Using LSA to Automatically Identify Givenness and Newness of Noun Phrases in Written Discourse

Christian F. Hempelmann, David Dufty, Philip M. McCarthy, Arthur C. Graesser, Zhiqiang Cai, and Danielle S. McNamara
({chmplmnn, ddufty, pmmccrth, a-graesser, zcai, dsmcnamr}@memphis.edu)

Institute for Intelligent Systems/FedEx Institute of Technology
University of Memphis
Memphis, TN 38152

Abstract

Identifying given and new information within a text has long been addressed as a research issue. However, there has previously been no accurate computational method for assessing the degree to which constituents in a text contain given versus new information. This study develops a method for automatically categorizing noun phrases into one of three categories of givenness/newness, using the taxonomy of Prince (1981) as the gold standard. The central computational technique used is span (Hu et al., 2003), a derivative of latent semantic analysis (LSA). We analyzed noun phrases from two expository and two narrative texts. Predictors of newness included span as well as pronoun status, determiners, and word overlap with previous noun phrases. Logistic regression showed that span was superior to LSA in categorizing noun-phrases, producing an increase in accuracy from 74% to 80%.

Introduction

Successive constituents in text, such as sentences or noun phrases (NPs), vary in how much *new* versus *given* information they contain. This distinction is not binary. For example, it is uncertain how to classify an idea that would have been inferred earlier in the text rather than explicitly stated, as will be discussed later. The aim of this paper is to assess the extent to which givenness and newness can be computed algorithmically from features of the text. Automatic assessment of givenness is useful for a variety of NLP applications, including the assessment of student responses to automatic tutoring systems, paragraph recognition, discourse feature identification, and recall scoring. The present application was devised for implementation in Coh-Metrix (Graesser et al., 2004b), a text-processing tool that provides new methods of automatically assessing text cohesion, readability, and difficulty.

When considering the dimension of familiarity, text constituents can be classified into three partitions: *given*, *partially given* (based on various types of inferential availability), or *not given* (that is, *new)*. When developing an automatic system, it is more natural to view new information as that information that is not given, rather than vice versa. So we would need to first need to compute how much given (old) information is in a constituent and then regard the remaining information as new. Therefore, any automated measure that describes how part of a text can be established as given by a reader is valuable as it will increase the amount of identified givenness.

To illustrate the basic distinctions of givenness, consider the following example.

(1) President Bush said on Friday he recognized that there were other solutions to bolster Social Security than his contentious proposal for personal retirement accounts, but they would be part of a broader overhaul of the country's largest entitlement program.

In this example, *Social Security* is new when it is first mentioned, while *the country's largest entitlement program* is coreferential with it. Thus, the constituent *the country's largest entitlement program* is given information even though there are lexical differences that have to be bridged inferentially. *Retirement accounts*, on the other hand, is only inferentially available from *Social Security*; that is, it is neither fully new nor unexpected in view of the previous mention of *Social Security*. Thus, *retirement accounts* is neither given nor new but somewhere in between.

We propose that any word in a text must be considered situated on a continuum between wholly given and wholly new. By extension, any phrase, clause or sentence analyzed in whole or part can be assessed for its degree of givenness. Our goal in this paper is thus to explore methods for automatically extracting these degrees of givenness for particular sections of text. However, before discussing computational measures of givenness in more detail, the theoretical basis for the relevant concepts will be addressed in the next section.

Theoretical accounts of the given/new dimension

Halliday (1967) defines given information as "recoverable either anaphorically or situationally" from the preceding discourse, and new information, conversely, as not recoverable. Chafe (1975, 1987) defines given information as "knowledge which the speaker assumes to be in the consciousness of the addressee" (1975: 30). In Chafe's initial binary framework of given and new, given information is previously activated, whereas new information is activated only by the current segment of text. Chafe then introduces a distinction between new, given, and a third category, 'quasi-given' (1977: 34). This third category is related to the inferential availability of information, and has been a central concept in modern approaches. Clark and Haviland (1977) extend the distinction using Gricean maxims, proposing a 'given-new contract' on the inferential processes involved in meaning construction. They argue that a speaker composes speech acts that have affiliated inferences and believes that the addressee has access to the inferences.

