The Parallel Meaning Bank

Annotation Manual

Version 0.3

Johan Bos Kilian Evang Johannes Bjerva Hessel Haagsma
Valerio Basile Noortje Venhuizen

April 18, 2016

Contents

Chapter 1

Introduction

The Parallel Meaning Bank (PMB) is a semantically annotated parallel corpus for English, Dutch, German and Italian. The PMB is a cross-lingual version of the Groningen Meaning Bank (Basile, Bos, Evang, and Venhuizen 2012a; Basile, Bos, Evang, and Venhuizen 2012b).

The annotations are automatically generated by a state-of-the-art natural language processing pipeline and then manually corrected. This document serves as a manual for linguistically annotating the documents of the PMB.

Chapter 2

Using the PMB Explorer

2.1 Exploring the PMB

The PMB is available online at http://pmb.let.rug.nl. It is organised in terms of document sets. Each document set contains an English text and one of more translations in Dutch, German, and Italian. These documents can be viewed with the **PMB explorer**. On the bottom of the screen one of the documents can be seen in the selected language (default: English). Tabs for other languages are shown just above the text if translations are available in the current document set. Other documents can be seen by browsing through the corpus using the arrows on the top left of the explorer. Several filters can be selected while browsing documents in the GMB.

2.2 Document identification

Every document set has a unique identifier XX/YYYY where XX is a part (ranging from 00–99) and YYYY the document within that part (ranging from 0000–9999, but not all slots are filled). Every document set has a **status** signifying its state of inclusion in the PMB: *accepted*, *postponed*, *uncategorized*, and *rejected*. These categories should be self-explanatory.

2.3 Overall Annotation Strategy

The PMB follows a stand-off annotation strategy. This means that the orignal, raw documents are preserved, and that all annotations are expressed in terms of operations on character offsets of the raw texts (see: **bits of wisdom**). The annotations are divided into various layers: *metadata*, *tokens*, *sentences*, and *discourse*.

2.4 Bits of Wisdom

Every document set has a possible empty set of bits of wisdom. A bit of wisdom (BOW) represents an annotation decision by a human, crowd, or machine. BOWs could be conflicting.

Chapter 3

Tokenization

Tokenization is the task of segmenting the text into tokens that are processable for following natural language processing tasks. Two different types of tokens are considered: word tokens and sentence tokens.

3.1 The token annotation scheme

In order to change the boundaries of words and sentences click on the **tokens** tab in the PMB explorer. You will see the current segmentation of the document: sentence tokens are separated by new lines, and word tokens by single spaces. Once you click on **[Edit]** it is possible to adjust the segmentation of both words and sentences. This is done by selecting one of four options for a character: start of a sentence; start of a token; inside a token; or outside a token. In other words, we use an IOB tagging type scheme on the character level, combining word with sentence tokenisation and hence killing two birds with one stone (Evang, Basile, Chrupała, and Bos 2013).

3.2 Sentence boundaries

Sentence tokens are determined by full stops, question marks and exclamation marks. But not every sentence ends with a punctuation symbol (they could be forgotten or not present, in case of titles, or a sentence could end with an abbreviation that contains a full stop). In quoted contexts, sentences can also finish with a quote. Some sentences end with an ellipsis symbol (a sequence of three dots). In case of haplographic full stops, the period is not separated from the abbreviation.

3.3 Word boundaries

3.3.1 Abbreviations

Abbreviations are not separated from periods. So `a.s.` is one word token, and so is `Prof.` In case of haplographic full stops, that may appear if a sentence ends with an abbreviation, the period is not separated from the abbreviation.

3.3.2 Multi-word expressions

Some multi-word expressions are treated as one word token by labelling the spaces as I (inside a token). These comprise names of locations such as `New York` and `San Francisco`, organizations such as `FC Groningen` and `National Basketball Association`, and titles such as `Prime Minister` and `Secretary of State`.

3.3.3 Contractions in English and Italian

Contractions of English modal and auxiliary verbs are split into two word tokens: `mustn't` becomes `must n't`, `can't` becomes `ca n't`, and so on. Contractions of Italian articles and prepositions are split into two word tokens: `l'Italia` becomes `l' Italia`, `un'auto` becomes `un' auto`, and `dall'Australia` becomes `dall' Australia`.

3.3.4 Italian Clitics

Can be attached to verbs: -ci, -mi, -vi, -gli, -ti, -le, -lo, -la, -li. For instance, `trovarci` becomes `trovar ci`.

3.3.5 Uniform Resource Locators

These are considered to be one token, including the prefix `http` or `https`. So `http://pmb.let.rug.nl` rather than `http : // pmb.let.rug.nl`

3.3.6 Scores and Units of Measurement

These are considered to be one token, so: `3-0` instead of `3 - 0` and `10 m/h` rather than `10 m / h`. The adverb `o'clock` in clocktime expressions is not split.

3.3.7 Ellipsis and Series of Punctiation Symbols

These are treated as one token: so `!!!` in an exclamation or `???` marking surprise are single tokens. Ellipsis (...) are also segmented as one word token.

