XXI. Data and Theory Hand in Hand: Determinants of Post-Auxiliary Ellipsis in Modern English[26]

Evelyn Gandón-Chapela and Javier Pérez-Guerra
Universidad de Cantabria and Universidade de Vigo
evelyn.gandon@unican.es, jperez@uvigo.gal

Abstract

This chapter analyses Post-Auxiliary Ellipsis in Modern English through the statistical modelling of the variation between its two subtypes: Verb Phrase Ellipsis (VPE), in *We don't want to postpone the conference, but due to the pandemic we will ~~postpone the conference~~*, and Pseudogapping (PG), as in *If you don't tell me, you will ~~tell~~ your mum*. VPE involves the ellipsis of the constituent following the licensor, whereas in PG a remnant is kept after the licensor. The research question addressed here is: what is the nature of the linguistic determinants that trigger either VPE or PG? Every example of VPE and PG in the Penn Parsed Corpus of Modern British English (PPCMBE) was analysed addressing a number of grammatical, semantic/discursive and processing linguistic predictors. A fixed-effects regression model, supported by Random Forests, determined the relative weight of the potential determinants of either VPE or PG in Modern English.

Keywords: *ellipsis, Post-Auxiliary Ellipsis, Verb Phrase Ellipsis, Pseudogapping, remnant, regression, corpus.*

1. INTRODUCTION

The focus in this chapter is on Post-Auxiliary Ellipsis (PAE) (Sag 1976; Warner 1993; Miller 2011; Miller and Pullum 2014), exemplified in (1) and (2), which respectively illustrate the PAE subtypes of Verb Phrase Ellipsis (VPE) and Pseudogapping (PG):

1. I have published a book but he hasn't ~~published a book.~~ [VPE]
2. John called Sarah, and Mary did ~~call~~ Jane. [PG]

In VPE the whole constituent following the licensor is ellipsed, whereas in PG a remnant is kept after the licensor in the ellipsis site.

The case study reported here explores these elliptical constructions through the statistical modelling of the variation between the two subtypes of PAE in a corpus of Modern English. The research question addressed in this investigation is: what is the nature of the linguistic determinants that trigger PAE (VPE versus PG) in English?

[26] We are grateful to the Spanish State Research Agency and the European Regional Development Fund (grant no. PID2020-117541GB-I00) and the Regional Government of Galicia (grant no. ED431C 2021/52).

This chapter is organised as follows. Section 2 defines PAE and its main linguistic characteristics. Section 3 describes the data, whose statistical modelling is accounted for in Section 4. Section 5 summarises the study, reports the results and provides the interpretation of the findings.

2. PAE: DEFINITION AND LINGUISTIC FEATURES

PAE involves the ellipsis of verb phrases, adjective phrases, (noun or) determiner phrases or prepositional phrases after main verb *be* or *have*, auxiliaries *be, have* or *do*, modal auxiliaries or after the infinitival marker *to* (also a defective non-finite auxiliary verb in, among others, Pullum 1982; Gazdar et al. 1985; Levine 2012; Miller and Pullum 2014). Such structural choices are illustrated in the following examples of the PAE subtype VPE (adapted from Gandón-Chapela, 2020), which, as pointed out in Section 1, involves the ellipsis of the whole phrasal constituent:

3. I have written a squib but he hasn't [~~written a squib~~]$_{VP}$.
4. John is tall but Sarah is not [~~tall~~]$_{AP}$.
5. John is a doctor and Anne is [~~a doctor~~]$_{DP}$ too.
6. Bill's son is on the beach, although he shouldn't be [~~on the beach~~]$_{PP}$ because he's allergic to the sun.

Unlike in VPE, in PG, the other subtype of VPE, a so-called remnant is left expressed after ellipsis, as in (7):

7. Paul invited Patrick, and Monica did ~~invite~~ Julia.

The first step taken in order to analyse the variation between the two subtypes of PAE was to identify potential factors from the literature. Firstly, as shown in (8), in Present-Day English infinitival marker *to* is licensed in VPE (Levin 1986; Bos and Spenader 2011; Miller 2014) and is not possible in PG. In fact, no instances of PG licensed by to have been reported in corpus studies on ellipsis, as Miller (2014) puts forward. This is instantiated in the following examples from Levin (1986, 54):

8. Speaker A: Van Gogh's work is beginning to impress me.
Speaker B: Well! It's finally starting to ø. [VPE]
*It's starting to ø me, too [PG]

Secondly, whereas VPE can be licensed by more than one auxiliary in Present-Day English (see (9)), as a general rule PG cannot (Levin 1986; Hoeksema 2006; Miller 2014). Levin (1986, 54) reports only one case of PG where there are two auxiliaries involved, here in (10).

