J Psycholinguist Res
DOI 10.1007/s10936-015-9398-7

Quantifying Semantic Linguistic Maturity in Children

Kristina Hansson[1] · **Rasmus Bååth**[2] · **Simone Löhndorf**[3] · **Birgitta Sahlén**[1] · **Sverker Sikström**[4]

Abstract We propose a method to quantify *semantic linguistic maturity* (SELMA) based on a high dimensional semantic representation of words created from the co-occurrence of words in a large text corpus. The method was applied to oral narratives from 108 children aged 4;0–12;10. By comparing the SELMA measure with maturity ratings made by human raters we found that SELMA predicted the rating of semantic maturity made by human raters over and above the prediction made using a child's age and number of words produced. We conclude that the semantic content of narratives changes in a predictable pattern with children's age and argue that SELMA is a measure quantifying semantic linguistic maturity. The study opens up the possibility of using quantitative measures for studying the development of semantic representation in children's narratives, and emphasizes the importance of word co-occurrences for understanding the development of meaning.

Keywords Semantic representation · Semantic development · Narratives · Child language · Semantic linguistic maturity

Introduction

Semantics is possibly the least understood aspect of language development. Most research on language development and language use focuses on form (e.g., phonology, utterance length, the use of different morpho-syntactic constructions), rather than on semantic content (i.e., meaning). Furthermore, semantics is usually studied in terms of vocabulary as an inventory,

✉ Kristina Hansson
kristina.hansson@med.lu.se

[1] Department of Clinical Sciences, Lund, Logopedics, Phoniatrics and Audiology, Lund University, Lund, Sweden

[2] Lund University Cognitive Science, Lund University, Lund, Sweden

[3] Department of Linguistics, Centre for Languages and Literature, Lund University, Lund, Sweden

[4] Department of Psychology, Lund University, Lund, Sweden

Published online: 06 October 2015

which has more to do with reference (the set of phenomena that a word applies to) than with sense (the concept associated with a word; Clark and Clark 1977). Thus, children's vocabulary has been addressed in numerous studies focussing on the development of the size and organization of vocabulary in experimental tasks (e.g., Dollaghan 1987; Koren et al. 2005; McGregor et al. 2002) as well as in parent reports on their children's vocabulary (Fenson et al. 2007). However, it is essential to explore semantic development beyond the single word level, that is, at the text level. The production of a text requires, for example, the ability to produce words that are specific to the context, and knowledge of how words co-occur (Vermeer 2001). Semantics is at the intersection between cognition and language (Kamhi 1992) and is strongly related to general aspects of cognitive development (Bloom 2001; Clark and Clark 1977). The semantic content we express reflects our knowledge about the world and about the objects, events and relations that it contains (Lahey 1988).

This study is a cross-scientific venture, bringing together methods and ideas from the fields of (computer) linguistics, cognitive science, psychology, and speech–language pathology. We rely on the theory that meaning is acquired from co-occurrences of words (Landauer and Dumais 1997) and propose that this approach to meaning can be used to understand how semantics mature during language development. This allows a quantificational approach, as opposed to, for example, the componential and procedural approaches (Clark and Clark 1977). In a number of publications methods for measuring semantic representations in adults have been developed (Arvidsson et al. 2011; Garcia and Sikström 2013; Gustafsson Sendén et al. 2014; Roll et al. 2011), where these representations can be generated by looking at the co-occurrence of words. These high dimensional representations measure similarities between words, so that words with similar meaning have similar representations (Sahlgren 2008). In this paper we are interested in studying how semantic representations develop as children mature in their linguistic skills. The theoretical assumption underlying these methods is that meaning is in the text, in the distribution of words, in their syntagmatic and paradigmatic relations, and that words are learnt from hearing or reading them in a context. Semantics is not an inventory; meaning is in contrasts and co-occurrences.

Here we apply a proposed semantic space method for the quantification of semantic content directly based on picture-elicited narratives actually generated by children. The method has a distributional approach, based on syntagmatic relations, that is, the co-occurrence of words. With semantic content we refer to the underlying meaning of words or narratives, rather than the specific words that are referenced. For example, the semantic representations of the words *adore* and *loving* are similar, whereas the non-semantic representations of these words in terms of word-length, word-class, word-frequency, phonetics, dictionary, letter-combination, etc. have little or no similarity. We further propose that semantic spaces (Arvidsson et al. 2011; Karlsson et al. 2013; Landauer and Dumais 1997; Sahlgren 2007) can be used to analyse the children's semantic representation of narratives as their linguistic skills mature. Based on information of the co-occurrence of words in large text corpora, the semantic representation of words can be quantified. The size of the context influences the semantic representation so that topical information arises for larger contexts, and lexical semantic information from smaller contexts (Sahlgren 2008). The resulting semantic representation identifies a location, for a given word, as a point in a high dimensional semantic space. The 'meaning' of a word is given by its distance to other words in this space, that is, how it differs from or contrasts with other words. Here we study how such a semantic representation of elicited narratives of children relates to semantic linguistic maturity by using the semantic representation as a predictor of the chronological age. The chronological age may be seen as a proxy variable of children's semantic maturity. Our proposed measure, *Semantic linguistic maturity* (SELMA), refers to how the semantic content, or meaning, of the words generated in a story, changes across childhood development. The purpose of the SELMA measure is to study the semantic representation in particular, and to disregard other aspects of linguistic maturity, such as morphology and syntax. We argue that the rich and high dimensional nature of the semantic representation can be used to produce a strong marker for semantic development, comparable to assessments made by human raters and will be an important complement to other unidimensional measures, such as lexical diversity (McKee et al. 2000), coherence (Halliday and Hasan 1976) or story grammar (Stein and Glenn 1979).

In the following we will give a short description of how semantic and narrative development and skills are usually investigated. Then we describe the principles and background of semantic spaces and present our aims. After that follows a description of the data analysed and a detailed description of how the SELMA measure was created. Finally, the results are presented and discussed.

