 22, 358-74.
- Rotaru, A. S., & Vigliocco, G., (2019). Modelling semantics by integrating linguistic, visual and affective information. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci)*, Montreal, Canada, 2681-2687.
- Ruder, S., (2016). An overview of gradient descent optimization algorithms. In *ArXiv:1609.04747*.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, Li. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115, 211-252.
- Saffran, J. R., (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, 12, 110-114.
- Salverda, A., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, 90, 51–89.
- Santus, E., Chersoni, E., Lenci, A., & Blache, P., (2017). Measuring Thematic Fit with Distributional Feature Overlap. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 659-669.
- Sayeed, A., & Demberg, V., (2014). Combining unsupervised syntactic and semantic models of thematic fit. In *Proceedings of the first Italian Conference on Computational Linguistics (CLiC-it)*.

- Sayeed, A., Demberg, V., & Shkadzko, P., (2015). An exploration of semantic features in an unsupervised thematic fit evaluation framework. *Italian Journal of Computational Linguistic, 1(1)*.
- Saylor, M., Baldwin, D., Baird, J., & LaBounty, J., (2007). Infants' on-line segmentation of dynamic human action. *Journal of Cognition and Development, 8*, 113-128.
- Scheepers, C., Keller, F., & Lapata, M., (2008). Evidence for serial coercion: A time course analysis using the visual-world paradigm. *Cognitive Psychology, 56(1)*, 1-29.
- Searle, J., (1984). *Minds, Brains and Science*. Harvard University Press, Cambridge, MA.
- Searle, J. R., (2019). *Il mistero della realtà*. Raffaello Cortina Editore, Milano.
- Shatzman, K. B., & McQueen, J. M., (2006). The modulation of lexical competition by segment duration. *Psychonomic Bulletin & Review, 13*, 966-971.
- Shatzman, K. B., & McQueen, J. M., (2006). Segment duration as a cue to word boundaries in spoken-word recognition. *Perception & Psychophysics, 68*, 1-16.
- Silberer, C., & Lapata, M., (2014). Learning grounded meaning representations with autoencoders. In *Proceedings of ACL*, Baltimore, Maryland, 721-732.
- Sirigu, A., Zalla, T., Pillson, B., Grafman, J., Agid, Y., & Dubois, B., (1995). Selective impairments in managerial knowledge following pre-frontal cortex damage. *Cortex, 31*, 301-316.
- Sirigu, A., Zalla, T., Pillson, B., Grafman, J., Agid, Y., & Dubois, B., (1996). Encoding of sequence and boundaries of scripts following prefrontal lesions. *Cortex, 32*, 297-310.
- Sommerville, J. A., & Woodward, A. L., (2005). Pulling out the intentional structure of action: The relation between action processing and action production in infancy. *Cognition, 95*, 1-30.
- Speer, N. K., Reynolds, J. R., & Zacks, J. M., (2007). Human brain activity time-locked to narrative event boundaries. *Psychological Science, 18*, 449-455.
- Spelke, E. S., & Kinzler, K. D., (2007). Core knowledge. *Developmental Science, 10*, 89-96.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*, 1929-1958.

- Swallow, K. M., Zacks, J. M., & Abrams, R. A., (2009). Event boundaries in perception affect memory encoding and updating. *Journal of Experimental Psychology: General*, *138*, 236-257.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., (2015). Going deeper with convolution. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, USA, 1-9.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C., (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632-1634.
- Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C., (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research*, *29*(6), 557-580.
- Taraban, R., and McClelland, J. L., (1988). Constituent attachment and thematic role assignment in sentence processing: Influences of content-based expectations. *Journal of Memory and Language*, *27*, 597-632.
- Tilk, O., Demberg, V., Sayeed, A., Klakow, D., & Thater, S., (2016). Event Participant Modelling with Neural Networks. In *Proceedings of EMNLP*.
- Tulving, E., (1985). How many memory systems are there? *American Psychologist*, *40*, 385-398.
- Turney, P. D., Pantel, P., (2010). From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141-188.
- Van Berkum, J. J. A., Brown, C. M., Hagoort, P., & Zwitserlood, P., (2003). Event-related brain potentials reflect discourse-referential ambiguity in spoken language comprehension. *Psychophysiology*, *40*(2), 235-248.
- Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P., (2005) Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443-46.
- Van Dijk, T. A., & Kintsch, W., (1983). *Strategies in Discourse Comprehension*. New York: Academic Press.

- Whitney, C., Huber, W., Klann, J., Weis, S., Krach, S., & Kircher, T., (2009). Neural correlates of narrative shifts during auditory story comprehension. *NeuroImage*, *47*, 360-366.
- Wynn, K., (1996). Infants' individuation and enumeration of actions. *Psychological Science*, *7*, 164-169.
- Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., et al., (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, *4*, 651-655.
- Zacks, J. M., Speer, N. K., Vettel, J. M., & Jacoby, L. L., (2006). Event understanding and memory in healthy aging and dementia of the Alzheimer type. *Psychology & Aging*, *21*, 466-482.
- Zacks, J. M., Swallow, K. M., Vettel, J. M., & McAvoy, M. P., (2006). Visual movement and the natural correlates of event perception. *Brain Research*, *1076*, 150-162.
- Zacks, J.M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R., (2007). Event perception: A mind/brain perspective. *Psychological Bulletin*, *133*, 273-293.
- Zacks, J. M., Speer, N. K., & Reynolds, J. R., (2009). Segmentation in reading and film comprehension. *Journal of Experimental Psychology: General*, *138*, 307-327.
- Zalla, T., Pradat-Diehl, P., & Sirigu, A., (2003). Perception of action boundaries in patients with frontal lobe damage. *Neuropsychologia*, *41*, 1619-1627.
- Zalla, T., Verlut, I., Franck, N., Puzenat, D., & Sirigu, A., (2004). Perception of dynamic action in patients with schizophrenia. *Psychiatry Research*, *128*, 39.
- Zwaan, R. A., Langston, M. C., & Graesser, A. C., (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, *6*, 292-297.
- Zwaan, R. A., Magliano, J. P., & Graesser, A. C., (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 386-397.
- Zwaan, R. A., & Radvansky, G. A., (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*, 162-185.
- Zwaan, R. A., (1999). Situation models: The mental leap into imagined worlds. *Current Directions in Psychological Science*, *8*, 15-18.

Zwaan, R. A., (1999). Five dimensions of narrative comprehension: The event-indexing model. In S. R., Goldman, A. C., Graesser, & P., van den Broek (Eds.), *Narrative Comprehension, Causality, and Coherence: Essay in Honor of Tom Trabasso*. Mahwah, NJ: Lawrence Erlbaum Associates, 93-110.

Appendix

Eye Tracking Experiment

Auditory Sentences

First List

1. The hiker fills the backpack
2. The olympian pulls on the boot
3. The graduate removes the cap
4. The librarian wears the glasses
5. The catcher tries on the glove
6. The swimmer adjusts the goggles
7. The witch puts on the hat
8. The biker loosens the helmet
9. The hiker ties the shoes
10. The cop tightens the vest
11. The letter carrier opens the box
12. The doctor uncaps the bottle
13. The quarterback throws the ball
14. The swimmer fits into the suit
15. The chef empties the cup
16. The cook fills up the pot
17. The photographer shines the light
18. The seamstress turns on the machine
19. The jockey sits on the saddle
20. The cyclist repairs the tire
21. The driver fastens the belt
22. The pilot inflates the balloon
23. The body builder lays back on the bench
24. The jeweller closes the box
25. The police records with the camera
26. The baby sits on the chair

27. The pupil tidies up the desk
28. The programmer presses the keyboard key
29. The kid takes a seat
30. The teacher cleans the board

Second List

31. The student fills the backpack
32. The fisherman pulls on the boot
33. The captain removes the cap
34. The lifeguard wears the glasses
35. The boxer tries on the glove
36. The snowboarder adjusts the goggles
37. The chef puts on the hat
38. The cyclist loosens the helmet
39. The soccer player ties the shoes
40. The sailor tightens the vest
41. The electrician opens the box
42. The bartender uncaps the bottle
43. The shortstop throws the ball
44. The judge fits into the suit
45. The toddler empties the cup
46. The gardener fills up the pot
47. The detective shines the light
48. The bartender turns on the machine
49. The cyclist sits on the saddle
50. The farmer repairs the tire
51. The handyman fastens the belt
52. The clown inflates the balloon
53. The homeless person lays back on the bench
54. The carpenter closes the box
55. The tourist records with the camera
56. The patient sits on the chair

57. The executive tidies up the desk
58. The musician presses the keyboard key
59. The spectator takes a seat
60. The chef cleans the board

Fillers

1. The man plays the banjo on the beach
2. Rebecca is rolling the barrel in the street
3. Susan bought a new bracelet for her nephew
4. The man doesn't like candies
5. Carl gave some cherries to his son
6. The kid is hidden in the chest
7. The woman broke the beach paddle
8. John wrote a letter to his friend
9. He has been a bus driver for ten years
10. Mary lost a button from her favorite sweater
11. He turned off the chandelier
12. She is not able to open the coconut
13. My grandmother visited a German castle during her holiday
14. The clock doesn't show the correct hour
15. A girl fell into the well last summer
16. Peter has learned to play the bass guitar
17. The mouse is eating the cheese
18. The man inserts the coins into the vending machine
19. The cat has slept all day on the pillow
20. My friends like using sticks to eat Japanese food
21. He crops the photo with the scissors
22. The man pours the drink into the glass
23. She has bought a new house for the birds
24. Karen made the tea with her new pot
25. Matthew and Molly have a picnic on the table
26. Linda will decorate the Christmas tree this year
27. My aunt will cook penne following a new recipe

28. The man is packing the suitcase for his trip
29. The boy received a new telescope for his birthday
30. She has put the sauce in the jar

Practise

1. Charlotte can hear the rain from the inside of the tent
2. The fish swims in the bowl
3. The kid prepared the Jack-O-Lantern
4. He is yelling in the megaphone

Results

Condition	Mean				Standard Deviation			
	Agent	Target	Action Related	Agent Related	Unrelated	Target	Action Related	Agent Related
Kid	0.4172	0.2340	0.1889	0.1403	0.4931	0.4234	0.3914	0.3474
Clown	0.4256	0.1238	0.1684	0.1131	0.4944	0.3294	0.3742	0.3167
Shortstop	0.4087	0.1643	0.1256	0.1253	0.4916	0.3705	0.3314	0.3311
Catcher	0.4451	0.1347	0.3385	0.0427	0.4970	0.3414	0.4732	0.2022
Swimmer	0.6167	0.0850	0.1352	0.1077	0.4862	0.2789	0.3420	0.3100
Bartender	0.4331	0.1443	0.2139	0.0611	0.4955	0.3514	0.4100	0.2394
Cyclist	0.4340	0.1831	0.1577	0.1094	0.4956	0.3867	0.3644	0.3121
Cyclist	0.3674	0.2601	0.1540	0.0976	0.4821	0.4387	0.3610	0.2968
Cyclist	0.4837	0.1373	0.2134	0.1204	0.4997	0.3442	0.4097	0.3254
Boxer	0.5102	0.1822	0.0861	0.0835	0.4999	0.3860	0.2805	0.2766
Cop	0.5114	0.1297	0.2267	0.0975	0.4999	0.3359	0.4187	0.2966
Teacher	0.5935	0.1164	0.1622	0.0860	0.4912	0.3207	0.3686	0.2803
Chef	0.3936	0.1331	0.1810	0.1457	0.4886	0.3397	0.3850	0.3528
Spectator	0.3988	0.2067	0.1796	0.1028	0.4897	0.4049	0.3838	0.3038
Programmer	0.5001	0.1986	0.2001	0.0629	0.5000	0.3989	0.4001	0.2429
Cook	0.4804	0.1612	0.2289	0.0850	0.4996	0.3678	0.4201	0.2789
Chef	0.4158	0.1336	0.1895	0.1220	0.4929	0.3403	0.3919	0.3273
Judge	0.3633	0.1469	0.1830	0.1533	0.4810	0.3540	0.3867	0.3603
Bartender	0.4965	0.1231	0.1634	0.0877	0.5000	0.3286	0.3697	0.2829
Body Builder	0.4811	0.1838	0.1981	0.0812	0.4997	0.3873	0.3986	0.2732
Detective	0.3541	0.2439	0.1408	0.1419	0.4783	0.4294	0.3478	0.3490
Quarterback	0.4726	0.0620	0.2803	0.1235	0.4993	0.2411	0.4492	0.3290
Electrician	0.4044	0.0992	0.1910	0.1072	0.4908	0.2990	0.3931	0.3094
Graduate	0.5213	0.1278	0.2524	0.0544	0.4996	0.3338	0.4344	0.2269
Baby	0.6455	0.1138	0.1379	0.0684	0.4784	0.3176	0.3448	0.2524
Hiker	0.5355	0.1301	0.1971	0.0942	0.4987	0.3364	0.3978	0.2921
Hiker	0.6169	0.1932	0.0900	0.0342	0.4861	0.3948	0.2862	0.1818
Jockey	0.5663	0.1401	0.0886	0.1284	0.4956	0.3471	0.2842	0.3346
Pilot	0.4960	0.2083	0.1744	0.0661	0.5000	0.4061	0.3794	0.2484
Sailor	0.4327	0.1527	0.1605	0.0798	0.4955	0.3597	0.3671	0.2709
Letter Carrier	0.3491	0.2177	0.3085	0.1046	0.4767	0.4127	0.4619	0.3061
Chef	0.5211	0.1862	0.1562	0.0858	0.4996	0.3893	0.3631	0.2801
Biker	0.3284	0.3278	0.2383	0.0721	0.4697	0.4694	0.4261	0.2586
Musician	0.5032	0.1443	0.1718	0.0618	0.5000	0.3514	0.3772	0.2409
Executive	0.3886	0.2185	0.1168	0.1496	0.4874	0.4132	0.3212	0.3567
Homeless Person	0.4294	0.1118	0.1408	0.1552	0.4950	0.3151	0.3479	0.3621
Doctor	0.5821	0.1610	0.1311	0.0617	0.4932	0.3676	0.3375	0.2406

Gardener	0.5019	0.0828	0.2060	0.0808	0.5000	0.2756	0.4044	0.2725
Pupil	0.4299	0.2030	0.2100	0.1215	0.4951	0.4022	0.4073	0.3267
Librarian	0.5694	0.1530	0.1824	0.0551	0.4952	0.3600	0.3862	0.2282
Jeweller	0.4797	0.1377	0.2185	0.0789	0.4996	0.3445	0.4132	0.2696
Fisherman	0.3891	0.1343	0.2611	0.1051	0.4876	0.3410	0.4393	0.3067
Student	0.4420	0.1403	0.1205	0.1214	0.4966	0.3473	0.3255	0.3267
Driver	0.5786	0.2029	0.0827	0.0713	0.4938	0.4022	0.2754	0.2573
Police	0.4693	0.2479	0.1545	0.0427	0.4991	0.4318	0.3614	0.2022
Seamstress	0.5285	0.1661	0.1013	0.0980	0.4992	0.3722	0.3018	0.2973
Toddler	0.4080	0.1425	0.1965	0.0928	0.4915	0.3495	0.3973	0.2902
Olympian	0.3754	0.1555	0.2733	0.1659	0.4842	0.3624	0.4457	0.3720
Snowboarder	0.4585	0.1280	0.1467	0.1278	0.4983	0.3341	0.3538	0.3338
Player	0.4530	0.1173	0.1538	0.1187	0.4978	0.3217	0.3607	0.3234
Photographer	0.4090	0.1654	0.2498	0.1084	0.4917	0.3716	0.4329	0.3108
Lifeguard	0.3465	0.1296	0.1524	0.1743	0.4759	0.3359	0.3594	0.3794
Swimmer	0.5505	0.1666	0.0989	0.1041	0.4975	0.3726	0.2985	0.3054
Handyman	0.3505	0.2240	0.1625	0.0829	0.4771	0.4169	0.3690	0.2758
Carpenter	0.4508	0.1299	0.1589	0.1299	0.4976	0.3362	0.3656	0.3362
Farmer	0.3790	0.1960	0.1176	0.1314	0.4852	0.3969	0.3222	0.3378
Captain	0.4243	0.1130	0.1958	0.0993	0.4943	0.3166	0.3968	0.2991
Tourist	0.3582	0.1436	0.2067	0.1406	0.4795	0.3507	0.4050	0.3476
Patient	0.4703	0.1868	0.1360	0.0630	0.4991	0.3898	0.3428	0.2429
Witch	0.5809	0.0827	0.1915	0.0816	0.4934	0.2754	0.3935	0.2738

Table 12 Mean and standard deviation of the eye fixations proportions for each agent.

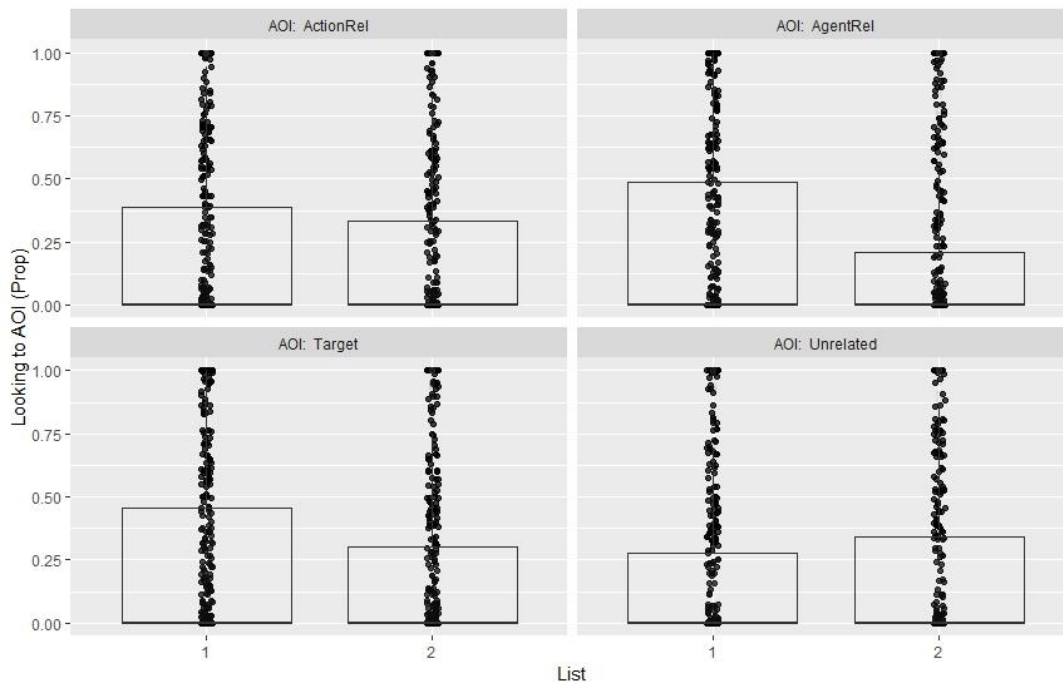


Figure 23 AOIs eye fixations proportions agent time window.

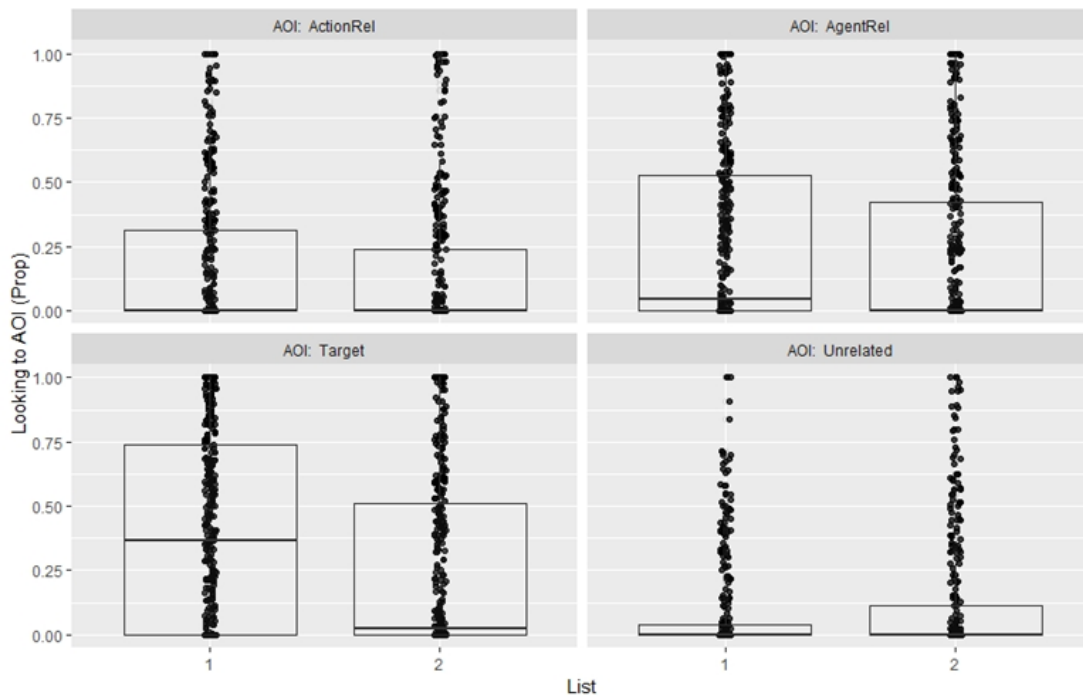


Figure 24 AOIs eye fixations proportions anticipatory (action) time window.

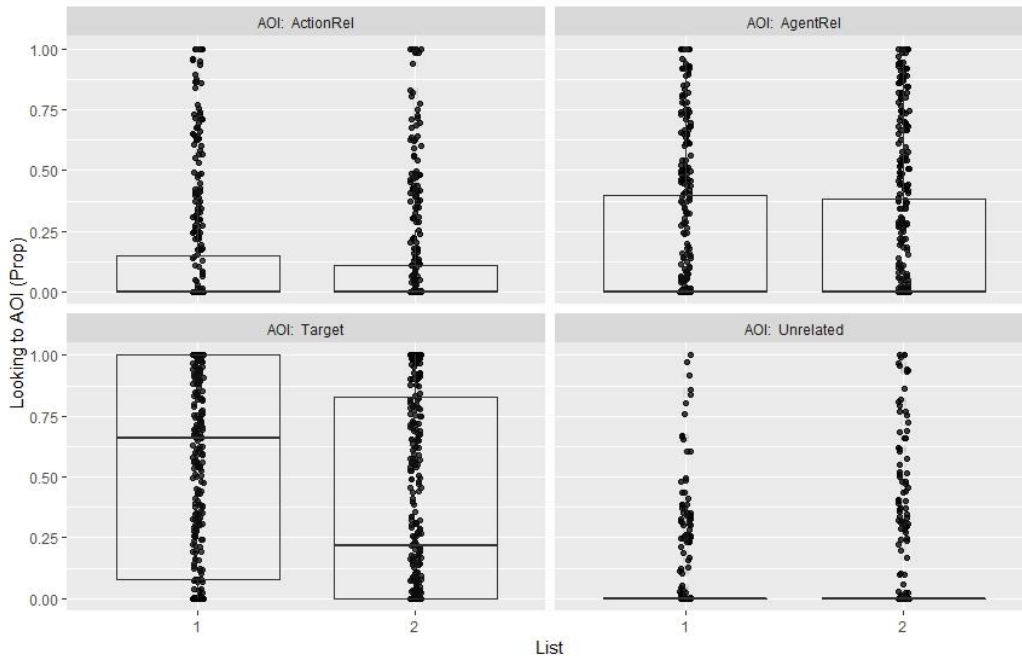


Figure 25 AOIs eye fixations proportions patient time window.

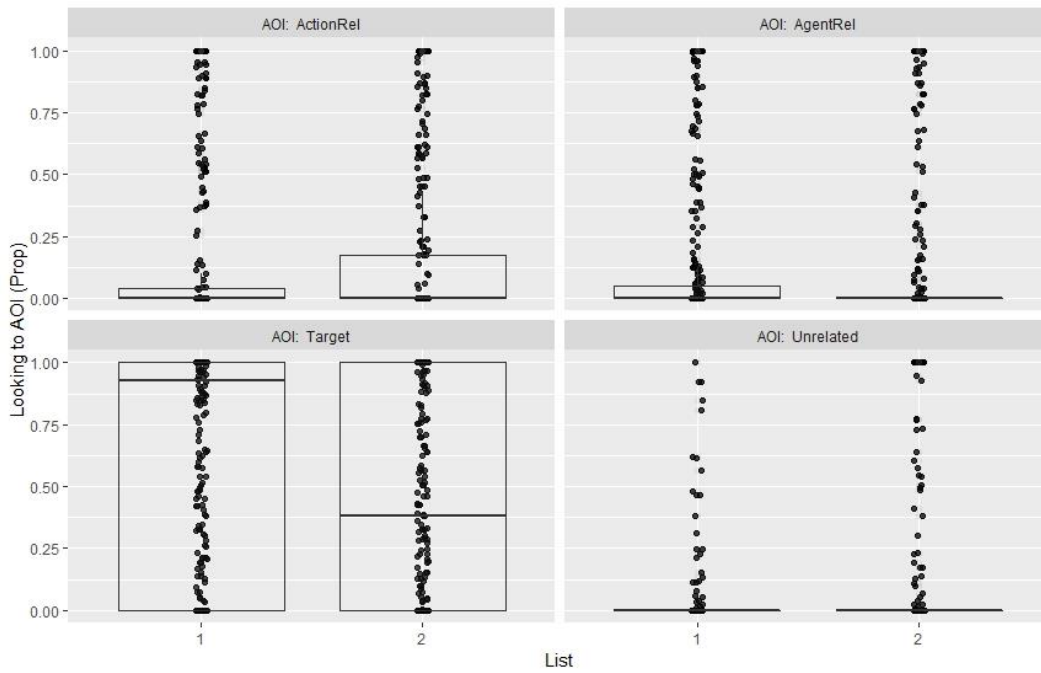


Figure 26 AOIs eye fixations proportions final silence.

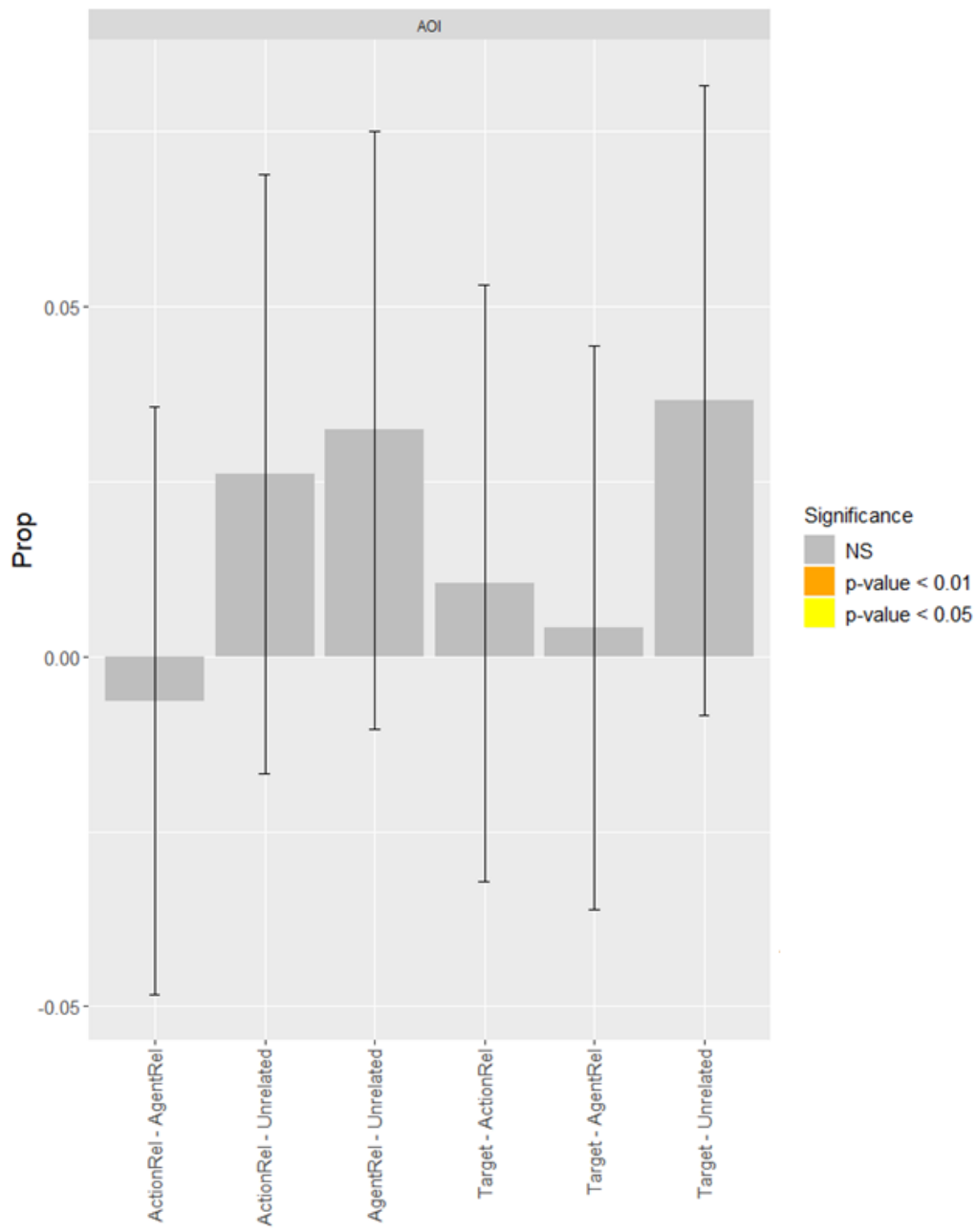


Figure 27 Agent time window.

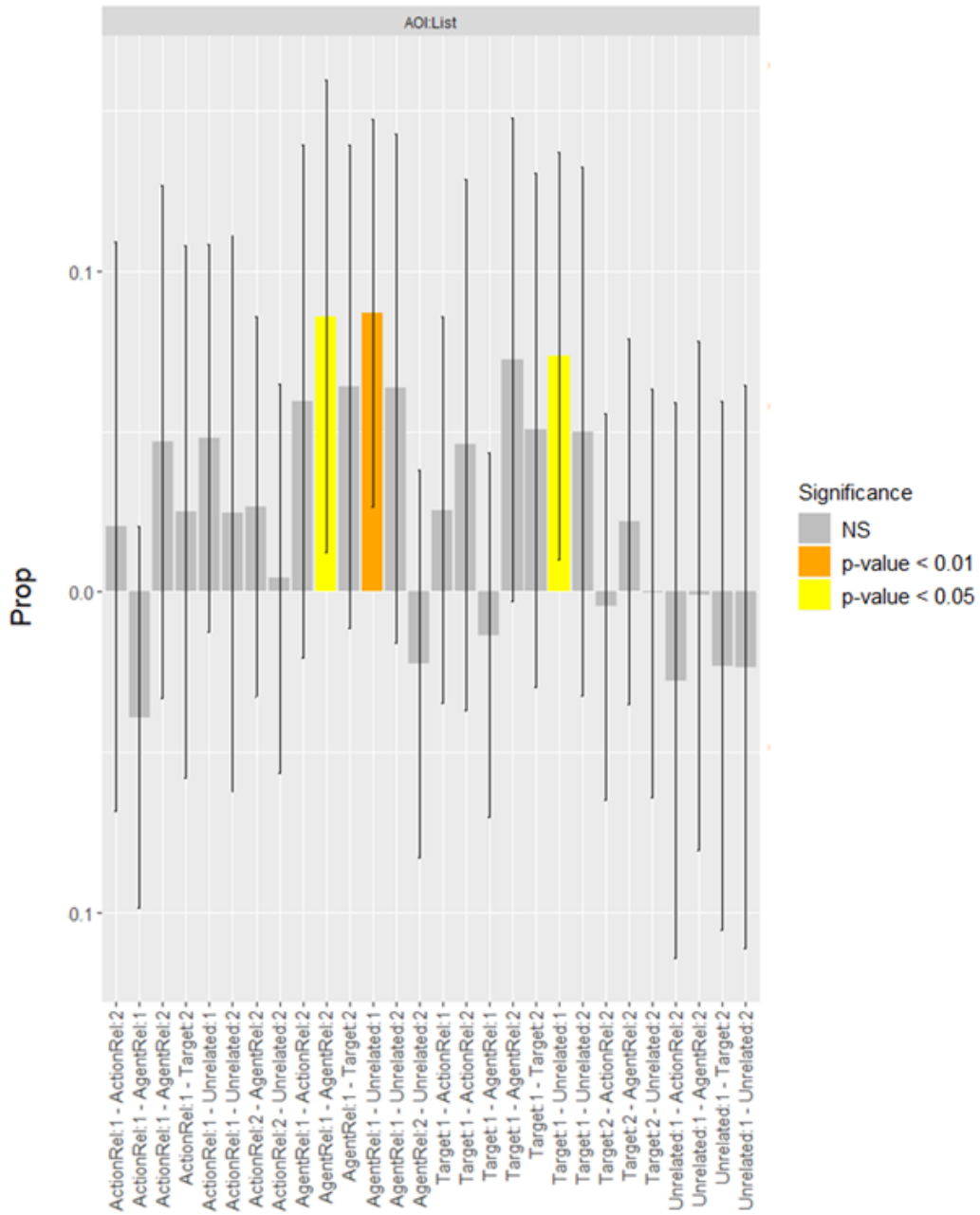


Figure 28 Between the two lists agent time window.