9. I saw it and obviously so did Arnold, but nobody else *could have*. [VPE]
10. Speaker A: Cream rinse makes my hair get dirty faster.
Speaker B: ??It may have mine once, too. [PG]

Thirdly, as regards the type of syntactic linking between the ellipsis site and the clause containing the antecedent, VPE has been claimed to have the capacity to occur in both coordination (in (12)) and subordination contexts (as in (13) – versus (11), where the ellipsis and the antecedent clauses are not formally linked). However, the distribution of PG seems to be much more restrictive and the literature on the topic has pointed out that this construction is favoured in adverbial (mostly comparative) contexts (Hardt and Rambow 2001; Nielsen 2005; Bos and Spenader 2011; Miller 2014), as shown in (14) and (15) (examples extracted from the PPCMBE):

11. I can recollect nothing more to say. When my letter is gone, I suppose I shall. [no linking/dependency – VPE]
12. That I had received such from Edward also I need not mention; but I do, you see, because it is a pleasure. [coordination – VPE]
13. that he would not look upon us as Enemies, but do us all the Service he could. [relative-clause subordination – VPE]
14. but did not admire the strain of its poetry in general, though I did its morality. [adverbial subordination – PG]
15. A skilled florist will produce a finer effect with a few inexpensive blossoms than an unskilled one will with a cartload of choice material. [comparative subordination – PG]

Fourthly, instances of voice mismatch have been considered among the determinants of PAE. The importance of checking this variable lies in the fact that in the literature it has been reported that while mismatches in voice between the antecedent and the ellipsis site are possible in VPE (Merchant 2008), as in (16), this is not true with PG, in (17).

16. I wish heartily, said Wyatt, it was in my power to entertainactive your honour as you ought to be ~~entertained~~$_{passive}$. (PPCMBE) [VPE]
17. *Klimt is admired by Abby$_{passive}$ more than anyone does admire Kleeactive. (Merchant [2008, 169–70]) [PG]

Fifthly, the ana- versus cataphoric connection between the target and its antecedent has been claimed to play a role in the variation between VPE and PG since the cataphoric connection is only possible in the former subtype (Levin 1986; Hardt 1993; Bos and Spenader 2011; Miller 2014), as demonstrated in examples (18) and (19), taken from Levin (1986, 54).

18. Although it doesn't always ø, it sometimes takes a long time to clean the hamster's cage. [VPE, cataphoric]
19. *Although it doesn't ø me, it takes Karen a long time to clean the hamster's cage. [PG, cataphoric]

Finally, the syntactic (in number of intervening Inflection Phrases or IPs) and the lexical (in number of words) distance between the antecedent and the target has also been used to characterise VPE (in (20) and (21)) versus PG (in (22)) (Hardt and Rambow 2001; Nielsen 2005; Gandón-Chapela 2020) – see Section 5.

20. I have written a squib but I think that Mary hasn't ~~written a squib~~. [VPE: lexical (or word) distance: 6 words; syntactic (or sentential) distance: 1 IP]
21. John is talkative but Sarah is not ~~talkative~~. [VPE: lexical distance: 4 words; syntactic distance: 0 IP]
22. Peter kissed Sonya, and Beth did kiss Jason. [PG: lexical distance: 4 words; syntactic distance: 0 IP]

