# The Discourse Structuring Potential of Definite Noun Phrases in Natural Discourse

Sofiana Chiriacescu[1], and Klaus von Heusinger[1]

[1] Institute of Linguistics, University of Stuttgart,
Keplerstr. 17, 70174 Stuttgart, Germany
{Sofiana.Chiriacescu, Klaus.vonHeusinger}@ling.uni-stuttgart.de

**Abstract.** This paper investigates an alternation found with definite noun phrases in direct object position in Romanian that represents a theoretical puzzle for current theories of Differential Object Marking or *pe*-marking (Dobrovie-Sorin 1994). When in direct object position and unmodified, definite noun phrases can be accompanied either by the differential object marker *pe*, or by the simple enclitic definite article. In light of the findings of a sentence-continuation experiment, we show that *pe*-marking is used for referents that display a high Discourse Structuring Potential (Chiriacescu 2011). We conclude that different types of definite noun phrases do not equally contribute to the Discourse Structuring Potential of their referents. We argue that *pe*-marked unmodified definite noun phrases display (i) a higher likelihood of subsequent mention (Givón 1983, Ariel 2001, Arnold 2010) and (ii) a higher topic shift potential (Givón 1983).

# 1 Introduction

Romanian is a language that displays *pe*-marking as an instance of Differential Object Marking, or DOM (Niculescu 1965, Cornilescu 2000). This means that direct objects are not realized in the same way: they are sometimes accompanied by the marker *pe* and in other contexts they are not. The presence or absence of the marker is generally assumed to depend on different factors, such as animacy, referentiality or specificity (Farkas 1978, Klein & de Swart 2011). Pronouns, proper names and modified definite noun phrases are generally accompanied by *pe*. Despite the long research tradition on DOM in general and Romanian in particular, the distribution of the marker with unmodified definite and indefinite noun phrases is still not entirely understood. In this paper we focus on variations with unmodified definite noun phrases like the one presented in (1), in which the direct object is either preceded by *pe* (e.g. the form ***pe băiat,*** '**PE** boy'), as in (1a), or remains unmarked by *pe* (e.g. the form *băiat**ul***, '**the** boy') in (1b).[1] Note that *pe*-marking and the definite article cannot co-occur whenever the noun phrase is not further modified. This incompatibility is reflected by the ungrammaticality of (1c).

Since modified definite noun phrases usually get the *pe*-marker, alternations as in (1), in which the direct object is unmodified, were either left unaccounted for (Gramatica Limbii Române 2005, Klein & de Swart 2011), or were only partially explained in terms of genericity (Dobrovie-Sorin 1994).

(1)   (a)   Doctorul   **îl**   examinează   **pe**   băiat.
               Doctor.DEF  **CL**  examines    **PE**   boy
               'The doctor examines the boy.'

       (b)   Doctorul     examinează   băiat**ul**.
               Doctor.DEF     examines    boy.**DEF**
               'The doctor examines the boy.'

       (c)   * Doctorul   **îl**   examinează   **pe**   băiat**ul**.
               Doctor.DEF  CL  examines    PE   boy.DEF
               'The doctor examines the boy.'

In this paper we extend the analysis of *in*definite noun phrases in direct object position in Romanian advanced in Chiriacescu & von Heusinger (2010) and motivate the variation in (1) in terms of the Discourse Structuring Potential of the different types of definite noun phrases. Specifically, in light of the findings of a story-continuation experiment, we show that the presence of the *pe*-marker in (1a)

---

[1] Note that in Romanian, *pe*-marked direct objects are usually accompanied by clitic pronouns, which are co-indexed weak pronouns (e.g. the 3. Pers. Masc. clitic *il* in ex. (1a)). Different semantic and syntactic factors are responsible for the presence or absence of clitics with *pe*-marked direct objects. In this paper, we cannot discuss clitic doubling, but focus instead solely on *pe*-marking (see Dobrovie-Sorin 1994, or Gramatica Limbii Române 2005 for a discussion on clitic doubling in Romanian).

correlates with a higher Discourse Structuring Potential (DSP) of the referents associated with this type of referring expression, compared to those associated with the non-*pe*-marked noun phrase in (1b). DSP is understood as the property of an expression that introduces a discourse referent to provide information about the discourse status of the referent in the subsequent discourse. The operational definition of DSP employed in this paper is the one developed in Chiriacescu & von Heusinger (2010) and Chiriacescu (2011), where we used two factors (i.e. referential persistence and topic shift potential) to determine the discourse status of a referent. DSP is measured by means of the two textual characteristics given above, as both metrics pertain to the following discourse.