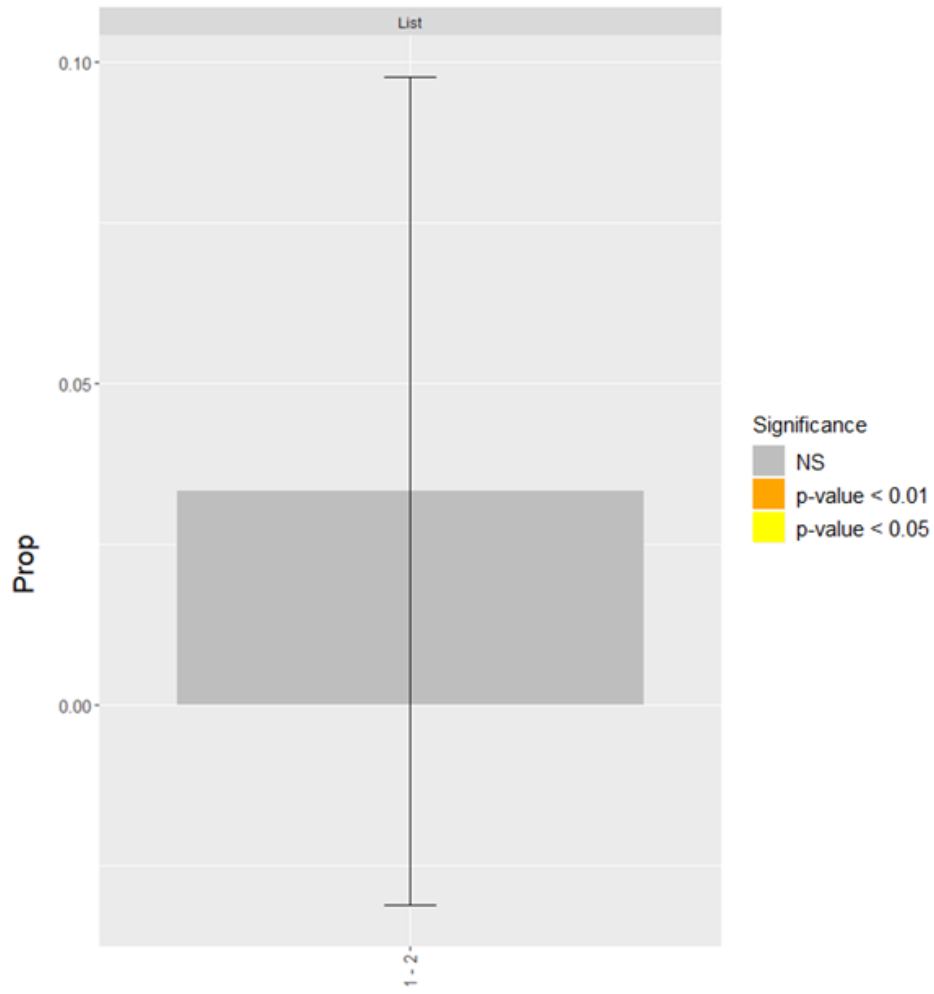


Figure 29 Between the two lists agent time window. Comparisons between the first and second lists.

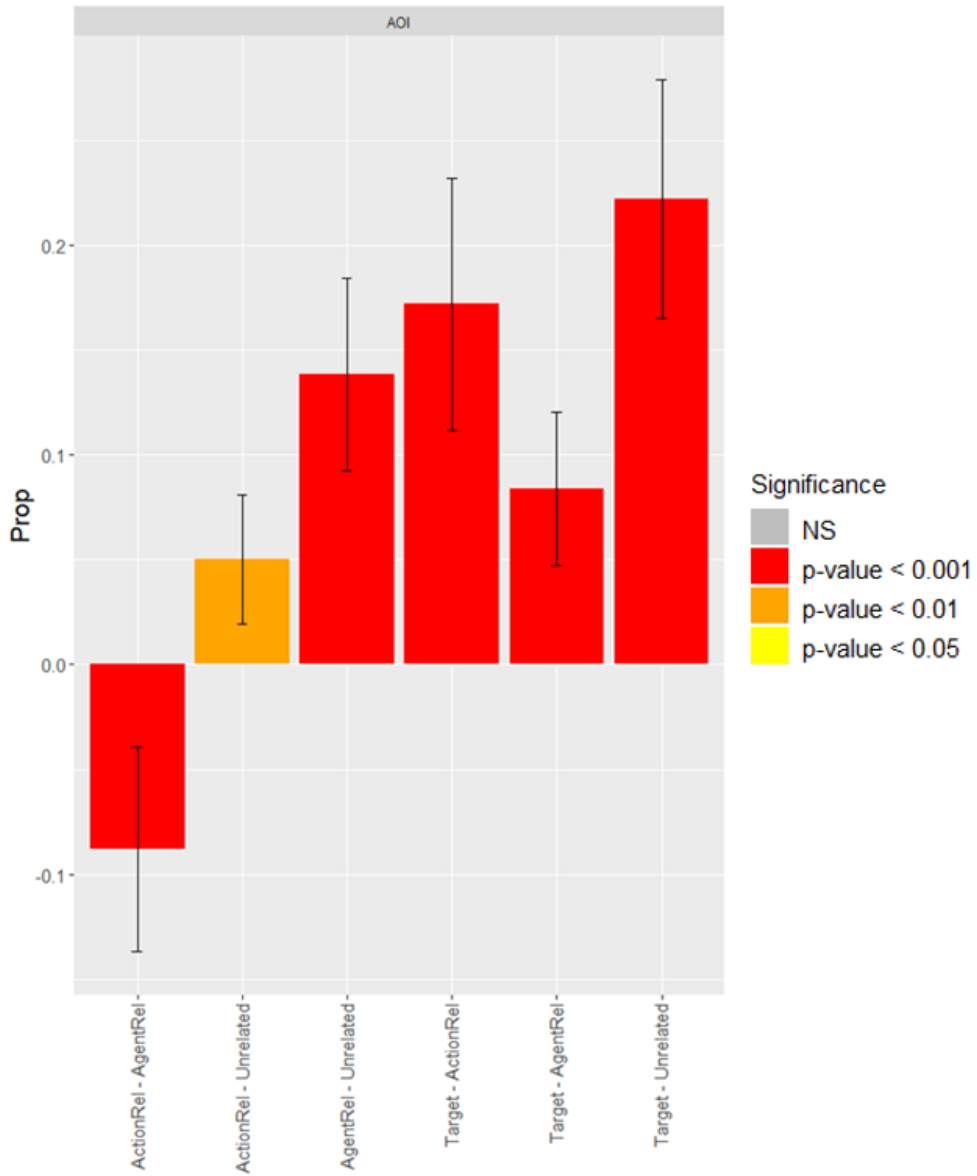


Figure 30 Anticipatory (action) time window.

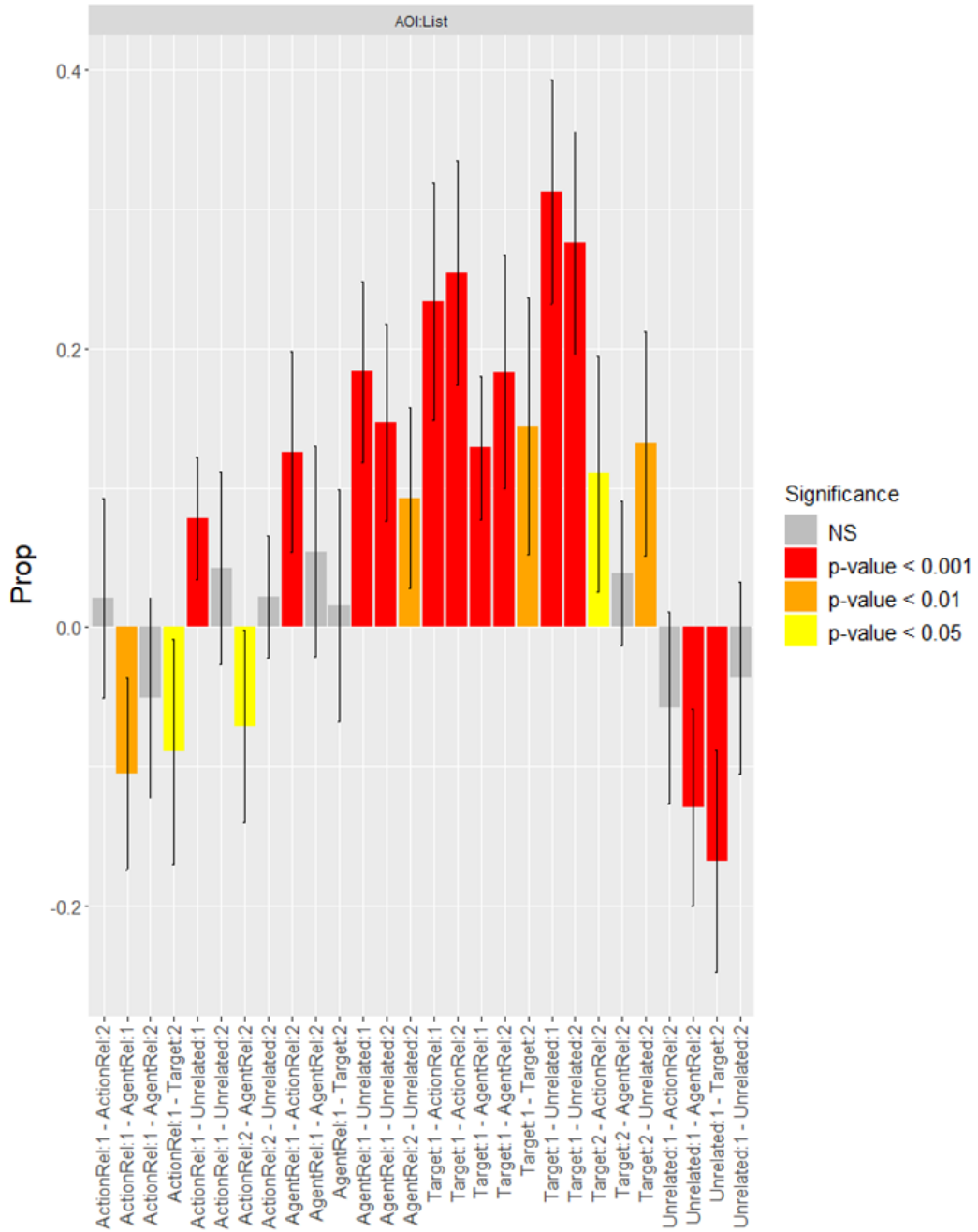


Figure 31 Between the two lists anticipatory (action) time window.

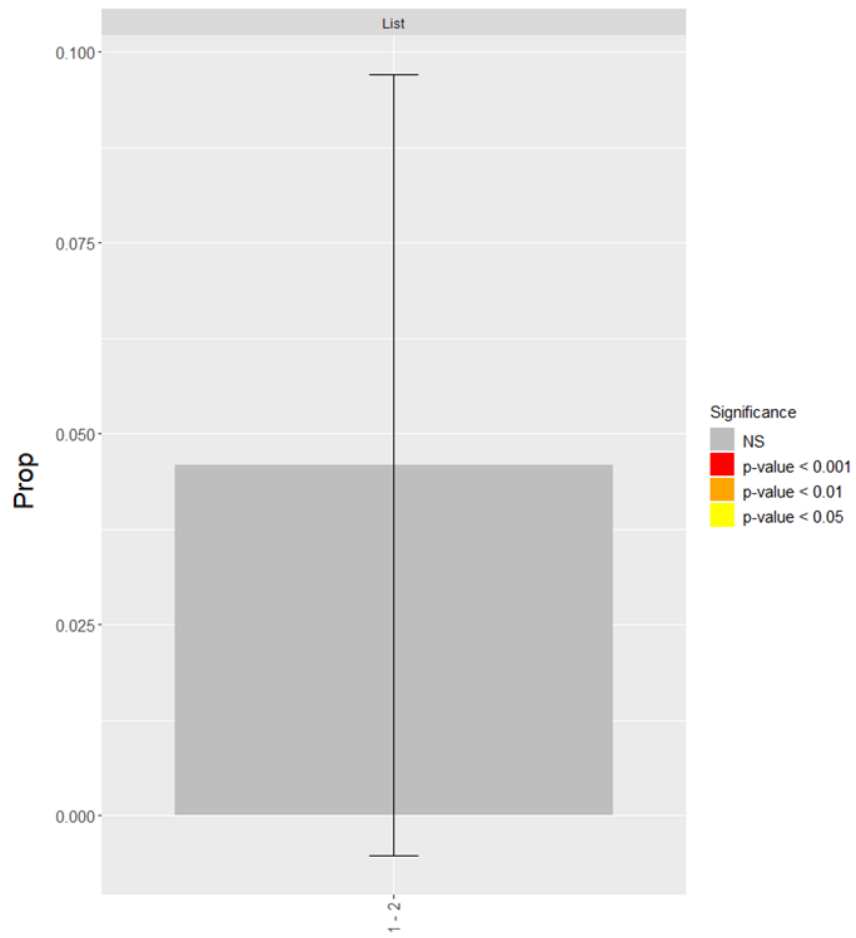


Figure 32 Between the two lists anticipatory (action) time window. Comparisons between the first and second lists.

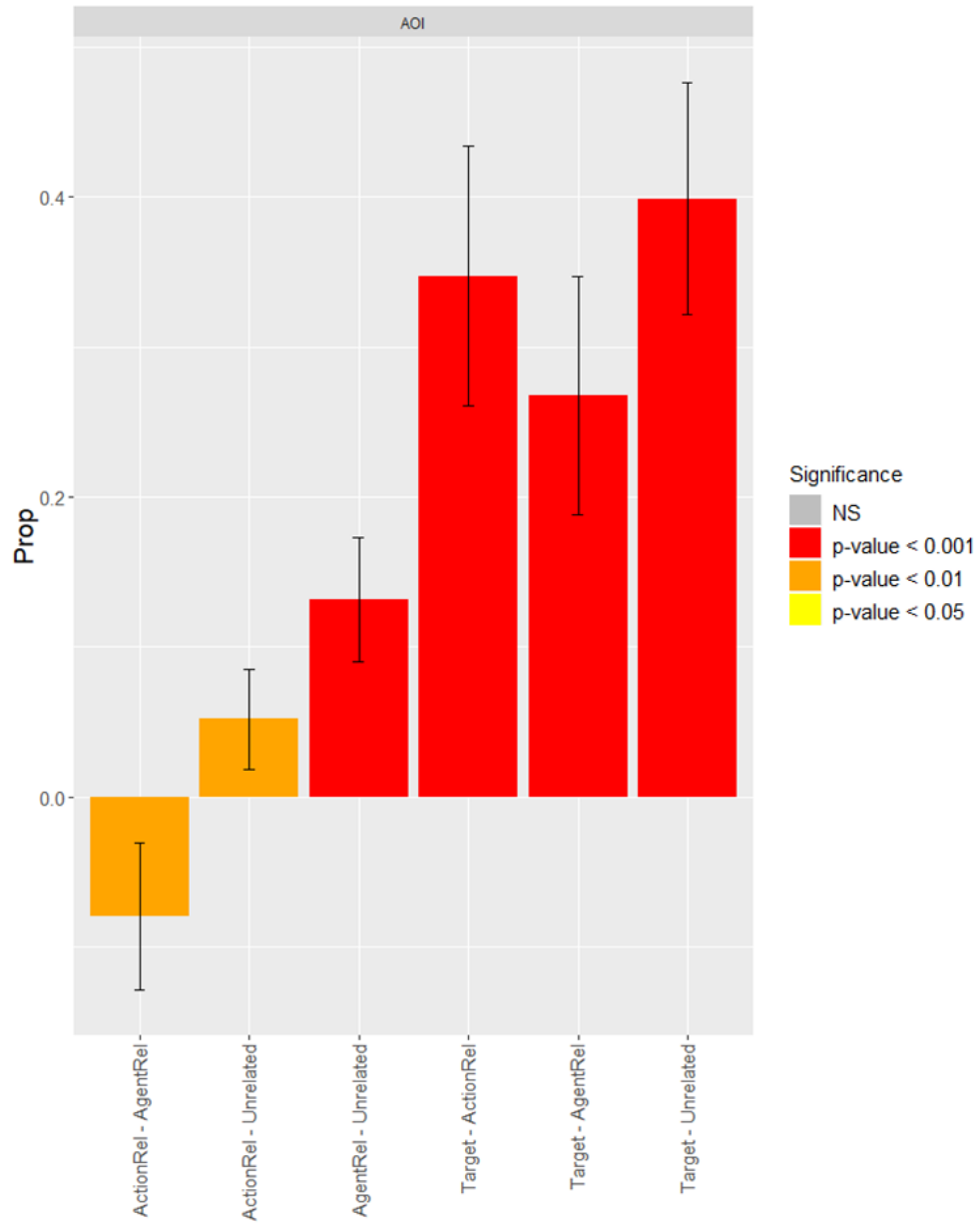


Figure 33 Patient time window.

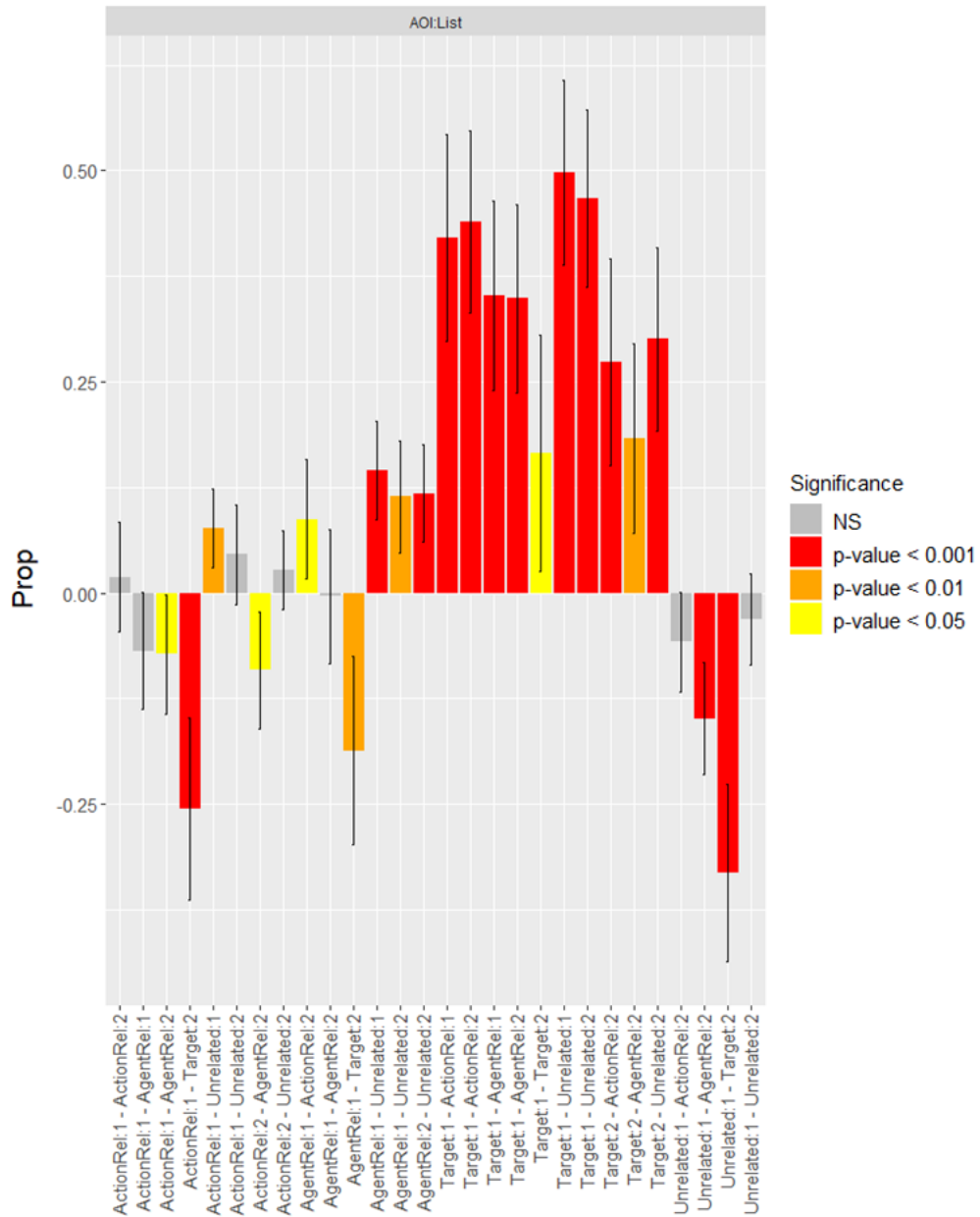


Figure 34 Between the two lists patient time window.

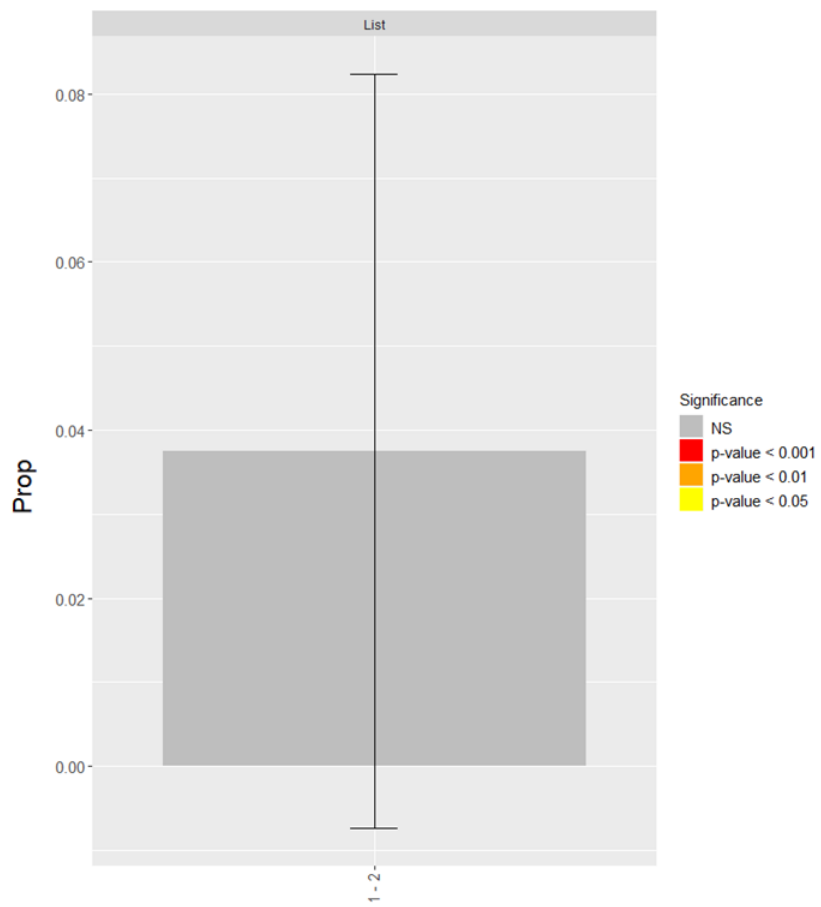


Figure 35 Between the two lists patient time window. Comparisons between the first and second lists.

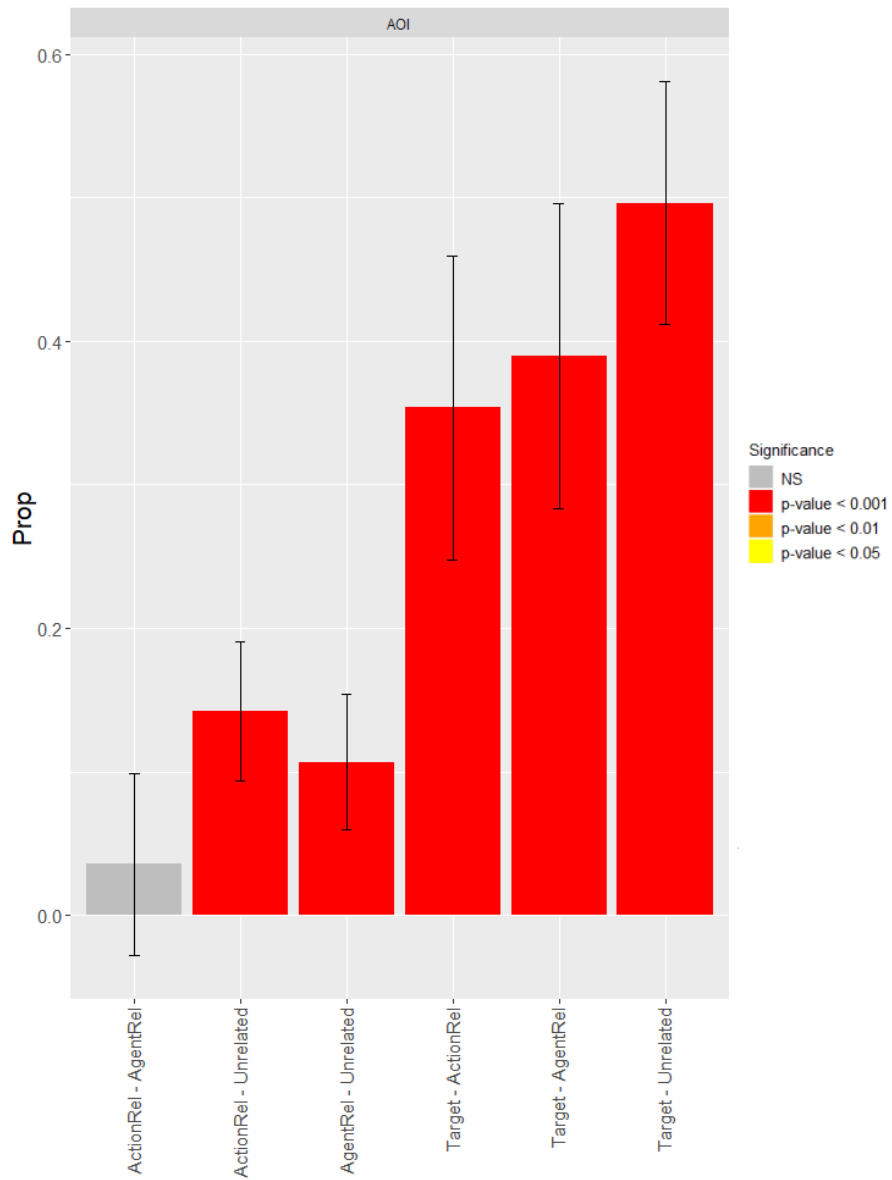


Figure 36 Final silence.

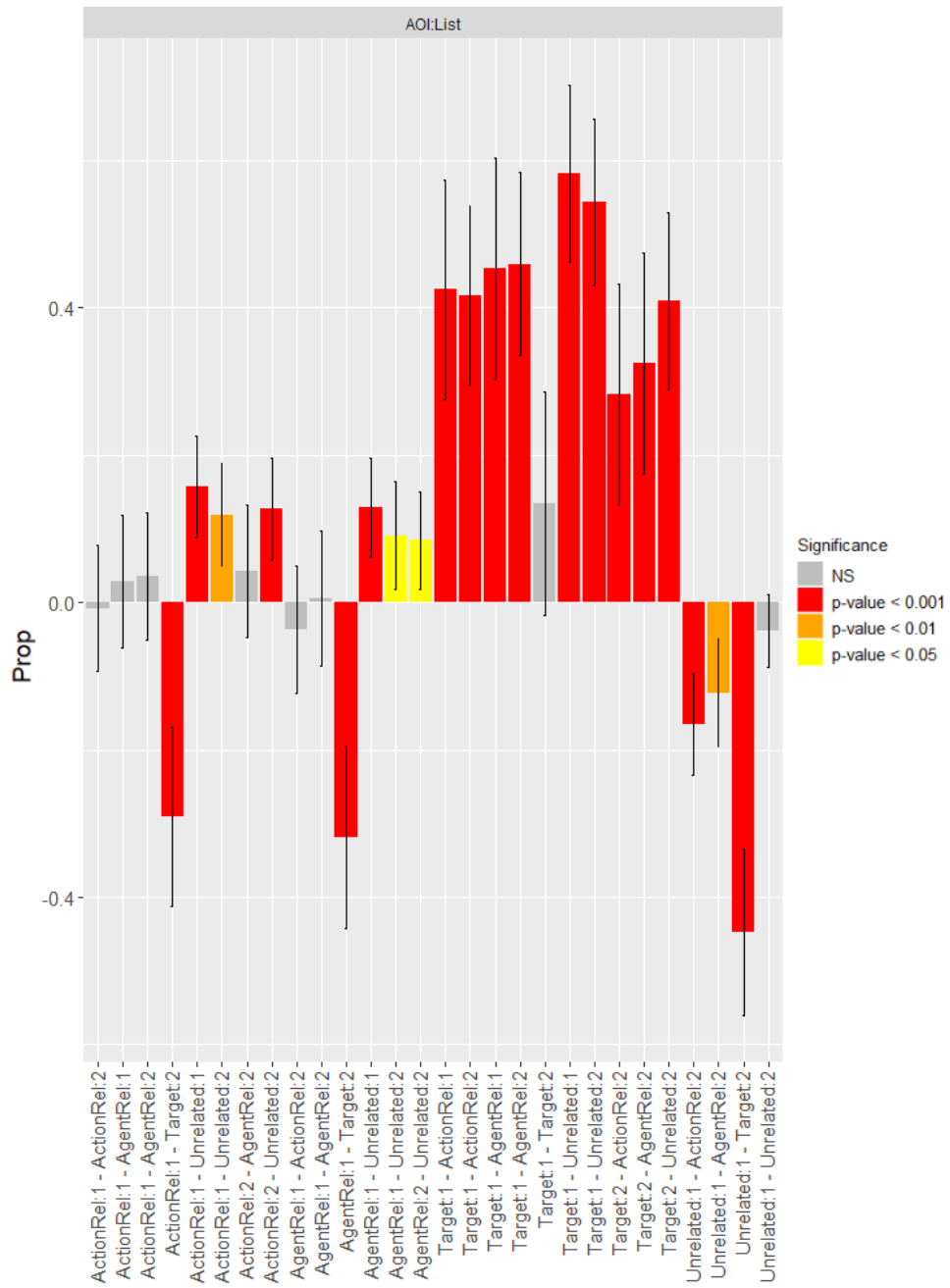


Figure 37 Between the two lists final silence.