Vocabulary Development and Narratives

Early learning and use of words is dependent on the physical context. As language skills and vocabulary grow, the learning of new words is decreasingly dependent on direct experiences of the referents of words. Instead, to a large extent, words and their meanings are learnt through how they co-occur with other words (as argued by for example Corrigan 2008; Landauer and Dumais 1997) and in what linguistic contexts they occur (Bloom 2000). Factors that are crucial for successful vocabulary growth are the amount and quality of the input and the level of the child's social and cognitive skills (Bloom 2000). As children hear and read known words in new contexts and new words together with known words their vocabulary grows in richness, which is determined by both its size and its depth. By depth we refer to the knowledge of how each word thematically, phonologically, morphologically, conceptually and sociolinguistically relates to other words (Vermeer 2001). Thus, there is a multidimensional source of knowledge connected to a word, for example its pronunciation, spelling, frequency, what other words it can be combined with, how its meaning is modulated in different combinations and contexts, which other words it can be replaced with and which other words often occur in the same context (Vermeer 2001).

Narratives in typical and disordered populations are an important area of research (e.g., Berman and Slobin 1994; Bishop and Edmundson 1987). The production of a narrative is dependent on linguistic, cognitive and social skills (Norbury and Bishop 2003) and can be considered "a meeting ground of developing linguistic knowledge and general cognitive growth" (Berman 2008: 736). Narrative tasks are also useful for studying and predicting later language development (Berman 2008). Manhart and Rescorla (2002) highlight story grammar structure, grammatical complexity and use of evaluative information as three important domains of narrative ability. Studies of the development of story grammar focus on episodes included in the narrative and how they are structured (Berman and Slobin 1994; Stein and Glenn 1979). These studies show that with age, an increasing number of story grammar elements are included. Studies of narrative development also show that with increasing age children use sentences with an increasing degree of syntactic complexity (e.g., Berman 2008). A further aspect of the development of narrative skills is cohesion, i.e., the ability to tie sentences or units together through the use of referencing and cohesive devices that link sentences together (Halliday and Hasan 1976; Liles 1985; Peterson and Dodsworth 1991; Shapiro and Hudson 1991). Evaluation, finally, refers to the use of comments that contribute to the interpretation of, for example, causes and consequences and the mental states of the characters involved in the narrative (e.g., Bamberg and Damrad-Frye 1991).

Lexical variation or diversity is another aspect that has been highlighted in studies of narratives in children (McKee et al. 2000). A related measure is lexical density (the ratio of content words to function words; Johansson 2009). These measures focus on the size and availability of the vocabulary used through the variation and the distribution of different word categories. Although lexical and story grammar analyses are clearly related to content, they do not well capture the semantic representations nor the semantic complexity.

A more qualitative way to assess narratives is to ask, for example, teachers to score or grade the quality of narratives. Many studies have found positive correlations (around r = .6–.8) between such holistic assessment by human raters and text length (e.g., Grandin and Lindskog 2007; Jafarpur 1991; Löfqvist 1990). It is also a common finding that older children produce longer texts than younger children (e.g., Asker-Árnason et al. 2008, 2010) and that children with typical language development produce longer texts than children in clinical groups (e.g., Reuterskiöld Wagner et al. 1999). McFadden and Gillam (1996) studied teachers' assessments of overall quality of written as well as spoken narratives from different groups of children. They found that longer narratives were rated higher than shorter narratives. Similarly, Newman and McGregor (2006) found that laypersons preferred longer oral narratives, in a comparison between children with language impairment and controls.

Semantic Spaces

Within a quantificatory perspective, the semantic content of text, in the present context elicited narratives, can be assessed by means of so called semantic spaces. Our research group has developed methods for the creation of semantic spaces and applied them in different contexts. For details of these methods see for example Andersson et al. (2012), Arvidsson et al. (2011), Garcia and Sikström (2012, 2013, 2014), Gustafsson et al. (2014), Karlsson et al. (2013), Kjell et al. (2013), Marklund et al. (2009), Roll et al. (2011), Rosenberg et al. (2013), and Sarwar et al. (2014). Here we summarize some of the main properties of semantic space methods, and for methodological details we refer the interested reader to references above.

Semantic spaces can be automatically generated by several computational methods applied to large text corpora. Latent semantic analysis (LSA) is one of the most prominent methods for creating semantic spaces. It was originally developed as a document retrieval method, where it was found to perform better than methods based on direct word matching (Deerwester et al. 1990). Later it was adopted in the cognitive science literature and computer assisted educational research (Landauer and Dumais 1997). The LSA algorithm, being distributional and descriptive, is highly data-driven and does not use syntactic information such as word order or what word class a word belongs to. LSA applied to a text corpus produces a high dimensional semantic space where each word is represented as a vector in this space. The number of dimensions can be automatically identified by choosing the number of dimensions that provides the highest quality of the space. The optimal value depends on the chosen corpus and the purpose of the semantic space, where one hundred dimensions is a good starting point (Dumais 1992). One way to evaluate the quality of the semantic space is to use it to automatically solve a synonym test. The rationale behind this is that the better the space can manage this task, the better the quality of the semantic representation (Landauer and Dumais 1997).

A semantic space can be used to measure the semantic distance between two words, simply by measuring how far the words are located from each other in the semantic space. For example, one would expect to find *puppy* close to *dog* but far from *puppet* (which, on the other hand, is a phonological neighbour; Stokes 2010). Documents can be compared in the semantic space following an aggregation of all semantic representations related to the words in the document.

The semantic representations of single words are best understood by how they are related to other words. Thus, the words that are most similar to a given word, tend to be synonyms that "defines" the meaning of these words. The dimensions of a semantic space are not normally assigned any specific meaning, and it is typically difficult to say what a specific dimension stands for.

LSA is one of several methods to generate and quantify semantic representations. Examples of other methods are random vectors (Sahlgren 2007), latent Dirichlet allocation (Blei et al. 2003) and the Hyperspace Analogue to Language model (Shaoul and Westbury 2010). Common to these methods are that all base the generation of the semantic representation on the co-occurrence of words; however, the mathematical foundation of how this occurs varies between the methods. Whereas LSA is based on syntagmatic relations, the Hyperspace Analogue to Language model is based on paradigmatic relations, that is, how words are exchangeable and share linguistic context.