This paper is structured as follows. In section 2 we provide a brief overview of the major factors discussed in the linguistic and psycholinguistic literature (Givón 1983, Kaiser & Trueswell 2004, Kehler, Kertz, Rohde & Elman 2008) as textual manifestations of accessibility, salience or discourse prominence. Instead of employing a backward-looking perspective on referent resolution, we will account for the distribution of DOM with unmodified definite noun phrases in Romanian from the perspective of production, by adopting a forward-looking perspective. In section 3 we discuss a sentence continuation study investigating the Discourse Structuring Potential of direct objects realized as unmodified definite noun phrases. The last section summarizes the findings of this paper and points out interesting directions for further research.

## 2    Accessibility and the Discourse Structuring Potential of Referents

A body of linguistic and psycholinguistic research has investigated various factors that influence the comprehension and production of different types of referring expressions (Givón 1983, Gundel, Hedberg & Zacharski 1993, Kehler, Kertz, Rohde & Elman 2008, Ariel 2001, Arnold 2010). The majority of these studies focused on anaphora resolution, as it is commonly assumed that lexically reduced referring expressions correlate with entities highly accessible or prominent in the preceding discourse. To determine the accessibility of a referent, researchers have generally employed a *backward-looking* perspective by analysing the factors that license the usage of a particular type of referring expression at a particular stage in the discourse. In other words, given a certain type of referring expression (e.g. a pronoun), the factors that contribute to its use in the present utterance were investigated.

In contrast to personal pronouns, which refer back to a previously mentioned entity, definite noun phrases do not always refer back to an explicitly introduced referent. Rather, definite noun phrases with descriptive material can refer to entities that lack an antecedent, but may be related to an already introduced referent by some bridging or inference relation (Hawkins 1978, Vieira & Poesio 2000). The way a definite noun phrase was introduced in the discourse is said to correspond to the

accessibility or degree of activation of the associated referent (von Heusinger 2003, 2007).

In this study we consider only those definite noun phrases that represent hearer-old and discourse-old information, in the sense that they refer back to a referent, which was explicitly mentioned in the preceding discourse. However, we do not explore the properties of the antecedents of definite noun phrases to determine the accessibility of their referents. We investigate two textual characteristics of the discourse referents of unmodified definite noun phrases that pertain to the following discourse, namely referential persistence and topic shift potential. We adopt a *forward-looking* perspective and test the effects of production-driven biases licensed by unmodified *pe*-marked nouns and simple definite nouns in Romanian.

The first metric, *referential persistence*, indicates the likelihood with which a particular referent will be picked up in the subsequent discourse (Givón 1983, Kehler et al. 2008). The second metric for DSP, *topic shift potential*, is defined in terms of the likelihood that a non-subject referent will be mentioned in grammatical subject position in the following discourse. The notion of topic is associated with the subject position because different linguistic and psycholinguistic studies (e.g. Crawley & Stevenson 1990) have shown that referents mentioned in syntactic subject position are often mentioned in topic position as well. The same referents are more salient or accessible in a given discourse than referents mentioned in other syntactic positions (e.g. as direct or indirect objects). For the sake of simplicity, the first instance in which a direct object referent becomes the grammatical subject in a matrix clause is treated as an instance of topic shift.

Despite being mentioned in a rather non-prominent grammatical position (i.e. as a direct object), we expect *pe*-marked definite noun phrases to show higher values for both metrics than their non-*pe*-marked counterparts. In the remainder of this paper, we present the sentence continuation study conducted to test the DSP of referents realized in direct object position.


## 3  The Experiment

The experiment presented in this section investigated the Discourse Structuring Potential of direct object referents realized as definite noun phrases. We tested two textual characteristics of the referents, namely: (i) referential persistence (i.e. the likelihood that the referent is picked up more often in the following discourse), and (ii) topic shift potential (i.e. the tendency of a referent to be mentioned as the grammatical subject in a subsequent matrix clause).