Terminology

The terms *given* and *new* are often used to refer to theme and rheme, respectively, as well as other similar dichotomies that adopt a functional sentence perspective (Mathesius 1947). Such issues, including foregrounding, topicality or saliency, interact with givenness, and for this reason the terms are often used synonymously (see, Steedman, 2000; Kruijff-Korbayová & Steedman, 2003. for a discussion of terminology and distinctions). While the theme is *usually* given, and the rheme is *usually* new, the theme sometimes contains new information. One example of this is when there is a change of subject, as in (2a). Similarly the rheme can also, and occasionally does, contain given information, as in the case of contrasts like (2b).

(2) a. Men work hard in order to be successful.
b Women work hard in order to be successful, too.

On the basis of sentence (2a), in sentence (2b) the theme *women* is new, while the rheme *work hard in order to be successful* is given. As can be seen from this example, it is entirely possible for a rheme to provide old information. We are primarily interested in the contextual and semantic aspects of the given/new distinction. Thus, we want to clearly distinguish the given/new dichotomy from theme/rheme, topic/comment, and so forth, rather than conflate them as does, for example, the BEAT system (Cassell, Vilhjálmsson, & Bickmore, 2001).

Another related line of research concerns notions such as primacy (Gernsbacher & Hargreaves, 1988) and recency effects (Caplan, 1972; Chang, 1980; von Eckhardt & Potter, 1983). Primacy effects are related to the assumption that words mentioned first in sentences, and sentences mentioned first in paragraphs, are more accessible in memory. Recency effects, on the other hand, are related to the assumption that words or sentences will be more accessible to memory when they have been more recently presented or when there are fewer words between them and the currently processed sentence.

These concepts also have implications for what can be considered given or new in a text. From a psychological perspective, a concept can only be considered given in any practical sense if the reader remembers it. Although we consider memory access relevant and fruitful avenues for research in relation to givenness, it is beyond the scope of the present research. Instead, our purpose is to operationalize the given/new distinction purely in terms of semantic recoverability. Eventually, it will be possible to compare given-new with other discourse structuring devices, such as theme-rheme, and recency-primacy.

Prince's (1981) taxonomy of given/new

A very influential definition of givenness is provided by Prince (1981). Prince developed a systematic taxonomy of given, inferable, and new information that can be used to hand-code written text for givenness (Donzel, 1994; Kruijff-Korbayová & Kruijff, 2004; Prince, 1988; Strube, 1998). This present paper is facilitated by three crucial advantages of Prince's approach. First, in contrast to other conceptualizations of givenness, she crafts her familiarity scale on a theoretical basis that integrates previous theoretical discussions (Chafe, 1975; Clark, 1967; Clark and Haviland, 1977; Halliday, 1967). Second, Prince does so without diluting givenness with other focusing and discourse structuring properties of text. Third, despite the complexity of the resulting model, she provides example analyses and a systematic methodology to apply her model. Because of the formal-theoretical nature, the clear focus of her approach, and the inclusion of a methodology, Prince's work can be applied to text analyses and ultimately implemented computationally. Prince's analysis is restricted to NPs, but we believe that a more version of Prince's theory that covers units other than NPs, prominently VPs, should be developed.