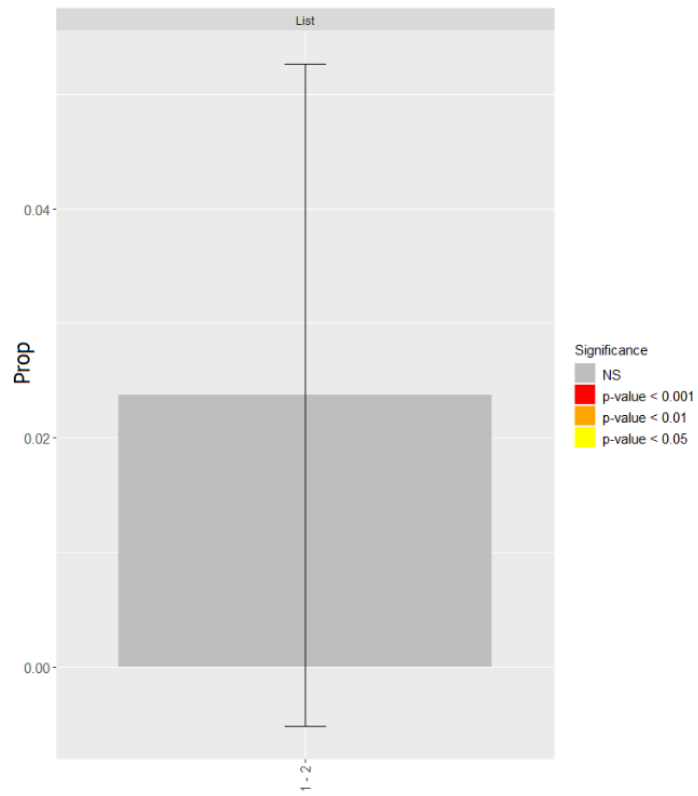


Figure 38 Between the two lists final silence. Comparisons between the first and second lists.

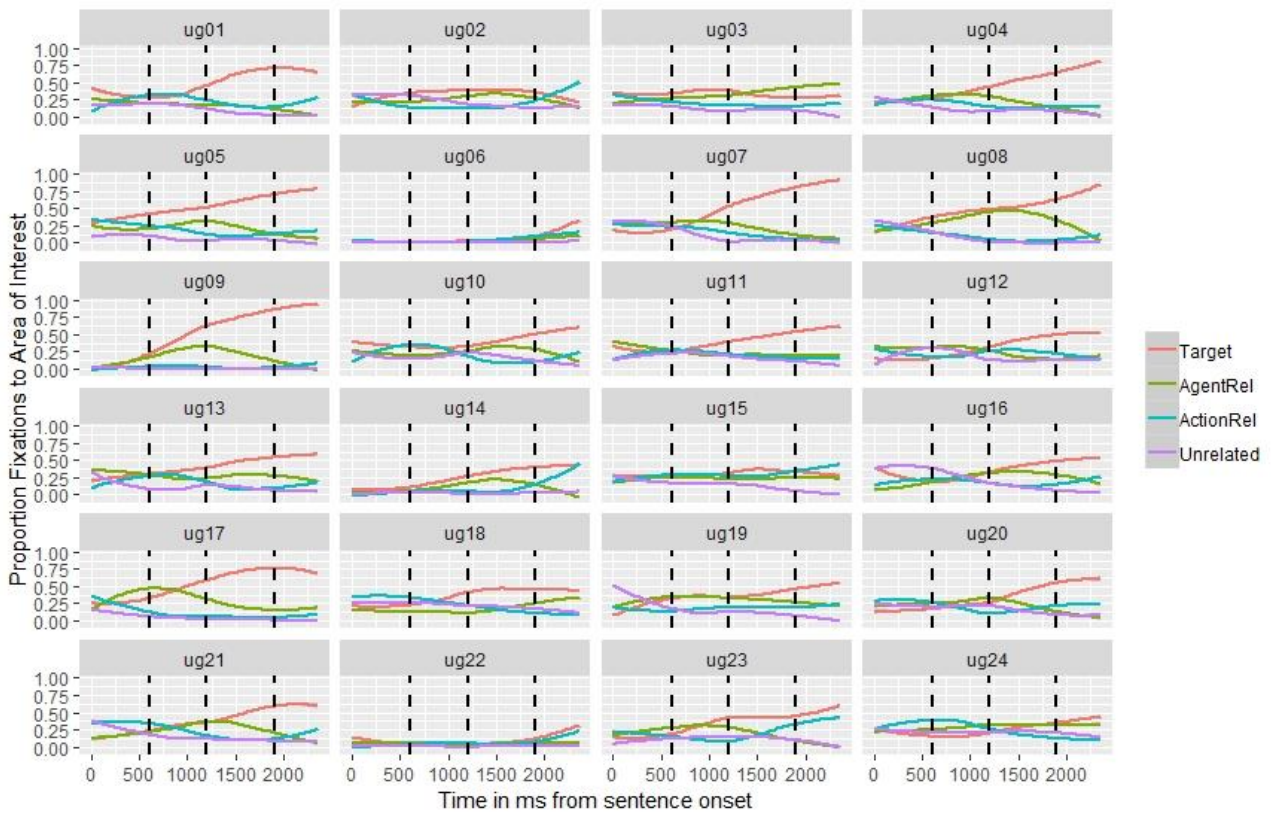


Figure 39 Time course AOIs eye fixations proportions for each participant.

List	Comparison	Estimate	SE	t-value	p-value
Between 1 & 2	Target_List1-Target_List2	0.05	0.04	1.3	0.21
	Target_List1-Action Related_List2	0.05	0.04	1.12	0.27
	Target_List1-Agent Related_List2	0.07	0.04	1.94	0.06
	Target_List1-Unrelated_List2	0.05	0.04	1.23	0.23
	Action Related_List1-Target_List2	0.02	0.04	0.61	0.55
	Action Related_List1-Action Related_List2	0.02	0.04	0.47	0.64
	Action Related_List1-Agent Related_List2	0.05	0.04	1.18	0.24
	Action Related_List1-Unrelated_List2	0.02	0.04	0.57	0.57
	Agent Related_List1-Target_List2	0.06	0.04	1.72	0.09
	Agent Related_List1-Action Related_List2	0.06	0.04	1.5	0.14
	Agent Related_List1-Agent Related_List2	0.09	0.04	2.41	0.02*
	Agent Related_List1-Unrelated_List2	0.06	0.04	1.62	0.11
	Unrelated_List1-Target_List2	-0.02	0.04	-0.57	0.57
	Unrelated_List1-Action Related_List2	-0.03	0.04	-0.65	0.52
	Unrelated_List1-Agent Related_List2	0	0.04	-0.03	0.98
Unrelated_List1-Unrelated_List2	-0.02	0.04	-0.55	0.59	

* p-value < 0.05

Table 13 Between the two lists agent time window.

List	Comparison	Estimate	SE	t-value	p-value
Between 1 & 2	Target_List1 - Target_List2	0.14	0.04	3.23	0.0036*
	Target_List1 - Action Related_List2	0.25	0.04	6.35	8.64e-08*
	Target_List1 - Agent Related_List2	0.18	0.04	4.48	8.34e-05*
	Target_List1 - Unrelated_List2	0.28	0.04	6.99	9.83e-09*
	Action Related_List1 - Target_List2	-0.09	0.04	-2.23	0.0309*
	Action Related_List1 - Action Related_List2	0.02	0.03	0.60	0.5568
	Action Related_List1 - Agent Related_List2	-0.05	0.04	-1.41	0.1640
	Action Related_List1 - Unrelated_List2	0.04	0.03	1.24	0.2214
	Agent Related_List1 - Target_List2	0.02	0.04	0.38	0.7033
	Agent Related_List1 - Action Related_List2	0.13	0.04	3.52	0.0010*
	Agent Related_List1 - Agent Related_List2	0.05	0.04	1.48	0.1507
	Agent Related_List1 - Unrelated_List2	0.15	0.04	4.20	0.0001*
	Unrelated_List1 - Target_List2	-0.17	0.04	-4.25	0.0001*
	Unrelated_List1 - Action Related_List2	-0.06	0.03	-1.70	0.0977
	Unrelated_List1 - Agent Related_List2	-0.13	0.04	-3.68	0.0006*
Unrelated_List1 - Unrelated_List2	-0.04	0.03	-1.08	0.2879	

* p-value < 0.05

Table 14 Between the two lists anticipatory (action) time window.

List	Comparison	Estimate	SE	t-value	p-value
Between 1 & 2	Target_List1 - Target_List2	0.17	0.07	2.45	0.0221*
	Target_List1 - Action Related_List2	0.44	0.05	8.31	1.50e-09*
	Target_List1 - Agent Related_List2	0.35	0.06	6.33	2.06e-07*
	Target_List1 - Unrelated_List2	0.47	0.05	9.06	2.92e-10*
	Action Related_List1 - Target_List2	-0.25	0.05	-4.82	3.26e-05*
	Action Related_List1 - Action Related_List2	0.02	0.03	0.59	0.5602
	Action Related_List1 - Agent Related_List2	-0.07	0.04	-2.06	0.0448*
	Action Related_List1 - Unrelated_List2	0.05	0.03	1.57	0.1239
	Agent Related_List1 - Target_List2	-0.19	0.06	-3.38	0.0017*
	Agent Related_List1 - Action Related_List2	0.09	0.04	2.48	0.0167*
	Agent Related_List1 - Agent Related_List2	0.00	0.04	-0.10	0.9194
	Agent Related_List1 - Unrelated_List2	0.11	0.03	3.47	0.0012*
	Unrelated_List1 - Target_List2	-0.33	0.05	-6.44	3.39e-07*
	Unrelated_List1 - Action Related_List2	-0.06	0.03	-1.99	0.0525
	Unrelated_List1 - Agent Related_List2	-0.15	0.03	-4.53	4.62e-05*
Unrelated_List1 - Unrelated_List2	-0.03	0.03	-1.17	0.2479	

* p-value < 0.05

Table 15 Between the two lists patient time window.

List	Comparison	Estimate	SE	t-value	p-value
Between 1 & 2	Target_List1 - Target_List2	0.13	0.07	1.82	8.14e-02
	Target_List1 - Action Related_List2	0.42	0.06	6.97	6.53e-08*
	Target_List1 - Agent Related_List2	0.46	0.06	7.54	8.56e-09*
	Target_List1 - Unrelated_List2	0.54	0.06	9.87	7.20e-11*
	Action Related_List1 - Target_List2	-0.29	0.06	-4.85	3.02e-05*
	Action Related_List1 - Action Related_List2	-0.01	0.04	-0.18	0.86
	Action Related_List1 - Agent Related_List2	0.04	0.04	0.81	0.42
	Action Related_List1 - Unrelated_List2	0.12	0.03	3.48	1.14e-03*
	Agent Related_List1 - Target_List2	-0.32	0.06	-5.24	8.24e-06*
	Agent Related_List1 - Action Related_List2	-0.04	0.04	-0.84	0.40
	Agent Related_List1 - Agent Related_List2	0.01	0.04	0.14	0.89
	Agent Related_List1 - Unrelated_List2	0.09	0.04	2.52	1.60e-02*
	Unrelated_List1 - Target_List2	-0.45	0.06	-8.13	5.01e-09*
	Unrelated_List1 - Action Related_List2	-0.17	0.03	-4.82	1.76e-05*
	Unrelated_List1 - Agent Related_List2	-0.12	0.04	-3.40	1.54e-03*
Unrelated_List1 - Unrelated_List2	-0.04	0.03	-1.53	0.13	

* p-value < 0.05

Table 16 Between the two lists final silence.

List	AOI	Estimate	SE	t-value	p-value
In 1 & 2	Target	0.22	0.02	11.12	5.97e-11*
	Action Related	0.21	0.02	9.56	1.17e-09*
	Agent Related	0.21	0.02	11.88	1.55e-11*
	Unrelated	0.18	0.02	8.44	1.21e-08*
1	Target	0.24	0.03	8.78	5.86e-09*
	Action Related	0.22	0.03	7.09	2.47e-07*
	Agent Related	0.25	0.03	10.10	4.05e-10*
	Unrelated	0.17	0.03	5.58	9.75e-06*
2	Target	0.19	0.03	6.94	3.51e-07*
	Action Related	0.20	0.03	6.43	1.20e-06*
	Agent Related	0.17	0.03	6.70	6.31e-07*
	Unrelated	0.19	0.03	6.36	1.43e-06*

* p-value < 0.05

Table 17 Agent time window.

List	AOI	Estimate	SE	t-value	p-value
In 1 & 2	Target	0.34	0.02	15.05	1.00e-13*
	Action Related	0.17	0.02	9.49	5.97e-10*
	Agent Related	0.25	0.02	13.80	6.46e-13*
	Unrelated	0.11	0.02	6.89	2.10e-07*
1	Target	0.41	0.03	12.93	2.64e-12*
	Action Related	0.18	0.02	7.13	1.39e-07*
	Agent Related	0.28	0.03	10.81	1.04e-10*
	Unrelated	0.10	0.02	4.10	0.000336*
2	Target	0.26	0.03	8.36	1.42e-08*
	Action Related	0.15	0.02	6.29	1.15e-06*
	Agent Related	0.23	0.03	8.71	6.75e-09*
	Unrelated	0.13	0.02	5.64	5.52e-06*

* p-value < 0.05

Table 18 Anticipatory (action) time window.

List	AOI	Estimate	SE	t-value	p-value
In 1 & 2	Target	0.48	0.03	14.12	4.00e-13*
	Action Related	0.13	0.02	8.33	8.63e-09*
	Agent Related	0.21	0.02	11.05	6.50e-11*
	Unrelated	0.08	0.01	6.07	4.82e-07*
1	Target	0.56	0.05	11.72	2.04e-11*
	Action Related	0.14	0.02	6.31	1.15e-06*
	Agent Related	0.21	0.03	7.74	5.51e-08*
	Unrelated	0.06	0.02	3.46	0.001354*
2	Target	0.40	0.05	8.26	1.80e-08*
	Action Related	0.12	0.02	5.48	9.86e-06*
	Agent Related	0.21	0.03	7.88	3.98e-08*
	Unrelated	0.10	0.02	5.12	9.38e-06*

* p-value < 0.05

Table 19 Patient time window.

List	AOI	Estimate	SE	t-value	p-value
In 1 & 2	Target	0.54	0.04	14.73	1.61e-13*
	Action Related	0.19	0.02	9.16	2.55e-09*
	Agent Related	0.15	0.02	6.95	3.31e-07*
	Unrelated	0.05	0.01	3.80	2.10e-04*
1	Target	0.61	0.05	11.70	2.10e-11*
	Action Related	0.19	0.03	6.35	1.43e-06*
	Agent Related	0.16	0.03	5.02	3.90e-05*
	Unrelated	0.03	0.02	1.61	0.11
2	Target	0.48	0.05	9.13	2.82e-09*
	Action Related	0.19	0.03	6.61	7.65e-07*
	Agent Related	0.15	0.03	4.81	6.55e-05*
	Unrelated	0.07	0.02	3.77	2.39e-04*

* p-value < 0.05

Table 20 Final silence.

Analyses

title: "Multimodality and Prediction in Sentence Comprehension"

author: "Valentina Benedettini"

date: "10 luglio 2018"

output: html_document

```
`` {r setup, include=FALSE}
```

```
knitr::opts_chunk$set(echo = TRUE)
```

```
rm(list=ls())
```

```
## Summary
```

- This was originally written by Arielle Borovsky to perform an analysis of the Familiar Frames study with Kids using the REyetracking tools code.

- It is elaborated for the "Multimodality and Expectations in Language Comprehension" study. It was performed at Western Ontario University thanks to the collaboration of Ken McRae, Mikayla Hall Bruce, Thea Lucille Knowles, Juweiriya Ahmed.

```
### Additional resources
```

- Check out the documentation for the [eyetrackingR package](<http://www.eyetracking-r.com/>)

- See lab protocols and tutorials used for the eyetracking studies in [Ken McRae's lab](<https://sites.google.com/site/kenmcraelab/lab-tutorials>)

```
## Load Libraries
```

```
`` {r_Load_Libraries}
```

```
library("eyetrackingR")
```

```
library("Matrix")
```

```
library("lme4")
```

```
library("ggplot2")
```

```
library("pbapply")
```

```
library("plyr")
```

```
library("LMERConvenienceFunctions")
```

```
library("lmerTest")
```

```
library("stringr")
```

```
library("sjPlot")
```

```

library("nlme")
## Set working directory
```{r_Set_Work_Directory}
setwd("C:/Users/Valentina/Documents/VerbSemantics_PhD_Thesis/PsycholinguisticExperiment/MyExperiment/SampleReport_200618/Output")
Set other environmental variables (means of "utteranceInfo_editedAudio")
```{r_Set_Environmental_Variables}
SampleReportName <- "SampleReport_200618.txt"
TrackLossThreshold <- 0.8
UtteranceStart <- 456.75
AgentOnset <- 610.318
AgentOffset <- 1200
VerbOnset <- 1200
VerbOffset <- 1898.81
ObjectOnset <- 1898.81
ObjectOffset <- 2523.725
UtteranceEnd <- 2823.725
####STEP 1:** LOAD A SAMPLE REPORT WITH AT LEAST THE FOLLOWING
COLUMNS (import Dataset)
1. RECORDING_SESSION_LABEL
+ participant number
2. TRIAL_LABEL
+ to determine the trial number
3. IP_START_TIME, IP_END_TIME
+ to determine the timing of sample within the trial
4. RIGHT_GAZE_X, RIGHT_GAZE_Y
####STEP 2:** READ SAMPLE REPORT INTO R
```{r_Read_Sample_Report}
ETData <- read.table(SampleReportName, header=TRUE, sep="\t")
#Remove practice trials with no data that were not part of the study
ETData <- subset(ETData, condition!="N/A")
Make a new condition column until we figure out where condition was stored

```

```

ETData$Cond <- as.factor(substr(ETData$referent_identifier,1,20))
levels(ETData$Cond)
Make a new list column until we figure out where list was stored
ETData$List <- as.factor(substr(ETData$counterbalance,0,2))
levels(ETData$List)
FIND PARTICIPANTS WHO HAD RECORDING ON THE LEFT EYE, AND
MERGE WITH RIGHT EYE
This is done because we do not think there should be any major differences in recording
of L or R eye
Left eye is sometimes recorded instead of right for various reasons like glasses glare, hair
over eye, etc.
```{r_Merging_Left_And_Right_Eyes}
ETData$RIGHT_INTEREST_AREA_LABEL[ETData$LEFT_INTEREST_AREA_L
ABEL!="."] <-
ETData$LEFT_INTEREST_AREA_LABEL[ETData$LEFT_INTEREST_AREA_LA
BEL!="."]
## SOME PARTICIPANT AND TRIAL INFORMATION
```{r}
N_subjs <-length(unique(ETData$RECORDING_SESSION_LABEL))
There are `r N_subjs` unique subjects in the loaded dataset file
```{r}
Tot_Trials<- ddply(ETData, c("RECORDING_SESSION_LABEL"), summarize,
N=length(unique(TRIAL_LABEL)))
There are a total `r sum(Tot_Trials$N)` unique trials in the raw dataset file
Each subject contributes the following number of trials in the raw dataset file:
`r knitr::kable(Tot_Trials)`
## TIMESTAMP
```{r_Timestamp}
ETData$TIMESTAMP_new <- str_sub(ETData$TIMESTAMP, 1, -4)
ETData$TIMESTAMP_new <- as.numeric(ETData$TIMESTAMP_new)
##**STEP 3:** MARK SAMPLES WHERE TRACKLOSS OCCURS, creating new
TrackLoss column

```

```

```{r_Marking_Trackloss}
ETData$TrackLoss<-          ETData$RIGHT_GAZE_X=="."          &
ETData$RIGHT_GAZE_Y=="."    &          ETData$LEFT_GAZE_X=="."    &
ETData$RIGHT_GAZE_Y=="."
###**STEP 4:** MAKE AOI COLUMNS
```{r_Make_AOI_Columns}
ETData$Target <- ETData$RIGHT_INTEREST_AREA_LABEL=="TARGET_IA"
ETData$ActionRel <- ETData$RIGHT_INTEREST_AREA_LABEL=="COMP_1_IA"
ETData$AgentRel <- ETData$RIGHT_INTEREST_AREA_LABEL=="COMP_2_IA"
ETData$Unrelated <- ETData$RIGHT_INTEREST_AREA_LABEL=="COMP_3_IA"
###**STEP 5:** READ DATA INTO MAKE_EYETRACKINGR_DATA
- Set treat_non_aoi_looks_as_missing to FALSE if you want to make the denominator
equal to total time looking
- Set treat_non_aoi_looks_as_missing to TRUE if you want to make the denominator
equal to looks only to other AOI
```{r_Make_Eye_Tracking_Data}
data <- make_eyetrackingr_data(ETData,
      participant_column = "RECORDING_SESSION_LABEL",
      trial_column = "TRIAL_LABEL",
      time_column = "TIMESTAMP_new",
      trackloss_column = "TrackLoss",
      aoi_columns = c("Target",'AgentRel', 'ActionRel', 'Unrelated'),
      treat_non_aoi_looks_as_missing = FALSE)
###**STEP 6:** Arrange time bins
Define Dataset within whole sentence time window (agent, verb, pronoun/art, object)
```{r_Time_Bins_Trial_Window}
trial_window <- subset_by_window(data, rezero = TRUE, remove = TRUE,
 window_start_col = "IP_START_TIME", window_end_col = "IP_END_TIME")
Define Dataset within Sentence time window
```{r_Time_Bins_Sentence_Window}
data$SentenceWindowStart <- (data$IP_START_TIME + UtteranceStart)
data$SentenceWindowEnd <- (data$IP_START_TIME + UtteranceEnd)

```



```

sentence_window <- subset_by_window(data, rezero = TRUE, remove = TRUE,
  window_start_col = "SentenceWindowStart", window_end_col =
  "SentenceWindowEnd")
Define Dataset within Agent time window
```{r_Time_Bins_Agent_Window}
data$AgentWindowStart <- (data$IP_START_TIME + AgentOnset)
data$AgentWindowEnd <- (data$IP_START_TIME + AgentOffset)
agent_window <- subset_by_window(data, rezero = TRUE, remove = TRUE,
 window_start_col = "AgentWindowStart", window_end_col = "AgentWindowEnd")
Define Dataset within Anticipatory/Verb+art time window
```{r_Time_Bins_Verb+art_Window}
data$VerbWindowStart <- (data$IP_START_TIME + VerbOnset)
data$VerbWindowEnd <- (data$IP_START_TIME + VerbOffset)
verb_art_window <- subset_by_window(data, rezero = TRUE, remove = TRUE,
  window_start_col = "VerbWindowStart", window_end_col = "VerbWindowEnd")
Define Dataset within Object time window
```{r_Time_Bins_Object_Window}
data$ObjectWindowStart <- (data$IP_START_TIME + ObjectOnset)
data$ObjectWindowEnd <- (data$IP_START_TIME + ObjectOffset)
object_window <- subset_by_window(data, rezero = TRUE, remove = TRUE,
 window_start_col = "ObjectWindowStart", window_end_col = "ObjectWindowEnd")
###STEP 7: deal with trackloss
Remove trials where fewer than `r TrackLossThreshold *100`% of samples are in
fixation
Remove excessive trackloss trials for sentence window:
Data:
```{r_Dealing_With_Trackloss_Sentence_Window}
#First, calculate the amount of trackloss by subjects and trials
trackloss_sentence <- trackloss_analysis(data=sentence_window)
knitr::kable(trackloss_sentence)
#Then remove trials with trackloss, and report how much data is left.
CurrentWindow <- "Sentence"

```

```

sentence_window_clean <- clean_by_trackloss(data = sentence_window,
      trial_prop_thresh = TrackLossThreshold)
#Total number of trials remaining per participant in sentence window:
Tot_Trials<- ddply(sentence_window_clean, c("RECORDING_SESSION_LABEL"),
  summarize, N=length(unique(TRIAL_LABEL)))
After Trackloss removal, there are a total `r sum(Tot_Trials$N)` unique trials in the `r
CurrentWindow` Window
Each subject contributes the following number of trials in the `r CurrentWindow`
Window
`r knitr::kable(Tot_Trials)`
Remove excessive trackloss trials for Agent window:
```{r_Dealing_With_Trackloss_Agent+art_Window}`
#First,calculate the amount of trackloss by subjects and trials
trackloss_agent <-trackloss_analysis(data=agent_window)
knitr::kable(trackloss_agent)
#Then remove trials with trackloss, and report how much data is left.
CurrentWindow <- "Agent"
agent_window_clean <- clean_by_trackloss(data = agent_window,
 trial_prop_thresh = TrackLossThreshold)
#Total number of trials remaining per participant in agent window:
Tot_Trials<- ddply(agent_window_clean, c("RECORDING_SESSION_LABEL"),
 summarize, N=length(unique(TRIAL_LABEL)))
Remove excessive trackloss trials for Anticipatory/Verb+art window:
```{r_Dealing_With_Trackloss_Verb+art_Window}`
#First,calculate the amount of trackloss by subjects and trials
trackloss_verb_art <-trackloss_analysis(data=verb_art_window)
knitr::kable(trackloss_verb_art)
#Then remove trials with trackloss, and report how much data is left.
CurrentWindow <- "VerbArt"
verb_art_window_clean <- clean_by_trackloss(data = verb_art_window,
      trial_prop_thresh = TrackLossThreshold)
#Total number of trials remaining per participant in verb window:

```

```

Tot_Trials<- dply(verb_art_window_clean, c("RECORDING_SESSION_LABEL"),
summarize, N=length(unique(TRIAL_LABEL)))
Remove excessive trackloss trials for Object window:
```{r_Dealing_With_Trackloss_Object_Window}
#First,calculate the amount of trackloss by subjects and trials
trackloss_object <-trackloss_analysis(data=object_window)
knitr::kable(trackloss_object)
#Then remove trials with trackloss, and report how much data is left.
CurrentWindow <- "Object"
object_window_clean <- clean_by_trackloss(data = object_window,
trial_prop_thresh = TrackLossThreshold)
#Total number of trials remaining per participant in object window:
Tot_Trials<- dply(object_window_clean, c("RECORDING_SESSION_LABEL"),
summarize, N=length(unique(TRIAL_LABEL)))
```{r_Describe_Data_For_AOI}
Data_Target <- describe_data(data, describe_column = "Target", group_columns =
"Cond")
Data_ActionRelated <- describe_data(data, describe_column = "ActionRel",
group_columns = "Cond")
Data_AgentRelated <- describe_data(data, describe_column = "AgentRel",
group_columns = "Cond")
Data_UnRelated <- describe_data(data, describe_column = "Unrelated", group_columns
= "Cond")
####STEP 8: add other data columns if needed
Do this here in case want to add any other data for analyses, like subject variables / offline
measures etc. (i.e. do this part before the analyses)
####STEP 9: ANALYSES OVER BROAD TIMEWINDOWS:
### Mixed effects model comparisons of conditions
- Aggregate data for conditions over entire sentence, and specify your predictor variables
(in this case List, which is the List I and List II)
```{r_Box_Plots}

```

```

#Aggregate data for Target and Competitor fixations over entire sentence and specify
predictor variables
sentence_window_agg <- make_time_window_data(sentence_window_clean,
predictor_columns = c("List"), summarize_by = c("RECORDING_SESSION_LABEL",
"TRIAL_INDEX"))
plot(sentence_window_agg, predictor_columns = "List", dv="Prop")
#Aggregate data for Target and Competitor fixations over agent window and specify
predictor variables
agent_window_agg <- make_time_window_data(agent_window_clean,
predictor_columns = c("List"), summarize_by = c("RECORDING_SESSION_LABEL",
"TRIAL_INDEX"))
plot(agent_window_agg, predictor_columns = "List", dv="Prop")
#Aggregate data for Target and Competitor fixations over anticipatory/verb+art window
and specify predictor variables
verb_art_window_agg <- make_time_window_data(verb_art_window_clean,
predictor_columns = c("List"), summarize_by = c("RECORDING_SESSION_LABEL",
"TRIAL_INDEX"))
plot(verb_art_window_agg, predictor_columns = "List", dv="Prop")
#Aggregate data for Target and Competitor fixations over object window and specify
predictor variables
object_window_agg <- make_time_window_data(object_window_clean,
predictor_columns = c("List"), summarize_by = c("RECORDING_SESSION_LABEL",
"TRIAL_INDEX"))
plot(object_window_agg, predictor_columns = "List", dv="Prop")
sum-code predictors
- Center variables to be entered as random effects (RECORDING_SESSION_LABEL
and TRIAL_INDEX)
```{r_My_Center_Function}
#Contrasts in Sentence Window
sentence_window_agg$AOI <-as.factor(sentence_window_agg$AOI)
contrasts(sentence_window_agg$AOI) = contr.sum(4)
levels(sentence_window_agg$AOI)

```

```

sentence_window_agg$AOI <-relevel(sentence_window_agg$AOI, "Target", first =
TRUE,xlevels=TRUE, nogroup=TRUE )
levels(sentence_window_agg$AOI)
contrasts(sentence_window_agg$List) = contr.sum(2)
sentence_window_agg$List <-relevel(sentence_window_agg$List, "1", first =
TRUE,xlevels=TRUE, nogroup=TRUE )
levels(sentence_window_agg$List)
#Constrasts in Agent Window
agent_window_agg$AOI <-as.factor(agent_window_agg$AOI)
contrasts(agent_window_agg$AOI) = contr.sum(4)
levels(agent_window_agg$AOI)
agent_window_agg$AOI <-relevel(agent_window_agg$AOI, "Target", first =
TRUE,xlevels=TRUE, nogroup=TRUE )
levels(agent_window_agg$AOI)
contrasts(agent_window_agg$List) = contr.sum(2)
agent_window_agg$List <-relevel(agent_window_agg$List, "1", first =
TRUE,xlevels=TRUE, nogroup=TRUE )
levels(agent_window_agg$List)
#Constrasts in Anticipatory/Verb+art Window
verb_art_window_agg$AOI <-as.factor(verb_art_window_agg$AOI)
contrasts(verb_art_window_agg$AOI) = contr.sum(4)
levels(verb_art_window_agg$AOI)
verb_art_window_agg$AOI <-relevel(verb_art_window_agg$AOI, "Target", first =
TRUE,xlevels=TRUE, nogroup=TRUE )
levels(verb_art_window_agg$AOI)
contrasts(verb_art_window_agg$List) = contr.sum(2)
verb_art_window_agg$List <-relevel(verb_art_window_agg$List, "1", first =
TRUE,xlevels=TRUE, nogroup=TRUE )
levels(verb_art_window_agg$List)
#Constrasts in Object Window
object_window_agg$AOI <-as.factor(object_window_agg$AOI)
contrasts(object_window_agg$AOI) = contr.sum(4)

```

```

levels(object_window_agg$AOI)
object_window_agg$AOI <-relevel(object_window_agg$AOI, "Target", first =
TRUE,xlevels=TRUE, nogroup=TRUE )
levels(object_window_agg$AOI)
contrasts(object_window_agg$List) = contr.sum(2)
object_window_agg$List <-relevel(object_window_agg$List, "1", first =
TRUE,xlevels=TRUE, nogroup=TRUE )
levels(object_window_agg$List)
#####
#####
###a nice general-purpose centering function from Florian Jaeger ###
###retrieved on 08/21/2016 from: ###
###https://hlplab.wordpress.com/2009/04/27/centering-several-variables/ ###
#####
#####
myCenter= function(x) {
  if (is.numeric(x)) { return(x - mean(x, na.rm=T)) }
  if (is.factor(x)) {
    x= as.numeric(x)
    return(x - mean(x, na.rm=T))
  }
  if (is.data.frame(x) || is.matrix(x)) {
    m= matrix(nrow=nrow(x), ncol=ncol(x))
    colnames(m)= paste("c", colnames(x), sep="")
    for (i in 1:ncol(x)) {
      m[,i]= myCenter(x[,i])
    }
    return(as.data.frame(m))
  }
}
#####