Several studies have compared automatic scoring of written texts, often using LSA. Landauer et al. (1997) compared LSA-based measures with human ratings of text quality and found high agreement, both between the human raters and between the human raters and LSA (r = around .70). Landauer et al. (2003) and Foltz et al. (1999) similarly found associations between automatic essay assessment and human raters to be as strong as between different human raters. Computer analysis is also used to measure text complexity (Landauer 2011). Graesser et al. (2011) discuss automated analysis of text complexity as a tool to select texts with an appropriate degree of difficulty for students at different levels, using a method that measures several different aspects, including measures of lexical diversity and LSA. In an overview McNamara (2011) concludes that semantic models, like LSA, are useful for extracting meaning, or semantic representations, from texts and that they can simulate human knowledge, but also that they need to be complemented with other approaches in order to catch the full meaning.

Purpose

The purpose of this study was to construct a measure of semantic linguistic maturity that is directly based on the narratives that children actually generate, and that does not rely on human subjective ratings of maturity. More specifically we wanted to explore SELMA in relation to chronological age, text length and ratings by humans.

A rich and high dimensional semantic representation of children's narratives is an essential component in semantic maturity and the hypothesis is that the semantic content of children's narratives, as generated from a specific set of stimuli, predicts their linguistic maturity. We use a machine learning algorithm to predict the chronological age of children given a semantic representation generated from the elicited narratives of the children, where the deviation between the chronological age and predicted age is indicative of whether a child is more or less semantically mature compared to the standard of his/her age. We hypothesize that SELMA will be more strongly associated with qualitative holistic rating by humans, compared to measures like chronological age and text length.

Method

Participants

Narratives were obtained from 108 Swedish-speaking children, 68 girls and 40 boys, in the age range 4:0 (years:months) to 12:10 (mean age 9:2). They were reported by parents and teachers to have typical development in all respects, including language, hearing and non-verbal IQ. These children were sampled from a larger data collection that also included children with hearing impairment and children with language impairment (Asker-Árnason et al. 2012; Reuterskiöld et al. 2010; Reuterskiöld Wagner et al. 1999, 2000). They were recruited from intermediate socio-economic status (SES) areas. The project was approved by the Regional Ethical Review Board, Lund, Sweden.

Collection of Narratives

The narratives were elicited using a selection of pictures from the story *One frog too many* (Mayer and Mayer 1975). The pictures were selected to represent six story grammar units suggested by Stein and Glenn (1979). The units were Setting, Initiating event, Response state, Response plan, Attempt, Consequence and Resolution/Reaction.

As a practice item, the test administrator first presented the children with another sequence of pictures *Frog on his own* (Mayer 1973). The examiner presented one picture at a time, asked the child to first look at all the pictures and then told the child a model story. Following the demonstration, the pictures from *One frog too many* were laid out, one at a time, and the child was asked to look carefully at each picture. The examiner pointed to the first picture and provided the following sentence: "This story is about a boy and his pets, who are going out on a raft", and asked the child to continue the story. The experimenter was instructed to avoid providing support apart from nodding and acknowledging by a "mhm" or a "yes", but in some cases asked a few questions when the child provided too little. The procedure was audio- and video recorded and later transcribed orthographically.

Analysis

Creation of Semantic Spaces

Semantic spaces provide an opportunity to measure the semantic distance between words. To generate a semantic representation with a reasonably high semantic quality, a very large text corpus is required. Furthermore, it is beneficial that the corpus consists of highly different semantic topics which makes it easier to differentiate between different semantic meanings/contexts. To generate a semantic representation for the present study we chose a corpus that fulfilled these criteria. It consists of more than 100,000 articles taken from the 100 largest Swedish newspapers in 2007. This corpus consists of text data different from the data that will be analysed (which in this case are narratives generated from *One frog too many*). The fact that the corpus used for generating the semantic representation is from a different source than the to-be-analysed text, may lead to less distinct, or less high quality representation, compared to if a more similar source would have been used. However, our extensive experience of working with semantic representation shows that it is more important to have a corpus of large size, than a corpus that is closely matched to the to-be-analysed topic (Garcia and Sikström 2014).

The space was created using the Infomap software (http://infomap-nlp.sourceforge.net/index.html), which applies the standard LSA algorithm (Landauer and Dumais 1997). The quality of the generated space was measured by a synonym test. This was conducted by looking at the rank order of the semantic closeness between two synonymous words (from a digital lexicon of synonyms for Swedish; http://folkets2.nada.kth.se/synlex.html). This rank order was divided by the total number of words, which generated a scale from 0 (perfect score) to 1, where .50 reflects random performance. We calculated this scale over 200 synonym pairs, and the median value was .03, which we interpret as indicative of a good quality space, see for example Arvidsson et al. (2011).

Quantifying Semantic Representations of Frog Stories

Further semantic analyses were then conducted using the Semanticexcel software, which is a web-based software for statistical analyses of semantic representations, that has been written by the last author of this paper (www.semanticexcel.com) and has been used in other studies (Andersson et al. 2012; Arvidsson et al. 2011; Garcia and Sikström 2012, 2013, 2014; Gustafsson et al. 2014; Karlsson et al. 2013; Kjell et al. 2013; Marklund et al. 2009; Roll et al. 2011; Rosenberg et al. 2013; Sarwar et al. 2014). The narratives generated by the subjects were summarized in the semantic representation generated by LSA. This was done by adding the semantic vectors representing each word in a narrative, so that each narrative was summarized in one vector (of the same number of dimensions as for single words). The length of this vector was normalized to a length of one (i.e., the same length as the representation of each word).