Prince identifies three different sources of givenness. First, *Predictability/Recoverability (Givenness$_P$)* is based on the speaker's assumption "that the hearer CAN PREDICT OR COULD HAVE PREDICTED that a PARTICULAR LINGUISTIC ITEM will or would occur in a particular position WITHIN A SENTENCE" (1981; emphases in the original). Second, *Saliency (Givenness)* is based on the speaker's assumption "that the hearer has or could appropriately have some particular thing/entity/… in his/her consciousness at the time of hearing the utterance." Third, *Shared Knowledge (Givenness$_K$)* is based on the speaker's assumption "that the hearer 'knows,' assumes, or can infer a particular thing (but is not necessarily thinking about it)." On the basis of these three types, Prince proposes the following taxonomy:

(3)	BN	brand-new
	BN_A[__]	brand-new anchored [Anchor^{type}]
	U	unused
	I(__)/__	inferrable (entity inferrable from^{type})/ inference-type
	I_C(__)/__	containing inferrable (containing. entity inferrable from^{type})/inference type
	E	(textually) evoked
	E_S	situationally evoked

In this taxonomy, BN indicates an item that is neither previously mentioned in the text nor readily and immediately available to the reader given the current situation. In the following example, *Heat can move from one object or place to another*, the NPs *heat*, *one object*, and *place* are all considered BNs. BN_A marked items are BN NPs that are tied to a given NP. For example, in the following sentence, *Chlorophyll traps the energy in sunlight*, the NP *energy in sunlight* is a BN_A: the NP *energy* being a BN anchored to the

NP *sunlight* given in a previous sentence. consider the following sentence: *People use thermometers to measure the temperature*. *People* in this sentence is considered *unused* (U) because the concept of humans in general is readily available to all participants regardless of textual context. Other concepts such as *the sun*, *the moon*, and *Genghis Khan*, would also count as unused items. Clearly, this element of the Prince taxonomy is open to some question due to the subjective judgment concerning concepts that people have available. That said, the raters of the texts in this study did not encounter any instances in which agreement could not be reached.

I_Cs differ from I_S in that they are inferences that can be made from inferences, in other words, two-word inferences. In this sense I_Cs are conceptually one step further removed from the textual item from which they are inferable. Consider the following sentence: *And he knew he would miss his home: the nights in the den watching sports, the barbecue parties in the backyard, his hideout in the attic, and of course, his room*. Both raters judged the NPs *the nights in the den watching sports, the barbecue parties in the backyard,* and *his hideout in the attic* as being I_C items. The head of the NP *the nights in the den watching sports* is *the nights,* which is not inferentially available from item such as *his home*. However, from *his home* we can infer that he would have *a den,* and from *den* we can infer that he might *spend nights there watching sports*. All other constituents are givens: An E has been previously mentioned, whereas an Es is situationally given. For example, the word *you* in a text is a given because you are in fact reading the text.

Prince's implied hierarchy can be represented in an explicit familiarity scale (4a below). The scale posits that higher items that are further to the left are more familiar to the hearer. Thus, the Gricean maxim of quantity can be applied: Speakers choose the most familiar method to refer to a constituent possible. If they choose one that is not as familiar to the hearer as they assume, the hearer will not understand (too little information). If they choose one that is too familiar to the hearer, they run the risk of sounding childish (too much information).

We adopted Prince's (1981) familiarity scale and translated it into values of newness from 0 (fully given) to 1 (fully new) as follows (4b):

(4)	a.	E/E_S	>	U	>	I	>	I_C	>	BN_A	>	BN
	b.	0		0.2		0.4		0.6		0.8		1

It should be noted that these numbers are only used for computational convenience. The scale is ordinal, not an interval or ratio scale. Type of scale affects the types of statistical analyses that can be conducted, as indicated later.

All NPs in the sample corpus described below were hand-scored according to the Prince taxonomy by two independent experts in linguistics. Inter-rater agreement produced kappa of .72. Differences occurred between raters because Prince's taxonomy is not unambiguous and frequently lead to a NP to be assigned to multiple categories (cf. Poesio & Vieira, 1998). Differences occurred in about 18% of cases and were resolved by consultation between the scorers. For an illustration of potential disagreements between judges, consider the following sentence from our corpus: *When some of his friends came to say good bye, tears flowed down his face*. One rater viewed the NP *tears* as a BN whereas the other viewed it as an I_C. Clearly there is a case for both. On the one hand, *tears* had not previously been mentioned (therefore *tears* is new); on the other hand, saying *goodbye* is often very sad, and sadness leads to *tears* (therefore *tears* is a containing inferable). Although these disagreements occurred, judges were able to resolve disputes after some discussions.