3. DATA

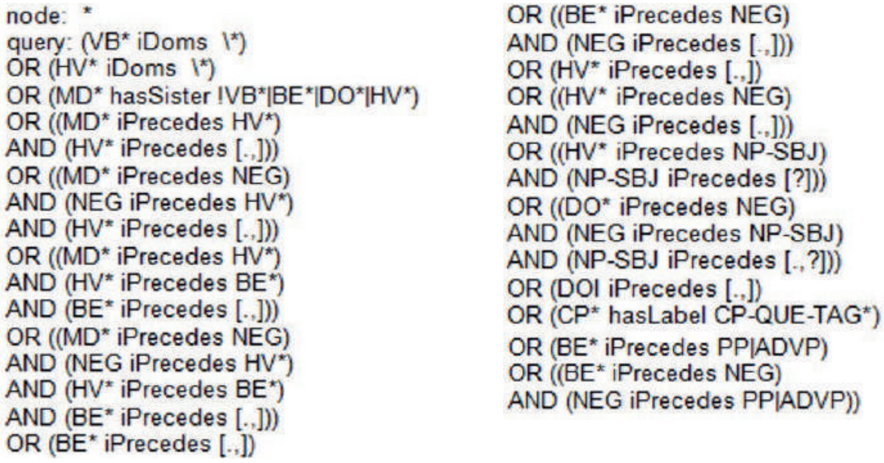
The examples of VPE and PG were retrieved from the 102 texts of the Penn Parsed Corpus of Modern British English (PPCMBE; Kroch et al. 2010), a corpus of 948,895 words of written Modern British English, dated 1700–1914. This parsed corpus adopts a syntactic tagset inspired by the Principles and Parameters model. Example (23) illustrates the way in which He *did* is parsed in the PPCMBE, where the covert object of *did* is linked to a ('*'-)trace in initial position.

23. He did.

```
(IP-SUB      (NP-OB1 *T*-1)
                   (NP-SBJ (PRO he))
                   (DOD did)
                   (VB *)))))
(. .))
```

In order to undertake the statistical analysis of every instance of PAE in our data, a query algorithm was designed that relies on the syntactic parsing of PPCMBE (in (24)). This algorithm led to the retrieval of the relevant set of examples of PAE, with successful recall/precision rates that were calculated on a (balanced) pilot subcorpus on 12 texts (112,347 words; see Gandón-Chapela 2020, 75–76). After manual pruning, the database consisted of 976 and 86 instances of VPE and PG, respectively.

24. Query



```
node:  *
query: (VB* iDoms  \*)
OR (HV* iDoms  \*)
OR (MD* hasSister !VB*|BE*|DO*|HV*)
OR ((MD* iPrecedes HV*)
AND (HV* iPrecedes [.,]))
OR ((MD* iPrecedes NEG)
AND (NEG iPrecedes HV*)
AND (HV* iPrecedes [.,]))
OR ((MD* iPrecedes HV*)
AND (HV* iPrecedes BE*)
AND (BE* iPrecedes [.,]))
OR ((MD* iPrecedes NEG)
AND (NEG iPrecedes HV*)
AND (HV* iPrecedes BE*)
AND (BE* iPrecedes [.,]))
OR (BE* iPrecedes [.,])
OR ((BE* iPrecedes NEG)
AND (NEG iPrecedes [.,]))
OR (HV* iPrecedes [.,])
OR ((HV* iPrecedes NEG)
AND (NEG iPrecedes [.,]))
OR ((HV* iPrecedes NP-SBJ)
AND (NP-SBJ iPrecedes [?]))
OR ((DO* iPrecedes NEG)
AND (NEG iPrecedes NP-SBJ)
AND (NP-SBJ iPrecedes [.,?]))
OR (DOI iPrecedes [.,])
OR (CP* hasLabel CP-QUE-TAG*)
OR (BE* iPrecedes PP|ADVP)
OR ((BE* iPrecedes NEG)
AND (NEG iPrecedes PP|ADVP))
```

4. MODELLING THE DATA

This section describes the statistical modelling of the database described in Section 3 and the grammatical, semantic/discursive and processing linguistic predictors listed in Section 2 that allowed us to analyse every example where variation between VPE and PG was potentially at work.

For technical reasons, not every determinant could enter the model. On the one hand, the low number of examples made us modify the predictor reflecting 'syntactic linking', in particular, the subordination subtypes relative-clause, adverbial and comparative subordination. Since we had few examples in some of these levels, we decided to group the subordination options together into one 'subordination' level. On the other hand, on other occasions the distribution of the examples per level was highly or fully categorical; an example of this is the infinitival marker *to* of the predictor 'licensor', which, as already pointed out in Section 2, is only possible with VPE. Since we did not have full variation here with respect to this variable, we had to get rid of the predictor as a whole. The definitive design of the database is given in (25).