In a previous study on the DSP of indefinite noun phrases in direct object position in Romanian (Chiriacescu & von Heusinger 2010), we concluded that the referents of the *pe*-marked indefinite noun phrases showed higher DSP values than the referents of their non-marked counterparts. In light of those findings, we predict that, if *pe*-marked definite noun phrases have a high potential to structure the discourse, then they will show high values for referential persistence and topic shift potential.

### 3.1 Method and Design

The methodology used was an open-ended sentence continuation task with four test items (Gernsbacher & Shroyer 1989, Chiriacescu & von Heusinger 2010). Participants (n=24) read 4 mini-discourses and were instructed to add 5 logical and natural-sounding sentence continuations to each of them. As participants' continuations were not restricted in any way (e.g. by pronouns, sentence connectives, etc.), their responses were made on the basis of the mental representations they developed during reading the stories. We manipulated the form of the direct objects in the critical sentence, which resulted in two conditions, i.e. one in which the direct object realized as an unmodified definite noun phrase is *pe*-marked, as in (2), and one in which the same direct object is unmarked, as in (3). This was a within-subjects manipulation, in that a participant completed a version of the study where all four targets contained simple definite noun phrases, and a version where all four targets contained *pe*-marked definite noun phrases. We will refer to these two versions as the "*pe*-condition" and the "non-*pe*-condition".

(2) *Sample experimental item for the pe-condition*

La petrecerea de aseara, Andrei$_1$ a cunoscut un politician$_2$ şi un cântăreţ$_3$ de renume. Astăzi **l$_2$-a intâlnit pe politician$_2$** în piaţă.
'At yesterday evening's party, Andrew$_1$ met a politician$_2$ and a famous singer$_3$. Today he$_1$ met **PE politician$_2$** at the market'.

(3) *Sample experimental item for the non-pe-condition*

La petrecerea de aseara, Andrei$_1$ a cunoscut un politician$_2$ şi un cântăreţ$_3$ de renume. Astăzi a intâlnit **politicianul$_2$** in piaţă.
'At yesterday evening's party, Andrew$_1$ met a politician$_2$ and a famous singer$_3$. Today he$_1$ met **the politician$_2$** at the market'.

### 3.2 Procedure and Data Analysis

The first five main sentences (including any associated subordinate clauses) provided by the participants were analyzed. We used subscript 1 for the entity that was the subject of the first and second sentence (e.g. *Andrei* in (2)). Subscript 2 was used for the target referent, i.e. the object whose form was manipulated in the critical sentence (e.g. *politician* in (2)). Subscript 3 was used for the referent of the singer introduced in the initial mini-discourse. Please note that the target referent was always mentioned with another conjoined argument in sentence 2. Example (4) presents an experimental item for the *pe*-condition and Table 1 presents one response given for this item and illustrates the coding methods used.

The first aspect under investigation was ***referential persistence***. We counted how often a referent was mentioned in the main and subordinate clauses of the subsequent

discourse and explored the way in which referential persistence relates to grammatical role (i.e. whether referents introduced in subject or object position display different referential persistence values). For this purpose, we calculated the referential persistence of all referents given in the mini-discourses. The number of times that a referent was mentioned in the continuations was added up to a sum representing the referential persistence of that referent at a particular stage in the discourse (i.e. cumulative value). For example, in the text provided in (4), the referent introduced by the direct object (the politician) is mentioned in S1 through S5 eight times, whereas the referent introduced by the subject is mentioned five times up to S5. The third referent was left unmentioned.

(4) *Coding methods for an experimental item*

La petrecerea de aseara, **Andrei**$_1$ a cunoscut **un politician**$_2$ şi **un cântăreţ de renume**$_3$. Astăzi **(pro)**$_1$ **l**-a întâlnit **pe politician**$_2$ în piaţă.
'At yesterday evening's party, Andrew$_1$ met a politician$_2$ and a famous singer$_3$. Today he$_1$ met PE politician$_2$ at the market.'