LSA-Based Automated Measures for Given/New

In earlier work (Dufty et al., 2005), we evaluated a range of computational measures for given/new, including constituent/lexical/stem/lemma overlap, a simplified version of coreference on the basis of ontological semantics (Nirenburg & Raskin, 2004), as well as measures based on Latent Semantic Analysis (LSA). In the present paper we further explore the capabilities of LSA in more detail.

LSA is a technique for computing the similarity of words by representing them in a vector space and computing the cosine of the angle between vectors for pairs of words (Landauer et al., 1998). Higher cosines represent greater similarity. The vector space is created by constructing a co-occurrence matrix out of a large corpus of texts. The space is then reduced using singular value decomposition, such that each word is represented in a space of approximately 300 dimensions. The dimensions themselves have no meaning, but are merely statistical constructs. Meaning is extracted by comparing the similarity of vectors in the space. LSA can be used to evaluate the similarity of text segments of any size through vector addition. For example, the similarity of two paragraphs can be calculated by adding all the vectors for words in the first paragraph to create a paragraph vector, adding the vectors for words in the second paragraph to create a second vector, then taking the cosine of the two paragraph vectors as an estimate of the similarity between them. LSA has been used for a variety of applications such as automated tutoring systems (Graesser et al., 2004a), essay grading (Foltz et al., 1999), and evaluating text coherence (Foltz et al., 1998).

LSA might seem at first glance to be the ideal candidate for evaluating the givenness of a segment of text. By comparing the vector of the current sentence with the vector for the preceding text, some estimate can be gained of the similarity of the current sentence with prior text. However, the concept of givenness, while related, is distinct from the concept of similarity. On the one hand, for a text item to be given, it need only be coreferential with one previous item. LSA captures overall similarity with the text, rather than a particular constituent. Thus, while the previous text may contain the very item that is being compared for its similarity, the measure takes all the other items in the preceding text into account as well. This dilutes the score considerably. On the other hand, a text item can be partially given on the basis of its inferential availability and world knowledge. LSA is not a symbolic approach, but it can only roughly approximate this.

Our second main measure, based on a variant of LSA, was developed for the specific purpose of detecting new information. The method is called *span* (Hu et al., 2003). It was formulated to test the accuracy of student answers in the automated tutoring system *AutoTutor* (Graesser et al., 2004a). Rather than simply adding vectors, span constructs a hyperplane out of all previous vectors. The comparison vector (in this case the current sentence in the text) is projected onto the hyperplane. The projection of the sentence vector on the hyperplane is considered to be the component of the vector that is shared with the previous text, or given (G). The component of the vector that is perpendicular to the hyperplane is considered to be the component of the sentence that is new (N). To calculate the newness of the information, a proportion score is then taken: Span(new information) = N/(N+G). N is the component of the vector that is perpendicular to the hyperplane and G is the projection of the vector along the hyperplane.

Span captures newness in a more sophisticated way than standard LSA. Standard LSA combines all previous text into a single composite vector and compares the sentence to that vector. In doing so, much of the information contained in vectors of individual sentences is lost, as the individual vectors can cancel each other out. Span constructs a hyperplane out of all the vectors of all the sentences, and compares the new sentence to that space. This method means that no information in the individual vectors is lost.

Materials, Method, and Results

We selected four texts of approximately equal size from 4th grade textbooks: two narrative texts, 'Moving' (McGraw-Hill Reading - TerraNova Test Preparation and Practice - Teacher's Edition) and 'Orlando' (Addison Wesley Phonics Take-Home Reader Grade 2), and two expository texts, 'The Needs of Plants' (McGraw-Hill Science) and 'Effects of Heat' (SRA Elementary Science). The texts contained 478 NPs in total, across 195 sentences.