3.4 Gold Standard

The first thousand document sets of Part 00 of the PMB have been manually checked and corrected by one annotator (for all available languages).

Chapter 4

Syntactic Analysis

4.1 CCG

We use Combinatory Categorial Grammar, CCG (Steedman 2001), as formalism to describing syntactic structure. The basic categories that we use are: N, S, NP, and PP. In addition we use CONJ for coordination.

4.2 Correcting

Lexical categories can be corrected via BOWs. Currently there is no interface for directly editing the tree structure. Lexical categories cannot disambiguate between all attachment decisions. E.g. although the CCG category of a preposition determines whether it attaches to an NP or to a VP, if, for example, more than one NP is available to attach to, then there is currently no way to disambiguate between the two.

Chapter 5

Thematic Roles

5.1 Role Inventory

The thematic roles used in the PMB are based on the VERBNET (Kipper, Korhonen, Ryant, and Palmer 2008) and LIRICS (Bonial, Corvey, Palmer, Petukhova, and Bunt 2011) inventories and are incorporated in the DRS by adopting a neo-Davidsonian semantics. They can be edited in the PMB explorer on the word level (Bos, Evang, and Nissim 2012): a word (verb/noun) are associated with a finite list of roles, depending on their number of arguments. The roles are organized in a hierarchy, with Participant subsuming all other roles (5.1).

5.2 Changes

In the PMB there are no roles for Co-Agent, Co-Theme and Co-Patients. In events where this is needed there are multiple participants with the same role. These and other changes in role names (wrt VERBNET) are shown in Table 5.2.

Table 5.1: Thematic Roles used in the Parallel Meaning Bank.

Role	**Super Role**	**Description**
Actor	Participant	participant that is an instigator
Undergoer	Participant	participant that is not an instigator
Time	Participant	participant that indicates an instance or interval of time
Place	Participant	participant that represents the place in which an entity exists
Abstract	Participant	participant that represents a goal, property or manner
Agent	Actor	intentional actor
Cause	Actor	unintentional actor (animate or inanimate)
Patient	Undergoer	undergoer that experiences a change of state, location or condition; exists independently of the event
Instrument	Undergoer	undergoer that is manipulated by an agent; exists independently of the event
Beneficiary	Undergoer	undergoer that is potentially advantaged or disadvantaged by the event
Theme	Undergoer	undergoer that is central to an event; not structurally changed by the event
Topic	Theme	theme with propositional information content
Pivot	Theme	theme that is more central than another theme in an event
Start	Time	time that indicates when an events begins or a state becomes true
Finish	Time	time that indicates when an events ends or a state becomes false
Duration	Time	length or extend of time
Frequency	Time	number of occurences of an event within a given time span
Location	Place	concrete place
Source	Place	concrete or abstract place that is starting point of action
Destination	Place	place that is an end point of an action
Path	Place	
Value	Place	place along a formal scale
Goal	Abstract	purpose of an (intentional) action
Manner	Abstract	the particular way an event unfolds
Attribute	Abstract	property of an entity
Extent	Value	value indicating the amount of measurable change to a participant
Asset	Value	value that is a concrete object
Recipient	Destination	animate destination
Experiencer	Patient	patient that is aware in perception events
Result	Goal	goal that comes into existence through the event
Stimulus	Cause	cause in perception event that elicits emotional or psychological response

Table 5.2: Renamed Thematic Roles in the Parallel Meaning Bank.

Old Role Name	**New Role Name**
Co-Agent	Agent
Co-Theme	Theme
Co-Patient	Patient
Initial-Time	Start
Final-Time	Finish
Initial-Location	Source
Final-Location	Destination
Result	Goal
Proposition	Topic
Trajectory	Path
Product	Result
Material	Source
Predicate	Attribute
Reflexive	Theme

Bibliography

Basile, V., J. Bos, K. Evang, and N. Venhuizen (2012a). Developing a large semantically annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, pp. 3196–3200.

Basile, V., J. Bos, K. Evang, and N. Venhuizen (2012b). A platform for collaborative semantic annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Avignon, France, pp. 92–96.

Bonial, C., W. J. Corvey, M. Palmer, V. Petukhova, and H. Bunt (2011). A hierarchical unification of LIRICS and verbnet semantic roles. In *Proceedings of the 5th IEEE International Conference on Semantic Computing (ICSC 2011)*, Palo Alto, CA, USA, pp. 483–489.

Bos, J., K. Evang, and M. Nissim (2012). Annotating semantic roles in a lexicalised grammar environment. In *Proceedings of the Eighth Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-8)*, Pisa, Italy, pp. 9–12.

Evang, K., V. Basile, G. Chrupała, and J. Bos (2013). Elephant: Sequence labeling for word and sentence segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1422–1426.

Kipper, K., A. Korhonen, N. Ryant, and M. Palmer (2008). A large-scale classification of English verbs. *Language Resources and Evaluation 42*(1), 21–40.

Steedman, M. (2001). *The Syntactic Process*. The MIT Press.