```

```

# Center TRIAL_INDEX (trial) and RECORDING_SESSION_LABEL (participant) in
Sentence Window
sentence_window_agg$TRIAL_INDEX <-
myCenter(sentence_window_agg$TRIAL_INDEX)
sentence_window_agg$RECORDING_SESSION_LABEL <-
myCenter(sentence_window_agg$RECORDING_SESSION_LABEL)
# Center TRIAL_INDEX (trial) and RECORDING_SESSION_LABEL (participant) in
Agent Window
agent_window_agg$TRIAL_INDEX <-
myCenter(agent_window_agg$TRIAL_INDEX)
agent_window_agg$RECORDING_SESSION_LABEL <-
myCenter(agent_window_agg$RECORDING_SESSION_LABEL)
# Center TRIAL_INDEX (trial) and RECORDING_SESSION_LABEL (participant) in
anticipaotory/Verb+art Window
verb_art_window_agg$TRIAL_INDEX <-
myCenter(verb_art_window_agg$TRIAL_INDEX)
verb_art_window_agg$RECORDING_SESSION_LABEL <-
myCenter(verb_art_window_agg$RECORDING_SESSION_LABEL)
# Center TRIAL_INDEX (trial) and RECORDING_SESSION_LABEL (participant) in
Object Window
object_window_agg$TRIAL_INDEX <-
myCenter(object_window_agg$TRIAL_INDEX)
object_window_agg$RECORDING_SESSION_LABEL <-
myCenter(object_window_agg$RECORDING_SESSION_LABEL)
### mixed-effects linear model
```{r_Mixed_Effect_Linear_Model}
#Sentence Window
names(sentence_window_agg)[1]<-"Subject"
names(sentence_window_agg)[2]<-"Trial"
sentence_model_prop_NoRel <- lmer(Prop ~ AOI + List + (1 +(AOI) | Subject) + (1 +
(List)| Trial), data=na.omit(sentence_window_agg), REML = FALSE)

```

```

sentence_model_prop_RelAOIList <- lmer(Prop ~ AOI * List + (1 +(AOI) | Subject) +
(1 + (List)| Trial), data=na.omit(sentence_window_agg), REML = FALSE)
anova(sentence_model_prop_NoRel, sentence_model_prop_RelAOIList)
cleanly show important parts of model (see `summary()` for more)
#(est <- broom::tidy(sentence_model_prop_RelAOIList, effects="fixed"))
summary_sentence_model<-summary(sentence_model_prop_RelAOIList)
tab_model(sentence_model_prop_RelAOIList)
use model comparison to attain p-values
ChiTest_sentence_model<-
drop1(sentence_model_prop_RelAOIList,"AOI:List",test="Chi")
plot the fitted vs. residual in our model using the plot method for lmer:
plot(sentence_model_prop_RelAOIList)
check model assumptions / see some model criticism plots
mcp.fnc(sentence_model_prop_RelAOIList)
Test whether Target differs from other distractors - note this is using the LSMEANS
function from the LMERTEST package, not the lsmeans package.
Do NOT load the LSMEANS package if you want this code snippet to work like this
means_sentence_model<-lsmeansLT(sentence_model_prop_RelAOIList,
test.offs="AOI")
comparison_sentence_model<-difflsmeans(sentence_model_prop_RelAOIList,
test.offs="AOI")
plot(difflsmeans(sentence_model_prop_RelAOIList, test.offs="AOI"))
write.table(means_sentence_model,"C:/Users/Valentina/Documents/VerbSemantics_Ph
D_Thesis/PsycholinguisticExperiment/MyExperiment/SampleReport_200618/Output/Fi
nal_Analyses/Sentence_Window/means_sentence_model.csv")
write.table(comparison_sentence_model,"C:/Users/Valentina/Documents/VerbSemantic
s_PhD_Thesis/PsycholinguisticExperiment/MyExperiment/SampleReport_200618/Out
put/Final_Analyses/Sentence_Window/comparison_sentence_model.csv")
#Agent Window
names(agent_window_agg)[1]<-"Subject"
names(agent_window_agg)[2]<-"Trial"

```



```

agent_model_prop_NoRel <- lmer(Prop ~ AOI + List + (1 +(AOI) | Subject) + (1 + (List)|
Trial), data=na.omit(agent_window_agg), REML = FALSE)
agent_model_prop_RelAOIList <- lmer(Prop ~ AOI * List + (1 +(AOI) | Subject) + (1 +
(List)| Trial), data=na.omit(agent_window_agg), REML = FALSE)
anova(agent_model_prop_NoRel, agent_model_prop_RelAOIList)
cleanly show important parts of model (see `summary()` for more)
#(est <- broom::tidy(agent_model_prop_RelAOIList, effects="fixed"))
summary_agent_model<-summary(agent_model_prop_RelAOIList)
tab_model(agent_model_prop_RelAOIList)
use model comparison to attain p-values
ChiTest_agent_model<-drop1(agent_model_prop_RelAOIList,"AOI:List",test="Chi")
plot the fitted vs. residual in our model using the plot method for lmer:
plot(agent_model_prop_RelAOIList)
check model assumptions / see some model criticism plots
mcp.fnc(agent_model_prop_RelAOIList)
Test whether Target differs from other distractors - note this is using the LSMEANS
function from the LMERTTEST package, not the lsmeans package.
Do NOT load the LSMEANS package if you want this code snippet to work like this
means_agent_model<-lsmeansLT(agent_model_prop_RelAOIList, test.offs="AOI")
comparison_agent_model<-difflsmeans(agent_model_prop_RelAOIList,
test.offs="AOI")
plot(difflsmeans(agent_model_prop_RelAOIList, test.offs="AOI"))
write.table(means_agent_model,"C:/Users/Valentina/Documents/VerbSemantics_PhD_
Thesis/PsycholinguisticExperiment/MyExperiment/SampleReport_200618/Output/Final
_Analyses/Agent_Window/means_agent_model.csv")
write.table(comparison_agent_model,"C:/Users/Valentina/Documents/VerbSemantics_
PhD_Thesis/PsycholinguisticExperiment/MyExperiment/SampleReport_200618/Output
/Final_Analyses/Agent_Window/comparison_agent_model.csv")
#Anticipatory/Verb+art Window
names(verb_art_window_agg)[1]<-"Subject"
names(verb_art_window_agg)[2]<-"Trial"

```

```

verb_art_model_prop_NoRel <- lmer(Prop ~ AOI + List + (1 | Subject) + (1 | Trial),
data=na.omit(verb_art_window_agg), REML = FALSE)
verb_art_model_prop_RelAOIList <- lmer(Prop ~ AOI * List + (1 +(AOI) | Subject) + (1
+ (List)| Trial), data=na.omit(verb_art_window_agg), REML = FALSE)
anova(verb_art_model_prop_NoRel, verb_art_model_prop_RelAOIList)
cleanly show important parts of model (see `summary()` for more)
#(est <- broom::tidy(verb_art_model_prop_RelAOIList, effects="fixed"))
summary_verb_art_model<-summary(verb_art_model_prop_RelAOIList)
tab_model(verb_art_model_prop_RelAOIList)
use model comparison to attain p-values
ChiTest_verb_art_model<-
drop1(verb_art_model_prop_RelAOIList,"AOI:List",test="Chi")
plot the fitted vs. residual in our model using the plot method for lmer:
plot(verb_art_model_prop_RelAOIList)
check model assumptions / see some model criticism plots
mcp.fnc(verb_art_model_prop_RelAOIList)
Test whether Target differs from other distractors - note this is using the LSMEANS
function from the LMERTTEST package, not the lsmeans package.
Do NOT load the LSMEANS package if you want this code snippet to work like this
means_verb_art_model<-lsmeansLT(verb_art_model_prop_RelAOIList,
test.effs="AOI")
comparison_verb_art_model<-diffsmeans(verb_art_model_prop_RelAOIList,
test.effs="AOI")
plot(diffsmeans(verb_art_model_prop_RelAOIList, test.effs="AOI"))
write.table(means_verb_art_model,"C:/Users/Valentina/Documents/VerbSemantics_Ph
D_Thesis/PsycholinguisticExperiment/MyExperiment/SampleReport_200618/Output/Fi
nal_Analyses/Anticipatory_Window_Verb+Art/means_verb_art_model.csv")
write.table(comparison_verb_art_model,"C:/Users/Valentina/Documents/VerbSemantic
s_PhD_Thesis/PsycholinguisticExperiment/MyExperiment/SampleReport_200618/Out
put/Final_Analyses/Anticipatory_Window_Verb+Art/comparison_verb_art_model.csv")
#Object Window
names(object_window_agg)[1]<-"Subject"

```

```

names(object_window_agg)[2]<-"Trial"
object_model_prop_NoRel <- lmer(Prop ~ AOI + List + (1 | Subject) + (1 | Trial),
data=na.omit(object_window_agg), REML = FALSE)
object_model_prop_RelAOIList <- lmer(Prop ~ AOI * List + (1 +(AOI) | Subject) + (1
+ (List)| Trial), data=na.omit(object_window_agg), REML = FALSE)
anova(object_model_prop_NoRel, object_model_prop_RelAOIList)
cleanly show important parts of model (see `summary()` for more)
#(est <- broom::tidy(object_model_prop_RelAOIList, effects="fixed"))
summary_object_model<-summary(object_model_prop_RelAOIList)
tab_model(object_model_prop_RelAOIList)
use model comparison to attain p-values
ChiTest_object_model<-drop1(object_model_prop_RelAOIList,"AOI:List",test="Chi")
plot the fitted vs. residual in our model using the plot method for lmer:
plot(object_model_prop_RelAOIList)
check model assumptions / see some model criticism plots
mcp.fnc(object_model_prop_RelAOIList)
Test whether Target differs from other distractors - note this is using the LSMEANS
function from the LMERTTEST package, not the lsmeans package.
Do NOT load the LSMEANS package if you want this code snippet to work like this
means_object_model<-lsmeansLT(object_model_prop_RelAOIList, test.offs="AOI")
comparison_object_model<-diffsmeans(object_model_prop_RelAOIList,
test.offs="AOI")
plot(diffsmeans(object_model_prop_RelAOIList, test.offs="AOI"))
write.table(means_object_model,"C:/Users/Valentina/Documents/VerbSemantics_PhD_
Thesis/PsycholinguisticExperiment/MyExperiment/SampleReport_200618/Output/Final
_Analyses/Object_Window/means_object_model.csv")
write.table(comparison_object_model,"C:/Users/Valentina/Documents/VerbSemantics_
PhD_Thesis/PsycholinguisticExperiment/MyExperiment/SampleReport_200618/Output
/Final_Analyses/Object_Window/comparison_object_model.csv")
##STEP 11: PLOT TIMECOURSE DATA
```{r_Plot_Timecourse_Data}
#Sentence Window

```

```

sentence_timecourse <- make_time_sequence_data(sentence_window_clean,
time_bin_size = 20,
                                aois = c("Target", "AgentRel", "ActionRel", "Unrelated"),
                                summarize_by = "RECORDING_SESSION_LABEL")
#Refactor the AOIs to appear in correct order in plot
sentence_timecourse$AOI <- factor(sentence_timecourse$AOI,
c("Target","AgentRel","ActionRel","Unrelated"))
#Means and variance of proportions for each AOI summarized by subjects and time
mean_Prop<-
tapply(sentence_timecourse$Prop,list(sentence_timecourse$Time,sentence_timecourse$
AOI),mean)
var_Prop<-
tapply(sentence_timecourse$Prop,list(sentence_timecourse$Time,sentence_timecourse$
AOI),var)
#Subset matrix by each 100ms
mean_Prop_100ms<-
mean_Prop[c(1,6,11,16,21,26,31,36,41,46,51,56,61,66,71,76,81,86,91,96,101,106,111,1
16),c(1,2,3,4)]
var_Prop_100ms<-
var_Prop[c(1,6,11,16,21,26,31,36,41,46,51,56,61,66,71,76,81,86,91,96,101,106,111,11
6),c(1,2,3,4)]
#Create a data frame for the plot and change columns names
mean_Prop_100ms<-as.data.frame(as.table(mean_Prop_100ms))
names(mean_Prop_100ms)[1]<-"Time"
names(mean_Prop_100ms)[2]<-"AOI"
names(mean_Prop_100ms)[3]<-"Prop"
mean_Prop_100ms$Time<-
as.numeric(levels(mean_Prop_100ms$Time))[mean_Prop_100ms$Time]
var_Prop_100ms<-as.data.frame(as.table(var_Prop_100ms))
names(var_Prop_100ms)[1]<-"Time"
names(var_Prop_100ms)[2]<-"AOI"
names(var_Prop_100ms)[3]<-"Var"

```

```

var_Prop_100ms$Time<-
as.numeric(levels(var_Prop_100ms$Time))[var_Prop_100ms$Time]
#Matrix for plot
Prop_100ms<-cbind(mean_Prop_100ms[1:3],var_Prop_100ms[3])
#Plot
pd <- position_dodge(0.1)
ggplot(Prop_100ms, aes(Time, Prop, shape=AOI))+ geom_errorbar(aes(ymin=Prop-Var,
ymax=Prop+Var), width=0.2, size= 1, position=pd) +
  geom_line(position=pd) +
  geom_point(position=pd,size=2.5)+
  geom_vline(xintercept=UtteranceStart, linetype="dashed", size=1) +
  geom_text(aes(x=UtteranceStart, label="Art", y=0.02), color="grey30", hjust=-.1,
fontface="italic") +
  geom_vline(xintercept=AgentOnset, linetype="dashed", size=1) +
  geom_text(aes(x=AgentOnset, label="Agent", y=0.02), color="grey30", hjust=-.1,
fontface="italic") +
  geom_vline(xintercept=VerbOnset, linetype="dashed", size=1) +
  geom_text(aes(x=VerbOnset, label="Verb", y=0.02), color="grey30", hjust=-.1,
fontface="italic") +
  geom_vline(xintercept=ObjectOnset, linetype="dashed", size=1) +
  geom_text(aes(x=ObjectOnset, label="Patient", y=0.02), color="grey30", hjust=-.1,
fontface="italic") +
  geom_vline(xintercept=UtteranceEnd, linetype="dashed", size=1) +
  geom_text(aes(x=UtteranceEnd, label="", y=0.02), color="grey30", hjust=-.1,
fontface="italic") +
  ggtitle ("Proportions of eye fixations to AOIs in sentence time course") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(y="Proportion Fixations to Area of Interest", x="Time in ms from sentence onset")
+
  theme(axis.title=element_text(size=14, face="bold"),
        axis.text=element_text(size=10),
        legend.text=element_text(size=12),

```

```

    legend.justification = c(0, 1),
    legend.position = c(0, 1),
    legend.title=element_blank()+
    scale_color_discrete(labels=c("Target","Agent-Related","Action-
Related","Unrelated"))
#plots averaged over all participants
#color version
#Sentence
ggplot(sentence_timecourse, aes(x=Time, y=Prop, color=AOI)) +
  coord_cartesian(xlim = c(456.75, 2366.98)) +
  geom_vline(xintercept=AgentOnset, linetype="dashed", size=1) +
  geom_text(aes(x=AgentOnset, label="Agent", y=0.02), color="grey30", hjust=-.1,
fontface="italic") +
  geom_vline(xintercept=VerbOnset, linetype="dashed", size=1) +
  geom_text(aes(x=VerbOnset, label="Verb", y=0.02), color="grey30", hjust=-.1,
fontface="italic") +
  geom_vline(xintercept=ObjectOnset, linetype="dashed", size=1) +
  geom_text(aes(x=ObjectOnset, label="Patient", y=0.02), color="grey30", hjust=-.1,
fontface="italic") +
  geom_vline(xintercept=2066.975, linetype="dashed", size=1) +
  geom_text(aes(x=2066.975, label="Sentence Offset", y=0.02), color="grey30",
hjust=-.1, fontface="italic") +
  ggtitle ("Proportions of eye fixations to AOIs in sentence time course") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(y="Proportion Fixations to Area of Interest", x="Time in ms from sentence onset")
+
  theme(axis.title=element_text(size=14, face="bold"),
    axis.text=element_text(size=12),
    legend.text=element_text(size=12),
    legend.justification = c(0, 1),
    legend.position = c(0, 1),
    legend.title=element_blank()+