The NPs in the texts were hand-coded according to the original categories postulated by Prince, conflating the two types of *evoked*, as they are both fully given, resulting in six categories. There was an inter-rater reliability of .74 given by kappa, with 88% of cases rated the same by both raters. The six categories ultimately had to be collapsed into three because of the sparseness of data. Two of the categories, unused and containing inferables, had very low counts (3 and 8 respectively), rendering them unsuitable for categories in a logistic regression. We therefore decided to reduce our number of categories, and decided to use the common three-category system: *given, new*, and *inferable*. Hence we collapsed these both into the category of *inferable*, and collapsed *brand new anchored* into *brand new*. Thus, the intermediate category between fully new (0) and fully given (1) subsumed all instances of NPs that were neither entirely given nor entirely new, such as unused, inferrable, containing-inferrable, and brandnew-anchored.

NPs were then coded for the following binary properties: whether the NP was a pronoun, whether the NP was preceded by the definite[1] article, and whether any content word in the NP had occurred in a previous NP (a modification of argument overlap; Kintsch & van Dijk, 1978). All binary variables were coded as 1=yes, 0=no.

Two computational measures were calculated based on LSA. The first was the LSA similarity between the NP and all previous noun-phrases in the text. The second was the span measure between the NP and all previous noun-phrases. Table 1 shows descriptive statistics for all predictor variables. The relative frequencies for the criteria variable, Prince category, across all 478 NPs were 317, 116, and 45, for given, inferable, and new observations, respectively.

Table 1: Descriptive statistics for all predictor variables

Binary variables	Yes	No
Pronoun	111	367
Definite article	71	407
Word overlap	141	347
Continuous variables	Mean	s.d.
LSA cosine with prev. NPs	.20	.27
Span with previous NPs	.29	.32

Two ordinal logistic regressions were performed with the hand-coded Prince categories as the dependant variable. An alpha level of .05 was used for all significance tests. The first analysis tested a predictive model consisting of the three binary variables (pronoun, definite article, and content word overlap), as well as LSA cosines as predictors. The second analysis tested the model in which span was added. The coefficients generated from both these analyses are shown in Table 2.

As can be seen from Table 2, LSA contributed to the categorization of NPs in the first model. As expected, pronouns and definite articles were more likely to reflect given information. Pronouns tend to refer to earlier entities in

[1] Since the class of NPs that surface as definite is not at all coextensive with those that speakers assume can be familiar to hearers, we will not focus on the notion of definiteness beyond its use as an auxiliary identifier for givenness. While every definite NP is given under our definition, but not every given NP is definite.

In general, ours is the opposite vantage point from that of existing work on definiteness (e.g., Fraurud, 1990; Vieira & Poesio 2000). They are interested in definiteness, which givenness and other semantic phenomena can help them account for. We are rather interested in a semantic phenomenon, givenness, that surface phenomena like definiteness can help to identify. A third type of related approaches may be looking for other classifications that surface partially as definiteness and are partially caused by givenness, e.g. Uryupina's (2003) *unique* and *discourse-new*.

the text or pragmatically available information. New nouns are typically (but not always) preceded by the indefinite article when they represent new information , and preceded by the definite article on subsequent mentions. Content word overlap showed a modest positive relationship with newness, which is the opposite direction to its theoretical relationship with newness. This is probably a suppression effect caused by LSA, since LSA and content word overlap attempt to capture a similar aspect of the text.

The addition of span into the second model produced an increase in predictive accuracy from 74% to 80%, which an incremental chi-square test showed to be significant, $\chi^2(1,478) = 183.07$, $p < .05$. Span also displaced both LSA and content word overlap as significant predictors from the second model.