Semantic Linguistic Maturity (SELMA)

SELMA measures language proficiency based on the notion that age correlates with linguistic maturity. The SELMA of the narratives was produced by linear regression between the semantic vectors of the narratives and the corresponding chronological age of the children. In order to avoid the situation where the age of a child is predicted using a data set that already includes that child's age we use a*leave-one-out cross validation* approach where the age of a child is predicted using a data set that includes all narratives *except* that from the child whose age is being predicted. When a child's age is predicted in this way we argue that this predicted age is a strong marker for semantic development and we call this measure SELMA. This method was repeated for all subjects.

Table 1 shows an example of how the narratives are summarized as semantic representations in the creation of SELMA. Narratives (column 2) are summarized as a high dimensional semantic representation using LSA (column 3) from children with a certain age (column 4)

Table 1 An example of the type of data used to calculate SELMA

Child	Narrative text	Semantic representation	Age	Train/test	SELMA
A	There was a frog…	.12, −.02, .03, .02, . . .	5	Test	7
B	A man and a frog	.06, −.05, .05, −.01, . . .	6	Train	4
C	There was a boat…	.04, −.02, .03, −.02, . . .	10	Train	9.5
D	On a sunny day…	.10, −.03, .04, −.03, . . .	15	Train	18

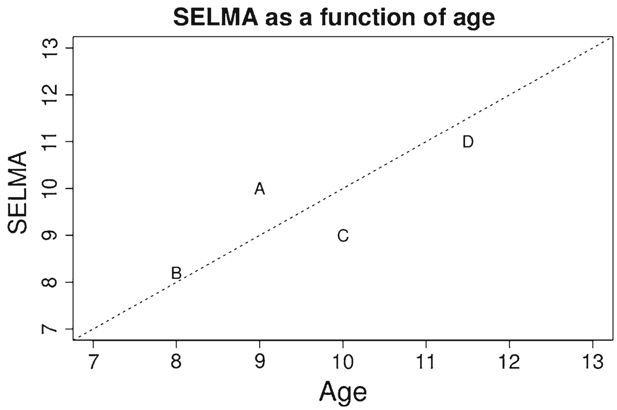

Fig. 1 SELMA as a function of age for the example data given in Table 1. The *dashed line* indicates what a perfect relation between Age and SELMA would look like

to predict SELMA (column 6) based on a one-leave-one out prediction method (column 5) where all narratives (labeled Train) are used to train the linear regression except the one SELMA is calculated for (labeled Test). This procedure is then repeated for all narratives with every narrative being left out from the linear regression once. Figure 1 shows SELMA as a function of age for the example data given in Table 1. Note that SELMA predicts age + "error", where positive "errors" (i.e. when chronological age is less than SELMA) indicates semantic linguistic maturity that is higher than normal for a given age (A and B), and vice versa for negative "errors" (C and D).

We chose to train our model on chronological age for two reasons. First, chronological age during childhood should serve as a predictor of semantic maturity. This idea comes from the fact that older children are more mature, on average, in their linguistic development compared to younger children. Second, chronological age is an objective and quantifiable measure. We wanted to construct a measure of semantic maturity that is fully data-driven, and that does not rely on subjective rankings. Therefore we chose not to train our measure on subjective ranking of semantic maturity, which would lead to a risk of biasing the measure towards properties of the texts that are salient to humans, but not representative of semantic quality, like for example text length or grammatical complexity. In other words, our intention is not to mimic human raters' performance, but rather to develop a new method for measuring semantic linguistic maturity and to capture the prototypical semantic representation for a given age.

However, variables other than chronological age may influence semantic maturity. Consistent with this, children that are of the same age may have different levels of semantic maturity. We argue that following training on a large number of children with different chronological ages, SELMA generalizes and is so sensitive that it can pick up differences in linguistic maturity, also in a group of individuals with the same chronological age. Thus, we do not expect, or seek, a perfect correlation between chronological age and SELMA; rather the discrepancy between these two variables indicates whether a child is more or less mature relative to his/her age.

SELMA is based on linear regression from the semantic representation of narratives to the chronological age of the children that produced the narratives. This approach has an implicit assumption that there is a linear relation between the semantic features and semantic linguistic maturity. We believe that this is a reasonable first approximation; however, further research is needed to validate this assumption.

Evaluation of SELMA

In order to evaluate SELMA as a measure of semantic linguistic maturity twelve human raters were recruited from the Lund University student population. Eight of the raters were students at the Speech- and Language-Pathology Programme, Lund University. All had completed courses in linguistics and in typical and impaired child language development. The remaining four raters were university students with no background in Linguistics or Speech- and Language-Pathology. The raters were in the age range 20–35 years.

A difference between the SELMA measure and what the human raters was asked to rate was the outcome scale. The outcome scale of SELMA is a continuous measure of linguistic maturity but rating such a measure from an isolated narrative was judged as a too difficult task for the raters. Therefore we asked the raters for judgments regarding a comparison between two texts. The transcribed narratives were numbered from 1 to 108 and the narratives were randomly paired to form 54 pairs. Three pairs had to be deleted, since the children in these pairs had exactly the same chronological age. The material presented to the raters was thus 51 pages with two narratives on each. Their task was to read each of the 51 pairs of narratives and to indicate which one of the two narratives in each pair was the more linguistically mature.

In order to use SELMA to generate ratings on the same scale as the human raters the SELMA measures of the narratives in each pair was compared. The narrative with the highest SELMA in each pair was then selected as the "choice" of the SELMA measure. Similarly, choices for each pair of narratives was also made using the actual ages of the children that produced the narratives and using the total number of words of each narrative. This measure was included, since it is a common finding that ratings of text quality correlates to text length, as discussed in the introduction. Thus our variables are SELMA rating (that is, which narrative in each pair had got the highest SELMA score), chronological age rating, text length rating and the judgements by the human raters.

Results

A Pearson's r showed a positive correlation between chronological age and SELMA ($r = .37$, $p < .001$, $N = 108$). Examples of narratives of low, medium, moderately high and very high SELMA scores are shown in "Appendix".