25.
```
  type     aux_licensor          linking
 vpe:976   no :1010     subord        :521
 pg : 86   yes:  52     coord         : 61
                        diff_sentence:480


  distance_ip       distance_word     phoric          voice
 Min.   : 0.0000   Min.   : 1.000   ana :1051   same    :1054
 1st Qu.: 0.0000   1st Qu.: 3.000   cata:  11   mismatch:   8
 Median : 0.0000   Median : 4.000
 Mean   : 0.3324   Mean   : 5.113
 3rd Qu.: 0.0000   3rd Qu.: 6.000
 Max.   :15.0000   Max.   :78.000
```

The database contains the response variable 'type', with the two levels of PAE: 'VPE' and 'PG', and a number of independent predictors:

- The dichotomous 'aux_licensor' (auxiliary before the licensor) predictor with two levels: 'yes' (presence of auxiliary plus licensor) and 'no'
- The different levels of syntactic 'linking': 'coord(ination)', 'subord(ination)' and 'diff_sentence' (no linking/dependency)
- The two distance variables: syntactic distance ('distance_ip'), measuring the number of IPs between the antecedent and the ellipsis site, and lexical distance ('distance_word'), reflecting the number of words between the target and the source of ellipsis
- The 'phoric' predictor expressing the 'ana(phoric)' or 'cata(phoric)' connection between the antecedent and the target of ellipsis
- The possibility of 'voice' mismatch, also dichotomous: 'same' voice and voice 'mismatch'

The data were modelled through fixed-effects binomial regression (functions 'glm'/'lrm' in R), which was responsible for detecting which predictors strongly explain the variation and which do not. Mutual collinearity (through the functions 'alias' and 'vif') proved not to be severe, only moderate (vif < 4.02) between the two 'distance' predictors. The backward-stepwise reduction of predictors, which was used to compare the AIC values of enriched and reduced models, allowed us to discard two variables that did not contribute significantly to the explanation of the variation, namely 'mismatch' and 'aux_licensor'. In other words, the explanatory power of the model with all the variables is not significantly different from the explanatory power of the model without these two variables. Besides, the C-index (0.813) of the definitive model in (26) indicates that the latter is very robust and explains the vast majority of the examples of the variation.