Table 2: Ordinal logistic regression analysis of Prince categorizations using pronouns, definite articles, and content word overlap, and comparing LSA and span

	β	*S.E.* β	Wald's χ^2	*df*
Model 1: Span not included as predictor[a]				
Threshold				
Prince= 0 (given)	4.74	1.05	20.24**	1
Prince = 1 (intermed.)	6.68	1.07	39.04**	1
Predictor				
LSA	-4.70	.67	49.77**	1
Pronoun	-5.17	1.03	25.14**	1
Content word overlap	.83	.34	6.10**	1
Definite article	-.79	.40	3.88*	1
Model 2: Span included as predictor[b]				
Threshold				
Prince= 0 (given)	8.94	1.12	56.89**	1
Prince = 1 (intermed.)	12.12	1.18	89.86**	1
Predictor				
LSA				1
Span	6.79	0.54	158.06**	1
Pronoun	-6.38	1.08	35.10**	1
Content word overlap	0.68	0.38	3.25	1
Definite article	-1.00	0.46	4.67*	1

* significant at .05 level

** significant at .01 level

a model $\chi^2(4,N=478) = 173.22$, p<.05. Accuracy 74%

b model $\chi^2(5,N=478) = 366.29$, p<.05. Accuracy 80%

In the second model, the largest contribution to prediction of newness category was made by span, followed by pronominalization. This demonstrates the different contributions that these variables make to predicting newness category. Span captures the semantic relationship between each NP and previous noun-phrases. This relationship is invisible when a pronoun is used, because of span's reliance on lexical-semantic relationships between content words. Conversely, pronouns capture indirect reference to earlier noun-phrases, which in turn is invisible to LSA and span.

For comparison, an ordinal regression was also performed without either span or LSA, but retaining the three binary variables, definite article, pronoun, and content word overlap. The resulting model achieved 66% accuracy, which, given that the most common category occurred 66% of the time, is no different than chance.

Discussion

We developed a multivariate model of givenness and newness using word repetition, pronominalization, articles, and a continuous measure of newness, span. The model allocated NPs to one of the three categories of newness with 80% accuracy, when compared to human ratings. Agreement between the human raters, in this case 88%, may be considered to be the benchmark of performance. Against this benchmark, span's performance, with an 8% difference, is very promising. Completely automatable measures were able to approximate hand-coded ratings by experts.

In a separate analysis, standard LSA was also a significant predictor of newness, although it was 6% less accurate than span. LSA was originally developed as a measure of similarity between two items of text, while span is specifically a measure of the newness of one text in comparison to another. The results confirm that span is a more appropriate measure when newness, rather than similarity, is the concept of interest in the text.

The analysis that only used simple algorithmic indicators such as whether the NP is a pronoun, whether the NP begins with *the,* or whether the NP repeats content words from an earlier NP, did no better than chance. This demonstrates the importance of similarity metrics such as span in determining linguistic and psycholinguistic properties of text.

The present results provide a bridge between theoretical linguistics and computational linguistics, they provide a reliable mapping between categories of newness as described by Prince (1981), and computable text-based variables.

Acknowledgements

This research was funded by Institute for Educations Science Grant IES R3056020018-02. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the IES.

References

Caplan, D. (1972). Clause boundaries and recognition latencies for words in sentences. *Perception and Psychophysics, 12*, 73-76

Cassell, J., Vilhjálmsson, J., and Bickmore, T. (2001). BEAT: the Behavior Expression Animation Toolkit. *Proceedings of SIGGRAPH.* New York: ACM. 477-486.

Chafe, W. L. (1975). Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View. In: C. N. Li (Ed.), *Subject and Topic* (pp 26-55). New York: Academic.

Chafe, W. L. (1987). Cognitive Constraints of Information Flow. In: R.S. Tomlin, (Ed.), *Coherence and Grounding in Discourse* (pp. 21-51). Amsterdam, Philadelphia: Benjamins.