The comparison between the human raters and the rating by SELMA shows that the mean percentage of agreement between each rater and SELMA (i.e., that a rater rated the narrative with the higher SELMA value as the more mature) was 74 % (SD = 5.4 % points) which is far above the chance level of 50 % agreement (One Sample *t* test, $t(11) = 16.3$, $p < .001$). There was no significant difference between the raters that had a background in linguistics or speech/language pathology compared to those that did not (Welch Two Sample *t* test, $t(6.6) = 0.56$, $p = 0.59$).

Figure 2 shows the distribution of the agreement between the human raters and the rating according to SELMA, age and text length. Table 2 shows the mean agreement of these measures. The figure and the table indicate that there is high agreement between the human raters and a high agreement between all raters and SELMA. The human raters tend to have higher agreement with SELMA than with chronological age and number of words, although for number of words (text length) this difference is small. The rating based on SELMA agreed with the human raters significantly more often than the ratings based on chronological age (paired samples *t* test, $t\ (11) = -5.81$, $p < .001$).

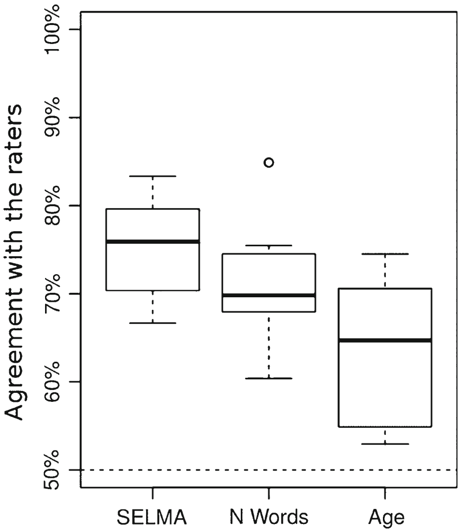

Fig. 2 The distributions of the percentage agreement between the human raters and the three data based rating methods. The *dashed line* indicates chance level agreement

Table 2 Agreement between the different methods of rating the narratives. For the human raters the agreement values show the mean agreement between each human rater and the eleven other raters

	SELMA (%)	Age (%)	No. of words (%)	Raters (%)
Age	55	–	–	–
No. of words	62	56	–	–
Raters	75	64	71	81

A mixed model logistic regression analysis (Baayen et al. 2008) was used to check if SELMA predicts the raters' responses when age and number of words are controlled for. The dependent variable is here the binary choice of the raters, whether they choose the first or the second story as being the most mature. The independent variables are the difference in age, number of words and SELMA between each of the pairs of narratives. A first model was fitted using age difference and difference in number of words as fixed factors and rater as a random factor. A second model was fitted with the addition of SELMA difference as a fixed factor. A likelihood ratio test using the first model as the null model and the second model as the alternative showed that the addition of SELMA difference as an independent variable was justified ($\chi^2 = 84$, $p < .001$). This implies that SELMA contributes to the prediction over and above what age and number of words do.

Discussion

The general purpose of this study was to investigate whether the semantic space model captures the semantic maturity of children in narratives. The results suggest that this is the case and the hypotheses were to a large extent confirmed.

A first hypothesis was that SELMA predicts linguistic maturity. The results show that SELMA is associated with chronological age. We argue that SELMA is a measure of the semantic maturity of the children. It should be emphasised that although SELMA is fitted to the chronological age, the outcome of this fit generalizes to a proxy for a typical semantic maturity of children at a certain age. The interesting aspect of the SELMA score is that it can be applied to a specific child, where the difference between the child's chronological age and the child's predicted age (SELMA) is telling us something about that specific child's level of semantic maturity. For example, the child in Example 4 in "Appendix" has an age of 8:9 and a SELMA score of 12:9. She thus shows a semantic maturity that is higher than what is expected from her age, whereas the same score in a child at an age of 16 may be considered as low semantic maturity.

With this in mind we can address the hypothesis that SELMA would be more strongly associated with maturity ratings made by humans than with the purely quantitative measures, such as number of words and age. The results supported this to a large extent. The mixed model logistic regression analysis that included SELMA as a fixed factor was significantly better than the model that included only age and number of words. Furthermore, SELMA agreed significantly better than chronological age with human ratings. In accordance with other studies (Foltz et al. 1999; Landauer et al. 2003) the mean agreement between SELMA and the human raters was about the same as the agreement between the different raters (around 75 %).

If SELMA was just a noisy estimate of chronological age we would expect the agreement between SELMA and the raters to be worse than the agreement between the raters and chronological age. The significant difference between maturity rating according to SELMA and according to chronological age indicates that SELMA is a better measure of semantic linguistic maturity than chronological age as the twelve human raters had a higher agreement with SELMA than with chronological age, as shown in Fig. 2. This stronger agreement between SELMA and the raters was present in spite of a relatively weak correlation between SELMA and chronological age. We interpret this as the discrepancy between SELMA and chronological age being not just noise but rather indicating a discrepancy between chronological age and actual semantic linguistic maturity, a discrepancy that SELMA is able to capture.

The raters were asked to account for what criteria they were using in order to assess the narratives. Most of them answered that they had found content coherence to be an important indicator. The raters further stated that inclusion of the two story grammar elements 'introduction' and 'conclusion' was of particular importance. The impression when reading the examples of narratives with low and high SELMA scores supports this. The narratives with high SELMA score are easier to follow and understand without the support of the pictures. The raters also found the choice of words to be of importance, that is, they acknowledged lexical variation and complexity. Finally, grammatical correctness was also an important factor. It should be noted that the semantic analysis conducted here, i.e. SELMA, is insensitive to several of these criteria reported by raters. Because our analysis is insensitive to word orders, it does not directly measure semantic coherence, grammar, lexical variation, and complexity. Although SELMA only accounts for the semantic content, whereas the human raters were asked to make holistic judgements, taking more aspects into account, there is still a strong correlation between the two rating methods. This could be taken as evidence that content and form are intimately interrelated and develop in tandem. The quality of the narrative is not only in the content itself, but in how the content is organized and expressed. This could be further explored by relating SELMA score to lexical and grammatical variables.