**Table 1.** Continuation sentences for an experimental item with coding methods

| | Continuation sentences | Ref$_1$ (Sum) | Ref$_2$ (Sum) | Ref$_3$ (Sum) | Topic |
|---|---|---|---|---|---|
| **S1** | **(pro)**$_1$ *ştia că acum e şansa **lui**$_1$.* <br><br> '**He**$_1$ knew that that's **his**$_1$ chance.' | 2 | 0 | 0 | Ref$_1$ |
| **S2** | **Politicianul**$_2$ *era un pic grizonat, slăbuţ cu accent baritonal.* <br><br> '**The politician**$_2$ had some greyish hair, was thin with baritone voice.' | 0 | 1 | 0 | **Ref$_2$** |
| **S3** | **Andrei**$_1$ *s-a dus spre **el**$_2$, şi **(pro)**$_1$ **i**$_2$-a cerut ajutorul să **(pro)**$_2$ aleagă un pepene bun.* <br><br> '**Andrei**$_1$ went towards **him**$_2$ and **he**$_1$ asked **(him**$_2$**)** for help to choose a tasty water melon.' | 4 | 4 | 0 | (Ref$_1$) |
| **S4** | **Politicianul**$_2$ *s-a întros şi **(pro)**$_2$ **i**$_1$-a răspuns cu un aer distrat.* <br><br> '**The politician**$_2$ turned around and **(pro)**$_2$ responded **him**$_1$ in a distracted voice.' | 5 | 6 | 0 | (Ref$_2$) |
| **S5** | **Îl**$_2$ *chema don Giuseppe şi **(pro)**$_2$ era inginer zootehnist de meserie.* <br><br> '**His**$_2$ name was don Giuseppe and **he**$_2$ was a zootechnician engineer.' | 5 | 8 | 0 | (Ref$_2$) |

***Topic shift*** was the second aspect tested and represented the first instance in which a non-subject referent was mentioned in grammatical subject position in a following matrix clause. We did not take into consideration whether this shift was maintained in the following discourse or not. In other words, we did not explore whether the referent of the target item was mentioned after this point in subject position or not. This is indicated by the round brackets surrounding the topic constituents starting with S3 in Table 1. Note that we considered topic shifts that occurred in matrix clauses alone, since different studies have shown that referents mentioned as subjects in subordinate clauses do not make good topics (Emonds 1970, among others).

## 3.3 Results

24 participants provided continuations for the initial mini-discourses. The results from the two metrics for DSP, referential persistence and topic shift potential indicate the preferential discourse status of the referents headed by *pe*, compared to those headed by the simple definite article. In the following subsections, we discuss the findings of the two textual characteristics in detail.

**Referential Persistence.** Figure 1 displays the mean values for referential persistence of all referents of the test items 1-4 (TI1-TI4). We found a highly significant difference between *pe*-marked and unmarked direct objects ($p < 0.0001$) with respect to referential persistence. Specifically, for the *pe*-condition, we notice a high likelihood for the direct object referents to be picked up often in the following discourse. In contrast, in the non-*pe*-marked condition, direct object referents are mentioned less often in the subsequent discourse. The results reported so far confirm our predictions regarding the higher DSP of *pe*-marked referents, as these referents display a strong tendency to be picked up in the subsequent discourse.
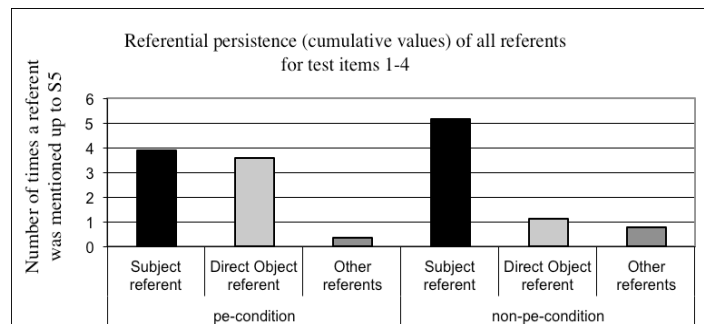


**Fig. 1.** Referential persistence of all referents introduced in the initial mini-discourses in all test items up to S5 (average per continuation story)

Furthermore, Figure 1 shows that the difference in referential persistence between subject and direct object referents is smaller in the *pe*-condition: The black subject bar

and the light grey direct object bar have similar heights in the *pe*-condition, but in the non-*pe*-condition, the subject bar is much higher that the direct object bar. These findings indicate that the referent associated with the *pe*-marked direct object becomes a better competitor for the subject referent in terms of referential persistence. In sum, we find that *pe*-marking reverses the general low expectancy of direct objects to be subsequently mentioned, since *pe*-marked direct object referents tend to be picked up frequently in the ensuing discourse.