26.

```
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)             -3.26311    0.25163 -12.968  < 2e-16 ***
linkingcoord            -0.61537    0.56072  -1.097 0.272435    
linkingdiff_sentence    -1.17075    0.33556  -3.489 0.000485 ***
distance_ip             -2.81724    0.48681  -5.787 7.16e-09 ***
distance_word            0.31249    0.04233   7.383 1.55e-13 ***
phoriccata             -15.26942  643.41257  -0.024 0.981066    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Random Forests was used to produce the conditional importance graph in Figure 1, with an outstanding C-index (0.852). In a visual way, this graph tells us that out of the four variables that eventually entered the model, two of them, syntactic linking (involving the levels of coordination, subordination and lack of dependency) and the ana-/cataphoric connection between the antecedent and the ellipsis site are close to the ‘0’ level, which means that they are not sufficiently explanatory. Therefore, the only variables that could strongly explain the variation were the distance variables reflecting syntactic distance, measured in number of intervening IPs, and lexical distance, computed by means of the number of words attested between the antecedent and the target of ellipsis.

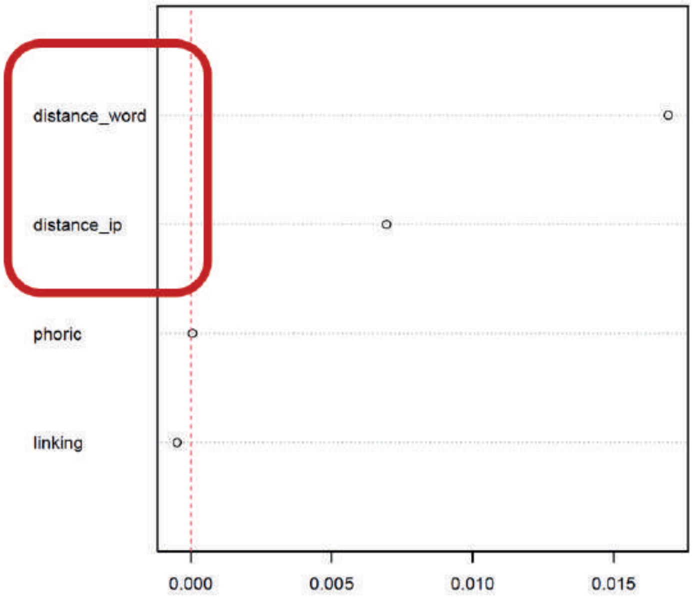


Figure 1: Conditional importance of variables

Let us now focus on, for example, the strong predictor of syntactic distance. The effects plot in Figure 2 shows the behaviour of the (95%-confidence) correlation between the types of PAE, that is, either VPE or PG, and syntactic distance. The vertical axis marshals the continuum between PG, in higher position, and VPE, in lower position. The horizontal axis reflects the number of IPs attested in our data occurring between the antecedent and the target of ellipsis. Moving on from PG (higher position) to VPE (lower position in this axis) is accompanied by an increase in syntactic distance, that is, in the number of IPs occurring between the antecedent and the ellipsis site.

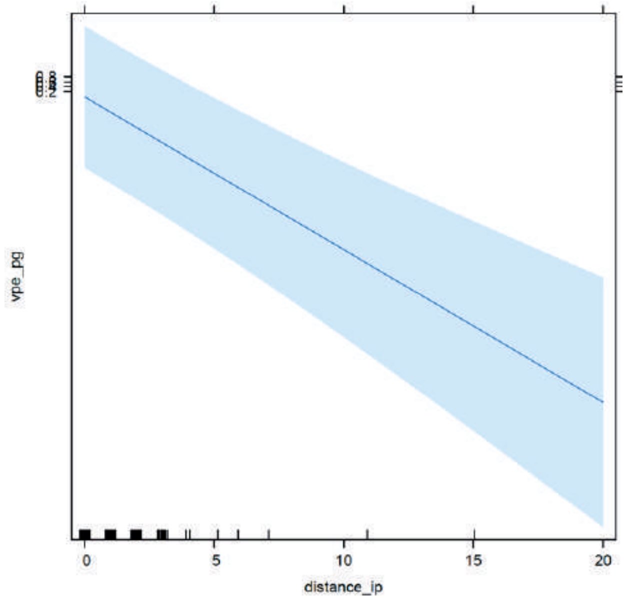


Figure 2: Effects plot of syntactic distance

5. DISCUSSION AND INTERPRETATION OF FINDINGS

Even though we are aware that the major linguistic factor accounting for the differences between VPE and PG is, respectively, the lack or the occurrence of a (commonly contrastive) remnant in the ellipsis site, which has consequences for the way in which, for example, information is structured in contexts of PAE, in this study we aimed at accounting for the contribution of other linguistic variables to the variation VPE versus PG in Modern English.

Adopting a number of quantifiable variables from the literature, we categorised the corpus examples and applied a widely-accepted statistical model to the data. Unfortunately, one of the predictors, licensor to, proved to be categorical with VPE, so it was not considered an alternative of the variation. Other variables were poor factors of the variation, namely the presence of an auxiliary before the licensor, voice mismatch, the ana-/cataphoric connection between the antecedent and the ellipsis site, and syntactic-linking choices (coordination, subordination or lack of formal dependency). Finally, the major contribution in this study was the significant weight of the distance predictors, that is, syntactic and lexical distance, which proved to be very strong contributors to the variation between VPE and PG.

The qualitative interpretation of the facts revealed by the statistical model allows us to claim that the variation between VPE and PG is not strongly subjected to grammatical, systematic, language-internal forces, and is not significantly conditioned by semantic factors either. On the contrary, VPE/PG choice is strongly explained by reference to processing demands. To illustrate this, the model demonstrated that when we have longer distance between the antecedent and the ellipsis site, we have more chances of VPE. When the distance is shorter, we have more chances of PG. Let us bear in mind that the main difference between VPE and PG is that we have a remnant in the latter. This remnant fulfils a specific