The current SELMA measure focuses on the 'absolute' semantic proficiency, as it increases with age. Another interesting focus would be to look at semantic proficiency relative to the age of the developing child. This relative semantic proficiency could be more clearly expressed by looking at the ratio between SELMA and the chronological age, and thus forming a "semantic quotient". A significant feature of this quotient would be that it could be expected to be relatively stable for a certain individual across childhood, similar to the IQ measure. Applying a semantic quotient to the girl presented in Example 4 in "Appendix" would yield a semantic quotient of 1.45 or a larger maturity compared to her age. In contrast, the girl of the same age in Example 1 has a semantic quotient of .94, suggesting a lower than average semantic maturity.

It is important to point out that the SELMA score differs fundamentally from several other related measures. This difference is not only related to what it measures (semantic maturity), but also to how this measure is constructed. Here we list a number of different qualities of SELMA and describe how it differs from other scores. We contrast it to text-length, as measured by number of words. However, the same arguments apply to many other related measures in the linguistic/psychological literature.

Training

The SELMA score is computed on a training set, where a known quantification (age) is required, on which the score is trained.

Task-Specificity

Training is done on data related to a specific task (i.e., generating statements from a specific set of pictures). Thus, once SELMA is trained on this type of task, there is no guarantee that it can be applied to another task without retraining (e.g., telling a story about your best friend). Although we cannot exclude the possibility that training of SELMA may generalize to related tasks, this is an empirical question that has not been investigated. However, once SELMA has been trained on a dataset (in this case picture-elicited frog stories in Swedish), the measure can readily and automatically be applied to new data on the same task and language, without a need for further training.

Quantification of Semantics

SELMA requires a quantification of the semantic representation, which is done by applying LSA to a huge text corpus (on an unrelated material).

Holistic

The term holistic, or the idea of weighting in a large number of factors in an evaluation, is often used in qualitative research. However, in our view SELMA is both a holistic and a quantitative measure in the sense that it weighs in a large number of semantic dimensions in common evaluation. We argue that a major advantage of SELMA is the fact that it allows us to quantify semantic representations, and information that otherwise requires qualitative analysis, which is time-consuming and subjective.

A potential problem is the fact that the data analyzed (picture generated oral narratives) were different from the type of text data in the corpus that was used to generate the semantic representation (newspaper text), which furthermore is a genre the children in the study have probably not yet been exposed to. However, as pointed out earlier, it seems that the quality of the semantic representation is more dependent on the size of the corpus than on close matching of the topic (Garcia and Sikström 2014).

More studies are needed in order to validate SELMA, for example to systematically manipulate the content of the story by providing different pictures as cues to the children. Most likely the training of the SELMA measure would have to be tailored to the specific stimuli material that is being presented in pictures. A next step will be to include clinical populations in order to see how well LSA can identify children with a diagnosis of language impairment. This is currently being explored by the authors.

Conclusions

We have presented a method to quantify semantic linguistic maturity (SELMA). This method is built on the theoretical assumption that meaning resides in how words co-occur. This SELMA score has been implemented in computerized software, Semantic, which can be acquired by contacting the last author. The results from the comparison with other variables indicate that SELMA contains additional information on semantic maturity. In particular the results suggest that the semantic representation of the narratives contains information on semantic maturity. Narratives with high SELMA scores are more likely to be rated as more mature by human raters and they also tend to be produced by children who are older, although this relationship is weaker. Several other methods (Blei et al. 2003; Sahlgren 2007; Shaoul and Westbury 2010) use semantic spaces to assess text complexity, text quality or other aspects of verbal data. The unique property of SELMA is that it is directly related to semantic development. This makes it a tool that is particularly relevant to use in studies of child language development. This type of method could also be very interesting to apply in clinical contexts, in the assessment of different populations with language problems. By introducing this method to research on child language development and child language disorders, we also highlight and confirm the theoretical basis it rests on, with relevance for the learning of words and word meaning. Words to a large extent get their meaning from the context they are used in. Awareness of the importance of hearing and reading words in context for learning their meaning will have theoretical implications for research in these areas and practical implications for intervention and advice to parents.

The analysis of narrative skills and the analysis of text complexity must include several domains (Graesser et al. 2011; Manhart and Rescorla 2002) with respect to form as well as to content. With a measure of semantic linguistic maturity we hope to contribute a new dimension to the analysis of content focussing on the meaning relations between words.

Acknowledgments This research was supported by the Linnaeus environment Thinking in Time: Cognition, Communication and Learning (financed by the Swedish Research Council, Grant No. 349-2007-8695), the Bank of Sweden Tercentenary Foundation (Grant No. 000-0171:01), the Swedish Council for Working Life and Social Research (Grant No. 2007-0248), the Stinger Foundation (Grant Nos. 2005-42, 2006-29), Sunnerdahls Disability Foundation, the VAAG project, funded by the European Science Foundation and a grant on "Semantic Spaces in Psychology" from the Swedish Research Council. We are also grateful to all the children who participated and students who collected parts of the data for their master's theses.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix: Examples of Narratives with Low, Medium, Moderately High and Very High SELMA, Approximate English Translations from Swedish

Example 1: Low SELMA

Girl, chronological age: 104 months, SELMA score: 98.20 'a boy sort of stands and sees something from a raft or something a dog a turtle and two small or two frogs that frog kicked away that other little frog and then the turtle woke up and that boy doesn't see it starts to get a little more mean or something the dog or the turtle tell that boy but he doesn't understand and doesn't care about that and that turtle looks a bit sly and mostly to or that boy notices it and he lets go of that stick that the raft with and the dog starts barking and the turtle is mad at the frog and yes then he gets scared the boy they started to look for him and they looked everywhere and they don't find him and then everyone has become disappointed with that frog and the dog get very angry at that frog and then they leave that place'

Example 2: Medium SELMA

Boy, chronological age: 126 months, SELMA score: 115.98 'it is a boy who is playing on a boat and then there are two frogs in the back and the big frog shoots away the small one with its/his foot then it's just looking so happy at the boy and then they are then the boy is totally surprised and wonders where the small frog is and then the dog is barking and the turtle is all mad at that frog and then they all start looking and looking and then the boy becomes sad and the turtle is also a little bit sad but the dog is mad at the frog'