```

```

    scale_color_discrete(labels=c("Target", "Agent-Related", "Action-
Related", "Unrelated")) +
    geom_smooth()
#plots for each participant, separately - use this to inspect if there are any problems with
individual participant data at this stage
ggplot(sentence_timecourse, aes(x=Time, y=Prop)) +
    geom_smooth(aes(color=AOI)) +
    geom_vline(xintercept=AgentOnset, linetype="dashed", size=1) +
    geom_vline(xintercept=VerbOnset, linetype="dashed", size=1) +
    geom_vline(xintercept=ObjectOnset, linetype="dashed", size=1) +
    labs(x="Time in ms from sentence onset", y="Proportion Fixations to Area of
Interest") +
    theme(legend.title=element_blank()) +
    facet_wrap( ~ RECORDING_SESSION_LABEL, ncol = 4)
#SAVE EVERYTHING
```{r_Save_Everything}
save.image("Multimodality and Prediction in Sentence Comprehension")

```

Script 1 Eye-tracking statistical analyses.

## **MEK Model**

### **Sequence Collections**

#### **LookAT**

1. hiker open backpack HIKING BACKPACK
2. hiker fill backpack HIKING BACKPACK
3. hiker close backpack HIKING BACKPACK
4. student open backpack SCHOOL BACKPACK
5. student fill backpack SCHOOL BACKPACK
6. student close backpack SCHOOL BACKPACK
7. olympian wear boot SKI BOOT
8. olympian adjust boot SKI BOOT
9. olympian remove boot SKI BOOT
10. fisherman wear boot RUBBER BOOT
11. fisherman adjust boot RUBBER BOOT
12. fisherman remove boot RUBBER BOOT
13. graduate try cap GRADUATION CAP
14. graduate remove cap GRADUATION CAP
15. graduate throw cap GRADUATION CAP
16. captain try cap UNIFORM CAP
17. captain remove cap UNIFORM CAP
18. captain throw cap UNIFORM CAP
19. librarian adjust glasses READING GLASSES
20. librarian remove glasses READING GLASSES
21. librarian clean glasses READING GLASSES
22. lifeguard adjust glasses SUNGLASSES
23. lifeguard remove glasses SUNGLASSES
24. lifeguard clean glasses SUNGLASSES
25. catcher fasten glove BASEBALL GLOVE
26. catcher try glove BASEBALL GLOVE
27. catcher throw glove BASEBALL GLOVE
28. boxer fasten glove BOXING GLOVE



29. boxer try glove BOXING GLOVE
30. boxer throw glove BOXING GLOVE
31. swimmer wear goggles SWIMMING GOGGLES
32. swimmer loosen goggles SWIMMING GOGGLES
33. swimmer tighten goggles SWIMMING GOGGLES
34. snowboarder wear goggles SKI GOGGLES
35. snowboarder loosen goggles SKI GOGGLES
36. snowboarder tighten goggles SKI GOGGLES
37. witch adjust hat WITCH HAT
38. witch remove hat WITCH HAT
39. witch throw hat WITCH HAT
40. chef adjust hat CHEF HAT
41. chef remove hat CHEF HAT
42. chef throw hat CHEF HAT
43. biker fasten helmet MOTORBIKE HELMET
44. biker loosen helmet MOTORBIKE HELMET
45. biker tighten helmet MOTORBIKE HELMET
46. cyclist fasten helmet BIKE HELMET
47. cyclist loosen helmet BIKE HELMET
48. cyclist tighten helmet BIKE HELMET
49. hiker try shoe HIKING SHOE
50. hiker loosen shoe HIKING SHOE
51. hiker tighten shoe HIKING SHOE
52. player try shoe SOCCER SHOE
53. player loosen shoe SOCCER SHOE
54. player tighten shoe SOCCER SHOE
55. cop wear vest BULLETPROOF VEST
56. cop fasten vest BULLETPROOF VEST
57. cop hold vest BULLETPROOF VEST
58. sailor wear vest LIFE VEST
59. sailor fasten vest LIFE VEST
60. sailor hold vest LIFE VEST

61. mailman open box MAILBOX
62. mailman empty box MAILBOX
63. mailman close box MAILBOX
64. electrician open box FUSE BOX
65. electrician empty box FUSE BOX
66. electrician close box FUSE BOX
67. doctor close bottle PILLS BOTTLE
68. doctor open bottle PILLS BOTTLE
69. doctor empty bottle PILLS BOTTLE
70. bartender close bottle BEER BOTTLE
71. bartender open bottle BEER BOTTLE
72. bartender empty bottle BEER BOTTLE
73. quarterback grab ball FOOTBALL BALL
74. quarterback hold ball FOOTBALL BALL
75. quarterback throw ball FOOTBALL BALL
76. shortstop grab ball BASEBALL BALL
77. shortstop hold ball BASEBALL BALL
78. shortstop throw ball BASEBALL BALL
79. swimmer wear suit BATHING SUIT
80. swimmer try suit BATHING SUIT
81. swimmer adjust suit BATHING SUIT
82. judge wear suit ELEGANT SUIT
83. judge try suit ELEGANT SUIT
84. judge adjust suit ELEGANT SUIT
85. chef clean cup MEASURING CUP
86. chef fill cup MEASURING CUP
87. chef empty cup MEASURING CUP
88. toddler clean cup SIPPY CUP
89. toddler fill cup SIPPY CUP
90. toddler empty cup SIPPY CUP
91. cook grab pot COOKING POT
92. cook fill pot COOKING POT

93. cook empty pot COOKING POT
94. gardener grab pot PLANT POT
95. gardener fill pot PLANT POT
96. gardener empty pot PLANT POT
97. photographer plug light SPOTLIGHT
98. photographer fix light SPOTLIGHT
99. photographer tune light SPOTLIGHT
100. detective plug light FLASH LIGHT
101. detective fix light FLASH LIGHT
102. detective tune light FLASH LIGHT
103. seamstress plug machine SEWING MACHINE
104. seamstress fix machine SEWING MACHINE
105. seamstress tune machine SEWING MACHINE
106. barista plug machine ESPRESSO MACHINE
107. barista fix machine ESPRESSO MACHINE
108. barista tune machine ESPRESSO MACHINE
109. jockey grab saddle HORSE SADDLE
110. jockey hold saddle HORSE SADDLE
111. jockey clean saddle HORSE SADDLE
112. cyclist grab saddle BIKE SADDLE
113. cyclist hold saddle BIKE SADDLE
114. cyclist clean saddle BIKE SADDLE
115. cyclist grab tire BIKE TIRE
116. cyclist hold tire BIKE TIRE
117. cyclist clean tire BIKE TIRE
118. farmer grab tire TRACTOR TIRE
119. farmer hold tire TRACTOR TIRE
120. farmer clean tire TRACTOR TIRE
121. driver fasten belt SEAT BELT
122. driver loosen belt SEAT BELT
123. driver tighten belt SEAT BELT
124. handyman fasten belt TOOL BELT

125. handyman loosen belt TOOL BELT
126. handyman tighten belt TOOL BELT
127. jeweller open box RING BOX
128. jeweller fill box RING BOX
129. jeweller close box RING BOX
130. carpenter open box TOOLBOX
131. carpenter fill box TOOLBOX
132. carpenter close box TOOLBOX
133. police plug camera SECURITY CAMERA
134. police fix camera SECURITY CAMERA
135. police tune camera SECURITY CAMERA
136. tourist plug camera VIDEO CAMERA
137. tourist fix camera VIDEO CAMERA
138. tourist tune camera VIDEO CAMERA
139. programmer plug keyboard COMPUTER KEYBOARD
140. programmer fix keyboard COMPUTER KEYBOARD
141. programmer tune keyboard COMPUTER KEYBOARD
142. musician plug keyboard MUSICKEY BOARD
143. musician fix keyboard MUSICKEY BOARD
144. musician tune keyboard MUSICKEY BOARD
145. hiker grab knife SWISS ARMY KNIFE
146. hiker clean knife SWISS ARMY KNIFE
147. student clean ruler TRIANGLE RULER
148. student turn ruler TRIANGLE RULER
149. olympian grab podium OLYMPIC PODIUM
150. olympian hold podium OLYMPIC PODIUM
151. fisherman shake rod FISHING ROD
152. fisherman hold rod FISHING ROD
153. graduate roll diploma GRADUATE DIPLOMA
154. graduate fold diploma GRADUATE DIPLOMA
155. captain strand ship CRUISE SHIP
156. captain moor ship CRUISE SHIP

- 157. librarian open book BOOK
- 158. librarian close book BOOK
- 159. lifeguard grab ring LIFE SAVER RING
- 160. lifeguard throw ring LIFE SAVER RING
- 161. catcher wear mask CATCHER MASK
- 162. catcher remove mask CATCHER MASK
- 163. boxer hit bag PUNCHING BAG
- 164. boxer kick bag PUNCHING BAG
- 165. swimmer grab slippers BATHING SLIPPERS
- 166. swimmer remove slippers BATHING SLIPPERS
- 167. snowboarder grab pole SKI POLE
- 168. snowboarder throw pole SKI POLE
- 169. witch ride broom WITCH BROOM
- 170. witch hold broom WITCH BROOM
- 171. chef shake pan FRYING PAN
- 172. chef grab pan FRYING PAN
- 173. biker ride motorbike MOTORBIKEP
- 174. biker clean motorbike MOTORBIKEP
- 175. cyclist empty pump BIKE PUMP
- 176. cyclist fill pump BIKE PUMP
- 177. hiker empty bottle WATER BOTTLE
- 178. hiker hold bottle WATER BOTTLE
- 179. player hit ball SOCCER BALL
- 180. player kick ball SOCCER BALL
- 181. cop adjust badge POLICE BADGE
- 182. cop exhibit badge POLICE BADGE
- 183. sailor tune compass MAGNETIC COMPASS
- 184. sailor turn compass MAGNETIC COMPASS
- 185. mailman paste stamp STAMP
- 186. mailman lick stamp STAMP
- 187. electrician run wire ELECTRIC WIRE
- 188. electrician plug wire ELECTRIC WIRE

189. doctor grab syringe SYRINGE
190. doctor fill syringe SYRINGE
191. bartender shake cocktail COCKTAIL
192. bartender garnish cocktail COCKTAIL
193. quarterback wear helmet FOOTBALL HELMET
194. quarterback adjust helmet FOOTBALL HELMET
195. shortstop shake bat BASEBALL BAT
196. shortstop clean bat BASEBALL BAT
197. swimmer fold towel TOWEL
198. swimmer spread towel TOWEL
199. judge exhibit hammer WOODEN HAMMER
200. judge shake hammer WOODEN HAMMER
201. chef lick spoon WOODEN SPOON
202. chef turn spoon WOODEN SPOON
203. toddler lick pacifier PACIFIER
204. toddler throw pacifier PACIFIER
205. cook adjust apron KITCHEN APRON
206. cook wear apron KITCHEN APRON
207. gardener clean rake GARDENING RAKE
208. gardener hold rake GARDENING RAKE
209. photographer clean camera PHOTO CAMERA
210. photographer hold camera PHOTO CAMERA
211. detective clean glass MAGNIFYING GLASS
212. detective hold glass MAGNIFYING GLASS
213. seamstress roll thread THREAD
214. seamstress spread thread THREAD
215. barista clean mug ESPRESSO MUG
216. barista fill mug ESPRESSO MUG
217. jockey remove horseshoe HORSESHOE
218. jockey fix horseshoe HORSESHOE
219. cyclist open lock BIKE LOCK
220. cyclist close lock BIKE LOCK

- 221. cyclist fix holder WATER BOTTLE HOLDER
- 222. cyclist remove holder WATER BOTTLE HOLDER
- 223. farmer shake shovel SHOVEL
- 224. farmer throw shovel SHOVEL
- 225. driver hit sign ROAD SIGN
- 226. driver turn sign ROAD SIGN
- 227. handyman grab screwdriver SCREWDRIVER
- 228. handyman turn screwdriver SCREWDRIVER
- 229. jeweller adjust necklace GOLD NECK LACE
- 230. jeweller remove necklace GOLD NECK LACE
- 231. carpenter grab wrench STUBBY WRENCH
- 232. carpenter turn wrench STUBBY WRENCH
- 233. police grab radio POLICE RADIO
- 234. police hold radio POLICE RADIO
- 235. tourist roll map TOURIST MAP
- 236. tourist fold map TOURIST MAP
- 237. programmer run cable ETHERNET CABLE
- 238. programmer collect cable ETHERNET CABLE
- 239. musician play flute FLUTE
- 240. musician clean flute FLUTE

### **WhoAct**

- 1. hiker open backpack
- 2. hiker fill backpack
- 3. hiker close backpack
- 4. student open backpack
- 5. student fill backpack
- 6. student close backpack
- 7. olympian wear boot
- 8. olympian adjust boot
- 9. olympian remove boot
- 10. fisherman wear boot
- 11. fisherman adjust boot

12. fisherman remove boot
13. graduate try cap
14. graduate remove cap
15. graduate throw cap
16. captain try cap
17. captain remove cap
18. captain throw cap
19. librarian adjust glasses
20. librarian remove glasses
21. librarian clean glasses
22. lifeguard adjust glasses
23. lifeguard remove glasses
24. lifeguard clean glasses
25. catcher fasten glove
26. catcher try glove
27. catcher throw glove
28. boxer fasten glove
29. boxer try glove
30. boxer throw glove
31. swimmer wear goggles
32. swimmer loosen goggles
33. swimmer tighten goggles
34. snowboarder wear goggles
35. snowboarder loosen goggles
36. snowboarder tighten goggles
37. witch adjust hat
38. witch remove hat
39. witch throw hat
40. chef adjust hat
41. chef remove hat
42. chef throw hat
43. biker fasten helmet



44. biker loosen helmet
45. biker tighten helmet
46. cyclist fasten helmet
47. cyclist loosen helmet
48. cyclist tighten helmet
49. hiker try shoe
50. hiker loosen shoe
51. hiker tighten shoe
52. player try shoe
53. player loosen shoe
54. player tighten shoe
55. cop wear vest
56. cop fasten vest
57. cop hold vest
58. sailor wear vest
59. sailor fasten vest
60. sailor hold vest
61. mailman open box
62. mailman empty box
63. mailman close box
64. electrician open box
65. electrician empty box
66. electrician close box
67. doctor close bottle
68. doctor open bottle
69. doctor empty bottle
70. bartender close bottle
71. bartender open bottle
72. bartender empty bottle
73. quarterback grab ball
74. quarterback hold ball
75. quarterback throw ball

76. shortstop grab ball
77. shortstop hold ball
78. shortstop throw ball
79. swimmer wear suit
80. swimmer try suit
81. swimmer adjust suit
82. judge wear suit
83. judge try suit
84. judge adjust suit
85. chef clean cup
86. chef fill cup
87. chef empty cup
88. toddler clean cup
89. toddler fill cup
90. toddler empty cup
91. cook grab pot
92. cook fill pot
93. cook empty pot
94. gardener grab pot
95. gardener fill pot
96. gardener empty pot
97. photographer plug light
98. photographer fix light
99. photographer tune light
100. detective plug light
101. detective fix light
102. detective tune light
103. seamstress plug machine
104. seamstress fix machine
105. seamstress tune machine
106. barista plug machine
107. barista fix machine

108. barista tune machine
109. jockey grab saddle
110. jockey hold saddle
111. jockey clean saddle
112. cyclist grab saddle
113. cyclist hold saddle
114. cyclist clean saddle
115. cyclist grab tire
116. cyclist hold tire
117. cyclist clean tire
118. farmer grab tire
119. farmer hold tire
120. farmer clean tire
121. driver fasten belt
122. driver loosen belt
123. driver tighten belt
124. handyman fasten belt
125. handyman loosen belt
126. handyman tighten belt
127. jeweller open box
128. jeweller fill box
129. jeweller close box
130. carpenter open box
131. carpenter fill box
132. carpenter close box
133. police plug camera
134. police fix camera
135. police tune camera
136. tourist plug camera
137. tourist fix camera
138. tourist tune camera
139. programmer plug keyboard

140. programmer fix keyboard
141. programmer tune keyboard
142. musician plug keyboard
143. musician fix keyboard
144. musician tune keyboard
145. hiker open HIKING BACKPACK
146. hiker fill HIKING BACKPACK
147. hiker close HIKING BACKPACK
148. student open SCHOOL BACKPACK
149. student fill SCHOOL BACKPACK
150. student close SCHOOL BACKPACK
151. olympian wear SKI BOOT
152. olympian adjust SKI BOOT
153. olympian remove SKI BOOT
154. fisherman wear RUBBER BOOT
155. fisherman adjust RUBBER BOOT
156. fisherman remove RUBBER BOOT
157. graduate try GRADUATION CAP
158. graduate remove GRADUATION CAP
159. graduate throw GRADUATION CAP
160. captain try UNIFORM CAP
161. captain remove UNIFORM CAP
162. captain throw UNIFORM CAP
163. librarian adjust READING GLASSES
164. librarian remove READING GLASSES
165. librarian clean READING GLASSES
166. lifeguard adjust SUNGLASSES
167. lifeguard remove SUNGLASSES
168. lifeguard clean SUNGLASSES
169. catcher fasten BASEBALL GLOVE
170. catcher try BASEBALL GLOVE
171. catcher throw BASEBALL GLOVE

172. boxer fasten BOXING GLOVE
173. boxer try BOXING GLOVE
174. boxer throw BOXING GLOVE
175. swimmer wear SWIMMING GOGGLES
176. swimmer loosen SWIMMING GOGGLES
177. swimmer tighten SWIMMING GOGGLES
178. snowboarder wear SKI GOGGLES
179. snowboarder loosen SKI GOGGLES
180. snowboarder tighten SKI GOGGLES
181. witch adjust WITCH HAT
182. witch remove WITCH HAT
183. witch throw WITCH HAT
184. chef adjust CHEF HAT
185. chef remove CHEF HAT
186. chef throw CHEF HAT
187. biker fasten MOTORBIKE HELMET
188. biker loosen MOTORBIKE HELMET
189. biker tighten MOTORBIKE HELMET
190. cyclist fasten BIKE HELMET
191. cyclist loosen BIKE HELMET
192. cyclist tighten BIKE HELMET
193. hiker try HIKING SHOE
194. hiker loosen HIKING SHOE
195. hiker tighten HIKING SHOE
196. player try SOCCER SHOE
197. player loosen SOCCER SHOE
198. player tighten SOCCER SHOE
199. cop wear BULLETPROOF VEST
200. cop fasten BULLETPROOF VEST
201. cop hold BULLETPROOF VEST
202. sailor wear LIFE VEST
203. sailor fasten LIFE VEST

204. sailor hold LIFE VEST
205. mailman open MAILBOX
206. mailman empty MAILBOX
207. mailman close MAILBOX
208. electrician open FUSE BOX
209. electrician empty FUSE BOX
210. electrician close FUSE BOX
211. doctor close PILLS BOTTLE
212. doctor open PILLS BOTTLE
213. doctor empty PILLS BOTTLE
214. bartender close BEER BOTTLE
215. bartender open BEER BOTTLE
216. bartender empty BEER BOTTLE
217. quarterback grab FOOTBALL BALL
218. quarterback hold FOOTBALL BALL
219. quarterback throw FOOTBALL BALL
220. shortstop grab BASEBALL BALL
221. shortstop hold BASEBALL BALL
222. shortstop throw BASEBALL BALL
223. swimmer wear BATHING SUIT
224. swimmer try BATHING SUIT
225. swimmer adjust BATHING SUIT
226. judge wear ELEGANT SUIT
227. judge try ELEGANT SUIT
228. judge adjust ELEGANT SUIT
229. chef clean MEASURING CUP
230. chef fill MEASURING CUP
231. chef empty MEASURING CUP
232. toddler clean SIPPY CUP
233. toddler fill SIPPY CUP
234. toddler empty SIPPY CUP
235. cook grab COOKING POT

- 236. cook fill COOKING POT
- 237. cook empty COOKING POT
- 238. gardener grab PLANT POT
- 239. gardener fill PLANT POT
- 240. gardener empty PLANT POT
- 241. photographer plug SPOTLIGHT
- 242. photographer fix SPOTLIGHT
- 243. photographer tune SPOTLIGHT
- 244. detective plug FLASHLIGHT
- 245. detective fix FLASHLIGHT
- 246. detective tune FLASHLIGHT
- 247. seamstress plug SEWING MACHINE
- 248. seamstress fix SEWING MACHINE
- 249. seamstress tune SEWING MACHINE
- 250. barista plug ESPRESSO MACHINE
- 251. barista fix ESPRESSO MACHINE
- 252. barista tune ESPRESSO MACHINE
- 253. jockey grab HORSE SADDLE
- 254. jockey hold HORSE SADDLE
- 255. jockey clean HORSE SADDLE
- 256. cyclist grab BIKE SADDLE
- 257. cyclist hold BIKE SADDLE
- 258. cyclist clean BIKE SADDLE
- 259. cyclist grab BIKE TIRE
- 260. cyclist hold BIKE TIRE
- 261. cyclist clean BIKE TIRE
- 262. farmer grab TRACTOR TIRE
- 263. farmer hold TRACTOR TIRE
- 264. farmer clean TRACTOR TIRE
- 265. driver fasten SEAT BELT
- 266. driver loosen SEAT BELT
- 267. driver tighten SEAT BELT

- 268. handyman fasten TOOL BELT
- 269. handyman loosen TOOL BELT
- 270. handyman tighten TOOL BELT
- 271. jeweller open RING BOX
- 272. jeweller fill RING BOX
- 273. jeweller close RING BOX
- 274. carpenter open TOOLBOX
- 275. carpenter fill TOOLBOX
- 276. carpenter close TOOLBOX
- 277. police plug SECURITY CAMERA
- 278. police fix SECURITY CAMERA
- 279. police tune SECURITY CAMERA
- 280. tourist plug VIDEO CAMERA
- 281. tourist fix VIDEO CAMERA
- 282. tourist tune VIDEO CAMERA
- 283. programmer plug COMPUTER KEYBOARD
- 284. programmer fix COMPUTER KEYBOARD
- 285. programmer tune COMPUTER KEYBOARD
- 286. musician plug MUSIC KEYBOARD
- 287. musician fix MUSIC KEYBOARD
- 288. musician tune MUSIC KEYBOARD
- 289. hiker grab SWISS ARMY KNIFE
- 290. hiker clean SWISS ARMY KNIFE
- 291. student clean TRIANGLE RULER
- 292. student turn TRIANGLE RULER
- 293. olympian grab OLYMPIC PODIUM
- 294. olympian hold OLYMPIC PODIUM
- 295. fisherman shake FISHING ROD
- 296. fisherman hold FISHING ROD
- 297. graduate roll GRADUATE DIPLOMA
- 298. graduate fold GRADUATE DIPLOMA
- 299. captain strand CRUISE SHIP



300. captain moor CRUISE SHIP
301. librarian open BOOK
302. librarian close BOOK
303. lifeguard grab LIFE SAVER RING
304. lifeguard throw LIFE SAVER RING
305. catcher wear CATCHER MASK
306. catcher remove CATCHER MASK
307. boxer hit PUNCHING BAG
308. boxer kick PUNCHING BAG
309. swimmer grab BATHING SLIPPERS
310. swimmer remove BATHING SLIPPERS
311. snowboarder grab SKI POLE
312. snowboarder throw SKI POLE
313. witch ride WITCH BROOM
314. witch hold WITCH BROOM
315. chef shake FRYING PAN
316. chef grab FRYING PAN
317. biker ride MOTORBIKE
318. biker clean MOTORBIKE
319. cyclist empty BIKE PUMP
320. cyclist fill BIKE PUMP
321. hiker empty WATER BOTTLE
322. hiker hold WATER BOTTLE
323. player hit SOCCER BALL
324. player kick SOCCER BALL
325. cop adjust POLICE BADGE
326. cop exhibit POLICE BADGE
327. sailor tune MAGNETIC COMPASS
328. sailor turn MAGNETIC COMPASS
329. mailman paste STAMP
330. mailman lick STAMP
331. electrician run ELECTRIC WIRE

- 332. electrician plug ELECTRIC WIRE
- 333. doctor grab SYRINGE
- 334. doctor fill SYRINGE
- 335. bartender shake COCKTAIL
- 336. bartender garnish COCKTAIL
- 337. quarterback wear FOOTBALL HELMET
- 338. quarterback adjust FOOTBALL HELMET
- 339. shortstop shake BASEBALL BAT
- 340. shortstop clean BASEBALL BAT
- 341. swimmer fold TOWEL
- 342. swimmer spread TOWEL
- 343. judge exhibit WOODEN HAMMER
- 344. judge shake WOODEN HAMMER
- 345. chef lick WOODEN SPOON
- 346. chef turn WOODEN SPOON
- 347. toddler lick PACIFIER
- 348. toddler throw PACIFIER
- 349. cook adjust KITCHEN APRON
- 350. cook wear KITCHEN APRON
- 351. gardener clean GARDENING RAKE
- 352. gardener hold GARDENING RAKE
- 353. photographer clean PHOTO CAMERA
- 354. photographer hold PHOTO CAMERA
- 355. detective clean MAGNIFYING GLASS
- 356. detective hold MAGNIFYING GLASS
- 357. seamstress roll THREAD
- 358. seamstress spread THREAD
- 359. barista clean ESPRESSO MUG
- 360. barista fill ESPRESSO MUG
- 361. jockey remove HORSESHOE
- 362. jockey fix HORSESHOE
- 363. cyclist open BIKE LOCK

364. cyclist close BIKE LOCK
365. cyclist fix WATER BOTTLE HOLDER
366. cyclist remove WATER BOTTLE HOLDER
367. farmer shake SHOVEL
368. farmer throw SHOVEL
369. driver hit ROAD SIGN
370. driver turn ROAD SIGN
371. handyman grab SCREWDRIVER
372. handyman turn SCREWDRIVER
373. jeweller adjust GOLD NECKLACE
374. jeweller remove GOLD NECKLACE
375. carpenter grab STUBBYWRENCH
376. carpenter turn STUBBYWRENCH
377. police grab POLICE RADIO
378. police hold POLICE RADIO
379. tourist roll TOURIST MAP
380. tourist fold TOURIST MAP
381. programmer run ethernet cable
382. programmer collect ethernet cable
383. musician play flute
384. musician clean flute
385. hiker grab knife
386. hiker clean knife
387. student clean ruler
388. student turn ruler
389. olympian grab podium
390. olympian hold podium
391. fisherman shake rod
392. fisherman hold rod
393. graduate roll diploma
394. graduate fold diploma
395. captain strand ship

- 396. captain moor ship
- 397. librarian open book
- 398. librarian close book
- 399. lifeguard grab ring
- 400. lifeguard throw ring
- 401. catcher wear mask
- 402. catcher remove mask
- 403. boxer hit bag
- 404. boxer kick bag
- 405. swimmer grab slippers
- 406. swimmer remove slippers
- 407. snowboarder grab pole
- 408. snowboarder throw pole
- 409. witch ride broom
- 410. witch hold broom
- 411. chef shake pan
- 412. chef grab pan
- 413. biker ride motorbike
- 414. biker clean motorbike
- 415. cyclist empty pump
- 416. cyclist fill pump
- 417. hiker empty bottle
- 418. hiker hold bottle
- 419. player hit ball
- 420. player kick ball
- 421. cop adjust badge
- 422. cop exhibit badge
- 423. sailor tune compass
- 424. sailor turn compass
- 425. mailman paste stamp
- 426. mailman lick stamp
- 427. electrician run wire

- 428. electrician plug wire
- 429. doctor grab syringe
- 430. doctor fill syringe
- 431. bartender shake cocktail
- 432. bartender garnish cocktail
- 433. quarterback wear helmet
- 434. quarterback adjust helmet
- 435. shortstop shake bat
- 436. shortstop clean bat
- 437. swimmer fold towel
- 438. swimmer spread towel
- 439. judge exhibit hammer
- 440. judge shake hammer
- 441. chef lick spoon
- 442. chef turn spoon
- 443. toddler lick pacifier
- 444. toddler throw pacifier
- 445. cook adjust apron
- 446. cook wear apron
- 447. gardener clean rake
- 448. gardener hold rake
- 449. photographer clean camera
- 450. photographer hold camera
- 451. detective clean glass
- 452. detective hold glass
- 453. seamstress roll thread
- 454. seamstress spread thread
- 455. barista clean mug
- 456. barista fill mug
- 457. jockey remove horseshoe
- 458. jockey fix horseshoe
- 459. cyclist open lock

- 460. cyclist close lock
- 461. cyclist fix holder
- 462. cyclist remove holder
- 463. farmer shake shovel
- 464. farmer throw shovel
- 465. driver hit sign
- 466. driver turn sign
- 467. handyman grab screwdriver
- 468. handyman turn screwdriver
- 469. jeweller adjust necklace
- 470. jeweller remove necklace
- 471. carpenter grab wrench
- 472. carpenter turn wrench
- 473. police grab radio
- 474. police hold radio
- 475. tourist roll map
- 476. tourist fold map
- 477. programmer run cable
- 478. programmer collect cable
- 479. musician play flute
- 480. musician clean flute

## Results

### Visual Representations Accuracy

Referents	Accuracy*	
	AlexNet CNN	GoogLeNet CNN
Target	88	93
Agent-Related	80	84

\* (%)

Table 21 AlexNet and GoogLeNet CNNs accuracy.

### Referents Report

#### Event

LookAT-GG					
N°	Referents	Precision	Recall	F-score	Support Samples
0	Baseball ball	1	1	1	15
1	Baseball bat	1	1	1	10
2	Baseball glove	1	1	1	15
3	Bathing slippers	1	1	1	10
4	Bathing suit	1	1	1	15
5	Beer bottle	1	1	1	15
6	Bike helmet	1	1	1	15
7	Bike lock	1	1	1	10
8	Bike pump	1	1	1	10
9	Bike saddle	1	1	1	15
10	Bike tire	1	1	1	15
11	book	1	1	1	10
12	Boxing glove	1	1	1	15
13	Bulletproof vest	1	1	1	15
14	Catcher mask	1	1	1	10
15	Chef hat	1	1	1	15
16	Cocktail	1	1	1	10
17	Computer keyboard	1	1	1	15
18	Cooking pot	1	1	1	15
19	Cruise ship	1	1	1	10
20	Electric wire	1	1	1	10
21	Elegant suit	1	1	1	15
22	Espresso machine	1	1	1	15
23	Espresso mug	1	1	1	10

24	Ethernet cable	1	1	1	10
25	Fishing rod	1	1	1	10
26	Flash light	1	1	1	15
27	Flute	1	1	1	10
28	Football ball	1	1	1	15
29	Football helmet	1	1	1	10
30	Frying pan	1	1	1	10
31	Fuse box	1	1	1	15
32	Gardening rake	1	1	1	10
33	Gold necklace	1	1	1	10
34	Graduate diploma	1	1	1	10
35	Graduation cap	1	1	1	15
36	Hiking backpack	1	1	1	15
37	Hiking shoe	1	1	1	15
38	Horse saddle	1	1	1	15
39	Horse shoe	1	1	1	10
40	Kitchen apron	1	1	1	10
41	Lifesaver ring	1	1	1	10
42	Life vest	1	1	1	15
43	Magnetic compass	1	1	1	10
44	Magnifying glass	1	1	1	10
45	Mailbox	1	1	1	15
46	Measuring cup	1	1	1	15
47	Motorbike helmet	1	1	1	15
48	Motorbike	1	1	1	10
49	Music keyboard	1	1	1	15
50	Olympic podium	1	1	1	10
51	Pacifier	1	1	1	10
52	Photo camera	1	1	1	10
53	Pill bottle	1	1	1	15
54	Plant pot	1	1	1	15
55	Police badge	1	1	1	10
56	Police radio	1	1	1	10
57	Punching bag	1	1	1	10
58	Reading glasses	1	1	1	15
59	Ring box	1	1	1	15
60	Road sign	1	1	1	10
61	Rubber boot	1	1	1	15
62	School backpack	1	1	1	15
63	Screwdriver	1	1	1	10
64	Seat belt	1	1	1	15



65	Security camera	1	1	1	15
66	Sewing machine	1	1	1	15
67	Shovel	1	1	1	10
68	Sippy cup	1	1	1	15
69	Ski boot	1	1	1	15
70	Ski goggles	1	1	1	15
71	Ski pole	1	1	1	10
72	Soccer ball	1	1	1	10
73	Soccer shoe	1	1	1	15
74	Spotlight	1	1	1	15
75	Stamp	1	1	1	10
76	Stubby wrench	1	1	1	10
77	Sunglasses	1	1	1	15
78	Swimming goggles	1	1	1	15
79	Swiss army knife	1	1	1	10
80	Syringe	1	1	1	10
81	Thread	1	1	1	10
82	Tool belt	1	1	1	15
83	Toolbox	1	1	1	15
84	Tourist map	1	1	1	10
85	Towel	1	1	1	10
86	Tractor tire	1	1	1	15
87	Triangle ruler	1	1	1	10
88	Uniform cap	1	1	1	15
89	Video camera	1	1	1	15
90	Water bottle	1	1	1	10
91	Water bottle holder	1	1	1	10
92	Witch broom	1	1	1	10
93	Witch hat	1	1	1	15
94	Wooden hammer	1	1	1	10
95	Wooden spoon	1	1	1	10
	<b>Accuracy</b>			1	1200
	<b>Macro Average</b>	1	1	1	1200
	<b>Weighted Average</b>	1	1	1	1200

Table 22 LookAT-GG.

<b>LookAT-WA</b>					
<b>N°</b>	<b>Referents</b>	<b>Precision</b>	<b>Recall</b>	<b>F-score</b>	<b>Support Samples</b>
0	Baseball ball	1	1	1	15
1	Baseball bat	1	1	1	10
2	Baseball glove	1	1	1	15
3	Bathing slippers	1	1	1	10
4	Bathing suit	1	1	1	15
5	Beer bottle	1	1	1	15
6	Bike helmet	1	1	1	15
7	Bike lock	1	1	1	10
8	Bike pump	1	1	1	10
9	Bike saddle	1	1	1	15
10	Bike tire	1	1	1	15
11	book	1	1	1	10
12	Boxing glove	1	1	1	15
13	Bulletproof vest	1	1	1	15
14	Catcher mask	1	1	1	10
15	Chef hat	1	1	1	15
16	Cocktail	1	1	1	10
17	Computer keyboard	1	1	1	15
18	Cooking pot	1	1	1	15
19	Cruise ship	1	1	1	10
20	Electric wire	1	1	1	10
21	Elegant suit	1	1	1	15
22	Espresso machine	1	1	1	15
23	Espresso mug	1	1	1	10
24	Ethernet cable	1	1	1	10
25	Fishing rod	1	1	1	10
26	Flash light	1	1	1	15
27	Flute	1	1	1	10
28	Football ball	1	1	1	15
29	Football helmet	1	1	1	10
30	Frying pan	1	1	1	10
31	Fuse box	1	1	1	15
32	Gardening rake	1	1	1	10
33	Gold necklace	1	1	1	10
34	Graduate diploma	1	1	1	10
35	Graduation cap	1	1	1	15
36	Hiking backpack	1	1	1	15
37	Hiking shoe	1	1	1	15
38	Horse saddle	1	1	1	15

39	Horse shoe	1	1	1	10
40	Kitchen apron	1	1	1	10
41	Lifesaver ring	1	1	1	10
42	Life vest	1	1	1	15
43	Magnetic compass	1	1	1	10
44	Magnifying glass	1	1	1	10
45	Mailbox	1	1	1	15
46	Measuring cup	1	1	1	15
47	Motorbike helmet	1	1	1	15
48	Motorbike	1	1	1	10
49	Music keyboard	1	1	1	15
50	Olympic podium	1	1	1	10
51	Pacifier	1	1	1	10
52	Photo camera	1	1	1	10
53	Pill bottle	1	1	1	15
54	Plant pot	1	1	1	15
55	Police badge	1	1	1	10
56	Police radio	1	1	1	10
57	Punching bag	1	1	1	10
58	Reading glasses	1	1	1	15
59	Ring box	1	1	1	15
60	Road sign	1	1	1	10
61	Rubber boot	1	1	1	15
62	School backpack	1	1	1	15
63	Screwdriver	1	1	1	10
64	Seat belt	1	1	1	15
65	Security camera	1	1	1	15
66	Sewing machine	1	1	1	15
67	Shovel	1	1	1	10
68	Sippy cup	1	1	1	15
69	Ski boot	1	1	1	15
70	Ski goggles	1	1	1	15
71	Ski pole	1	1	1	10
72	Soccer ball	1	1	1	10
73	Soccer shoe	1	1	1	15
74	Spotlight	1	1	1	15
75	Stamp	1	1	1	10
76	Stubby wrench	1	1	1	10
77	Sunglasses	1	1	1	15
78	Swimming goggles	1	1	1	15
79	Swiss army knife	1	1	1	10

80	Syringe	1	1	1	10
81	Thread	1	1	1	10
82	Tool belt	1	1	1	15
83	Toolbox	1	1	1	15
84	Tourist map	1	1	1	10
85	Towel	1	1	1	10
86	Tractor tire	1	1	1	15
87	Triangle ruler	1	1	1	10
88	Uniform cap	1	1	1	15
89	Video camera	1	1	1	15
90	Water bottle	1	1	1	10
91	Water bottle holder	1	1	1	10
92	Witch broom	1	1	1	10
93	Witch hat	1	1	1	15
94	Wooden hammer	1	1	1	10
95	Wooden spoon	1	1	1	10
	<b>Accuracy</b>			1	1200
	<b>Macro Average</b>	1	1	1	1200
	<b>Weighted Average</b>	1	1	1	1200

Table 23 LookAT-WA.

<b>WhoAct-GG</b>					
<b>N°</b>	<b>Referents</b>	<b>Precision</b>	<b>Recall</b>	<b>F-score</b>	<b>Support Samples</b>
0	Baseball ball	1	1	1	15
1	Baseball bat	1	1	1	10
2	Baseball glove	1	1	1	15
3	Bathing slippers	1	1	1	10
4	Bathing suit	1	1	1	15
5	Beer bottle	1	1	1	15
6	Bike helmet	1	1	1	15
7	Bike lock	1	1	1	10
8	Bike pump	1	1	1	10
9	Bike saddle	1	1	1	15
10	Bike tire	1	1	1	15
11	book	1	1	1	10
12	Boxing glove	1	1	1	15
13	Bulletproof vest	1	1	1	15
14	Catcher mask	1	1	1	10
15	Chef hat	1	1	1	15
16	Cocktail	1	1	1	10
17	Computer keyboard	1	1	1	15
18	Cooking pot	1	1	1	15
19	Cruise ship	1	1	1	10
20	Electric wire	1	1	1	10
21	Elegant suit	1	1	1	15
22	Espresso machine	1	1	1	15
23	Espresso mug	1	1	1	10
24	Ethernet cable	1	1	1	10
25	Fishing rod	1	1	1	10
26	Flash light	1	1	1	15
27	Flute	1	1	1	10
28	Football ball	1	1	1	15
29	Football helmet	1	1	1	10
30	Frying pan	1	1	1	10
31	Fuse box	1	1	1	15
32	Gardening rake	1	1	1	10
33	Gold necklace	1	1	1	10
34	Graduate diploma	1	1	1	10
35	Graduation cap	1	1	1	15
36	Hiking backpack	1	1	1	15
37	Hiking shoe	1	1	1	15
38	Horse saddle	1	1	1	15

39	Horse shoe	1	1	1	10
40	Kitchen apron	1	1	1	10
41	Lifesaver ring	1	1	1	10
42	Life vest	1	1	1	15
43	Magnetic compass	1	1	1	10
44	Magnifying glass	1	1	1	10
45	Mailbox	1	1	1	15
46	Measuring cup	1	1	1	15
47	Motorbike helmet	1	1	1	15
48	Motorbike	1	1	1	10
49	Music keyboard	1	1	1	15
50	Olympic podium	1	1	1	10
51	Pacifier	1	1	1	10
52	Photo camera	1	1	1	10
53	Pill bottle	1	1	1	15
54	Plant pot	1	1	1	15
55	Police badge	1	1	1	10
56	Police radio	1	1	1	10
57	Punching bag	1	1	1	10
58	Reading glasses	1	1	1	15
59	Ring box	1	1	1	15
60	Road sign	1	1	1	10
61	Rubber boot	1	1	1	15
62	School backpack	1	1	1	15
63	Screwdriver	1	1	1	10
64	Seat belt	1	1	1	15
65	Security camera	1	1	1	15
66	Sewing machine	1	1	1	15
67	Shovel	1	1	1	10
68	Sippy cup	1	1	1	15
69	Ski boot	1	1	1	15
70	Ski goggles	1	1	1	15
71	Ski pole	1	1	1	10
72	Soccer ball	1	1	1	10
73	Soccer shoe	1	1	1	15
74	Spotlight	1	1	1	15
75	Stamp	1	1	1	10
76	Stubby wrench	1	1	1	10
77	Sunglasses	1	1	1	15
78	Swimming goggles	1	1	1	15
79	Swiss army knife	1	1	1	10

80	Syringe	1	1	1	10
81	Thread	1	1	1	10
82	Tool belt	1	1	1	15
83	Toolbox	1	1	1	15
84	Tourist map	1	1	1	10
85	Towel	1	1	1	10
86	Tractor tire	1	1	1	15
87	Triangle ruler	1	1	1	10
88	Uniform cap	1	1	1	15
89	Video camera	1	1	1	15
90	Water bottle	1	1	1	10
91	Water bottle holder	1	1	1	10
92	Witch broom	1	1	1	10
93	Witch hat	1	1	1	15
94	Wooden hammer	1	1	1	10
95	Wooden spoon	1	1	1	10
	<b>Accuracy</b>			1	1200
	<b>Macro Average</b>	1	1	1	1200
	<b>Weighted Average</b>	1	1	1	1200

Table 24 WhoAct-GG.

<b>WhoAct-WA</b>					
<b>N°</b>	<b>Referents</b>	<b>Precision</b>	<b>Recall</b>	<b>F-score</b>	<b>Support Samples</b>
0	Baseball ball	1	1	1	15
1	Baseball bat	1	1	1	10
2	Baseball glove	1	1	1	15
3	Bathing slippers	1	1	1	10
4	Bathing suit	1	1	1	15
5	Beer bottle	1	1	1	15
6	Bike helmet	1	1	1	15
7	Bike lock	1	1	1	10
8	Bike pump	1	1	1	10
9	Bike saddle	1	1	1	15
10	Bike tire	1	1	1	15
11	book	1	1	1	10
12	Boxing glove	1	1	1	15
13	Bulletproof vest	1	1	1	15
14	Catcher mask	1	1	1	10
15	Chef hat	1	1	1	15
16	Cocktail	1	1	1	10
17	Computer keyboard	1	1	1	15
18	Cooking pot	1	1	1	15
19	Cruise ship	1	1	1	10
20	Electric wire	1	1	1	10
21	Elegant suit	1	1	1	15
22	Espresso machine	1	1	1	15
23	Espresso mug	1	1	1	10
24	Ethernet cable	1	1	1	10
25	Fishing rod	1	1	1	10
26	Flash light	1	1	1	15
27	Flute	1	1	1	10
28	Football ball	1	1	1	15
29	Football helmet	1	1	1	10
30	Frying pan	1	1	1	10
31	Fuse box	1	1	1	15
32	Gardening rake	1	1	1	10
33	Gold necklace	1	1	1	10
34	Graduate diploma	1	1	1	10
35	Graduation cap	1	1	1	15
36	Hiking backpack	1	1	1	15
37	Hiking shoe	1	1	1	15
38	Horse saddle	1	1	1	15



39	Horse shoe	1	1	1	10
40	Kitchen apron	1	1	1	10
41	Lifesaver ring	1	1	1	10
42	Life vest	1	1	1	15
43	Magnetic compass	1	1	1	10
44	Magnifying glass	1	1	1	10
45	Mailbox	1	1	1	15
46	Measuring cup	1	1	1	15
47	Motorbike helmet	1	1	1	15
48	Motorbike	1	1	1	10
49	Music keyboard	1	1	1	15
50	Olympic podium	1	1	1	10
51	Pacifier	1	1	1	10
52	Photo camera	1	1	1	10
53	Pill bottle	1	1	1	15
54	Plant pot	1	1	1	15
55	Police badge	1	1	1	10
56	Police radio	1	1	1	10
57	Punching bag	1	1	1	10
58	Reading glasses	1	1	1	15
59	Ring box	1	1	1	15
60	Road sign	1	1	1	10
61	Rubber boot	1	1	1	15
62	School backpack	1	1	1	15
63	Screwdriver	1	1	1	10
64	Seat belt	1	1	1	15
65	Security camera	1	1	1	15
66	Sewing machine	1	1	1	15
67	Shovel	1	1	1	10
68	Sippy cup	1	1	1	15
69	Ski boot	1	1	1	15
70	Ski goggles	1	1	1	15
71	Ski pole	1	1	1	10
72	Soccer ball	1	1	1	10
73	Soccer shoe	1	1	1	15
74	Spotlight	1	1	1	15
75	Stamp	1	1	1	10
76	Stubby wrench	1	1	1	10
77	Sunglasses	1	1	1	15
78	Swimming goggles	1	1	1	15
79	Swiss army knife	1	1	1	10

80	Syringe	1	1	1	10
81	Thread	1	1	1	10
82	Tool belt	1	1	1	15
83	Toolbox	1	1	1	15
84	Tourist map	1	1	1	10
85	Towel	1	1	1	10
86	Tractor tire	1	1	1	15
87	Triangle ruler	1	1	1	10
88	Uniform cap	1	1	1	15
89	Video camera	1	1	1	15
90	Water bottle	1	1	1	10
91	Water bottle holder	1	1	1	10
92	Witch broom	1	1	1	10
93	Witch hat	1	1	1	15
94	Wooden hammer	1	1	1	10
95	Wooden spoon	1	1	1	10
	<b>Accuracy</b>			1	1200
	<b>Macro Average</b>	1	1	1	1200
	<b>Weighted Average</b>	1	1	1	1200

Table 25 WhoAct-WA.

**Agent**

<b>LookAT-GG</b>					
<b>N°</b>	<b>Referents</b>	<b>Precision</b>	<b>Recall</b>	<b>F-score</b>	<b>Support Samples</b>
0	Baseball ball	0.6	1	0.75	15
1	Baseball bat	0	0	0	10
2	Baseball glove	0.6	1	0.75	15
3	Bathing slippers	0	0	0	10
4	Bathing suit	0	0	0	15
5	Beer bottle	0.6	1	0.75	15
6	Bike helmet	0	0	0	15
7	Bike lock	0	0	0	10
8	Bike pump	0	0	0	10
9	Bike saddle	0	0	0	15
10	Bike tire	0.2	1	0.33	15
11	book	0	0	0	10
12	Boxing glove	0.6	1	0.75	15
13	Bulletproof vest	0.6	1	0.75	15
14	Catcher mask	0	0	0	10
15	Chef hat	0	0	0	15
16	Cocktail	0	0	0	10
17	Computer keyboard	0.6	1	0.75	15
18	Cooking pot	0.6	1	0.75	15
19	Cruise ship	0	0	0	10
20	Electric wire	0	0	0	10
21	Elegant suit	0.6	1	0.75	15
22	Espresso machine	0.6	1	0.75	15
23	Espresso mug	0	0	0	10
24	Ethernet cable	0	0	0	10
25	Fishing rod	0	0	0	10
26	Flash light	0.6	1	0.75	15
27	Flute	0	0	0	10
28	Football ball	0.6	1	0.75	15
29	Football helmet	0	0	0	10
30	Frying pan	0	0	0	10
31	Fuse box	0.6	1	0.75	15
32	Gardening rake	0	0	0	10
33	Gold necklace	0	0	0	10
34	Graduate diploma	0	0	0	10
35	Graduation cap	0.6	1	0.75	15
36	Hiking backpack	0.3	1	0.46	15

37	Hiking shoe	0	0	0	15
38	Horse saddle	0.6	1	0.75	15
39	Horse shoe	0	0	0	10
40	Kitchen apron	0	0	0	10
41	Lifesaver ring	0	0	0	10
42	Life vest	0.6	1	0.75	15
43	Magnetic compass	0	0	0	10
44	Magnifying glass	0	0	0	10
45	Mailbox	0.6	1	0.75	15
46	Measuring cup	0.3	1	0.46	15
47	Motorbike helmet	0.6	1	0.75	15
48	Motorbike	0	0	0	10
49	Music keyboard	0.6	1	0.75	15
50	Olympic podium	0	0	0	10
51	Pacifier	0	0	0	10
52	Photo camera	0	0	0	10
53	Pill bottle	0.6	1	0.75	15
54	Plant pot	0.6	1	0.75	15
55	Police badge	0	0	0	10
56	Police radio	0	0	0	10
57	Punching bag	0	0	0	10
58	Reading glasses	0.6	1	0.75	15
59	Ring box	0.6	1	0.75	15
60	Road sign	0	0	0	10
61	Rubber boot	0.6	1	0.75	15
62	School backpack	0.6	1	0.75	15
63	Screwdriver	0	0	0	10
64	Seat belt	0.6	1	0.75	15
65	Security camera	0.6	1	0.75	15
66	Sewing machine	0.6	1	0.75	15
67	Shovel	0	0	0	10
68	Sippy cup	0.6	1	0.75	15
69	Ski boot	0.6	1	0.75	15
70	Ski goggles	0.6	1	0.75	15
71	Ski pole	0	0	0	10
72	Soccer ball	0	0	0	10
73	Soccer shoe	0.6	1	0.75	15
74	Spotlight	0.6	1	0.75	15
75	Stamp	0	0	0	10
76	Stubby wrench	0	0	0	10
77	Sunglasses	0.6	1	0.75	15

78	Swimming goggles	0.3	1	0.46	15
79	Swiss army knife	0	0	0	10
80	Syringe	0	0	0	10
81	Thread	0	0	0	10
82	Tool belt	0.6	1	0.75	15
83	Toolbox	0.6	1	0.75	15
84	Tourist map	0	0	0	10
85	Towel	0	0	0	10
86	Tractor tire	0.6	1	0.75	15
87	Triangle ruler	0	0	0	10
88	Uniform cap	0.6	1	0.75	15
89	Video camera	0.6	1	0.75	15
90	Water bottle	0	0	0	10
91	Water bottle holder	0	0	0	10
92	Witch broom	0	0	0	10
93	Witch hat	0.6	1	0.75	15
94	Wooden hammer	0	0	0	10
95	Wooden spoon	0	0	0	10
	<b>Accuracy</b>			0.54	1200
	<b>Macro Average</b>	0.26	0.45	0.32	1200
	<b>Weighted Average</b>	0.31	0.54	0.39	1200

Table 26 LookAT-GG.

<b>LookAT-WA</b>					
<b>N°</b>	<b>Referents</b>	<b>Precision</b>	<b>Recall</b>	<b>F-score</b>	<b>Support Samples</b>
0	Baseball ball	0.6	1	0.75	15
1	Baseball bat	0	0	0	10
2	Baseball glove	0.6	1	0.75	15
3	Bathing slippers	0	0	0	10
4	Bathing suit	0	0	0	15
5	Beer bottle	0.6	1	0.75	15
6	Bike helmet	0	0	0	15
7	Bike lock	0	0	0	10
8	Bike pump	0	0	0	10
9	Bike saddle	0	0	0	15
10	Bike tire	0.2	1	0.33	15
11	book	0	0	0	10
12	Boxing glove	0.6	1	0.75	15
13	Bulletproof vest	0.6	1	0.75	15
14	Catcher mask	0	0	0	10
15	Chef hat	0	0	0	15
16	Cocktail	0	0	0	10
17	Computer keyboard	0.6	1	0.75	15
18	Cooking pot	0.6	1	0.75	15
19	Cruise ship	0	0	0	10
20	Electric wire	0	0	0	10
21	Elegant suit	0.6	1	0.75	15
22	Espresso machine	0.6	1	0.75	15
23	Espresso mug	0	0	0	10
24	Ethernet cable	0	0	0	10
25	Fishing rod	0	0	0	10
26	Flash light	0.6	1	0.75	15
27	Flute	0	0	0	10
28	Football ball	0.6	1	0.75	15
29	Football helmet	0	0	0	10
30	Frying pan	0	0	0	10
31	Fuse box	0.6	1	0.75	15
32	Gardening rake	0	0	0	10
33	Gold necklace	0	0	0	10
34	Graduate diploma	0	0	0	10
35	Graduation cap	0.6	1	0.75	15
36	Hiking backpack	0.3	1	0.46	15
37	Hiking shoe	0	0	0	15
38	Horse saddle	0.6	1	0.75	15

39	Horse shoe	0	0	0	10
40	Kitchen apron	0	0	0	10
41	Lifesaver ring	0	0	0	10
42	Life vest	0.6	1	0.75	15
43	Magnetic compass	0	0	0	10
44	Magnifying glass	0	0	0	10
45	Mailbox	0.6	1	0.75	15
46	Measuring cup	0.3	1	0.46	15
47	Motorbike helmet	0.6	1	0.75	15
48	Motorbike	0	0	0	10
49	Music keyboard	0.6	1	0.75	15
50	Olympic podium	0	0	0	10
51	Pacifier	0	0	0	10
52	Photo camera	0	0	0	10
53	Pill bottle	0.6	1	0.75	15
54	Plant pot	0.6	1	0.75	15
55	Police badge	0	0	0	10
56	Police radio	0	0	0	10
57	Punching bag	0	0	0	10
58	Reading glasses	0.6	1	0.75	15
59	Ring box	0.6	1	0.75	15
60	Road sign	0	0	0	10
61	Rubber boot	0.6	1	0.75	15
62	School backpack	0.6	1	0.75	15
63	Screwdriver	0	0	0	10
64	Seat belt	0.6	1	0.75	15
65	Security camera	0.6	1	0.75	15
66	Sewing machine	0.6	1	0.75	15
67	Shovel	0	0	0	10
68	Sippy cup	0.6	1	0.75	15
69	Ski boot	0.6	1	0.75	15
70	Ski goggles	0.6	1	0.75	15
71	Ski pole	0	0	0	10
72	Soccer ball	0	0	0	10
73	Soccer shoe	0.6	1	0.75	15
74	Spotlight	0.6	1	0.75	15
75	Stamp	0	0	0	10
76	Stubby wrench	0	0	0	10
77	Sunglasses	0.6	1	0.75	15
78	Swimming goggles	0.3	1	0.46	15
79	Swiss army knife	0	0	0	10

80	Syringe	0	0	0	10
81	Thread	0	0	0	10
82	Tool belt	0.6	1	0.75	15
83	Toolbox	0.6	1	0.75	15
84	Tourist map	0	0	0	10
85	Towel	0	0	0	10
86	Tractor tire	0.6	1	0.75	15
87	Triangle ruler	0	0	0	10
88	Uniform cap	0.6	1	0.75	15
89	Video camera	0.6	1	0.75	15
90	Water bottle	0	0	0	10
91	Water bottle holder	0	0	0	10
92	Witch broom	0	0	0	10
93	Witch hat	0.6	1	0.75	15
94	Wooden hammer	0	0	0	10
95	Wooden spoon	0	0	0	10
	<b>Accuracy</b>			0.54	1200
	<b>Macro Average</b>	0.26	0.45	0.32	1200
	<b>Weighted Average</b>	0.31	0.54	0.39	1200

Table 27 LookAT-WA.



<b>WhoAct-GG</b>					
<b>N°</b>	<b>Referents</b>	<b>Precision</b>	<b>Recall</b>	<b>F-score</b>	<b>Support Samples</b>
0	Baseball ball	0.6	1	0.75	30
1	Baseball bat	0	0	0	20
2	Baseball glove	0.6	1	0.75	30
3	Bathing slippers	0	0	0	20
4	Bathing suit	0	0	0	30
5	Beer bottle	0.6	1	0.75	30
6	Bike helmet	0	0	0	30
7	Bike lock	0	0	0	20
8	Bike pump	0	0	0	20
9	Bike saddle	0.2	1	0.33	30
10	Bike tire	0	0	0	30
11	book	0	0	0	20
12	Boxing glove	0.6	1	0.75	30
13	Bulletproof vest	0.6	1	0.75	30
14	Catcher mask	0	0	0	20
15	Chef hat	0.3	1	0.46	30
16	Cocktail	0	0	0	20
17	Computer keyboard	0.6	1	0.75	30
18	Cooking pot	0.6	1	0.75	30
19	Cruise ship	0	0	0	20
20	Electric wire	0	0	0	20
21	Elegant suit	0.6	1	0.75	30
22	Espresso machine	0.6	1	0.75	30
23	Espresso mug	0	0	0	20
24	Ethernet cable	0	0	0	20
25	Fishing rod	0	0	0	20
26	Flash light	0.6	1	0.75	30
27	Flute	0	0	0	20
28	Football ball	0.6	1	0.75	30
29	Football helmet	0	0	0	20
30	Frying pan	0	0	0	20
31	Fuse box	0.6	1	0.75	30
32	Gardening rake	0	0	0	20
33	Gold necklace	0	0	0	20
34	Graduate diploma	0	0	0	20
35	Graduation cap	0.6	1	0.75	30
36	Hiking backpack	0	0	0	30
37	Hiking shoe	0.3	1	0.46	30
38	Horse saddle	0.6	1	0.75	30

39	Horse shoe	0	0	0	20
40	Kitchen apron	0	0	0	20
41	Lifesaver ring	0	0	0	20
42	Life vest	0.6	1	0.75	30
43	Magnetic compass	0	0	0	20
44	Magnifying glass	0	0	0	20
45	Mailbox	0.6	1	0.75	30
46	Measuring cup	0	0	0	30
47	Motorbike helmet	0.6	1	0.75	30
48	Motorbike	0	0	0	20
49	Music keyboard	0.6	1	0.75	30
50	Olympic podium	0	0	0	20
51	Pacifier	0	0	0	20
52	Photo camera	0	0	0	20
53	Pill bottle	0.6	1	0.75	30
54	Plant pot	0.6	1	0.75	30
55	Police badge	0	0	0	20
56	Police radio	0	0	0	20
57	Punching bag	0	0	0	20
58	Reading glasses	0.6	1	0.75	30
59	Ring box	0.6	1	0.75	30
60	Road sign	0	0	0	20
61	Rubber boot	0.6	1	0.75	30
62	School backpack	0.6	1	0.75	30
63	Screwdriver	0	0	0	20
64	Seat belt	0.6	1	0.75	30
65	Security camera	0.6	1	0.75	30
66	Sewing machine	0.6	1	0.75	30
67	Shovel	0	0	0	20
68	Sippy cup	0.6	1	0.75	30
69	Ski boot	0.6	1	0.75	30
70	Ski goggles	0.6	1	0.75	30
71	Ski pole	0	0	0	20
72	Soccer ball	0	0	0	20
73	Soccer shoe	0.6	1	0.75	30
74	Spotlight	0.6	1	0.75	30
75	Stamp	0	0	0	20
76	Stubby wrench	0	0	0	20
77	Sunglasses	0.6	1	0.75	30
78	Swimming goggles	0.3	1	0.46	30
79	Swiss army knife	0	0	0	20

80	Syringe	0	0	0	20
81	Thread	0	0	0	20
82	Tool belt	0.6	1	0.75	30
83	Toolbox	0.6	1	0.75	30
84	Tourist map	0	0	0	20
85	Towel	0	0	0	20
86	Tractor tire	0.6	1	0.75	30
87	Triangle ruler	0	0	0	20
88	Uniform cap	0.6	1	0.75	30
89	Video camera	0.6	1	0.75	30
90	Water bottle	0	0	0	20
91	Water bottle holder	0	0	0	20
92	Witch broom	0	0	0	20
93	Witch hat	0.6	1	0.75	30
94	Wooden hammer	0	0	0	20
95	Wooden spoon	0	0	0	20
	<b>Accuracy</b>			0.54	2400
	<b>Macro Average</b>	0.26	0.45	0.32	2400
	<b>Weighted Average</b>	0.31	0.54	0.39	2400

Table 28 WhoAct-GG.

<b>WhoAct-WA</b>					
<b>N°</b>	<b>Referents</b>	<b>Precision</b>	<b>Recall</b>	<b>F-score</b>	<b>Support Samples</b>
0	Baseball ball	0.6	1	0.75	30
1	Baseball bat	0	0	0	20
2	Baseball glove	0.6	1	0.75	30
3	Bathing slippers	0	0	0	20
4	Bathing suit	0	0	0	30
5	Beer bottle	0.6	1	0.75	30
6	Bike helmet	0.2	1	0.33	30
7	Bike lock	0	0	0	20
8	Bike pump	0	0	0	20
9	Bike saddle	0	0	0	30
10	Bike tire	0	0	0	30
11	book	0	0	0	20
12	Boxing glove	0.6	1	0.75	30
13	Bulletproof vest	0.6	1	0.75	30
14	Catcher mask	0	0	0	20
15	Chef hat	0	0	0	30
16	Cocktail	0	0	0	20
17	Computer keyboard	0.6	1	0.75	30
18	Cooking pot	0.6	1	0.75	30
19	Cruise ship	0	0	0	20
20	Electric wire	0	0	0	20
21	Elegant suit	0.6	1	0.75	30
22	Espresso machine	0.6	1	0.75	30
23	Espresso mug	0	0	0	20
24	Ethernet cable	0	0	0	20
25	Fishing rod	0	0	0	20
26	Flash light	0.6	1	0.75	30
27	Flute	0	0	0	20
28	Football ball	0.6	1	0.75	30
29	Football helmet	0	0	0	20
30	Frying pan	0	0	0	20
31	Fuse box	0.6	1	0.75	30
32	Gardening rake	0	0	0	20
33	Gold necklace	0	0	0	20
34	Graduate diploma	0	0	0	20
35	Graduation cap	0.6	1	0.75	30
36	Hiking backpack	0	0	0	30
37	Hiking shoe	0.3	1	0.46	30
38	Horse saddle	0.6	1	0.75	30

39	Horse shoe	0	0	0	20
40	Kitchen apron	0	0	0	20
41	Lifesaver ring	0	0	0	20
42	Life vest	0.6	1	0.75	30
43	Magnetic compass	0	0	0	20
44	Magnifying glass	0	0	0	20
45	Mailbox	0.6	1	0.75	30
46	Measuring cup	0.3	1	0.46	30
47	Motorbike helmet	0.6	1	0.75	30
48	Motorbike	0	0	0	20
49	Music keyboard	0.6	1	0.75	30
50	Olympic podium	0	0	0	20
51	Pacifier	0	0	0	20
52	Photo camera	0	0	0	20
53	Pill bottle	0.6	1	0.75	30
54	Plant pot	0.6	1	0.75	30
55	Police badge	0	0	0	20
56	Police radio	0	0	0	20
57	Punching bag	0	0	0	20
58	Reading glasses	0.6	1	0.75	30
59	Ring box	0.6	1	0.75	30
60	Road sign	0	0	0	20
61	Rubber boot	0.6	1	0.75	30
62	School backpack	0.6	1	0.75	30
63	Screwdriver	0	0	0	20
64	Seat belt	0.6	1	0.75	30
65	Security camera	0.6	1	0.75	30
66	Sewing machine	0.6	1	0.75	30
67	Shovel	0	0	0	20
68	Sippy cup	0.6	1	0.75	30
69	Ski boot	0.6	1	0.75	30
70	Ski goggles	0.6	1	0.75	30
71	Ski pole	0	0	0	20
72	Soccer ball	0	0	0	20
73	Soccer shoe	0.6	1	0.75	30
74	Spotlight	0.6	1	0.75	30
75	Stamp	0	0	0	20
76	Stubby wrench	0	0	0	20
77	Sunglasses	0.6	1	0.75	30
78	Swimming goggles	0.3	1	0.46	30
79	Swiss army knife	0	0	0	20

80	Syringe	0	0	0	20
81	Thread	0	0	0	20
82	Tool belt	0.6	1	0.75	30
83	Toolbox	0.6	1	0.75	30
84	Tourist map	0	0	0	20
85	Towel	0	0	0	20
86	Tractor tire	0.6	1	0.75	30
87	Triangle ruler	0	0	0	20
88	Uniform cap	0.6	1	0.75	30
89	Video camera	0.6	1	0.75	30
90	Water bottle	0	0	0	20
91	Water bottle holder	0	0	0	20
92	Witch broom	0	0	0	20
93	Witch hat	0.6	1	0.75	30
94	Wooden hammer	0	0	0	20
95	Wooden spoon	0	0	0	20
	<b>Accuracy</b>			0.54	2400
	<b>Macro Average</b>	0.26	0.45	0.32	2400
	<b>Weighted Average</b>	0.31	0.54	0.39	2400

Table 29 WhoAct-GG.

## Agent and Verb

LookAT-GG					
N°	Referents	Precision	Recall	F-score	Support Samples
0	Baseball ball	1	1	1	15
1	Baseball bat	1	1	1	10
2	Baseball glove	1	1	1	15
3	Bathing slippers	1	1	1	10
4	Bathing suit	1	0.67	0.8	15
5	Beer bottle	1	1	1	15
6	Bike helmet	1	1	1	15
7	Bike lock	1	1	1	10
8	Bike pump	1	1	1	10
9	Bike saddle	0	0	0	15
10	Bike tire	0.5	1	0.67	15
11	book	1	1	1	10
12	Boxing glove	1	1	1	15
13	Bulletproof vest	1	1	1	15
14	Catcher mask	1	1	1	10
15	Chef hat	1	1	1	15
16	Cocktail	1	1	1	10
17	Computer keyboard	1	1	1	15
18	Cooking pot	1	1	1	15
19	Cruise ship	1	1	1	10
20	Electric wire	1	1	1	10
21	Elegant suit	1	1	1	15
22	Espresso machine	1	1	1	15
23	Espresso mug	1	1	1	10
24	Ethernet cable	1	1	1	10
25	Fishing rod	1	1	1	10
26	Flash light	1	1	1	15
27	Flute	1	1	1	10
28	Football ball	1	1	1	15
29	Football helmet	1	1	1	10
30	Frying pan	1	1	1	10
31	Fuse box	1	1	1	15
32	Gardening rake	1	1	1	10
33	Gold necklace	1	1	1	10
34	Graduate diploma	1	1	1	10
35	Graduation cap	1	1	1	15
36	Hiking backpack	1	1	1	15

37	Hiking shoe	1	1	1	15
38	Horse saddle	1	1	1	15
39	Horse shoe	1	1	1	10
40	Kitchen apron	1	1	1	10
41	Lifesaver ring	1	1	1	10
42	Life vest	1	1	1	15
43	Magnetic compass	1	1	1	10
44	Magnifying glass	1	1	1	10
45	Mailbox	1	1	1	15
46	Measuring cup	1	1	1	15
47	Motorbike helmet	1	1	1	15
48	Motorbike	1	1	1	10
49	Music keyboard	1	1	1	15
50	Olympic podium	1	1	1	10
51	Pacifier	1	1	1	10
52	Photo camera	1	1	1	10
53	Pill bottle	1	1	1	15
54	Plant pot	1	1	1	15
55	Police badge	1	1	1	10
56	Police radio	1	1	1	10
57	Punching bag	1	1	1	10
58	Reading glasses	1	1	1	15
59	Ring box	1	1	1	15
60	Road sign	1	1	1	10
61	Rubber boot	1	1	1	15
62	School backpack	1	1	1	15
63	Screwdriver	1	1	1	10
64	Seat belt	1	1	1	15
65	Security camera	1	1	1	15
66	Sewing machine	1	1	1	15
67	Shovel	1	1	1	10
68	Sippy cup	1	1	1	15
69	Ski boot	1	1	1	15
70	Ski goggles	1	1	1	15
71	Ski pole	1	1	1	10
72	Soccer ball	1	1	1	10
73	Soccer shoe	1	1	1	15
74	Spotlight	1	1	1	15
75	Stamp	1	1	1	10
76	Stubby wrench	1	1	1	10
77	Sunglasses	1	1	1	15



78	Swimming goggles	0.75	1	0.86	15
79	Swiss army knife	1	1	1	10
80	Syringe	1	1	1	10
81	Thread	1	1	1	10
82	Tool belt	1	1	1	15
83	Toolbox	1	1	1	15
84	Tourist map	1	1	1	10
85	Towel	1	1	1	10
86	Tractor tire	1	1	1	15
87	Triangle ruler	1	1	1	10
88	Uniform cap	1	1	1	15
89	Video camera	1	1	1	15
90	Water bottle	1	1	1	10
91	Water bottle holder	1	1	1	10
92	Witch broom	1	1	1	10
93	Witch hat	1	1	1	15
94	Wooden hammer	1	1	1	10
95	Wooden spoon	1	1	1	10
	<b>Accuracy</b>			0.98	1200
	<b>Macro Average</b>	0.98	0.99	0.98	1200
	<b>Weighted Average</b>	0.98	0.98	0.98	1200

Table 30 LookAT-GG.

<b>LookAT-WA</b>					
<b>N°</b>	<b>Referents</b>	<b>Precision</b>	<b>Recall</b>	<b>F-score</b>	<b>Support Samples</b>
0	Baseball ball	1	1	1	15
1	Baseball bat	1	1	1	10
2	Baseball glove	1	1	1	15
3	Bathing slippers	1	1	1	10
4	Bathing suit	1	0.67	0.8	15
5	Beer bottle	1	1	1	15
6	Bike helmet	1	1	1	15
7	Bike lock	1	1	1	10
8	Bike pump	1	1	1	10
9	Bike saddle	0	0	0	15
10	Bike tire	0.5	1	0.67	15
11	book	1	1	1	10
12	Boxing glove	1	1	1	15
13	Bulletproof vest	1	1	1	15
14	Catcher mask	1	1	1	10
15	Chef hat	1	1	1	15
16	Cocktail	1	1	1	10
17	Computer keyboard	1	1	1	15
18	Cooking pot	1	1	1	15
19	Cruise ship	1	1	1	10
20	Electric wire	1	1	1	10
21	Elegant suit	1	1	1	15
22	Espresso machine	1	1	1	15
23	Espresso mug	1	1	1	10
24	Ethernet cable	1	1	1	10
25	Fishing rod	1	1	1	10
26	Flash light	1	1	1	15
27	Flute	1	1	1	10
28	Football ball	1	1	1	15
29	Football helmet	1	1	1	10
30	Frying pan	1	1	1	10
31	Fuse box	1	1	1	15
32	Gardening rake	1	1	1	10
33	Gold necklace	1	1	1	10
34	Graduate diploma	1	1	1	10
35	Graduation cap	1	1	1	15
36	Hiking backpack	1	1	1	15
37	Hiking shoe	1	1	1	15
38	Horse saddle	1	1	1	15

39	Horse shoe	1	1	1	10
40	Kitchen apron	1	1	1	10
41	Lifesaver ring	1	1	1	10
42	Life vest	1	1	1	15
43	Magnetic compass	1	1	1	10
44	Magnifying glass	1	1	1	10
45	Mailbox	1	1	1	15
46	Measuring cup	1	1	1	15
47	Motorbike helmet	1	1	1	15
48	Motorbike	1	1	1	10
49	Music keyboard	1	1	1	15
50	Olympic podium	1	1	1	10
51	Pacifier	1	1	1	10
52	Photo camera	1	1	1	10
53	Pill bottle	1	1	1	15
54	Plant pot	1	1	1	15
55	Police badge	1	1	1	10
56	Police radio	1	1	1	10
57	Punching bag	1	1	1	10
58	Reading glasses	1	1	1	15
59	Ring box	1	1	1	15
60	Road sign	1	1	1	10
61	Rubber boot	1	1	1	15
62	School backpack	1	1	1	15
63	Screwdriver	1	1	1	10
64	Seat belt	1	1	1	15
65	Security camera	1	1	1	15
66	Sewing machine	1	1	1	15
67	Shovel	1	1	1	10
68	Sippy cup	1	1	1	15
69	Ski boot	1	1	1	15
70	Ski goggles	1	1	1	15
71	Ski pole	1	1	1	10
72	Soccer ball	1	1	1	10
73	Soccer shoe	1	1	1	15
74	Spotlight	1	1	1	15
75	Stamp	1	1	1	10
76	Stubby wrench	1	1	1	10
77	Sunglasses	1	1	1	15
78	Swimming goggles	0.75	1	0.86	15
79	Swiss army knife	1	1	1	10

80	Syringe	1	1	1	10
81	Thread	1	1	1	10
82	Tool belt	1	1	1	15
83	Toolbox	1	1	1	15
84	Tourist map	1	1	1	10
85	Towel	1	1	1	10
86	Tractor tire	1	1	1	15
87	Triangle ruler	1	1	1	10
88	Uniform cap	1	1	1	15
89	Video camera	1	1	1	15
90	Water bottle	1	1	1	10
91	Water bottle holder	1	1	1	10
92	Witch broom	1	1	1	10
93	Witch hat	1	1	1	15
94	Wooden hammer	1	1	1	10
95	Wooden spoon	1	1	1	10
	<b>Accuracy</b>			0.98	1200
	<b>Macro Average</b>	0.98	0.99	0.98	1200
	<b>Weighted Average</b>	0.98	0.98	0.98	1200

Table 31 LookAT-WA.

<b>WhoAct-GG</b>					
<b>N°</b>	<b>Referents</b>	<b>Precision</b>	<b>Recall</b>	<b>F-score</b>	<b>Support Samples</b>
0	Baseball ball	1	1	1	30
1	Baseball bat	1	1	1	20
2	Baseball glove	1	1	1	30
3	Bathing slippers	1	1	1	20
4	Bathing suit	0.75	1	0.86	30
5	Beer bottle	1	1	1	30
6	Bike helmet	1	1	1	30
7	Bike lock	1	1	1	20
8	Bike pump	1	1	1	20
9	Bike saddle	0.5	0.67	0.57	30
10	Bike tire	0.5	0.33	0.4	30
11	book	1	1	1	20
12	Boxing glove	1	1	1	30
13	Bulletproof vest	1	1	1	30
14	Catcher mask	1	1	1	20
15	Chef hat	1	1	1	30
16	Cocktail	1	1	1	20
17	Computer keyboard	1	1	1	30
18	Cooking pot	1	1	1	30
19	Cruise ship	1	1	1	20
20	Electric wire	1	1	1	20
21	Elegant suit	1	1	1	30
22	Espresso machine	1	1	1	30
23	Espresso mug	1	1	1	20
24	Ethernet cable	1	1	1	20
25	Fishing rod	1	1	1	20
26	Flash light	1	1	1	30
27	Flute	1	1	1	20
28	Football ball	1	1	1	30
29	Football helmet	1	1	1	20
30	Frying pan	1	1	1	20
31	Fuse box	1	1	1	30
32	Gardening rake	1	1	1	20
33	Gold necklace	1	1	1	20
34	Graduate diploma	1	1	1	20
35	Graduation cap	1	1	1	30
36	Hiking backpack	1	1	1	30
37	Hiking shoe	1	1	1	30
38	Horse saddle	1	1	1	30

39	Horse shoe	1	1	1	20
40	Kitchen apron	1	1	1	20
41	Lifesaver ring	1	1	1	20
42	Life vest	1	1	1	30
43	Magnetic compass	1	1	1	20
44	Magnifying glass	1	1	1	20
45	Mailbox	1	1	1	30
46	Measuring cup	1	1	1	30
47	Motorbike helmet	1	1	1	30
48	Motorbike	1	1	1	20
49	Music keyboard	1	1	1	30
50	Olympic podium	1	1	1	20
51	Pacifier	1	1	1	20
52	Photo camera	1	1	1	20
53	Pill bottle	1	1	1	30
54	Plant pot	1	1	1	30
55	Police badge	1	1	1	20
56	Police radio	1	1	1	20
57	Punching bag	1	1	1	20
58	Reading glasses	1	1	1	30
59	Ring box	1	1	1	30
60	Road sign	1	1	1	20
61	Rubber boot	1	1	1	30
62	School backpack	1	1	1	30
63	Screwdriver	1	1	1	20
64	Seat belt	1	1	1	30
65	Security camera	1	1	1	30
66	Sewing machine	1	1	1	30
67	Shovel	1	1	1	20
68	Sippy cup	1	1	1	30
69	Ski boot	1	1	1	30
70	Ski goggles	1	1	1	30
71	Ski pole	1	1	1	20
72	Soccer ball	1	1	1	20
73	Soccer shoe	1	1	1	30
74	Spotlight	1	1	1	30
75	Stamp	1	1	1	20
76	Stubby wrench	1	1	1	20
77	Sunglasses	1	1	1	30
78	Swimming goggles	1	0.67	0.8	30
79	Swiss army knife	1	1	1	20

80	Syringe	1	1	1	20
81	Thread	1	1	1	20
82	Tool belt	1	1	1	30
83	Toolbox	1	1	1	30
84	Tourist map	1	1	1	20
85	Towel	1	1	1	20
86	Tractor tire	1	1	1	30
87	Triangle ruler	1	1	1	20
88	Uniform cap	1	1	1	30
89	Video camera	1	1	1	30
90	Water bottle	1	1	1	20
91	Water bottle holder	1	1	1	20
92	Witch broom	1	1	1	20
93	Witch hat	1	1	1	30
94	Wooden hammer	1	1	1	20
95	Wooden spoon	1	1	1	20
	<b>Accuracy</b>			0.98	2400
	<b>Macro Average</b>	0.99	0.99	0.99	2400
	<b>Weighted Average</b>	0.98	0.98	0.98	2400

Table 32 WhoAct-GG.

WhoAct-WA					
N°	Referents	Precision	Recall	F-score	Support Samples
0	Baseball ball	1	1	1	30
1	Baseball bat	1	1	1	20
2	Baseball glove	1	1	1	30
3	Bathing slippers	1	1	1	20
4	Bathing suit	1	0.67	0.8	30
5	Beer bottle	1	1	1	30
6	Bike helmet	1	1	1	30
7	Bike lock	1	1	1	20
8	Bike pump	1	1	1	20
9	Bike saddle	0.5	0.33	0.4	30
10	Bike tire	0.5	0.67	0.57	30
11	book	1	1	1	20
12	Boxing glove	1	1	1	30
13	Bulletproof vest	1	1	1	30
14	Catcher mask	1	1	1	20
15	Chef hat	1	1	1	30
16	Cocktail	1	1	1	20
17	Computer keyboard	1	1	1	30
18	Cooking pot	1	1	1	30
19	Cruise ship	1	1	1	20
20	Electric wire	1	1	1	20
21	Elegant suit	1	1	1	30
22	Espresso machine	1	1	1	30
23	Espresso mug	1	1	1	20
24	Ethernet cable	1	1	1	20
25	Fishing rod	1	1	1	20
26	Flash light	1	1	1	30
27	Flute	1	1	1	20
28	Football ball	1	1	1	30
29	Football helmet	1	1	1	20
30	Frying pan	1	1	1	20
31	Fuse box	1	1	1	30
32	Gardening rake	1	1	1	20
33	Gold necklace	1	1	1	20
34	Graduate diploma	1	1	1	20
35	Graduation cap	1	1	1	30
36	Hiking backpack	1	1	1	30
37	Hiking shoe	1	1	1	30
38	Horse saddle	1	1	1	30



39	Horse shoe	1	1	1	20
40	Kitchen apron	1	1	1	20
41	Lifesaver ring	1	1	1	20
42	Life vest	1	1	1	30
43	Magnetic compass	1	1	1	20
44	Magnifying glass	1	1	1	20
45	Mailbox	1	1	1	30
46	Measuring cup	1	1	1	30
47	Motorbike helmet	1	1	1	30
48	Motorbike	1	1	1	20
49	Music keyboard	1	1	1	30
50	Olympic podium	1	1	1	20
51	Pacifier	1	1	1	20
52	Photo camera	1	1	1	20
53	Pill bottle	1	1	1	30
54	Plant pot	1	1	1	30
55	Police badge	1	1	1	20
56	Police radio	1	1	1	20
57	Punching bag	1	1	1	20
58	Reading glasses	1	1	1	30
59	Ring box	1	1	1	30
60	Road sign	1	1	1	20
61	Rubber boot	1	1	1	30
62	School backpack	1	1	1	30
63	Screwdriver	1	1	1	20
64	Seat belt	1	1	1	30
65	Security camera	1	1	1	30
66	Sewing machine	1	1	1	30
67	Shovel	1	1	1	20
68	Sippy cup	1	1	1	30
69	Ski boot	1	1	1	30
70	Ski goggles	1	1	1	30
71	Ski pole	1	1	1	20
72	Soccer ball	1	1	1	20
73	Soccer shoe	1	1	1	30
74	Spotlight	1	1	1	30
75	Stamp	1	1	1	20
76	Stubby wrench	1	1	1	20
77	Sunglasses	1	1	1	30
78	Swimming goggles	0.75	1	0.86	30
79	Swiss army knife	1	1	1	20

80	Syringe	1	1	1	20
81	Thread	1	1	1	20
82	Tool belt	1	1	1	30
83	Toolbox	1	1	1	30
84	Tourist map	1	1	1	20
85	Towel	1	1	1	20
86	Tractor tire	1	1	1	30
87	Triangle ruler	1	1	1	20
88	Uniform cap	1	1	1	30
89	Video camera	1	1	1	30
90	Water bottle	1	1	1	20
91	Water bottle holder	1	1	1	20
92	Witch broom	1	1	1	20
93	Witch hat	1	1	1	30
94	Wooden hammer	1	1	1	20
95	Wooden spoon	1	1	1	20
	<b>Accuracy</b>			0.98	2400
	<b>Macro Average</b>	0.99	0.99	0.99	2400
	<b>Weighted Average</b>	0.98	0.98	0.98	2400

Table 33 WhoAct-WA.

## Perceptually Underspecified Noun

LookAT-GG					
N°	Referents	Precision	Recall	F-score	Support Samples
0	Baseball ball	1	1	1	15
1	Baseball bat	1	1	1	10
2	Baseball glove	1	1	1	15
3	Bathing slippers	1	1	1	10
4	Bathing suit	1	1	1	15
5	Beer bottle	1	1	1	15
6	Bike helmet	1	1	1	15
7	Bike lock	1	1	1	10
8	Bike pump	1	1	1	10
9	Bike saddle	1	1	1	15
10	Bike tire	1	1	1	15
11	book	1	1	1	10
12	Boxing glove	1	1	1	15
13	Bulletproof vest	1	1	1	15
14	Catcher mask	1	1	1	10
15	Chef hat	1	1	1	15
16	Cocktail	1	1	1	10
17	Computer keyboard	1	1	1	15
18	Cooking pot	1	1	1	15
19	Cruise ship	1	1	1	10
20	Electric wire	1	1	1	10
21	Elegant suit	1	1	1	15
22	Espresso machine	1	1	1	15
23	Espresso mug	1	1	1	10
24	Ethernet cable	1	1	1	10
25	Fishing rod	1	1	1	10
26	Flash light	1	1	1	15
27	Flute	1	1	1	10
28	Football ball	1	1	1	15
29	Football helmet	1	1	1	10
30	Frying pan	1	1	1	10
31	Fuse box	1	1	1	15
32	Gardening rake	1	1	1	10
33	Gold necklace	1	1	1	10
34	Graduate diploma	1	1	1	10
35	Graduation cap	1	1	1	15
36	Hiking backpack	1	1	1	15

37	Hiking shoe	1	1	1	15
38	Horse saddle	1	1	1	15
39	Horse shoe	1	1	1	10
40	Kitchen apron	1	1	1	10
41	Lifesaver ring	1	1	1	10
42	Life vest	1	1	1	15
43	Magnetic compass	1	1	1	10
44	Magnifying glass	1	1	1	10
45	Mailbox	1	1	1	15
46	Measuring cup	1	1	1	15
47	Motorbike helmet	1	1	1	15
48	Motorbike	1	1	1	10
49	Music keyboard	1	1	1	15
50	Olympic podium	1	1	1	10
51	Pacifier	1	1	1	10
52	Photo camera	1	1	1	10
53	Pill bottle	1	1	1	15
54	Plant pot	1	1	1	15
55	Police badge	1	1	1	10
56	Police radio	1	1	1	10
57	Punching bag	1	1	1	10
58	Reading glasses	1	1	1	15
59	Ring box	1	1	1	15
60	Road sign	1	1	1	10
61	Rubber boot	1	1	1	15
62	School backpack	1	1	1	15
63	Screwdriver	1	1	1	10
64	Seat belt	1	1	1	15
65	Security camera	1	1	1	15
66	Sewing machine	1	1	1	15
67	Shovel	1	1	1	10
68	Sippy cup	1	1	1	15
69	Ski boot	1	1	1	15
70	Ski goggles	1	1	1	15
71	Ski pole	1	1	1	10
72	Soccer ball	1	1	1	10
73	Soccer shoe	1	1	1	15
74	Spotlight	1	1	1	15
75	Stamp	1	1	1	10
76	Stubby wrench	1	1	1	10
77	Sunglasses	1	1	1	15

78	Swimming goggles	1	1	1	15
79	Swiss army knife	1	1	1	10
80	Syringe	1	1	1	10
81	Thread	1	1	1	10
82	Tool belt	1	1	1	15
83	Toolbox	1	1	1	15
84	Tourist map	1	1	1	10
85	Towel	1	1	1	10
86	Tractor tire	1	1	1	15
87	Triangle ruler	1	1	1	10
88	Uniform cap	1	1	1	15
89	Video camera	1	1	1	15
90	Water bottle	1	1	1	10
91	Water bottle holder	1	1	1	10
92	Witch broom	1	1	1	10
93	Witch hat	1	1	1	15
94	Wooden hammer	1	1	1	10
95	Wooden spoon	1	1	1	10
	<b>Accuracy</b>			1	1200
	<b>Macro Average</b>	1	1	1	1200
	<b>Weighted Average</b>	1	1	1	1200

Table 34 LookAT-GG.

<b>LookAT-WA</b>					
<b>N°</b>	<b>Referents</b>	<b>Precision</b>	<b>Recall</b>	<b>F-score</b>	<b>Support Samples</b>
0	Baseball ball	1	1	1	15
1	Baseball bat	1	1	1	10
2	Baseball glove	1	1	1	15
3	Bathing slippers	1	1	1	10
4	Bathing suit	1	1	1	15
5	Beer bottle	1	1	1	15
6	Bike helmet	1	1	1	15
7	Bike lock	1	1	1	10
8	Bike pump	1	1	1	10
9	Bike saddle	1	1	1	15
10	Bike tire	1	1	1	15
11	book	1	1	1	10
12	Boxing glove	1	1	1	15
13	Bulletproof vest	1	1	1	15
14	Catcher mask	1	1	1	10
15	Chef hat	1	1	1	15
16	Cocktail	1	1	1	10
17	Computer keyboard	1	1	1	15
18	Cooking pot	1	1	1	15
19	Cruise ship	1	1	1	10
20	Electric wire	1	1	1	10
21	Elegant suit	1	1	1	15
22	Espresso machine	1	1	1	15
23	Espresso mug	1	1	1	10
24	Ethernet cable	1	1	1	10
25	Fishing rod	1	1	1	10
26	Flash light	1	1	1	15
27	Flute	1	1	1	10
28	Football ball	1	1	1	15
29	Football helmet	1	1	1	10
30	Frying pan	1	1	1	10
31	Fuse box	1	1	1	15
32	Gardening rake	1	1	1	10
33	Gold necklace	1	1	1	10
34	Graduate diploma	1	1	1	10
35	Graduation cap	1	1	1	15
36	Hiking backpack	1	1	1	15
37	Hiking shoe	1	1	1	15
38	Horse saddle	1	1	1	15

39	Horse shoe	1	1	1	10
40	Kitchen apron	1	1	1	10
41	Lifesaver ring	1	1	1	10
42	Life vest	1	1	1	15
43	Magnetic compass	1	1	1	10
44	Magnifying glass	1	1	1	10
45	Mailbox	1	1	1	15
46	Measuring cup	1	1	1	15
47	Motorbike helmet	1	1	1	15
48	Motorbike	1	1	1	10
49	Music keyboard	1	1	1	15
50	Olympic podium	1	1	1	10
51	Pacifier	1	1	1	10
52	Photo camera	1	1	1	10
53	Pill bottle	1	1	1	15
54	Plant pot	1	1	1	15
55	Police badge	1	1	1	10
56	Police radio	1	1	1	10
57	Punching bag	1	1	1	10
58	Reading glasses	1	1	1	15
59	Ring box	1	1	1	15
60	Road sign	1	1	1	10
61	Rubber boot	1	1	1	15
62	School backpack	1	1	1	15
63	Screwdriver	1	1	1	10
64	Seat belt	1	1	1	15
65	Security camera	1	1	1	15
66	Sewing machine	1	1	1	15
67	Shovel	1	1	1	10
68	Sippy cup	1	1	1	15
69	Ski boot	1	1	1	15
70	Ski goggles	1	1	1	15
71	Ski pole	1	1	1	10
72	Soccer ball	1	1	1	10
73	Soccer shoe	1	1	1	15
74	Spotlight	1	1	1	15
75	Stamp	1	1	1	10
76	Stubby wrench	1	1	1	10
77	Sunglasses	1	1	1	15
78	Swimming goggles	1	1	1	15
79	Swiss army knife	1	1	1	10

80	Syringe	1	1	1	10
81	Thread	1	1	1	10
82	Tool belt	1	1	1	15
83	Toolbox	1	1	1	15
84	Tourist map	1	1	1	10
85	Towel	1	1	1	10
86	Tractor tire	1	1	1	15
87	Triangle ruler	1	1	1	10
88	Uniform cap	1	1	1	15
89	Video camera	1	1	1	15
90	Water bottle	1	1	1	10
91	Water bottle holder	1	1	1	10
92	Witch broom	1	1	1	10
93	Witch hat	1	1	1	15
94	Wooden hammer	1	1	1	10
95	Wooden spoon	1	1	1	10
	<b>Accuracy</b>			1	1200
	<b>Macro Average</b>	1	1	1	1200
	<b>Weighted Average</b>	1	1	1	1200

Table 35 LookAT-WA.



WhoAct-GG					
N°	Referents	Precision	Recall	F1_score	Support Samples
0	Baseball ball	0.38	1	0.55	15
1	Baseball bat	1	1	1	10
2	Baseball glove	0.5	1	0.67	15
3	Bathing slippers	1	1	1	10
4	Bathing suit	0	0	0	15
5	Beer bottle	0.38	1	0.55	15
6	Bike helmet	0.38	1	0.55	15
7	Bike lock	1	1	1	10
8	Bike pump	1	1	1	10
9	Bike saddle	0.5	1	0.67	15
10	Bike tire	0.5	1	0.67	15
11	book	1	1	1	10
12	Boxing glove	0	0	0	15
13	Bulletproof vest	0	0	0	15
14	Catcher mask	1	1	1	10
15	Chef hat	0	0	0	15
16	Cocktail	1	1	1	10
17	Computer keyboard	0	0	0	15
18	Cooking pot	0	0	0	15
19	Cruise ship	1	1	1	10
20	Electric wire	1	1	1	10
21	Elegant suit	0.5	1	0.67	15
22	Espresso machine	0	0	0	15
23	Espresso mug	1	1	1	10
24	Ethernet cable	1	1	1	10
25	Fishing rod	1	1	1	10
26	Flash light	0.5	1	0.67	15
27	Flute	1	1	1	10
28	Football ball	0	0	0	15
29	Football helmet	0	0	0	10
30	Frying pan	1	1	1	10
31	Fuse box	0	0	0	15
32	Gardening rake	1	1	1	10
33	Gold necklace	1	1	1	10
34	Graduate diploma	1	1	1	10
35	Graduation cap	0.5	1	0.67	15
36	Hiking backpack	0.5	1	0.67	15
37	Hiking shoe	0.5	1	0.67	15
38	Horse saddle	0	0	0	15

39	Horse shoe	1	1	1	10
40	Kitchen apron	1	1	1	10
41	Lifesaver ring	1	1	1	10
42	Life vest	0.5	1	0.67	15
43	Magnetic compass	1	1	1	10
44	Magnifying glass	1	1	1	10
45	Mailbox	0	0	0	15
46	Measuring cup	0	0	0	15
47	Motorbike helmet	0	0	0	15
48	Motorbike	1	1	1	10
49	Music keyboard	0.5	1	0.67	15
50	Olympic podium	1	1	1	10
51	Pacifier	1	1	1	10
52	Photo camera	0	0	0	10
53	Pill bottle	0	0	0	15
54	Plant pot	0.5	1	0.67	15
55	Police badge	1	1	1	10
56	Police radio	1	1	1	10
57	Punching bag	1	1	1	10
58	Reading glasses	0.5	1	0.67	15
59	Ring box	0	0	0	15
60	Road sign	1	1	1	10
61	Rubber boot	0	0	0	15
62	School backpack	0	0	0	15
63	Screwdriver	1	1	1	10
64	Seat belt	0	0	0	15
65	Security camera	0.38	1	0.55	15
66	Sewing machine	0.5	1	0.67	15
67	Shovel	1	1	1	10
68	Sippy cup	0.5	1	0.67	15
69	Ski boot	0.5	1	0.67	15
70	Ski goggles	0	0	0	15
71	Ski pole	1	1	1	10
72	Soccer ball	0	0	0	10
73	Soccer shoe	0	0	0	15
74	Spotlight	0	0	0	15
75	Stamp	1	1	1	10
76	Stubby wrench	1	1	1	10
77	Sunglasses	0	0	0	15
78	Swimming goggles	0.5	1	0.67	15
79	Swiss army knife	1	1	1	10

80	Syringe	1	1	1	10
81	Thread	1	1	1	10
82	Tool belt	0.5	1	0.67	15
83	Toolbox	0.25	1	0.4	15
84	Tourist map	1	1	1	10
85	Towel	1	1	1	10
86	Tractor tire	0	0	0	15
87	Triangle ruler	1	1	1	10
88	Uniform cap	0	0	0	15
89	Video camera	0	0	0	15
90	Water bottle	0	0	0	10
91	Water bottle holder	1	1	1	10
92	Witch broom	1	1	1	10
93	Witch hat	0.5	1	0.67	15
94	Wooden hammer	1	1	1	10
95	Wooden spoon	1	1	1	10
	<b>Accuracy</b>			0.65	1200
	<b>Macro Average</b>	0.57	0.7	0.61	1200
	<b>Weighted Average</b>	0.5	0.65	0.55	1200

Table 36 WhoAct-GG.

<b>WhoAct-WA</b>					
<b>N°</b>	<b>Referents</b>	<b>Precision</b>	<b>Recall</b>	<b>F-score</b>	<b>Support Samples</b>
0	Baseball ball	0	0	0	15
1	Baseball bat	1	1	1	10
2	Baseball glove	0	0	0	15
3	Bathing slippers	1	1	1	10
4	Bathing suit	0.5	1	0.67	15
5	Beer bottle	0.38	1	0.55	15
6	Bike helmet	0	0	0	15
7	Bike lock	1	1	1	10
8	Bike pump	1	1	1	10
9	Bike saddle	0	0	0	15
10	Bike tire	0.5	1	0.67	15
11	book	1	1	1	10
12	Boxing glove	0.5	1	0.67	15
13	Bulletproof vest	0	0	0	15
14	Catcher mask	1	1	1	10
15	Chef hat	0.5	1	0.67	15
16	Cocktail	1	1	1	10
17	Computer keyboard	0.5	1	0.67	15
18	Cooking pot	0.5	1	0.67	15
19	Cruise ship	1	1	1	10
20	Electric wire	1	1	1	10
21	Elegant suit	0	0	0	15
22	Espresso machine	0	0	0	15
23	Espresso mug	1	1	1	10
24	Ethernet cable	1	1	1	10
25	Fishing rod	1	1	1	10
26	Flash light	0.5	1	0.67	15
27	Flute	1	1	1	10
28	Football ball	0.38	1	0.55	15
29	Football helmet	0	0	0	10
30	Frying pan	1	1	1	10
31	Fuse box	0	0	0	15
32	Gardening rake	1	1	1	10
33	Gold necklace	1	1	1	10
34	Graduate diploma	1	1	1	10
35	Graduation cap	0.5	1	0.67	15
36	Hiking backpack	0.5	1	0.67	15
37	Hiking shoe	0.5	1	0.67	15
38	Horse saddle	0.5	1	0.67	15

39	Horse shoe	1	1	1	10
40	Kitchen apron	1	1	1	10
41	Lifesaver ring	1	1	1	10
42	Life vest	0.5	1	0.67	15
43	Magnetic compass	1	1	1	10
44	Magnifying glass	1	1	1	10
45	Mailbox	0	0	0	15
46	Measuring cup	0	0	0	15
47	Motorbike helmet	0.38	1	0.55	15
48	Motorbike	1	1	1	10
49	Music keyboard	0	0	0	15
50	Olympic podium	1	1	1	10
51	Pacifier	1	1	1	10
52	Photo camera	0	0	0	10
53	Pill bottle	0	0	0	15
54	Plant pot	0	0	0	15
55	Police badge	1	1	1	10
56	Police radio	1	1	1	10
57	Punching bag	1	1	1	10
58	Reading glasses	0.5	1	0.67	15
59	Ring box	0	0	0	15
60	Road sign	1	1	1	10
61	Rubber boot	0	0	0	15
62	School backpack	0	0	0	15
63	Screwdriver	1	1	1	10
64	Seat belt	0	0	0	15
65	Security camera	0	0	0	15
66	Sewing machine	0.5	1	0.67	15
67	Shovel	1	1	1	10
68	Sippy cup	0.5	1	0.67	15
69	Ski boot	0.5	1	0.67	15
70	Ski goggles	0.5	1	0.67	15
71	Ski pole	1	1	1	10
72	Soccer ball	0	0	0	10
73	Soccer shoe	0	0	0	15
74	Spotlight	0	0	0	15
75	Stamp	1	1	1	10
76	Stubby wrench	1	1	1	10
77	Sunglasses	0	0	0	15
78	Swimming goggles	0	0	0	15
79	Swiss army knife	1	1	1	10

80	Syringe	1	1	1	10
81	Thread	1	1	1	10
82	Tool belt	0.5	1	0.67	15
83	Toolbox	0.25	1	0.4	15
84	Tourist map	1	1	1	10
85	Towel	1	1	1	10
86	Tractor tire	0	0	0	15
87	Triangle ruler	1	1	1	10
88	Uniform cap	0	0	0	15
89	Video camera	0.38	1	0.55	15
90	Water bottle	0	0	0	10
91	Water bottle holder	1	1	1	10
92	Witch broom	1	1	1	10
93	Witch hat	0	0	0	15
94	Wooden hammer	1	1	1	10
95	Wooden spoon	1	1	1	10
	<b>Accuracy</b>				
	<b>Macro Average</b>			0.65	1200
	<b>Weighted Average</b>	0.57	0.7	0.61	1200

Table 37 WhoAct-WA.

## Event

<b>Matrix</b>	<b>Loss*</b>	<b>Accuracy *</b>	<b>Precision*</b>	<b>Recall*</b>	<b>F-score</b>
GG	0	100	100	99,82	1
WA	0	100	100	99,80	1

\* (%)

Table 38 LookAT loss, accuracy, precision, recall and F-score measures.

<b>Matrix</b>	<b>Loss*</b>	<b>Accuracy *</b>	<b>Precision*</b>	<b>Recall*</b>	<b>F-score</b>
GG	0	100	100	99,69	1
WA	0	100	100	99,65	1

\* (%)

Table 39 WhoAct loss, accuracy, precision, recall and F-score measures.

## Agent

<b>Matrix</b>	<b>Loss*</b>	<b>Accuracy *</b>	<b>Precision*</b>	<b>Recall*</b>	<b>F-score</b>
<b>GG</b>	1,61	99,13	57,86	47,08	0,54
<b>WA</b>	1,61	99,13	57,47	46,96	0,54

\* (%)

Table 40 LookAT loss, accuracy, precision, recall and F-score measures.

<b>Matrix</b>	<b>Loss*</b>	<b>Accuracy *</b>	<b>Precision*</b>	<b>Recall*</b>	<b>F-score</b>
GG	1,61	99,13	57,26	47,01	0,54
WA	1,61	99,13	56,70	47,10	0,54

\* (%)

Table 41 WhoAct loss, accuracy, precision, recall and F-score measures.

### Agent and Verb

<b>Matrix</b>	<b>Loss*</b>	<b>Accuracy *</b>	<b>Precision*</b>	<b>Recall*</b>	<b>F-score</b>
GG	2,31	98,33	98,26	98,15	0,98
WA	2,31	98,33	98,26	98,13	0,98

\* (%)

Table 42 LookAT loss, accuracy, precision, recall and F-score measures.

<b>Matrix</b>	<b>Loss*</b>	<b>Accuracy *</b>	<b>Precision*</b>	<b>Recall*</b>	<b>F-score</b>
GG	2,31	98,33	98,29	98,10	0,98
WA	2,31	98,33	98,29	98,10	0,98

\* (%)

Table 43 WhoAct loss, accuracy, precision, recall and F-score measures.

### Perceptually Underspecified Noun

<b>Matrix</b>	<b>Loss*</b>	<b>Accuracy *</b>	<b>Precision*</b>	<b>Recall*</b>	<b>F-score</b>
GG	0	100	97,13	99,71	1
WA	0	100	98,94	99,92	1

\* (%)

Table 44 LookAT loss, accuracy, precision, recall and F-score measures.

<b>Matrix</b>	<b>Loss*</b>	<b>Accuracy *</b>	<b>Precision*</b>	<b>Recall*</b>	<b>F-score</b>
GG	1,03	99,34	69,40	58,96	0,65
WA	1,03	99,34	71,06	59,14	0,65

\* (%)

Table 45 WhoAct loss, accuracy, precision, recall and F-score measures.



## Analyses