**Topic Shift.** The second textual characteristic investigated was the topic shift potential of direct object referents. Recall that the first time the referent of a direct object is mentioned in grammatical subject position in a matrix clause is counted as an instance of topic shift (Givón 1983). The counts for the topic shift potential are cumulative, which means that we added up to a sum all instances in which the target referent (i.e. the non-subject referent) of a test item became the new topic in S1 through S5. The findings condensed in Figure 2 reveal several patterns. We found a significant difference between *pe*-marked direct objects and non-*pe*-marked objects with respect to topic shift ($p < 0.021$). Specifically, the referent of the *pe*-marked direct object displays a stronger preference to become a subject in the continuation sentences (S1-S5) than the referent of the non-*pe*-marked direct object referent. Participants mentioned the referent of the *pe*-marked direct object sooner or later in subject position and did so in less than 25% of the cases with unmarked direct objects.
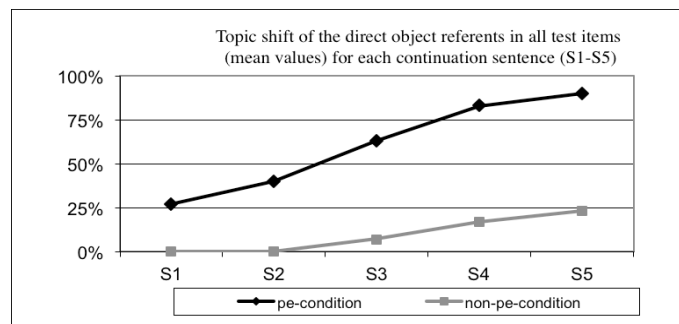


**Fig. 2.** The topic shift potential of referents mentioned in both conditions, measured as the cumulations of the *first* occurrence of the direct object in subject position in the continuation sentences S1-S5.

Figure 2 shows that the referent of the unmarked direct object was never picked up in subject position in the first two continuation sentences (S1 and S2) provided by the participants. In contrast, when we look at the *pe*-marked conditions, we see that the referent of the *pe*-marked direct object was occasionally mentioned in subject position in the first two continuation sentences, even though the rate was not very high (around 35%).

Another very interesting observation is the fact that the *pe*-marked referents show a particular latency with respect to their topic shift potential. This means that these referents are more likely to be mentioned in subject position after continuation

sentence 2 and before sentence continuation 3. After this point, the topic shift potential remains constant (i.e. it is not expected to drastically increase). This observation can be accounted for only in a model in which hierarchical recency, as operationalized by Grosz & Sidner (1986) is abandoned in favour of a more flexible discourse model. Such a model has the advantage of predicting which referent is likely to show a high Discourse Structuring Potential, despite being introduced in a non-prominent syntactic or semantic position and despite being sometimes not linearly recent.

The findings concerning the topic shift potential of direct objects realized as unmodified definite noun phrases confirmed our prediction that *pe*-marked referents are expected to be more prone to become the topic in the subsequent discourse (i.e. to become the grammatical subject) than the unmarked referents.

All in all, the results of the sentence-continuation study presented in this section underlined the preferential discourse status of the *pe*-marked referents and parallel similar finings reported in Chiriacescu & von Heusinger (2010) and Chiriacescu (2011) about the discourse behaviour of indefinite noun phrases in direct object position. We explained the distribution of *pe* with unmodified definite direct objects by showing that *pe*-marking contributes to the higher DSP of the noun phrases it precedes.


## 4 Conclusion

The findings of the study described in this paper extend an expectancy-driven approach to language processing (e.g. Kehler et al. 2008, Arnold 2010). We have shown that referents realized in positions that are otherwise low in accessibility (i.e. referents realized as definite noun phrases in direct object position) will show a high DSP, whenever this potential is morphologically encoded by a special marker, such as *pe* in Romanian. The consequences of *pe*-marking on unmodified definite noun phrases, as exemplified in (1), are explained by showing that the relevant discourse contribution of *pe* is to signal to the addressee that further information about the referent marked in this way will follow and that the same referent is more likely to be picked up in grammatical subject position (i.e. as a topic constituent) in the following discourse.

In light of the present findings, we expect different markers of noun phrases to cross-linguistically vary with respect to the Discourse Structuring Potential of the referents they are associated with. A confirmation of this hypothesis comes from Gernsbacher & Shroyer (1989), Chiriacescu (2011) on the discourse behaviour of German *so'n* and English indefinite *this,* and Deichsel & von Heusinger (2011) on German *dieser*, among others.