Example 3: Moderately High SELMA

Boy, chronological age: 129 months, SELMA score: 128.41 'the dad is looking angrily at the frog then he kicks away the small one and then he sits on the wooden board on the boat all by himself and the person discovers the dog or the turtle him and the boat runs on ground is he falling into the water and he is running around and looking for his turtle and the frog jumps away or the frog wants to go with them then the dog looks angrily at the frog while the person is walking around crying'

Example 4: Very High SELMA

Girl, chronological age: 105 months, SELMA score: 152,66 'there is a boy standing on a raft and on the raft there are two frogs a/one turtle a/one dog the boy is pointing at something the dog is also looking at the thing he is pointing at and then there is water and reed the big frog kicks away the small one from the raft and the turtle looks surprised when the boy is pointing the boy looks tired the dog looks sad the turtle looks scared or frightened and that frog looks content sort of the frog says that it pushed the small frog into the water and then everybody is sort of surprised they look for the small frog the boy looks underneath a water-lily leaf the turtle looks behind stone the frog looks behind a stick then a mosquito flies or a little the boy is sad and the dog is cross with the frog and the turtle is a bit surprised the frog is sad'

References

Andersson, R., Bååth, R., & Sikström, S. (2012). Visually mediated valence effects in dialogue: An explorative study. *Lund University Cognitive Studies*, *151*, 1.

Arvidsson, D., Werbart, A., & Sikström, S. (2011). Psychotherapy causes changes in object representations: A computational and theory-free semantic-space method. *Psychotherpahy Methods*, *21*, 430–446. doi:10.1080/10503307.2011.577824.

Asker-Árnason, L., Åkerlund, V., Skoglund, C., Ek Lagergren, I., Wengelin, Å., & Sahlén, B. (2012). Spoken and written narratives in children and adolescents with hearing impairment. *Communication Disorders Quarterly*, *33*, 131–145.

Asker-Árnason, L., Ibertsson, T., Wass, M., Wengelin, Å., & Sahlén, B. (2010). Picture-elicited written narratives, process and product, in 18 children with cochlear implants. *Communication Disorders Quarterly*, *31*, 195–121.

Asker-Árnason, L., Wengelin, Å., & Sahlén, B. (2008). Process and product in writing: A methodological contribution to the assessment of written narratives in 8–12-year-old Swedish children using ScriptLog. *Logopedics Phoniatrics Vocology*, *33*, 143–152.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Bamberg, M., & Damrad-Frye, R. (1991). On the ability to provide evaluative comments: Further explorations of children's narrative competencies. *Journal of Child Language*, *18*, 689–710.

Berman, R. (2008). The psycholinguistics of developing text construction. *Journal of Child Language*, *35*, 735–771.

Berman, R., & Slobin, D. (1994). *Relating events in narrative: A crosslinguistic developmental study*. Hillsdale, NJ: Lawrence Erlbaum.

Bishop, D., & Edmundson, A. (1987). Language-impaired 4-year-olds: Distinguishing transient from persistent impairment. *Journal of Speech and Hearing Disorders*, *52*, 156–173.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, Mass.: The MIT Press.

Bloom, P. (2001). Roots of word learning. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 159–181). Cambridge: Cambridge University Press.

Clark, H., & Clark, E. (1977). *Psychology and language. An introduction to psycholinguistics*. New York: Harcourt Brace Jovanovich.

Corrigan, R. (2008). Beyond the obvious: Constructing meaning from subtle patterns in the language environment. *Communication Disorders Quarterly*, *29*, 109–124.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Dollaghan, C. (1987). Fast mapping in normal and language-impaired children. *Journal of Speech and Hearing Disorders*, *52*, 218–222.

Dumais, S. (1992). Enhancing performance in latent semantic indexing (LSI) retrieval. Technical Report, Bellcore (now Telcordia Technologies), Morristown, NJ (September 1992).

Fenson, L., Marchman, V., Thal, D., Dale, P., Bates, E., & Reznick, J. (2007). *The MacArthur–Bates communicative development inventories: User's guide and technical manual* (2nd ed.). Baltimore: Paul H. Brookes.

Foltz, P., Laham, D., & Landauer, T. (1999). The Intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Cumputer-Enhanced Learning*, *1*(2). http://imej.wfu.edu/articles/1999/2/04/

Garcia, D., & Sikström, S. (2012). Quantifying the semantic representations of adolescents' memories of positive and negative life events. *Journal of Happiness studies*, *14*(4), 1309–1323. doi:10.1007/s10902-012-9385-8.

Garcia, D., & Sikström, S. (2013). A collective theory of happiness: Words related to the word happiness in Swedish online newspapers. *Cyberpsychology, Behavior, and Social Networking*, *16*, 469.

Garcia, D., & Sikström, S. (2014). The dark side of Facebook: Semantic representations of status updates predict the Dark Triad of personality. *Personality and Individual Differences (PAID)*,. doi:10.1016/j.paid.2013.10.001.

Graesser, A., McNamara, D., & Kulikowich, J. (2011). Och-Metrix: Providing multilevel analysis of text characteristics. *Educational Researcher*, *40*, 223–234.

Grandin, S., & Lindskog, M. (2007). Logopeders bedömning av textkvalitet. Master's thesis in Speech–Language Pathology, Department of Logopedics, Phoniatrics and Audiology, Clinical Sciences, Lund, Lund University.

Gustafsson Sendén, M., Lindholm, T., & Sikström, S. (2014). Selection bias as reflected by choice of words: The evaluations of "I" and "We" differ between communication contexts, but "They" are always worse. *Journal of Language and Social Psychology*. doi:10.1177/0261927X13495856.

Gustafsson, M., Sikström, S., & Lindholm, T. (2014). Biases in news media as reflected by personal pronouns in evaluative contexts. *Social Psychology*, *45*, 103–111.