```
import csv
import pandas as pd
import matplotlib.pyplot as plt
corpus = []
Load the corpus
with open('LookAT', 'r') as csvfile: # WhoAct
 csvfile = csv.reader(csvfile, delimiter=' ')
 for s in csvfile:
 " ".join(s)
 row = [str(w) for w in s]
 corpus.append(row)
len_corpus = len(corpus)
Visualize train and test data statistics
df = pd.DataFrame(corpus, columns=['Agent', 'Verb', 'Patient', 'Referent'])
print(df)
print("There are {} observations and {} features in this dataset. \n".
 format(df.shape[0],df.shape[1]))
print("There are {} types of agent in this dataset such as {}... \n".
 format(len(df.Agent.unique()),
 ", ".join(df.Agent.unique()[0:5])))
print("There are {} types of verb in this dataset such as {}... \n".
 format(len(df.Verb.unique()),
 ", ".join(df.Verb.unique()[0:5])))
print("There are {} types of patient in this dataset such as {}... \n".
 format(len(df.Patient.unique()),
 ", ".join(df.Patient.unique()[0:5])))
print("There are {} types of referent in this dataset such as {}... \n".
 format(len(df.Referent.unique()),
 ", ".join(df.Referent.unique()[0:5])))
df[['Agent', 'Verb', 'Patient', 'Referent']].head()
agents = df.groupby('Agent')
```