## References

1. Ariel, M.: Referring and Accessibility. Journal of Linguistics 24, 65–87 (1988)
2. Arnold, J.: How Speakers Refer: The Role of Accessibility. Language and Linguistic Compass 4, 187–203 (2010)
3. Chiriacescu, S.: The discourse structuring potential of indefinite noun phrases. Special markers in Romanian, German and English. Doctoral dissertation. University of Stuttgart (2011)
4. Chiriacescu, S., von Heusinger, K.: Discourse prominence and *pe*-marking in Romanian. International Review of Pragmatics. 2(2), 298-332 (2010)
5. Cornilescu, A.: Notes on the Interpretation of the Prepositional Accusative in Romanian. In: Bucharest Working Papers in Linguistics (1), pp. 91−106. Bucharest, Bucharest (2000)
6. Crawley, R., Stevenson, R., Kleinman, D.: The use of heuristic strategies in the interpretation of pronouns. Journal of Psycholinguistic Research (19), 245-264 (1990)
7. Deichsel, A., von Heusinger, K.: The Cataphoric Potentia of Indefinites in German. In: S.L. Devi, A. Branco, R. Mitkov. (eds.). Selected papers from the 9th Discourse Anaphora and Anaphor Resolution Conference. Springer, Heidelberg (2011)
8. Dobrovie-Sorin, C.: The Syntax of Romanian. Mouton de Gruyter, Berlin (1994)
9. Emonds, J.: Root and structure-preserving transformations. Doctoral dissertation. MIT, Cambridge, Mass (1970)
10. Farkas, D.: Direct and indirect object reduplication in Rumanian. In: D. Farkas, Jacobsen, W.M., Todrys, K.W. (eds.). Papers from the Seventeenth Regional Meetithe Chicago Linguistics Society. CLS vol. 14, pp. 88-97. Chicago: Chicago Linguistic Society (1978)
11. Gernsbacher, M., Shroyer, S.: The cataphoric use of The indefinite 'this' in spoken narratives. Memory and Cognition. 17, 536 – 540 (1989)
12. Givón, T.: Topic continuity in discourse: An introduction. In: T. Givón. (ed.). Topic Continuity in Discourse: A Quantitative Cross-Language Study, pp. 109-136. John Benjamins, Amsterdam, Philadelphia (1983)
13. Limbii Române: Vol I. Cuvantul. Editura Academiei, Bucuresti (2005)
14. Grosz, B., Sidner, C: Attentions, intentions and the structure of discourse. Computational Linguistics 12, 175-204 (1986)
15. Gundel, J., Hedland, N., Zacharski, R.: Cognitive Status and the form of referring expressions in discourse. Language. 69, 274-307 (1993)
16. Hawkins, J.: Definiteness and Indefiniteness- A Study in the Reference and Grammaticality Prediction. Croom Helm, London (1987)

17. von Heusinger, K.: The Double Dynamics of Definite Descriptions. In: J. Peregrin. (ed.). Meaning in the Dynamic Turn, pp. 150-168. Elsevier, Amsterdam (2003)

18. von Heusinger, K.: Accessibility and Definite Noun Phrases. In: M. Schwarz Friesel, M. Consten, M.Knees (eds.). Anaphors in Text: Cognitive, Formal and Applied Approaches to Anaphoric Reference, pp. 123-144. John Benjamins: Amsterdam (2007)

19. Kaiser, E., Trueswell, J. C.: The role of discourse context in the processing of a flexible word order language. Cognition. 94, 113–147 (2004)

20. Kehler, A., Kertz, L., Rohde, H., Elman, J.: Coherence and Coreference Revised. Journal of Semantics. 25, 1-44 (2008)

21. Klein, U., de Swart, P.: Case and Referential Properties. Lingua. 122 (1), 3-19 (2011)

22. Niculescu, A.: Obiectul direct prepozitional in limbile romanice. In: Individualitatea limbii romane intre limbile romanice. Bucuresti: Editura Stiintifica (1965)

23. Vieira, R., Poesio, M.: An empirically-based system for processing definite descriptions. Computational Linguistics. 26(4), 539–593 (2000)