syntactic function in its clause. In order to reconstruct the syntax (and meaning) of the ellipsis site, that is, in order to associate the remnant with its correct syntactic function, we need to retain the syntactic structure of the antecedent clause. This is specifically urgent in the case of PG. In VPE we omit the whole predicate, so reconstruction or identification of the syntatic roles of VP constituents is not so urgent as it is in those cases of PG, in which the function of the remnants is mandatory. So, and this is our claim, the preference for shorter distance between the antecedent and the target of ellipsis, in particular in those cases of PG, can be justified by the fact that PG disprefers the insertion of IPs between the antecedent and the target clause. The insertion of IPs is considerably disruptive from the point of view of processing, so we need to ease processing and one way of doing this is by shortening this syntactic distance between the antecedent and the ellipsis site.

As already pointed out, attention needs to be paid to information structure in further research. We should also pinpoint text-type differences, undertake a more fine-grained analysis of turns and of clause type/clause mode (differences between declarative, interrogative, tags, exclamative sentences).

REFERENCES

- Bos, Johan and Jennifer Spenader. 2011. "An Annotated Corpus for the Analysis of VP Ellipsis." *Language Resources and Evaluation* 45 (4): 463-94.

- Gandón-Chapela, Evelyn. 2020. *On Invisible Language in Modern English. A Corpus-based Approach to Ellipsis.* London: Bloomsbury Academic.

- Gazdar, Gerald, Ewan Klein, Geoffrey K. Pullum and Ivan A. Sag 1985. *Generalized Phrase Structure Grammar*. Oxford: Basil Blackwell.

- Hardt, Daniel. 1993. "Verb Phrase Ellipsis: Form, Meaning, and Processing." PhD diss., University of Pennsylvania.

- Hardt, Daniel and Owen Rambow. 2001. "Generation of VP Ellipsis: A Corpus-based Approach." In Webber 2001, 290-97.

- Hendrickx, Iris, Sobha Lalitha Devi, António Branco and Ruslan Mitkov, eds. 2011. *Anaphora Processing and Applications: 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), Lecture Notes in Artificial Intelligence*, vol. 7099. Berlin: Springer.

- Hoeksema, Jack. 2006. "Pseudogapping: Its Syntactic Analysis and Cumulative Effects on Acceptability." *Research on Language and Computation* 4: 335-352.

- Hofmeister, Philip and Elisabeth Norcliffe, eds. 2014. *The Core and the Periphery: Data-driven Perspectives on Syntax Inspired by Ivan A. Sag*. Stanford, CA: CSLI Publications.

- Kroch, Anthony, Beatrice Santorini and Ariel Diertani. 2010. Penn Parsed Corpus of Modern British English. *http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCMBE2-RELEASE-1.*

- Levin, Nancy S. 1986. *Main Verb Ellipsis in Spoken English*. New York: Garland.

- Levine, Robert D. 2012. "Auxiliaries: *To's* company." *Journal of Linguistics* 48 (1): 187-203.

- Merchant, Jason. 2008. "An Asymmetry in Voice Mismatches in VP-ellipsis and Pseudogapping." *Linguistic Inquiry* 39 (1): 169-179.

- Miller, Philip. 2011. "The Choice between Verbal Anaphors in Discourse." In Hendrickx et al. 2011, 82-95.

- Miller, Philip. 2014. "A Corpus Study of Pseudogapping and its Theoretical Consequences." In Piñón 2011, 73-90.

- Miller, Philip and Geoffrey K. Pullum. 2014. "Exophoric VP Ellipsis." In Hofmeister and Norcliffe 2014, 5-32.

- Nielsen, Lief Arda. 2005. "A Corpus-based Study of Verb Phrase Ellipsis Identification and Resolution." PhD diss., University of London – King's College London.

- Piñón, Christopher, ed. 2011. *Empirical Issues in Syntax and Semantics* 10. *http://www.cssp.cnrs.fr/eiss10/ [Accessed March 28, 2023].*

- Pullum, Geoffrey K. 1982. "Syncategorematicity and English Infinitival *to.*" *Glossa* 16 (2): 181-215.

- Sag, Ivan A. 1976. "Deletion and Logical Form." PhD diss., MIT.

- Warner, Anthony R. 1993. English Auxiliaries: *Structure and History*. Cambridge: Cambridge UP.

- Webber, Bonnie Lynn, ed. 2001. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Toulouse, France, 9-11 July 2001. Stroudsburg*, PA: Association for Computational Linguistics.