```

agents.describe().head()
verbs = df.groupby('Verb')
verbs.describe().head()
patients = df.groupby('Patient')
patients.describe().head()
referents = df.groupby('Referent')
referents.describe().head()
plt.figure(figsize=(15,10))
agents.size().sort_values(ascending=False).plot.bar()
plt.xticks(rotation=50)
plt.xlabel("Agents")
plt.ylabel("Referents")
plt.show()
plt.figure(figsize=(15,10))
patients.size().sort_values(ascending=False).plot.bar()
plt.xticks(rotation=50)
plt.xlabel("Patients")
plt.ylabel("Referents")
plt.show()
plt.figure(figsize=(15,10))
patients.size().sort_values(ascending=False).plot.bar()
plt.xticks(rotation=50)
plt.xlabel("Patients")
plt.ylabel("Referents")
plt.show()
plt.figure(figsize=(15,10))
verbs.size().sort_values(ascending=False).plot.bar()
plt.xticks(rotation=50)
plt.xlabel("Verbs")
plt.ylabel("Patients")
plt.show()
plt.figure(figsize=(15,10))

```

```

verbs.size().sort_values(ascending=False).plot.bar()
plt.xticks(rotation=50)
plt.xlabel("Verbs")
plt.ylabel("Agents")
plt.show()
plt.figure(figsize=(15,10))
verbs.size().sort_values(ascending=False).plot.bar()
plt.xticks(rotation=50)
plt.xlabel("Verbs")
plt.ylabel("Referents")
plt.show()

```

Script 2 Statistical analyses sequences.

## Visual Representations

```

imds = imageDatastore('targets_classes', ... # agent_classes or agentRelated_classes
 'IncludeSubfolders',true, ...
 'LabelSource','foldernames');
[imdsTrain,imdsTest] = splitEachLabel(imds,0.7,'randomized');
numImagesTrain = numel(imdsTrain.Labels);
idx = randperm(numImagesTrain,16);
for i = 1:16
 I{i} = readimage(imdsTrain,idx(i));
end
figure
imshow(imtile(I))
net = alexnet; # googlenet
net.Layers
inputSize = net.Layers(1).InputSize
augimdsTrain = augmentedImageDatastore(inputSize(1:2),imdsTrain);
For agent related pictures:
augimdsTrain=augmentedImageDatastore(inputSize(1:3),imdsTrain,'ColorPreprocessin
g','gray2rgb')

```

```

augimdsTest = augmentedImageDatastore(inputSize(1:2),imdsTest);
For agent related pictures:
augimdsTest = augmentedImageDatastore(inputSize(1:2),imdsTest)
layer = 'fc7'; # pool5-7x7_s1
featuresTrain = activations(net,augimdsTrain,layer,'OutputAs','rows');
featuresTest = activations(net,augimdsTest,layer,'OutputAs','rows');
YTrain = imdsTrain.Labels;
YTest = imdsTest.Labels;
mdl = fitcecoc(featuresTrain,YTrain);
YPred = predict(mdl,featuresTest);
idx = [1 2 3 4]; # Visualize predictions. Until the end.
figure
for i = 1:numel(idx)
 subplot(2,2,i)
 I = readimage(imdsTest,idx(i));
 label = YPred(idx(i));
 imshow(I)
 title(char(label))
end
accuracy = mean(YPred == YTest)

```

Script 3 Transfer Learning.

## MEK

```
import csv
from keras.utils import plot_model
from keras_preprocessing.text import Tokenizer
import numpy as np
from matplotlib import pyplot
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelBinarizer
from sklearn.model_selection import KFold
from numpy import zeros, mean
from keras.initializers import Constant
from keras.layers import Embedding, Input, Bidirectional, LSTM, Dense, Dropout
from keras.models import Model
import random
import tensorflow as tf
from sklearn.metrics import f1_score
from sklearn.metrics import confusion_matrix, classification_report
import seaborn as sns

Define functions

def compare(a, b):
 if (a == b).all():
 return 1
 else:
 return 0

Define variables

t = Tokenizer()
encoder = LabelBinarizer()
embeddings_index = dict()
```

```

corpus = []
sequences = []
labels = []
X = []
repeats = random.sample(range(0, 30), 5)
scores_MEK = []
kf = KFold(n_splits=2)

Load vocabulary embeddings

f = open('GloVeGoogLeNet') # Word2vecAlexNet
for row in f:
 values = row.split()
 words = values[0]
 coefs = np.asarray(values[1:], dtype='float32')
 embeddings_index[words] = coefs
f.close()

Load the corpus

with open('LookAT', 'r') as csvfile: # WhoAct
 csvfile = csv.reader(csvfile, delimiter=' ')
 for s in csvfile:
 " ".join(s)
 row = [str(w) for w in s]
 corpus.append(row)

len_corpus = len(corpus)

Define sequences and referents

for line in corpus:

```

```

seq = line[:3]
lab = line[-1]
sequences.append(seq)
labels.append(lab)

Tokenize the corpus

t.fit_on_texts(corpus)
word_to_id = t.word_index
vocab_size = len(t.word_index) + 1
token_sequences = t.texts_to_sequences(corpus) # sequences

X = np.asarray(token_sequences)
print(X.shape)

Binary referents

Y = encoder.fit_transform(labels)
print(Y.shape)
Out = encoder.inverse_transform(Y)

Prepare embedding layer

embedding_matrix = zeros((vocab_size, 300))
for word, i in t.word_index.items():
 embedding_vector = embeddings_index.get(word)
 if embedding_vector is not None:
 embedding_matrix[i] = embedding_vector
data_dim = len(embedding_matrix[0])

transfer_learning_1 = Embedding(vocab_size,
 data_dim,

```

```

 embeddings_initializer=Constant(embedding_matrix),
 input_length=4, # 3
 trainable=True)

MEK

input_MEK = Input(shape=(4,), dtype='int32') # 3
embedding_MEK = transfer_learning_1(input_MEK)
activation_MEK = Bidirectional(LSTM(vocab_size, recurrent_dropout=0.2),
 merge_mode='ave')(embedding_MEK)
classification_MEK = Dense(Y.shape[1], activation='softmax')(activation_MEK)
MEK = Model(input_MEK, classification_MEK)

MEK.compile(loss='categorical_crossentropy',
 optimizer='adam',
 metrics=['accuracy',
 tf.keras.metrics.Precision(),
 tf.keras.metrics.Recall(),
 tf.keras.metrics.TruePositives(),
 tf.keras.metrics.TrueNegatives(),
 tf.keras.metrics.FalsePositives(),
 tf.keras.metrics.FalseNegatives()])

print(MEK.summary())

Check Overfitting

X_train_eval, X_test_eval = X[:2160], X[2160:] # 4321
Y_train_eval, Y_test_eval = Y[:2160], Y[2160:] # 4321
history_MEK_eval = MEK.fit(X_train_eval, Y_train_eval,
 batch_size=240, # 480
 validation_data=(X_test_eval, Y_test_eval),

```



```
epochs=500)
```

```
predictions_MEK_eval = MEK.predict(X_test_eval)
```

```
train_MEK = MEK.evaluate(X_train_eval, Y_train_eval, verbose=0)
```

```
print("MEK Loss Train: %.2f%%" % (train_MEK[0] * 100))
```

```
print("MEK Accuracy Train: %.2f%%" % (train_MEK[1] * 100))
```

```
test_MEK = MEK.evaluate(X_test_eval, Y_test_eval, verbose=0)
```

```
print("MEK Loss: %.2f%%" % (test_MEK[0] * 100))
```

```
print("MEK Accuracy: %.2f%%" % (test_MEK[1] * 100))
```

```
print("MEK Precision: %.2f%%" % (test_MEK[2] * 100))
```

```
print("MEK Recall: %.2f%%" % (test_MEK[3] * 100))
```

```
print("MEK TruePositives:", test_MEK[4])
```

```
print("MEK TrueNegatives:", test_MEK[5])
```

```
print("MEK FalsePositives:", test_MEK[6])
```

```
print("MEK FalseNegatives:", test_MEK[7])
```

```
f1_MEK = f1_score(Y_test_eval.argmax(axis=1),
```

```
predictions_MEK_eval.argmax(axis=1),
```

```
average='micro')
```

```
print('F1 score: %f' % f1_MEK)
```

```
pyplot.plot(history_MEK_eval.history['loss'])
```

```
pyplot.plot(history_MEK_eval.history['val_loss'])
```

```
pyplot.title('model train vs validation loss')
```

```
pyplot.ylabel('loss')
```

```
pyplot.xlabel('epoch')
```

```
pyplot.legend(['train', 'validation'], loc='upper right')
```

```
pyplot.show()
```

```
pyplot.plot(history_MEK_eval.history['accuracy'])
```

```
pyplot.plot(history_MEK_eval.history['val_accuracy'])
```

```
pyplot.title('model train vs validation accuracy')
```

```

pyplot.ylabel('accuracy')
pyplot.xlabel('epoch')
pyplot.legend(['train', 'validation'], loc='upper right')
pyplot.show()

```

```

pyplot.plot(history_MEK_eval.history['recall'])
pyplot.plot(history_MEK_eval.history['val_recall'])
pyplot.title('model train vs validation recall')
pyplot.ylabel('recall')
pyplot.xlabel('epoch')
pyplot.legend(['train', 'validation'], loc='upper right')
pyplot.show()

```

# Cross-validation

```

kf.get_n_splits(X)
for i in range(len(repeats)):
 run_scores_MEK = list()
 for train_index, test_index in kf.split(X):
 X_train, X_test = X[train_index], X[test_index]
 Y_train, Y_test = Y[train_index], Y[test_index]
 history_sequence = MEK.fit(X_train, Y_train,
 batch_size=240, # 480
 validation_data=(X_test, Y_test),
 epochs=500)
 predictions_MEK = MEK.predict(X_test)
 skill_MEK = compare(Y_test, predictions_MEK)
 run_scores_MEK.append(skill_MEK)
 scores_MEK.append(mean(run_scores_MEK))

```

# Print results

```

for s in scores_MEK:
 train_MEK = MEK.evaluate(X_train, Y_train, verbose=0)
 print("MEK Loss Train: %.2f%%" % (train_MEK[0] * 100))
 test_MEK = MEK.evaluate(X_test, Y_test, verbose=0)
 print("MEK Accuracy Test: %.2f%%" % (test_MEK[1] * 100))
 print("MEK Loss: %.2f%%" % (test_MEK[0] * 100))
 print("MEK Accuracy: %.2f%%" % (test_MEK[1] * 100))
 print("MEK Precision: %.2f%%" % (test_MEK[2] * 100))
 print("MEK Recall: %.2f%%" % (test_MEK[3] * 100))
 print("MEK TruePositives:", test_MEK[4])
 print("MEK TrueNegatives:", test_MEK[5])
 print("MEK FalsePositives:", test_MEK[6])
 print("MEK FalseNegatives:", test_MEK[7])
 f1_MEK = f1_score(Y_test.argmax(axis=1), predictions_MEK.argmax(axis=1),
average='micro')
 print('F1 score: %f' % f1_MEK)