Chang, F. R. (1980). Active memory processes in visual sentence comprehension: Clause effects and pronominal references. *Memory and Cognition, 8,* 58-64

Clark, H. H. & Haviland, S. E. (1977). Comprehension and the Given-New Contrast. In: R.O. Freedle (Ed.), *Discourse Production and Comprehension* (pp. 1-40). Norwood, NJ: Ablex.

Dufty, D.F., Graesser, A.C., Hempelmann, C.F., & McCarthy, P.M.(2005). Givenness and Newness of Information: Automated Identification in Written Discourse. Manuscript submitted for publication.

Foltz, P.W., Kintsch, W., and Landauer, T.K. (1998). The measurement of textual Coherence with Latent Semantic Analysis. *Discourse Processes, 25,* 285-307.

Foltz, P.W., Laham, D., & Landauer,T.K. (1999). The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning* 1-2. Available at: http://imej.wfu.edu/articles/1999/2/04/printver.asp.

Fraurud, K. (1990). Definiteness and the processing of NPs in natural discourse. *Journal of Semantics*, 7, 395-433.

Gernsbacher, M.A. & Hargreaves, D. (1988). Assessing sentence participants: The advantage of first mention. *Journal of Memory and Language, 27*, 699-717.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H., Ventura, M., Olney, A., &. Louwerse, M. M. (2004a). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers, 36*, 180-193.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004b). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers 36,* 193-202.

Halliday. M. A. K. (1967). Notes on Transitivity and Theme in English. *Journal of Linguistics, 3*, 199-244.

Hu, X., Cai, Z., Louwerse, M. M., A. Olney, P. Penumatsa, A. C. Graesser, & TRG. (2003). A revised algorithm for Latent Semantic Analysis. *Proceedings of the 2003 International Joint Conference on Artificial Intelligence*: 1489-1491.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363-394.

Kruijff-Korbayová, I. & Kruijff, G. J. M. (2004). Discourse-Level Annotation for Investigating Information Structure. Paper presented at the ACL 2004 Workshop on Discourse Annotation, Barcelona, Spain, July 25-26, 2004.

Kruijff-Korbayová, I. & M. Steedman. (2003). Discourse and Information Structure. *Journal of Logic, langauge and Information: Special Issue on Discourse and Information Structure, 12,* 249-259.

Landauer, T.K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

Mathesius, V. (1947). O tak zvaném aktuálnim clenení vetném. /On the so-called actual articulation of the sentence./ In: *Cestina a obecny jazykozpyt.* Prague: Melantrich.

Nirenburg, S and Raskin, V. (2004). *Ontological Semantics.* Cambridge: MIT Press.

Poesio, M. & Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics,* 24-2, 183-216.

Prince, E. (1981). Towards a Taxonomy of Given—New Information. In: Cole, Peter. Ed. *Radical Pragmatics*. New York: Academic: 223-255.

Prince, E. (1992). The ZPG Letter: Subjects, Definiteness, and Information-status. In: S. Thompson, & W. Mann (Eds.), *Discourse Description: Diverse Analyses of a Fund Raising Text.* Philadelphia/Amsterdam: Benjamin's. 295-325.

Steedman, M. (2000). Information Structure and the Syntax-Phonology Interface. *Linguistic Inquiry 34,* 649-689.

Strube, M. (1998). Never look back: An alternative to centering. In: *Proceedings of the 36th Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics,* 1251-1257.

Uryupina, O. (2003). High-precision Identification of Discourse New and Unique Noun Phrases. In: *Proceedings of the ACL Student Workshop, Sapporo, 2003.*

van Donzel, M. E. (1994). How to specify focus without using acoustic features. *Proceedings,* 18. Institute of Phonetic Sciences, University of Amsterdam: 1-17.

Vieira R., & Poesio, M. (2000). An Empirically-Based System for Processing Definite Descriptions", *Computational Linguistics*, 26-.4, 539-593.

von Eckhardt, B. & Potter., M. C. (1985). Clauses and the semantic representation of words. *Memory and Cognition* 13, 371-376.