Halliday, M., & Hasan, T. (1976). *Cohesion in English*. London: Longman.

Hills, T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks. *Psychological Science*, *20*, 729–739.

Jafarpur, A. (1991). Cohesiveness as a basis for evaluating compositions. *System*, *19*, 459–465.

Johansson, V. (2009). *Developmental aspects of text production in writing and speech*. Travaux de L'Institut de Linguistique de Lund, 48.

Kamhi, A. (1992). Three perspectives on language processing: Interactionism, modularity, and holism. In R. Chapman (Ed.), *Processes in language acquisition and disorders* (pp. 45–64). Missouri: Mosby-Year Book.

Karlsson, K., Sikström, S., & Willander, J. (2013). The semantic representation of event information depends on the cue-modality: The organization and selection of event information revisited. *PLoS One*, *8*(10), e73378. doi:10.1371/journal.pone.0073378.

Kjell, O., Nima, A., Sikström, S., Archer, T., & Garcia, D. (2013). Iranian and Swedish adolescents: Differences in personality traits and well-being. *PeerJ*, *1*, e197. doi:10.7717/peerj.197.

Koren, R., Kofman, O., & Berger, A. (2005). Analysis of word clustering in verbal fluency of school-aged children. *Archives of Clinical Neuropsychology*, *20*, 1087–1104.

Lahey, M. (1988). *Language disorders and language development*. Needham Heights, MA: Allyn & Bacon.

Landauer, T. (2011). Pearson's text complexity measure. http://kt.pearsonassessments.com/download/PearsonsTextComplexity-May2011.pdf.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Landauer, T., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education*, *10*, 295–308.

Landauer, T., Laham, D., Rehder, B., & Scheriner, M. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. *Proceedings of the* 19th *annual conference of the cognitive science society* (pp. 412–417). Mahwah, NJ: Erlbaum.

Liles, B. (1985). Cohesion in the narratives of normal and language disordered children. *Journal of Speech and Hearing Research*, *28*, 123–133.

Löfqvist, G. (1990). The IEA study of written composition in Sweden. *Studia Psychologica et Paedagogica—Series Altera*, no. 93, Lund.

Manhart, J., & Rescorla, L. (2002). Oral narrative skills of late talkers at ages 8 and 9. *Applied Psycholinguistics*, *23*, 1–21.

Marklund, P., Sikström, S., Bååth, R., & Nilsson, L.-G. (2009). *Age effects on semantic coherence: Latent semantic analysis applied to letter fluency data.* Paper presented at the confereence proceeding SEMAPRO 09, October 11–16, Malta.

Mayer, M. (1973). *Frog on his own*. USA: Penguin Books Inc.

Mayer, M., & Mayer, M. (1975). *One frog too many*. USA: Penguin Books Inc.

McFadden, T., & Gillam, R. (1996). An examination of the quality of narratives produced by children with language disorders. *Language, Speech, and Hearing Services in Schools*, *27*, 48–56.

McGregor, K., Friedman, R., Reilly, R., & Newman, R. (2002). Semantic representation and naming in young children. *Journal of Speech, Language, and Hearing Research*, *45*, 332–346.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistics Computing*, *15*, 323–337.

McNamara, D. (2011). Computational methods to extract meaning from text and advance theories of human cognition. *Topics in Cognitive Science*, *3*, 3–17.

Newman, R., & McGregor, K. (2006). Teachers and laypersons discern quality differences between narratives produced by children with or without SLI. *Journal of Speech, Language, and Hearing Research*, *49*, 1022–1036.

Norbury, C., & Bishop, D. (2003). Narrative skills of children with communication impairments. *International Journal of Language and Communication Disorders*, *38*, 287–313.

Peterson, C., & Dodsworth, P. (1991). A longitudinal analysis of young children's cohesion and noun specification in narratives. *Journal of Child Language*, *18*, 397–415.

Reuterskiöld, C., Ibertsson, T., & Sahlén, B. (2010). Venturing beyond the sentence level. Narrative skills in children with hearing loss. *Volta Review*, *110*, 389–406.

Reuterskiöld Wagner, C., Nettelbladt, U., Sahlén, B., & Nilholm, C. (2000). Conversation versus narration in children with language impairment. *International Journal of Language & Communication Disorders*, *35*, 83–93.

Reuterskiöld Wagner, C., Sahlén, B., & Nettelbladt, U. (1999). What's the story? Narration and comprehension in Swedish preschool children with language impairment. *Child Language Teaching and Therapy*, *15*, 113–137.

Roll, M., Mårtensson, F., Sikström, S., Apt, P., Arnling-Bååth, R., & Horne, M. (2011). Atypical associations to abstract words in Broca's aphasia. *Cortex*, *48*, 1068–1072.

Rosenberg, P., Sikström, S., & Garcia, D. (2013). The difference between living biblically and just imagining it: A study on experiential-based learning among Swedish adolescents. *School Psychology International Journal*, *34*(5), 566–572.

Sahlgren, M. (2007). An Introduction to Random Indexing. SICS2007, Stockholm University: Stockholm, pp. 1–9

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, *20*, 33–53.

Sarwar, F., Sikström, S., Allwood, C. M., & Kerr-Innes, Å. (2014). Predicting correctness of eyewitness statements using the semantic evaluation method (SEM). *Quality & Quantity*,. doi:10.1007/s11135-014-9997-7.

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, *42*, 393–413.

Shapiro, L., & Hudson, J. (1991). Tell me a make-believe story: Coherence and cohesion in young children's picture-elicited narratives. *Developmental Psychology*, *27*, 960–974.

Stein, N., & Glenn, C. (1979). An analysis of story comprehension in elementary school children. In R. Freedle (Ed.), *New directions in discourse processing* (pp. 53–120). Norwood, NJ: Ablex.

Stokes, S. (2010). Neighbourhood density and word frequency predict vocabulary size in toddlers. *Journal of Speech, Language, and Hearing Research*, *53*, 670–683.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, *22*, 217–234.