MEK_confusion_matrix = confusion_matrix(Y_test.argmax(axis=1),
 predictions_MEK.argmax(axis=1))
print(MEK_confusion_matrix)
print(classification_report(Y_test.argmax(axis=1), predictions_MEK.argmax(axis=1)))

figure = plt.figure(figsize=(8, 8))
sns.heatmap(MEK_confusion_matrix, annot=False, cmap=plt.cm.Blues)
plt.tight_layout()
plt.ylabel('Actual classes')
plt.xlabel('Predicted classes')
plt.show()

Extract weights

word_embeddings_MEK_0 = MEK.layers[2].get_weights()[0]

```

```

print(word_embeddings_MEK_0.shape)
word_embeddings_MEK_1 = MEK.layers[2].get_weights()[1]
print(word_embeddings_MEK_1.shape)
word_embeddings_MEK_2 = MEK.layers[2].get_weights()[2]
print(word_embeddings_MEK_2.shape)
word_embeddings_MEK_3 = MEK.layers[2].get_weights()[3]
print(word_embeddings_MEK_3.shape)

embeddings_names = {w: word_embeddings_MEK_3[idx] for w, idx in
word_to_id.items()}
embedding_matrix_MEK = zeros((vocab_size, word_embeddings_MEK_3.shape[1]))
for word, i in t.word_index.items():
 embedding_vector_MEK = embeddings_names.get(word)
 if embedding_vector_MEK is not None:
 embedding_matrix_MEK[i] = embedding_vector_MEK
data_dim_MEK = len(embedding_matrix_MEK[0])
print(embedding_matrix_MEK)
print(embedding_matrix_MEK.shape)

```

Script 4 Training.

### Event

```

[...]*

Define sequence input

XS = []
for x in X:
 s = x[:3]
 XS.append(s)

XS = np.asarray(XS)
print(XS.shape)

```

```

transfer_learning_sequence = Embedding(vocab_size,
 data_dim_MEK,
 embeddings_initializer=Constant(embedding_matrix_MEK),
 input_length=3,
 trainable=False)

Model_sequence

input_sequence = Input(shape=(3,))
embedding_sequence = transfer_learning_sequence(input_sequence)
activation_sequence = Bidirectional(LSTM(vocab_size, recurrent_dropout=0.2),
 merge_mode='ave')(embedding_sequence)
classification_sequence = Dense(Y.shape[1], activation='softmax')(activation_sequence)

model_sequence = Model(input_sequence, classification_sequence)

model_sequence.compile(loss='categorical_crossentropy',
 optimizer='adam',
 metrics=['accuracy',
 tf.keras.metrics.Precision(),
 tf.keras.metrics.Recall(),
 tf.keras.metrics.TruePositives(),
 tf.keras.metrics.TrueNegatives(),
 tf.keras.metrics.FalsePositives(),
 tf.keras.metrics.FalseNegatives()])

print(model_sequence.summary())

Check Overfitting

XS_train_eval, XS_test_eval = XS[:2160], XS[2160:] # 4321

```

```

Y_train_eval, Y_test_eval = Y[:2160], Y[2160:] # 4321
history_sequence_eval = model_sequence.fit(XS_train_eval, Y_train_eval,
 batch_size=240, # 480
 validation_data=(XS_test_eval, Y_test_eval),
 epochs=300)

predictions_sequence_eval = model_sequence.predict(XS_test_eval)

train_sequence = model_sequence.evaluate(XS_train_eval, Y_train_eval, verbose=0)
print("sequence Loss Train: %.2f%%" % (train_sequence[0] * 100))
print("sequence Accuracy Train: %.2f%%" % (train_sequence[1] * 100))
test_sequence = model_sequence.evaluate(XS_test_eval, Y_test_eval, verbose=0)
print("sequence Loss: %.2f%%" % (test_sequence[0] * 100))
print("sequence Accuracy: %.2f%%" % (test_sequence[1] * 100))
print("sequence Precision: %.2f%%" % (test_sequence[2] * 100))
print("sequence Recall: %.2f%%" % (test_sequence[3] * 100))
print("sequence TruePositives:", test_sequence[4])
print("sequence TrueNegatives:", test_sequence[5])
print("sequence FalsePositives:", test_sequence[6])
print("sequence FalseNegatives:", test_sequence[7])
f1_sequence = f1_score(Y_test_eval.argmax(axis=1),
 predictions_sequence_eval.argmax(axis=1), average='micro')
print('F1 score: %f' % f1_sequence)

pyplot.plot(history_sequence_eval.history['loss'])
pyplot.plot(history_sequence_eval.history['val_loss'])
pyplot.title('model train vs validation loss')
pyplot.ylabel('loss')
pyplot.xlabel('epoch')
pyplot.legend(['train', 'validation'], loc='upper right')
pyplot.show()

```

```

pyplot.plot(history_sequence_eval.history['accuracy'])
pyplot.plot(history_sequence_eval.history['val_accuracy'])
pyplot.title('model train vs validation accuracy')
pyplot.ylabel('accuracy')
pyplot.xlabel('epoch')
pyplot.legend(['train', 'validation'], loc='upper right')
pyplot.show()

```

```
Cross-validation sequence
```

```

scores_sequence = []
kf.get_n_splits(XS)
for i in range(len(repeats)):
 run_scores_sequence = list()
 for train_index, test_index in kf.split(XS):
 XS_train, XS_test = XS[train_index], XS[test_index]
 Y_train, Y_test = Y[train_index], Y[test_index]
 history_sequence = model_sequence.fit(XS_train, Y_train,
 batch_size=240, # 480
 validation_data=(XS_test, Y_test),
 epochs=300)
 predictions_sequence = model_sequence.predict(XS_test)
 skill_sequence = compare(Y_test, predictions_sequence)
 run_scores_sequence.append(skill_sequence)
 scores_sequence.append(mean(run_scores_sequence))

```

```
Print results
```

```

for s in scores_sequence:
 train_sequence = model_sequence.evaluate(XS_train, Y_train, verbose=0)
 print("Sequence Loss Train: %.2f%%" % (train_sequence[0] * 100))
 print("Sequence Accuracy Train: %.2f%%" % (train_sequence[1] * 100))

```

```

test_sequence = model_sequence.evaluate(XS_test, Y_test, verbose=0)
print("Sequence Loss: %.2f%%" % (test_sequence[0] * 100))
print("Sequence Accuracy: %.2f%%" % (test_sequence[1] * 100))
print("Sequence Precision: %.2f%%" % (test_sequence[2] * 100))
print("Sequence Recall: %.2f%%" % (test_sequence[3] * 100))
print("Sequence TruePositives:", test_sequence[4])
print("Sequence TrueNegatives:", test_sequence[5])
print("Sequence FalsePositives:", test_sequence[6])
print("Sequence FalseNegatives:", test_sequence[7])
f1_sequence = f1_score(Y_test.argmax(axis=1),
predictions_sequence.argmax(axis=1),
 average='micro')
print('F1 score: %f' % f1_sequence)

sequence_confusion_matrix = confusion_matrix(Y_test.argmax(axis=1),
 predictions_sequence.argmax(axis=1))
print(sequence_confusion_matrix)

figure = plt.figure(figsize=(8, 8))
sns.heatmap(sequence_confusion_matrix, annot=False, cmap=plt.cm.Blues)
plt.tight_layout()
plt.ylabel('Actual referents')
plt.xlabel('Predicted referents')
plt.show()

print(classification_report(Y_test.argmax(axis=1),
predictions_sequence.argmax(axis=1)))

```

Script 5 Event input.

**Agent**

[ ... ]\*



```

YA = Y.reshape(len_corpus, 1, 96)
print(YA.shape)

Define the input agent

XA = []
for x in X:
 a = x[0]
 XA.append(a)

XA = np.asarray(XA)
print(XA.shape)

transfer_learning_agent = Embedding(vocab_size,
 data_dim_MEK,
 embeddings_initializer=Constant(embedding_matrix_MEK),
 input_length=1,
 trainable=False)

Model_agent

input_agent = Input(shape=(1,))
embedding_agent = transfer_learning_agent(input_agent)
dropout = Dropout(0.2)(embedding_agent)
classification_agent = Dense(Y.shape[1], activation='sigmoid')(dropout)

model_agent = Model(input_agent, classification_agent)

model_agent.compile(loss='binary_crossentropy',
 optimizer='adam',
 metrics=['accuracy',
 tf.keras.metrics.Precision()],

```

```

 tf.keras.metrics.Recall(),
 tf.keras.metrics.TruePositives(),
 tf.keras.metrics.TrueNegatives(),
 tf.keras.metrics.FalsePositives(),
 tf.keras.metrics.FalseNegatives())

print(model_agent.summary())

Check Overfitting

XA_train_eval, XA_test_eval = XA[:2160], XA[2160:] # 4321
YA_train_eval, YA_test_eval = YA[:2160], YA[2160:] # 4321
history_agent_eval = model_agent.fit(XA_train_eval, YA_train_eval,
 batch_size=240, # 480
 validation_data=(XA_test_eval, YA_test_eval),
 epochs=1000)

predictions_agent_eval = model_agent.predict(XA_test_eval)

train_agent = model_agent.evaluate(XA_train_eval, YA_train_eval, verbose=0)
print("agent Loss Train: %.2f%%" % (train_agent[0] * 100))
print("agent Accuracy Train: %.2f%%" % (train_agent[1] * 100))
test_agent = model_agent.evaluate(XA_test_eval, YA_test_eval, verbose=0)
print("agent Loss: %.2f%%" % (test_agent[0] * 100))
print("agent Accuracy: %.2f%%" % (test_agent[1] * 100))
print("agent Precision: %.2f%%" % (test_agent[2] * 100))
print("agent Recall: %.2f%%" % (test_agent[3] * 100))
print("agent TruePositives:", test_agent[4])
print("agent TrueNegatives:", test_agent[5])
print("agent FalsePositives:", test_agent[6])
print("agent FalseNegatives:", test_agent[7])

```

```
f1_agent = f1_score(YA_test_eval.argmax(axis=2),
predictions_agent_eval.argmax(axis=2),
 average='micro')
print('F1 score: %f' % f1_agent)
```

```
pyplot.plot(history_agent_eval.history['loss'])
pyplot.plot(history_agent_eval.history['val_loss'])
pyplot.title('model train vs validation loss')
pyplot.ylabel('loss')
pyplot.xlabel('epoch')
pyplot.legend(['train', 'validation'], loc='upper right')
pyplot.show()
```

```
pyplot.plot(history_agent_eval.history['accuracy'])
pyplot.plot(history_agent_eval.history['val_accuracy'])
pyplot.title('model train vs validation accuracy')
pyplot.ylabel('accuracy')
pyplot.xlabel('epoch')
pyplot.legend(['train', 'validation'], loc='upper right')
pyplot.show()
```

# Cross-validation agent

```
scores_agent = []
kf.get_n_splits(XA)
for i in range(len(repeats)):
 run_scores_agent = list()
 for train_index, test_index in kf.split(XA):
 XA_train, XA_test = XA[train_index], XA[test_index]
 YA_train, YA_test = YA[train_index], YA[test_index]
 history_agent = model_agent.fit(XA_train, YA_train,
 batch_size=240, # 480
```

```

 validation_data=(XA_test, YA_test),
 epochs=1000)

 predictions_agent = model_agent.predict(XA_test)
 skill_agent = compare(YA_test, predictions_agent)
 run_scores_agent.append(skill_agent)
 scores_agent.append(mean(run_scores_agent))

Print results

for s in scores_agent:
 train_agent = model_agent.evaluate(XA_train, YA_train, verbose=0)
 print("Agent Loss Train: %.2f%%" % (train_agent[0] * 100))
 print("Agent Accuracy Train: %.2f%%" % (train_agent[1] * 100))
 test_agent = model_agent.evaluate(XA_test, YA_test, verbose=0)
 print("Agent Loss: %.2f%%" % (test_agent[0] * 100))
 print("Agent Accuracy: %.2f%%" % (test_agent[1] * 100))
 print("Agent Precision: %.2f%%" % (test_agent[2] * 100))
 print("Agent Recall: %.2f%%" % (test_agent[3] * 100))
 print("Agent TruePositives:", test_agent[4])
 print("Agent TrueNegatives:", test_agent[5])
 print("Agent FalsePositives:", test_agent[6])
 print("Agent FalseNegatives:", test_agent[7])
 f1_agent = f1_score(YA_test.argmax(axis=2), predictions_agent.argmax(axis=2),
 average='micro')
 print('F1 score: %f' % f1_agent)

agent_confusion_matrix = confusion_matrix(YA_test.argmax(axis=2),
 predictions_agent.argmax(axis=2))

print(agent_confusion_matrix)

figure = plt.figure(figsize=(8, 8))
sns.heatmap(agent_confusion_matrix, annot=False, cmap=plt.cm.Blues)

```

```

plt.tight_layout()
plt.ylabel('Actual referents')
plt.xlabel('Predicted referents')
plt.show()

print(classification_report(YA_test.argmax(axis=2),
predictions_agent.argmax(axis=2)))

```

Script 6 Agent input.

### Agent-Verb

```

[...]*

transfer_learning_agent_verb = Embedding(vocab_size,
 data_dim_MEK,
 embeddings_initializer=Constant(embedding_matrix_MEK),
 input_length=2,
 trainable=False)

Define the input agent-verb

XAV = []
for x in X:
 av = x[:2]
 XAV.append(av)

XAV = np.asarray(XAV)
print(XAV.shape)

Model_agent_verb

input_agent_verb = Input(shape=(2,))
embedding_agent_verb = transfer_learning_agent_verb(input_agent_verb)

```

```

activation_agent_verb = Bidirectional(LSTM(vocab_size,
 recurrent_dropout=0.2))(embedding_agent_verb)
classification_agent_verb = Dense(Y.shape[1],
activation='softmax')(activation_agent_verb)

model_agent_verb = Model(input_agent_verb, classification_agent_verb)

model_agent_verb.compile(loss='categorical_crossentropy',
 optimizer='adam',
 metrics=['accuracy',
tf.keras.metrics.Precision(),
tf.keras.metrics.Recall(),
tf.keras.metrics.TruePositives(),
tf.keras.metrics.TrueNegatives(),
tf.keras.metrics.FalsePositives(),
tf.keras.metrics.FalseNegatives()])

print(model_agent_verb.summary())

Check Overfitting

XAV_train_eval, XAV_test_eval = XAV[:2160], XAV[2160:] # 4321
Y_train_eval, Y_test_eval = Y[:2160], Y[2160:] # 4321
history_agent_verb_eval = model_agent_verb.fit(XAV_train_eval, Y_train_eval,
 batch_size=240, # 480
 validation_data=(XAV_test_eval, Y_test_eval),
 epochs=500)

predictions_agent_verb_eval = model_agent_verb.predict(XAV_test_eval)

train_agent_verb = model_agent_verb.evaluate(XAV_train_eval, Y_train_eval,
 verbose=0)

```

```

print("agent_verb Loss Train: %.2f%%" % (train_agent_verb[0] * 100))
print("agent_verb Accuracy Train: %.2f%%" % (train_agent_verb[1] * 100))
test_agent_verb = model_agent_verb.evaluate(XAV_test_eval, Y_test_eval, verbose=0)
print("agent_verb Loss: %.2f%%" % (test_agent_verb[0] * 100))
print("agent_verb Accuracy: %.2f%%" % (test_agent_verb[1] * 100))
print("agent_verb Precision: %.2f%%" % (test_agent_verb[2] * 100))
print("agent_verb Recall: %.2f%%" % (test_agent_verb[3] * 100))
print("agent_verb TruePositives:", test_agent_verb[4])
print("agent_verb TrueNegatives:", test_agent_verb[5])
print("agent_verb FalsePositives:", test_agent_verb[6])
print("agent_verb FalseNegatives:", test_agent_verb[7])
f1_agent_verb = f1_score(Y_test_eval.argmax(axis=1),
 predictions_agent_verb_eval.argmax(axis=1),
 average='micro')
print('F1 score: %f' % f1_agent_verb)

pyplot.plot(history_agent_verb_eval.history['loss'])
pyplot.plot(history_agent_verb_eval.history['val_loss'])
pyplot.title('model train vs validation loss')
pyplot.ylabel('loss')
pyplot.xlabel('epoch')
pyplot.legend(['train', 'validation'], loc='upper right')
pyplot.show()

pyplot.plot(history_agent_verb_eval.history['accuracy'])
pyplot.plot(history_agent_verb_eval.history['val_accuracy'])
pyplot.title('model train vs validation accuracy')
pyplot.ylabel('accuracy')
pyplot.xlabel('epoch')
pyplot.legend(['train', 'validation'], loc='upper right')
pyplot.show()

```

```

Cross-validation agent-verb

scores_agent_verb = []
kf.get_n_splits(XAV)
for i in range(len(repeats)):
 run_scores_agent_verb = list()
 for train_index, test_index in kf.split(XAV):
 XAV_train, XAV_test = XAV[train_index], XAV[test_index]
 Y_train, Y_test = Y[train_index], Y[test_index]
 history_agent_verb = model_agent_verb.fit(XAV_train, Y_train,
 batch_size=240, # 480
 validation_data=(XAV_test, Y_test),
 epochs=500)
 predictions_agent_verb = model_agent_verb.predict(XAV_test)
 skill_agent_verb = compare(Y_test, predictions_agent_verb)
 run_scores_agent_verb.append(skill_agent_verb)
 scores_agent_verb.append(mean(run_scores_agent_verb))

Print results

for s in scores_agent_verb:
 train_agent_verb = model_agent_verb.evaluate(XAV_train, Y_train, verbose=0)
 print("Agent_verb Loss Train: %.2f%%" % (train_agent_verb[0] * 100))
 print("Agent_verb Accuracy Train: %.2f%%" % (train_agent_verb[1] * 100))
 test_agent_verb = model_agent_verb.evaluate(XAV_test, Y_test, verbose=0)
 print("Agent-verb Loss: %.2f%%" % (test_agent_verb[0] * 100))
 print("Agent-verb Accuracy: %.2f%%" % (test_agent_verb[1] * 100))
 print("Agent-verb Precision: %.2f%%" % (test_agent_verb[2] * 100))
 print("Agent-verb Recall: %.2f%%" % (test_agent_verb[3] * 100))
 print("Agent-verb TruePositives:", test_agent_verb[4])
 print("Agent-verb TrueNegatives:", test_agent_verb[5])
 print("Agent-verb FalsePositives:", test_agent_verb[6])

```



```

print("Agent-verb FalseNegatives:", test_agent_verb[7])
f1_agent_verb = f1_score(Y_test.argmax(axis=1),
predictions_agent_verb.argmax(axis=1), average='micro')
print('F1 score: %f % f1_agent_verb)

agent_verb_confusion_matrix = confusion_matrix(Y_test.argmax(axis=1),
predictions_agent_verb.argmax(axis=1))
print(agent_verb_confusion_matrix)

figure = plt.figure(figsize=(8, 8))
sns.heatmap(agent_verb_confusion_matrix, annot=False, cmap=plt.cm.Blues)
plt.tight_layout()
plt.ylabel('Actual referents')
plt.xlabel('Predicted referents')
plt.show()

print(classification_report(Y_test.argmax(axis=1),
predictions_agent_verb.argmax(axis=1)))

```

Script 7 Agent-verb input.

### **Perceptually Underspecified Noun**

```

[...]*

YP = Y.reshape(len_corpus, 1, 96)
print(YP.shape)

Define the input agent

XP = []
for x in X:
 p = x[-1]
 XP.append(p)

```

```

XP = np.asarray(XP)
print(XP.shape)

transfer_learning_agent = Embedding(vocab_size,
 data_dim_MEK,
 embeddings_initializer=Constant(embedding_matrix_MEK),
 input_length=1,
 trainable=False)

Model_patient

input_patient = Input(shape=(1,))
embedding_patient = transfer_learning_agent(input_patient)
dropout = Dropout(0.2)(embedding_patient)
classification_patient = Dense(Y.shape[1], activation='sigmoid')(dropout)

model_patient = Model(input_patient, classification_patient)

model_patient.compile(loss='binary_crossentropy',
 optimizer='adam',
 metrics=['accuracy',
 tf.keras.metrics.Precision(),
 tf.keras.metrics.Recall(),
 tf.keras.metrics.TruePositives(),
 tf.keras.metrics.TrueNegatives(),
 tf.keras.metrics.FalsePositives(),
 tf.keras.metrics.FalseNegatives()])

print(model_patient.summary())

Check Overfitting

```

```

XP_train_eval, XP_test_eval = XP[:2160], XP[2160:] # 4321
YP_train_eval, YP_test_eval = YP[:2160], YP[2160:] # 4321
history_patient_eval = model_patient.fit(XP_train_eval, YP_train_eval,
 batch_size=240, # 480
 validation_data=(XP_test_eval, YP_test_eval),
 epochs=1000)

predictions_patient_eval = model_patient.predict(XP_test_eval)

train_patient = model_patient.evaluate(XP_train_eval, YP_train_eval, verbose=0)
print("patient Loss Train: %.2f%%" % (train_patient[0] * 100))
print("patient Accuracy Train: %.2f%%" % (train_patient[1] * 100))
test_patient = model_patient.evaluate(XP_test_eval, YP_test_eval, verbose=0)
print("patient Loss: %.2f%%" % (test_patient[0] * 100))
print("patient Accuracy: %.2f%%" % (test_patient[1] * 100))
print("patient Precision: %.2f%%" % (test_patient[2] * 100))
print("patient Recall: %.2f%%" % (test_patient[3] * 100))
print("patient TruePositives:", test_patient[4])
print("patient TrueNegatives:", test_patient[5])
print("patient FalsePositives:", test_patient[6])
print("patient FalseNegatives:", test_patient[7])
f1_patient = f1_score(YP_test_eval.argmax(axis=2),
 predictions_patient_eval.argmax(axis=2),
 average='micro')
print('F1 score: %f' % f1_patient)

pyplot.plot(history_patient_eval.history['loss'])
pyplot.plot(history_patient_eval.history['val_loss'])
pyplot.title('model train vs validation loss')
pyplot.ylabel('loss')
pyplot.xlabel('epoch')

```

```
pyplot.legend(['train', 'validation'], loc='upper right')
pyplot.show()
```

```
pyplot.plot(history_patient_eval.history['accuracy'])
pyplot.plot(history_patient_eval.history['val_accuracy'])
pyplot.title('model train vs validation accuracy')
pyplot.ylabel('accuracy')
pyplot.xlabel('epoch')
pyplot.legend(['train', 'validation'], loc='upper right')
pyplot.show()
```

```
Cross-validation perceptually underspecified patient
```

```
scores_patient = []
kf.get_n_splits(XP)
for i in range(len(repeats)):
 run_scores_patient = list()
 for train_index, test_index in kf.split(XP):
 XP_train, XP_test = XP[train_index], XP[test_index]
 YP_train, YP_test = YP[train_index], YP[test_index]
 history_patient = model_patient.fit(XP_train, YP_train,
 batch_size=240, # 480
 validation_data=(XP_test, YP_test),
 epochs=1000)
 predictions_patient = model_patient.predict(XP_test)
 skill_patient = compare(YP_test, predictions_patient)
 run_scores_patient.append(skill_patient)
 scores_patient.append(mean(run_scores_patient))
```

```
Print results
```

```
for s in scores_patient:
```

```

train_patient = model_patient.evaluate(XP_train, YP_train, verbose=0)
print("patient Loss Train: %.2f%%" % (train_patient[0] * 100))
print("patient Accuracy Train: %.2f%%" % (train_patient[1] * 100))
test_patient = model_patient.evaluate(XP_test, YP_test, verbose=0)
print("patient Loss: %.2f%%" % (test_patient[0] * 100))
print("patient Accuracy: %.2f%%" % (test_patient[1] * 100))
print("patient Precision: %.2f%%" % (test_patient[2] * 100))
print("patient Recall: %.2f%%" % (test_patient[3] * 100))
print("patient TruePositives:", test_patient[4])
print("patient TrueNegatives:", test_patient[5])
print("patient FalsePositives:", test_patient[6])
print("patient FalseNegatives:", test_patient[7])
f1_patient = f1_score(YP_test.argmax(axis=2), predictions_patient.argmax(axis=2),
 average='micro')
print('F1 score: %f' % f1_patient)

patient_confusion_matrix = confusion_matrix(YP_test.argmax(axis=2),
 predictions_patient.argmax(axis=2))

print(patient_confusion_matrix)

figure = plt.figure(figsize=(8, 8))
sns.heatmap(patient_confusion_matrix, annot=False, cmap=plt.cm.Blues)
plt.tight_layout()
plt.ylabel('Actual referents')
plt.xlabel('Predicted referents')
plt.show()

print(classification_report(YP_test.argmax(axis=2),
 predictions_patient.argmax(axis=2)))

```

Script 8 Perceptually underspecified noun input.

## Figure Index

Figure 1 Example of first and second lists trials. ....	60
Figure 2 Example of the norming study task. For each type of agent, we asked to the participants to judge the degree of relatedness between the person and the depicted Target object. The relatedness had been judged based on the probability of the two entities to appear in the same situations. ....	63
Figure 3 The procedure includes the calibration period followed by the preview period of the visual scene lasting 1000ms. After the preview period the auditory sentence was presented while the pictures remained on the visual scene. ....	66
Figure 4 Time course of the AOIs eye fixations proportions. The plot shows the agent, the verb and the patient onsets, and the sentence offset. ....	77
Figure 5 Plant pot. ....	86
Figure 6 Rake. ....	87
Figure 7 Cooking pot. ....	87
Figure 8 LookAT Architecture. ....	91
Figure 9 WhoAct Architecture. ....	93
Figure 10 Football ball and baseball ball. ....	95
Figure 11 Textual Event and Target Picture. ....	102
Figure 12 Multimodal thematic fit. ....	103
Figure 13 Multimodal verb selectional restrictions. ....	104
Figure 14 Relationships between a perceptually underspecified noun (hypernym) and its referents. ....	106
Figure 15 LookAT Confusion Matrix. ....	109
Figure 16 WhoAct Confusion Matrix. ....	110
Figure 17 LookAT Confusion Matrix. ....	111
Figure 18 WhoAct Confusion Matrix. ....	112
Figure 19 LookAT Confusion Matrix. ....	114
Figure 20 WhoAct Confusion Matrix. ....	115
Figure 21 LookAT Confusion Matrix. ....	117
Figure 22 WhoAct Confusion Matrix. ....	118
Figure 23 AOIs eye fixations proportions agent time window. ....	145
Figure 24 AOIs eye fixations proportions anticipatory (action) time window. ....	145

Figure 25 AOIs eye fixations proportions patient time window. ....	146
Figure 26 AOIs eye fixations proportions final silence.....	146
Figure 27 Agent time window.....	147
Figure 28 Between the two lists agent time window.....	148
Figure 29 Between the two lists agent time window. Comparisons between the first and second lists.....	149
Figure 30 Anticipatory (action) time window. ....	150
Figure 31 Between the two lists anticipatory (action) time window.....	151
Figure 32 Between the two lists anticipatory (action) time window. Comparisons between the first and second lists. ....	152
Figure 33 Patient time window.....	153
Figure 34 Between the two lists patient time window.....	154
Figure 35 Between the two lists patient time window. Comparisons between the first and second lists.....	155
Figure 36 Final silence. ....	156
Figure 37 Between the two lists final silence.....	157
Figure 38 Between the two lists final silence. Comparisons between the first and second lists.....	158
Figure 39 Time course AOIs eye fixations proportions for each participant. ....	159

## Table Index

Table 1 Critical time points of the auditory sentence time course. ....	68
Table 2 Analyses in the two lists.....	70
Table 3 Analyses in the first and second lists.....	70
Table 4 Analyses in the two lists.....	71
Table 5 Analyses in the first and second lists.....	72
Table 6 Analyses in the two lists.....	73
Table 7 Analyses in the first and second lists.....	73
Table 8 Analyses in the two lists.....	75
Table 9 Analyses in the first and second lists.....	75

Table 10 Comparisons among eye fixations proportions toward the Target AOI with respect to eye fixations proportions toward the other AOIs in the agent, the action and the patient time windows.....	77
Table 11 Text-Picture relationships in MEK training data. ....	89
Table 12 Mean and standard deviation of the eye fixations proportions for each agent. ....	144
Table 13 Between the two lists agent time window. ....	160
Table 14 Between the two lists anticipatory (action) time window. ....	160
Table 15 Between the two lists patient time window. ....	161
Table 16 Between the two lists final silence. ....	161
Table 17 Agent time window. ....	162
Table 18 Anticipatory (action) time window. ....	163
Table 19 Patient time window. ....	163
Table 20 Final silence. ....	163
Table 21 AlexNet and GoogLeNet CNNs accuracy. ....	206
Table 22 LookAT-GG. ....	208
Table 23 LookAT-WA. ....	211
Table 24 WhoAct-GG. ....	214
Table 25 WhoAct-WA. ....	217
Table 26 LookAT-GG. ....	220
Table 27 LookAT-WA. ....	223
Table 28 WhoAct-GG. ....	226
Table 29 WhoAct-GG. ....	229
Table 30 LookAT-GG. ....	232
Table 31 LookAT-WA. ....	235
Table 32 WhoAct-GG. ....	238
Table 33 WhoAct-WA. ....	241
Table 34 LookAT-GG. ....	244
Table 35 LookAT-WA. ....	247
Table 36 WhoAct-GG. ....	250
Table 37 WhoAct-WA. ....	253
Table 38 LookAT loss, accuracy, precision, recall and F-score measures. ....	254



Table 39 WhoAct loss, accuracy, precision, recall and F-score measures.....	254
Table 40 LookAT loss, accuracy, precision, recall and F-score measures.....	254
Table 41 WhoAct loss, accuracy, precision, recall and F-score measures.....	254
Table 42 LookAT loss, accuracy, precision, recall and F-score measures.....	255
Table 43 WhoAct loss, accuracy, precision, recall and F-score measures.....	255
Table 44 LookAT loss, accuracy, precision, recall and F-score measures.....	255
Table 45 WhoAct loss, accuracy, precision, recall and F-score measures.....	255

## **Script Index**

Script 1 Eye-tracking statistical analyses. ....	182
Script 2 Statistical analyses sequences. ....	258
Script 3 Transfer Learning.....	259
Script 4 Training.....	267
Script 5 Event input.....	271
Script 6 Agent input.....	276
Script 7 Agent-verb input. ....	280
Script 8 Perceptually underspecified noun input.....	284

## Acknowledgements

Many people contributed to the creation of this thesis. Everyone gave me something that is behind the words of this work.

Thanks to Professor Pier Marco Bertinetto, who made me discover the charm of clear and precise language. Thanks to Professor Alessandro Lenci, who checked that I continued on the right path while I entered the wood of the computational tools. Thanks to Irene, Chiara and all people that were part of the Linguistics Laboratory G. Nencioni of Scuola Normale Superiore (SNS) of Pisa for the invaluable support and encouragement during every day spent in the classrooms of the SNS. A special thanks to Professor Ken McRae of the Western Ontario University (OWU) because he provided me invaluable assistance during my visitor student period in London (Canada). I would like to say thank you also to Dr. Tea Knowles and Makayla Hall Bruce for their precious help in carrying on my researches at OWU.

Thanks to Alessandro, my love, my lighthouse, my inestimable treasure, for being there always and everywhere. Thanks to my mum and dad for always being my most loyal fans and the solid support to my every dream. Thanks to Marilena e Ottaviano for never have stopped believing in us, Alessandro e Valentina.

Thanks to Marco, Caterina and all colleagues with whom I lived unforgettable moments in Pisa. Thanks to all my friends because this thesis would not have been possible without their friendship and encouragement.

Speaking of language, there are always many ways to express what do you want to tell. We have only to find the best alternative for those who listen to us.